

AI-LINGUISTICA

Linguistic Studies on AI-Generated
Texts and Discourses



Special Issue

The Notion of Authenticity in Hybrid Human/AI Productions

Eds. Sophia Burnett & Silvia Lima Gonçalves Araújo

AILing

Vol. 4 No. 1

DOI: 10.62408/ai-ling.v4i1
CC-BY-NC-SA 4.0

AI-Linguistica
Linguistic Studies on AI-Generated Texts and

ISSN: 2943-0070

AI-Linguistica

Linguistic Studies on AI-Generated Texts and Discourses

Special Issue

**The Notion of Authenticity
in Hybrid Human/AI Productions**

Eds. Sophia Burnett & Silvia Lima Gonçalves Araújo

Vol. 4, No. 1

DOI: 10.62408/ai-ling.v4.i1

ISSN: 2943-0070

AILing

CC-BY-NC-SA 4.0

About the Journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is an international peer-reviewed journal devoted to the linguistic study of generated texts and discourses. It accepts original articles devoted to generated texts and discourses in all areas of linguistic research. *AI-Linguistica* publishes studies on all languages, with a special focus on Romance and Germanic languages.

The journal is aimed at all scholars interested in the intersection between linguistics (across all theoretical frameworks), as well as neighbouring disciplines, and generative artificial intelligence (GenAI).

Editorial Team

Editor-in-Chief:

Anna-Maria De Cesare, *Technische Universität Dresden, Germany*

Co-Editor:

Mirko Tavosanis, *Università di Pisa, Italy*

Editorial Board:

Valentina Bambini, *University School for Advanced Studies, IUSS, Pavia, Italy*

Noah Bubenhofer, *Universität Zürich, Switzerland*

Kerstin Fischer, *University of Southern Denmark, Denmark*

Alexander Lasch, *Technische Universität Dresden, Germany*

Alessandro Panunzi, *Università degli Studi di Firenze, Italy*

Rachele Raus, *Università di Bologna, Italy*

Britta Schneider, *Universität Wien, Austria*

Maria Grazia Sindoni, *Università degli Studi di Messina, Italy*

Eva Vanmassenhove, *Tilburg University, Netherlands*

Editorial Office:

Michela Gargiulo, *Technische Universität Dresden, Germany*

Giulia Mantovani, *Technische Universität Dresden, Germany*

Tom Weidensdorfer, *Technische Universität Dresden, Germany*

ISSN: 2943-0070

DOI: 10.62408/ai-ling

Table of Contents

Sophia Burnett

Introduction

PART I

IDEOLOGIES, REPRODUCTIONS, AND CIRCULATIONS

Andrew Frank Bradley and María del Mar Palomares Marín

Theorising authenticity vis-à-vis, not versus, syntheticity in the age of AI:
Introducing a continuum-based framework

Perrine Maurel

Style hybride d'une chimère. Authentifier le style dans une production textuelle
hybride

Nina Kalwa

AI imitating individuals: Deepfakes and the unexpected in digital communication

Alejandro Acero Ayuda and Carlos Yebra López

Halis Sefaradi: Critical authenticity and the voice of the first Ladino chatbot
Estreyika

PART II

COGNITION, SOCIAL INTERACTION, AND THE CO-CONSTRUCTION OF MEANING

Louise Robert and Jérémie Sauvage

Écrire des textes cliniques simplifiés pour les familles aux urgences pédiatriques
avec l'IA générative : tensions entre lisibilité, auctorialité et charge de travail

Luca Pallanti and Rawad Chaker

Humans vs. Machines: Can enaction theory help?

Laura Margarita Merino Hernández and Francisco Ariano Cifuentes

Perceived authenticity of positive politeness strategies in human-computer
interactions

Sophia Burnett

Habeas corpus: An exploration of authentic embodiment and ecology in a French
contrastive corpus

Acknowledgements

Author Biographies

Introduction to Special Issue: The notion of authenticity in hybrid human/AI productions

Sophia Burnett (Université de Toulouse & CY Cergy Paris)

sophia.burnett(at)cyu.fr

This special issue of *AI-Linguistica* took root amidst the fulgurant and seemingly ubiquitous adoption of commercial generative artificial intelligence (GenAI) and large language models (LLMs) by the general public, the advent of which has forced linguists and other scholars to revisit our knowledge of meaning production, reception, and perception, in particular via text. In these pages we examine how the notion of authenticity is maintained, transformed, or redefined within hybrid human/AI textual and multimodal productions. Why this collection? Firstly, such scrutiny allows us to reposition “AI” to its correct place, that of a powerful, but lifeless assembly of algorithms. The term AI, *Artificial Intelligence* has been widely adopted as a catch-all term by both the general public and governmental institutions, however, GenAI and LLMs do not possess the quality, nor the complexity of thought and cognition of the sort promised by the still elusive *Artificial General Intelligence*. Secondly, it is a lens which provides pertinent evaluations of these advanced tools, and how we perceive and apprehend them; and as linguists, it is a heuristic that renders visible otherwise intangible processes, providing new insights into questions surrounding intention, function, authorship, style, perception, cognition, didactics, etc., questions that have perdured since long before these models even existed. Finally, the notion of authenticity has always been linked to the written word, not only as authorship in the sense of intellectual property of say a poem or a recipe, but as an artifact of the person who created the semiosis, a spatio-temporal marker of living breathing human existence via a sign etched into a rock, a hand against a cave wall, a beautifully penned signature, all signifying an individual, and that signification’s legible authenticity.

Authenticity is etymologically grounded in the Greek *authentikos*, from *authentēs* (‘doer’, ‘master’), itself composed of *autos* (‘self’) and *hentes* (‘one who acts’). The term thus encodes, at its root, a relation between selfhood and action that continues to structure contemporary debates. Beyond simple authorship, the semantic field of authenticity is existentially rich: truth, genuineness, originality, sincerity, integrity, identity, credibility, trustworthiness, legitimacy, and heritage. It is a fundamental notion tackled since antiquity, and there exist several definitions of authenticity that have come to us through the ages, each serving slightly different purposes.

As per Tomas and Das (2022), three main interpretations of authenticity emerge across disciplines. The first, rooted in philosophical traditions from Ancient

AILing

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses
CC-BY-NC-SA 4.0

Burnett, Sophia. 2026. Introduction.
Special Issue: *The notion of authenticity*. Vol. 4 No.1
DOI: 10.62408/ai-ling.v4i1.89

ISSN: 2943-0070

Greece to twentieth-century critical theory, defines authenticity as *the alignment between an entity's internal ideals and its external expression*. The second, drawing on cognitive science categorization, schemas and socio-institutional categories, defines authenticity as *adherence to the norms of a social category*. The third, grounded in psychological essentialism and semiotics, defines authenticity as *a relation linking an entity to a person, place, or time*. These interpretations inform many of the frameworks developed in the following papers.

The eight articles published in this special issue tackle the notion of authenticity through the lens of *hybridity*. The term *hybrid* here refers to the dynamic interaction between human agents and artificial intelligence systems, in which the production of language results from distributed co-agency between human cognition and algorithmic computation, whether in language sciences (Beguš et al. 2023; De Cesare 2023; Dynel 2023; Meier 2024; Weissweiler 2025) translation studies (Li, Rao, Shang et al. 2025; Xu et al. 2025) literary/narratology studies (Beguš 2024; Chakrabarty, Laban, Agarwal et al. 2024; Koivisto and Grassini 2024) discursivity/discourse analysis (Merton 1968; Yoo et al. 2024; Liu et al. 2025; Cohn, Snyder, Montenegro et al. 2024) NLP/computational linguistics (Liu et al. 2024) didactics (Petkovic and Alrahabi 2022; Ifelebuegu 2024; Werdiningsih, Marzuki, Inda et al. 2024) or cognition (Carrasco-Farre 2024; Grindrod 2024; Wang, Zhang and Seedhouse 2025).

The need for a dedicated special issue is also explained by the layered and complex nature of the problem. During a fittingly hybrid *Journée d'Etudes* of the same name at the Université de Lorraine on 13/03/2025¹, which brought together scholars from several countries and fields, it soon became apparent that questions surrounding authenticity in hybrid human/AI productions are inherently multifaceted, if not interdisciplinary, and that an appropriate scholarly response therefore needs to engage to some degree with interdisciplinarity. Within linguistics in particular, these questions connect with work on discourse practices, minority languages, pragmatics, corpus analysis, second-language acquisition, embodiment, etc., which is why a dedicated call for this special issue was launched, allowing us to receive dozens of propositions, and to retain only eight.

The notion of authenticity is linked to those of identity and truth. Today, it is being reconfigured in the context of hybrid productions between humans and artificial intelligence. While the general public tends to perceive artificial intelligence as a whole greater than the sum of its parts, it is important to recall that this gestalt—referring merely to the generative capacity of computational models LLMs—does not produce language grounded in embodied experience, but rather draws on billions of tokens from disparate sources (Zhao et al. 2023) often collected without authorization (Baack et al. 2025). In other words, these productions are amalgams of signs, symbols, or images, resulting from statistical calculations and not from lived experiences or embodied reflections. The

authenticity of hybrid productions is an issue that brings together cognitivists and generativists. Lakoff (1986) contrasted computational production and human production, rejecting the computer-brain model, and according to Chomsky, Roberts, and Watumull (2023) “we know from the science of linguistics and the philosophy of knowledge that they differ profoundly from how humans reason and use language”.

In order to apply any analysis of authenticity to the examination of hybrid productions, we must first critically interrogate the very meaning of authenticity. To do so, we draw on an epistemic framework that predates the emergence of LLMs. For a more comprehensive introduction to the notion of authenticity, we suggest Lindholm (2013). Trilling (1978), frames the evolution of authenticity as a derivation from sincerity. Handler (1986) argues that authenticity is not an innate property, but a discursive construction mobilized, for example, by nationalists to assuage anxieties around continuity and legitimation. Linnekin (1991), drawing on a Maori case-study, shows that so-called “authentic” traditions are in fact dynamic, interpreted, and politically invested: authenticity becomes a narrative rather than a reproduction of empirical reality. Lindholm (2013), addressing authenticity in the digital era—prior to the emergence of LLMs—in the context of early online banking and other official forms of digital authentication, writes: “Anxiety about the validity of experience and about the maintenance of personal identity is at the core of this computerized definition.” Drawing on linguistic and semiotic frameworks, van Leeuwen (2001) offers several responses to the question, “What is authenticity?” He observes that media tend to reproduce and reinforce the idea that authenticity is concealed behind masks, only to be revealed in order to produce an effect of realism within a saturated media landscape. This highlights the paradox of authenticity in late modernity: it must appear spontaneous, even though it is often carefully assembled [...]

Excerpt from the call for papers, 14/04/2025
https://ailing.publia.org/ai_ling/announcement/view/2

The contributions are organised into two thematic sections that approach authenticity in hybrid productions from complementary perspectives.

Part I, *Ideologies, reproductions, and circulations*, opens with Bradley and Palomares, who propose a conceptual framework reconciling authenticity and *syntheticity* by positioning them along a continuum rather than treating them as binary opposites. Their model distinguishes between the technological origin of a production and its social reception, thereby providing an analytical tool for understanding the complex validation states that characterise hybrid texts. Maurel then turns to the question of style, examining how authorial legitimacy is negotiated within hybrid writing processes involving large language models. By proposing a categorisation of hybrid textual productions based on the origin of textual material, the direction of stylistic modification, and the distribution of authorial intervention, the article highlights the role of human editorial endorsement in the authentication of style. Kalwa approaches authenticity through the lens of digital communication, analysing discourse surrounding deepfake videos circulated on social media. Through close analysis of online commentary, their study demonstrates how

audiences mobilise expectations of authenticity and respond to perceived anomalies when evaluating the credibility of mediated representations. The first section concludes with Acero Ayuda and Yebra López, who examine the Ladino-speaking chatbot *Estreyika* as a sociolinguistic site where authenticity is negotiated among heritage speakers, learners, and cultural outsiders. Their analysis highlights how authenticity operates not as a fixed property of linguistic production but as a positional and ideological construct shaped by community practices and language revitalisation efforts.

Part II, *Cognition, social interaction, and the co-construction of meaning*, opens with Robert and Sauvage, who investigate the use of generative AI in the production of simplified clinical texts intended for families in paediatric emergency care. Drawing on ethnographic observation, interviews with medical professionals, and corpus analysis, the study reveals how the perceived authenticity of generated texts depends on their alignment with established institutional discourse practices and on the corrective labour required to appropriate machine-generated output. Pallanti and Chaker approach the distinction between human and AI-generated discourse through the framework of enaction theory, proposing sentence-length variance as a quantitative indicator of the irregular perception–action coupling characteristic of human cognition. Their findings suggest measurable structural differences between human and machine-produced texts, while also revealing the limits of current generative models in reproducing human linguistic variability. Merino Hernández and Ariano Cifuentes examine perceptions of authenticity in human–computer interaction through an experimental study of positive politeness strategies in Spanish-language chatbot interactions. Their results show that conversational cues such as opening comments and follow-up questions significantly shape perceptions of authenticity, while emojis may reduce the perceived naturalness of interaction. The section concludes with Burnett’s investigation of embodiment in French human, AI-generated, and hybrid texts. By analysing linguistic markers associated with *ecognition*, the study proposes an embodiment scale designed to measure the proximity between textual production and embodied human experience, offering a novel methodological approach for examining humanity in hybrid discourse.

Taken together, the contributions in this special issue highlight the importance of authenticity in hybrid human/AI productions as an object of study; one that requires careful apprehension in order to locate and to gauge its behaviour within hybridity, which is not some neat median solid and fixed state between opposing humans and machines. Rather, it flows through more or less condensed and diluted states *borne upon and bearing* linguistic, cognitive, and social material. Such a convergence of theoretical models and applied analyses across diverse

linguistic and social contexts affords this issue a unique analytical vantage point from which to grasp authenticity as a means and an end.

References

- Baack, Stefan & Biderman, Stella & Odrozek, Kasia & Skowron, Aviya & Bdeir, Ayah & Bommarito, Jillian & Ding, Jennifer et al. 2025. Towards best practices for open datasets for LLM training. <https://doi.org/10.48550/arXiv.2501.08365>
- Beguš, Gašper & Dąbkowski, Maksymilian & Rhodes, Ryan. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of LLMs. <https://doi.org/10.1109/TAI.2025.3575745>
- Beguš, Nina. 2024. Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. *Humanities and Social Sciences Communications* 11(1): 1–22. <https://doi.org/10.1057/s41599-024-03868-8>
- Carrasco-Farré, Carlos. 2024. Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. <https://doi.org/10.48550/arXiv.2404.09329>
- Chakrabarty, Tuhin & Laban, Philippe & Agarwal, Divyansh & Muresan, Smaranda & Wu, Chien-Sheng. 2024. Art or artifice? Large language models and the false promise of creativity. In Kyburz, Penny & Mueller, Florian Floyd (eds.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–34. <https://doi.org/10.48550/arXiv.2309.14556>
- Chomsky, Noam & Roberts, Ian & Watumull, Jeffrey. 2023. Noam Chomsky: The false promise of ChatGPT. *The New York Times* 8, no. 4: 177–179.
- Cohn, Clayton & Snyder, Caitlin & Montenegro, Justin & Biswas, Gautam. 2024. Towards a human-in-the-loop LLM approach to collaborative discourse analysis. In Cristea, Alexandra I. & Walker, Erin & Lu, Yu & Santos, Olga C. & Isotani, Seiji (eds.), *Proceedings of the International Conference on Artificial Intelligence in Education*. 11–19. Cham: Springer Nature Switzerland.
- De Cesare, Anna-Maria. 2023. Assessing the quality of ChatGPT's generated output in light of human-written texts: A corpus study based on textual parameters. *CHIMERA. Romance Corpora and Linguistic studies* 10: 179–210. <https://revistas.uam.es/chimera/article/view/17979>. (last accessed 23/03/26)
- Dynel, Marta. 2023. Lessons in linguistics with ChatGPT: Metapragmatics, metacommunication, metadiscourse and metalanguage in human-AI interactions. *Language & Communication* 93: 107–124. <https://doi.org/10.1016/j.langcom.2023.09.002>
- Grindrod, Jumbly. 2024. Large language models and linguistic intentionality. *Synthese* 204(2). 71–95. <https://doi.org/10.48550/arXiv.2404.09576>
- Handler, Richard. 1986. Authenticity. *Anthropology Today* 21: 2–4.
- Ifelebuegu, Augustine Osamor. 2024. Rise of the robots: What it means for educators. *Journal of Applied Learning & Teaching* 7(1): 413–420.

- Koivisto, Mika & Grassini, Simone. 2024. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports* 13(1): 13601. <https://doi.org/10.1038/s41598-023-40858-3>
- Lakoff, George. 1988. Cognitive semantics. *Meaning and Mental Representations* 119: 154.
- Li, Zongyao & Rao, Zhiqiang & Shang, Hengchao & Guo, Jiabin & Li, Shaojun & Wei, Daimeng & Yang, Hao. 2025. Enhancing Large Language Models for Document-Level Translation Post-Editing Using Monolingual Data. In Rambow, Owen & Wanner, Leo & Apidianaki, Marianna & Al-Khalifa, Hend & Di Eugenio, Barbara & Schockaert, Steven (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*. 8830–8840. Abu Dhabi, UAE. Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.591.pdf> (last accessed 23/03/26)
- Lindholm, Charles. 2013. The rise of expressive authenticity. *Anthropological Quarterly* 86(2): 361–395. <https://doi.org/10.1353/anq.2013.0020>
- Linnekin, Jocelyn. 1991. Cultural invention and the dilemma of authenticity. *American Anthropologist* 93(2): 446–449.
- Liu, Xiao & Lei, Xuanyu & Wang, Shengyuan & Huang, Yue & Feng, Andrew & Wen, Bosi & Cheng, Jiale et al. 2024. Alignbench: Benchmarking chinese alignment of large language models. In Ku, Lun-Wei & Martins, André & Srikumar, Vivek (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*. 11621–11640. <https://doi.org/10.48550/arXiv.2311.18743>
- Meier, Raphael. 2024. LLM-aided social media influence operations. *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*. 105–112. <https://doi.org/10.48550/arXiv.2309.03670>
- Merton, Robert King. *Social theory and social structure*. New York, Simon and Schuster.
- Petkovic, Ljudmila & Alrahabi, Motasem. 2022. Impact de la correction automatique de l'OCR/HTR sur la reconnaissance d'entités nommées dans un corpus bruité. *Journal of Information Sciences* 21(2). 42–57. <https://doi.org/10.34874/IMIST.PRSM/jis-v21i2.36599>
- Qin, Libo & Chen, Qiguang & Zhou, Yuhang & Chen, Zhi & Li, Yinghui & Liao, Lizi & Li, Min & Che, Wanxiang & Yu, Philip S. 2025. A survey of multilingual large language models. *Patterns* 6(1): 1–30. <https://doi.org/10.1016/j.patter.2024.101118>
- Thomas, Pius V. & Das, Bappa Ranjan. 2022. The Notion of Authenticity and Jean-Paul Sartre: A Conceptual Inquiry. *International Journal of Scientific Development and Research IJSDR* 7 11: 1178–1185. <https://ijsdr.org/papers/IJSDR2211175.pdf> (last accessed 23/03/26)
- Trilling, Lionel & Howe, Irving & Farber, Leslie H. & Hamilton, William & Orrill, Robert & Boyers, Robert. 1978. Sincerity and Authenticity: A symposium. *Salmagundi* 41: 87–110.
- Van Leeuwen, Theo. 2001. What is authenticity? *Discourse Studies* 34, 392–397. <https://doi.org/10.1515/9783110347012.1>

- Wang, Yiyin & Zhang, Tiancheng & Yao, Le & Seedhouse, Paul. 2025. A scoping review of empirical studies on generative artificial intelligence in language education. *Innovation in Language Learning and Teaching*, 1–28. <https://doi.org/10.1080/17501229.2025.2509759>
- Weissweiler, Leonie & Mahowald, Kyle & Goldberg, Adele. 2025. *Linguistic Generalizations are not Rules: Impacts on Evaluation of LMs*. In Bonial, Claire & Torgbi, Melissa & Weissweiler, Leonie & Blodgett, Austin & Beuls, Katrien & Van Eecke, Paul & Tayyar Madabushi, Harish (eds.), *Proceedings of the Second International Workshop on Construction Grammars and NLP*, 61–74. Düsseldorf, Germany: Association for Computational Linguistics. <https://doi.org/10.48550/arXiv.2502.13195>
- Werdiningsih, Indah & Marzuki & Indrawati, Inda & Rusdin, Diyenti & Ivone, Francisca Maria & Basthomi, Yazid & Zulfahreza. 2024. Revolutionizing EFL writing: Unveiling the strategic use of ChatGPT by Indonesian master's students. *Cogent Education* 11(1). <https://doi.org/10.1080/2331186X.2024.2399431>
- Xi, Zhiheng & Chen, Wenxiang & Guo, Xin & He, Wei & Ding, Yiwen & Hong, Boyang & Zhang, Ming et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68(2): 121101. <https://doi.org/10.48550/arXiv.2309.07864>
- Xu, Xiuming & Zhao, Wenjuan. 2025. Design of Intelligent Translation Optimization Algorithm Based on Large Language Model. In *Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (ICDSIC '25)*. Association for Computing Machinery, New York, NY, USA, 243–247. <https://doi.org/10.1145/3788910.3788946>
- Yang, Liu & Jiahuan, Cao & Liu, Chongyu & Ding, Kai & Jin, Lianwen. 2025. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review* 58(12):403. <https://doi.org/10.1007/s10462-025-11403-7>
- Yoo, Youngjin & Henfridsson, Ola & Kallinikos, Jannis & Gregory, Robert & Burtch, Gordon & Chatterjee, Sutirtha & Sarker, Suprateek. 2024. The next frontiers of digital innovation research. *Information Systems Research* 35(4): 1507–1523. <https://doi.org/10.1287/isre.2024.editorial.v35.n4>
- Zhao, Wayne Xin & Zhou, Kun & Li, Junyi & Tang, Tianyi & Wang, Xiaolei & Hou, Yupeng & Min, Yingqian et al. 2023. *A survey of large language models*. ArXiv preprint arXiv:2303.18223. <https://doi.org/10.48550/arXiv.2303.18223>

Theorising authenticity vis-à-vis, not versus, syntheticity in the age of AI: Introducing a continuum-based framework

Andrew Frank Bradley (Open University of Catalonia) & María del Mar Palomares Marín (University of Limerick)

afrank(at)uoc.edu, maria.palomares(at)ul.ie

Abstract

This article addresses contemporary debates surrounding authenticity in the age of Generative Artificial Intelligence (GenAI), where hybrid human-machine productions have given rise to a dual crisis of origin and reception. As GenAI-human co-authored content becomes more ubiquitous, the range of reactions it elicits – from dismissal to acceptance – highlights the need to reassess the relationship between the authentic and the synthetic within modern technologically-mediated and hybrid human-AI ecosystems. To this end, the article introduces the concept of syntheticity as both the degree of technological mediation in production (synthetic origin) and its social perception (synthetic attribution), which may coincide or diverge. Furthermore, the article advances a continuum-based model that reconceptualises authenticity and syntheticity as relational rather than oppositional constructs. This model positions authenticity along two axes: synthetic origin (human to artificial/generated) and reception (authenticated to unauthenticated), yielding four configurations of *Human authenticity*, *Human inauthenticity*, *Synthetic authenticity*, *Synthetic inauthenticity*. Intermediary zones account for hybrid productions (*Authenticated* and *Unauthenticated hybridity*) and contested validation states (*Liminal authentication*). By decoupling origin from reception in the evaluation of authenticity and syntheticity, the model positions hybridity as a legitimate locus of authentic expression. As a conceptual toolkit, the framework offers a new lens for analysing co-authored productions in which human and machine agency coexist and facilitates the critical evaluation of hybrid content.

Keywords

authenticity, syntheticity, artificial intelligence, human-AI hybrid productions, Generative Artificial Intelligence

1 Introduction

Generative Artificial Intelligence (GenAI) involves the use of complex algorithmic computation trained on pre-existing datasets to produce “new” content (Chan and Hu 2023). In practice, this allows users to input a variety of prompts and, in return, generate texts, images, sounds, videos, or a combination of these in mere seconds. Despite its relative infancy, GenAI has already had a profoundly transformative impact on shaping human experiences (Kudina and Verbeek 2019) in and across increasingly diverse contexts.

While Artificial Intelligence (AI) has existed in many forms for decades now (Thakur, Barker and Pathan 2024), the development of GenAI stands out as a revolutionary new phase in its evolution, particularly in terms of its accessibility,



Bradley, Andrew Frank & Palomares Marín, María del Mar. 2026.
Theorising authenticity vis-à-vis syntheticity.
Special Issue: *The notion of authenticity*. Vol. 4 No.1
DOI: 10.62408/ai-ling.v4i1.39

ease of use, and reach among technologically proficient and non-expert users alike. The public release of OpenAI's ChatGPT in November 2022 can be considered to be a pivotal moment that not only accelerated the mainstream adoption of AI, but also democratised GenAI to millions of users. Since then, the GenAI landscape has expanded substantially with the emergence of multiple competing systems and a dramatic increase in global user uptake, with current industry leaders estimating that 1 in 6 people worldwide use GenAI tools (Microsoft 2026).

In the new post-GenAI landscape, where users have the ability to effortlessly generate multimodal outputs on demand, the explosive growth of users is mirrored by a corresponding surge in AI-generated content. For instance, some provisional analyses suggest that at least 30% of text on active websites derives, in some way, from GenAI (Spennemann 2025), while other large-scale studies report that nearly three-quarters of newly created webpages now include AI-generated content (Law, Guan and Soulo 2025). Similarly, in the audio-visual sphere, providers such as YouTube have reported that up to 92% of content creators use GenAI to empower their creative process on its video platform (YouTube 2024). Beyond text and audiovisual media, AI-generated images have also become ubiquitous across a wide range of physical and digital domains (Rapp et al. 2025). Therefore, the exponential growth of AI-generated content, whether fully automated or hybridised through human-machine collaboration, suggests that such productions are likely to become even more pervasive, and, in some contexts, functionally unavoidable.

In terms of verisimilitude and fidelity, early GenAI artefacts exhibited notable limitations, especially in image and video generation, and were often criticised or delegitimised for their conspicuous markers of artificiality (Yang et al. 2022; Borji 2023), or “synthetic” qualities (Ghiurău and Popescu 2024). However, each successive model iteration is able to produce incrementally higher-quality outputs. This poses a significant epistemological challenge when coupled with the sheer volume of generated content now readily available, as contextual cues that once underpinned judgements of origin and credibility are no longer reliable indicators of authorship. To this end, Knott et al. note that “as the differences between AI-generated and human-generated content decrease, it becomes intrinsically harder to adjudicate individual cases” (2024: 62). Floridi (2024) similarly highlights how GenAI's impact on content creation calls into question conventional boundaries surrounding authorship, challenging established distinctions between human and machine contributions, as well as how such productions are understood and categorised.

In this context, authenticity emerges as a central, if contested, concept through which users or observers attempt to negotiate validity and legitimacy amid proliferating hybrid forms of co-production. These dynamics invite renewed reflection on what it means for something to be “authentic”; how “authorship” or attribution is determined; the authentication processes through which “authenticity” is assessed, and even what it means to be “human” itself (Beerends and Aydin

2025). Traditionally, authenticity has been theorised as a multidimensional and complex process of verification (Newman and Smith 2016) that is tied to humans and their creations (Dutton 2003). Although the expansion of digital and online communicative practices has complicated these associations (Mishan 2017), authenticity remains pervasive in contemporary culture (Martínez 2020) and is a central concern in daily life (Guignon 2004). Moreover, it continues to feature prominently in academic discussions, particularly in relation to digitalisation and the Internet (Mercier and Abidin 2025).

However, little attention has been paid to how AI-generated and hybrid productions intersect with non-naturalist and post-GenAI formulations of authenticity, particularly at the level of conceptual modelling. In fact, there is a tendency in both popular and academic discourses for the “authentic” to be coupled with human-origin productions versus the “inauthentic” with AI-generated or hybrid creations (cf. discussions in Bui, Filimonau and Sezerel 2024; Beerends and Aydin 2025; Erdocia, Schneider and Migge 2025; Wu and Lingel 2025). In this paper, we argue that such binary categorisations fail to capture the complexity of AI-mediated content creation. In response, we propose a conceptual, continuum-based model for theorising authenticity in the era of human-machine collaboration, with a specific interest in GenAI co-productions and ecologies. The framework advances two central claims: first, that the authentic should be understood in relation to the synthetic, as opposed to its inherent counterpart; and by extension, that hybrid human-machine (including GenAI) productions can, under certain conditions, be legitimately experienced and evaluated as authentic. By doing so, this paper aims to contribute to ongoing debates concerning authenticity across different disciplines (e.g., Newman and Smith 2016; Pinner 2016; Kalpokas 2024; Lavazza 2025; Lee 2025, among many others) and serve as a tool for assisting in the navigation of current and future AI-human experiences and landscapes.

This article functions as a conceptual and theoretical synthesis that draws on illustrative empirical examples to develop a theory-building framework. The paper is structured as follows. Section 2 offers a conceptual synthesis of existing frameworks of authenticity across different fields, drawing in particular on Newman and Smith’s (2016) seminal review of the literature, and extends their analysis to highlight definitional ambiguities and theoretical tensions that are sharpened in the age of GenAI and hybrid human-machine production. The empirical studies cited throughout are drawn selectively from existing scholarship and function as illustrative cases, rather than as an exhaustive review of specific domains. Section 3 examines how authenticity is renegotiated in contemporary mediated communication and interaction, introducing two interrelated crises of post-GenAI authenticity: a crisis of origin, in which provenance becomes increasingly indeterminate, and a crisis of perception, in which authenticity judgements are shaped by attribution and reception rather than by content alone. This section situates these crises within communication studies, AI-mediated interaction, and human-AI collaboration research. Section 4 introduces the

proposed theory-building framework and advances a structured conceptual model for analysing how authenticity and syntheticity are validated, negotiated, and contested across contexts of human-machine and GenAI collaboration. Section 5 concludes the article by discussing the framework's applications, limitations, and implications for future research.

2 Authenticity: a brief overview

Authenticity is a multidimensional concept with no static or singular meaning, whose interpretation(s) evolve over time, vary across disciplines, and circulate in public discourses to serve different social and evaluative aims. Traditionally, authenticity has been associated with qualities such as truthfulness (Williams 2002), originality (Bialystok 2014), sincerity (Trilling 1972), nativeness (Woolard 2016), and (social) virtue (Guignon 2008), among many others. While authenticity has frequently been theorised in relation to human subjects and lived experiences, it has also been extended to non-human referents such as institutions, objects and actions (Bialystok 2014), and, in more recent decades, to digital artefacts and environments (Mishan 2017; Mercier and Abidin 2025). This breadth of application positions authenticity as a polysemous concept that resists a universally applicable framework. At the same time, however, such conceptual flexibility does not imply complete indeterminacy but instead draws attention to whether authenticity has any stable points of reference.

Addressing this, in their seminal examination of existing research, Newman and Smith (2016) successfully outline four areas of convergence across multiple different definitional typologies of authenticity. The authors find that discussions about or theorisations of authenticity tend to gravitate towards one of four distinct categories: historical authenticity, categorical authenticity, value authenticity, and self-authenticity. Rather than proposing a universal definition, their framework offers an analytic synthesis that identifies recurring themes while documenting how authenticity is evaluated across different disciplines. In what follows, we discuss and expand upon these in turn.

2.1 Historical authenticity

In relation to historical authenticity, Newman and Smith describe this as involving “the evaluation of an object’s unique spatiotemporal history” (2016: 612). Through an external process of validation (i.e. by an expert), an original item (e.g. the Mona Lisa) is deemed either legitimate or not, and thus historically (in)authentic. Historical authenticity appears under different labels and closely corresponds to several related constructs, such as indexical authenticity in marketing (Grayson and Martinec 2004), nominal authenticity in aesthetics (Dutton 2003), objective authenticity in tourism (Wang 1999) and pure authenticity in advertising (Beverland, Lindgreen and Vink 2008). Despite their disciplinary specificity, these

constructs converge around a shared emphasis on provenance and historical linkage. For instance, Grayson and Martinec (2004) invoke indexicality to describe the traceable link between an object and a specific time and place in the past, through which evaluations of authenticity are formed in relation to consumer products. In the context of artistic production, Dutton (2003) similarly refers to an item's nominal authenticity as the genuine identity of an item, distinguishing it from plagiarism and forgery. Likewise, Wang (1999) emphasises objective authenticity through the uniqueness of objects with a verifiable history, or those considered one of a kind. Finally, pure authenticity is proposed by Beverland, Lindgreen and Vink (2008) to explain the verifiable link between traditional practices and historical cues that confirm if an item has remained unchanged over time and remains true to its origin. Across fields, and within this conceptual lens, historical cues constitute an integral part of the authentication process, which is typically both binary and rooted in origin.

Moreover, Groth, Block and Newman (2023) affirm that authenticity judgements about objects are contingent upon their perceived connection to a person, time, or place. This, in turn, highlights three crucial aspects of an item's origin: essence transfer, spatiotemporal connection, and belief. First, the connection between essence and authenticity stems from the perception that an object can "absorb" an immaterial essence through contact with people, especially the original author or through processes of contagion (Newman and Bloom 2012: 558), thereby enhancing its perceived authenticity. Second, the authenticity of an item is anchored in a specific time and/or place, such that a "real" spatiotemporal connection is understood to exist between an object, its author, and its time or place of creation (Lehman et al. 2019). Third, authenticity judgements may also involve beliefs about whether, and to what degree, a work embodies the essence of its author and its historical context. This is particularly widespread in literary domains where authenticity is typically grounded in the conviction that a work is genuinely (i.e. authentically) the product of a specific author (Martínez 2020).

Historical authenticity can also be subject to degrees of evaluation, particularly in contexts involving mass-produced goods, where distinctions between first editions or signed copies acquire heightened value through their perceived proximity to an author or moment in time (Newman 2019). As Grayson and Martinec (2004: 299) argue, perceptions of indexicality (i.e. what distinguishes the real and valued object from its copies) may vary according to individual interpretations of essence. For example, a book signed by an author may be viewed as less indexical compared to a book personally handed by the writer, which in turn may be experienced as carrying a stronger trace of the author's essence. In contrast, in cases involving unique artworks such as the Mona Lisa, judgements of authenticity and value extend beyond individual evaluation or historical cues, and instead rely on accepted forms of nominal authentication.

2.2 Categorical authenticity

While historical authenticity is frequently conceptualised through provenance, scholars have also highlighted the role of interpretative variation in shaping (historical) authenticity evaluations. In such cases, subjectivity arises from individually formed perceptions of an object's connection to a specific person, time, or place (cf. the notion of *essence transfer* discussed above). This form of subjectivity, however, continues to rest on the assumption of a verifiable historical link and therefore cannot account for other kinds of authenticity where such a link is fragile, secondary, or nonessential. Under these conditions, authenticity is better understood as categorical (Newman and Smith 2016) and grounded in a subjective, socially constructed epistemological framework (Wang 1999; Carroll and Kovács 2021), where judgements are not necessarily based on an authenticated historical connection but instead can fluctuate across time, space, or community. As Newman and Smith (2016) remark, these various approaches converge in their emphasis on the degree to which an entity is perceived to be true to its associated social type, genre, or category. This category-based authenticity relies on perceived typicality (Busselle and Bilandzic 2008) or the cognitive process of internal validation that occurs between the observer's prior mental expectations and the item or concept under observation itself. As noted by Lee (2020), when there are deviations from this expectancy, claims of inauthenticity often arise. In other words, categorical authenticity concerns the construction of the "authentic" according to the pre-conceived expectations of an observer (or observers) about how something is, or ought to be, and the extent to which something is perceived to be real and true.

Carroll (2015; cf. Carroll and Kovács 2021) similarly refers to this as type authenticity, whereby an attribution of authenticity indicates that a person or other entity fits appropriately into the social category to which it has been assigned or that it has claimed for itself (Lehman, O'Connor and Carroll 2019: 21). In a related formulation, Grayson and Martinec (2004) draw on social semiotics in their exploration of iconic authenticity. This describes evaluations of authenticity that are not inherent to an object (cf. indexical relationships above) but are instead conferred through an evaluator's contextual and personal judgement, even when an entity is not itself materially "real". Other theorists employ different nomenclature to reflect similarly category-based interpretations, such as *expressive authenticity* (Dutton 2003), *constructed authenticity* (Wang 1999), and *approximate authenticity* (Beverland, Lindgreen and Vink 2008).

Categorical authenticity is therefore central to many aspects of contemporary human socialisation, such as how cultural products are consumed, how identities are performed, as well as how languages are evaluated and used. As Carroll and Kovács state, the external negotiation of socially constructed authenticity relies significantly on moments when "someone verbalizes something that invokes the concept of authenticity or perhaps uses the word 'authentic' or any host of closely associated words" (2021: 2). In this sense, categorical authenticity

becomes observable not only through implicit judgements and evaluation, but also through explicit (metalinguistic) labelling and everyday interactions.

In the case of languages, this is evident with the kinds of categories language users often ascribe to certain linguistic varieties or speakers. One ubiquitous language ideological belief documented by sociolinguists, language educators, and linguistic anthropologists concerns the construction of the “native speaker”. This category of speaker is often idealised as a model language user on account of having acquired a language “naturally” through intergenerational transmission, instead of learning it deliberately at school or later in life (Pinner 2014; Llorca and Mocanu 2024). This analytic conflation of authenticity and naturalism (Woolard 2016) extends to the construction of “authentic speakers” (cf. Eckert 2003; Coupland 2003), who are categorised as local(ised) actors rooted in a particular place, who produce language(s) that belong to that specific location or culture. Underpinning such categories of both “native” and “authentic” speakers is a naturalist assumption that the natural has a positive valence and is therefore valued or valuable, while that which is inauthentic (i.e. the mechanical and artificial) is spurious and suspect (Coupland 2014; Woolard 2016: 32). Within this ideological framing of linguistic authenticity, so-called “authentic language” cannot easily be appropriated by outsiders as it is seen to belong to an exclusive in-group. Delegitimising social labels such as “non-native speaker” or “new speaker” (O’Rourke and Ramallo 2013) are typically mobilised in instances of perceived “inauthentic” language or other illegitimate linguistic practices such as having an L2 accent. Since these judgements are contextual interpretations of the world (Mishan 2017), categorical authenticity is necessarily multiple, negotiable, and layered.

2.3 Value and Self authenticity

The next two categories that Newman and Smith (2016) identify in their convergence analysis involve forms of authenticity that are highly agentive and rooted in deeply personal, subjective assessments, making them less amenable to fixed or externally verifiable criteria. The first is value authenticity, for which “observers can place varying degrees of importance on the extent to which one ought to conform to norms or moral values” (*ibid*: 613), and the second is self-authenticity, which relates to one’s own individual sense of self. The question of whether attending Midnight Mass on Christmas Eve constitutes an authentic expression of (certain) Christian faith(s) would be an example of the former, while acting in accordance with, or in opposition to, one’s true, or authentic, self (however this may be understood) would exemplify the latter.

The role of observers is key for value authenticity as they proceed to assess the alignment between a person’s actions (or an object or experience) and the expression of an “implicitly or explicitly claimed set of values” (Newman 2019: 10). On the other hand, self-authenticity emphasises a complex, and difficult to define ideation of the self. Popular discourse on authenticity often entails “being

true to what someone (or something) truly is” (Guignon 2008: 277). As Taylor (1991) observes, within the Western tradition, the Romantic era introduced an expressivist understanding of the self that paved the way for the “age of authenticity” of the 1960s. Since then, being true to oneself has revolved around the principle of originality, that is, the idea of having something inner, true, and unique, such as an inner voice or a particular way of being human. This understanding is commonly positioned in contrast to imitation or conformity with models imposed by religious, political, societal, or intergenerational pressures (McEvoy 2009).

Other theorists, including Taylor (1991) himself, have challenged this understanding of the self, particularly the notion of a stable inner-self or core. Instead, Taylor (*ibid*) suggests that authenticity is dialogic, while for Guignon (2004) it constitutes a social virtue. More recently, discourses of self-realisation and personal fulfilment have revolved around the idea of expressing one’s own inner (true) self (Guignon 2004), especially in online settings (Mercier and Abidin 2025), where both value and self-authenticity have become important for mediated authenticity. In this context, Bailey et al. (2020) posit that authenticity is the unobstructed expression of one’s self, which is measured by the degree of congruence between an individual’s sense of self and their observable expressions of self, and which also must remain consistent across online-offline spaces and different social media or internet platforms.

2.4 Definitional challenges

As Newman and Smith’s (2016) convergence analysis demonstrates, authenticity operates through multiple and context-dependent evaluative processes. Therefore, outlining the nuances of this typological framework is challenging, and as the authors remark, “there are many instances that blur the lines between these distinctions” (2016: 614). This sentiment is echoed by others, such as Dammann et al. (2021) who acknowledge the disadvantageous semantic heterogeneity of the term, or Bialystok who doubts “the possibility of generating a complete philosophical account of authenticity” (2014: 1). However, convergent analyses of authenticity research indicate that authenticity functions primarily as a process of verification across its different manifestations (Newman and Smith 2016: 613-614, cf. Lee 2020; Beerends and Aydin 2021). Newman and Smith (*ibid*) find that the authentication necessary for authenticity judgements is organised along two dimensions: the type of entity that is evaluated (i.e. object vs. agent) and the source of information that is consulted (i.e. external vs. internal). Assessments of authenticity may therefore rely on either objective or subjective criteria, or on combinations of both, at varying degrees in different contexts. Authenticity is thus not fixed nor absolute, but fluid and multifaceted. Furthermore, authenticity is ubiquitous and highly influential, as it informs not only how people interpret the world but also how they behave in different social or cultural settings.

Notwithstanding the plurality of ways in which authenticity can be expressed, interpreted, or reified, we agree with Carroll and Kovács (2021) and others (Pinner 2014; Lee 2020) that authenticity is a positive attribution at its core. For instance, consumers prefer authentic products (Lehman, O'Connor and Carroll 2019), while language learners and teachers find that authentic materials and assessments are better for enhancing critical thinking, problem-solving, and collaboration skills (Vlachopoulos and Makri 2024). However, if authenticity in its various conceptualisations is understood as a markedly positive experience or interpretation of the world, then its opposite is, explicitly or by implication, subtractive, inferior, less genuine, untrue, or false in some way (Wilt, Thomas and McAdams 2019; see Michael and Shuttleworth 2021 for an example in the context of ethics).

This evaluative asymmetry becomes particularly salient in cases where authenticity is associated with behaviours that challenge moral or societal ideas and standards. Several authors have expressed concerns about the different interpretations of what it means to be authentic. Taylor (1991) describes the “culture of authenticity” as an expression of soft relativism and moral individualism, driven by the goal of self-fulfilment reflecting a self-absorbed and narcissist view of personal realisation. Something that Taylor denounced as part of the “most degraded, absurd or trivialized forms” [of the culture of authenticity] (1991: 29). While authenticity is critiqued here in relation to how humans try to be authentic, this interpretation closely aligns with Newman and Smith’s (2016) concept of self-authenticity. For Guignon this ideal “makes a very heavy demand on you” (2004: 76), criticising the common assumption that self-fulfilment is achieved only through the expression of one’s true self, while also challenging the widely held belief that one’s goal in life is to become authentic without further discussion or questioning.

In digital and social media contexts, these pressures are further intensified. Haimson et al. describe what they term the *online authenticity paradox*, whereby efforts to achieve online authenticity are “often unreachable or is possible at great personal cost, especially for those with marginalized identities and difficult life experiences” (2021: 1). Feelings of inauthenticity appear when one fails to gain social and virtual acceptance. Users are often penalised for showing excessive authenticity (e.g. “trying too hard”) or too little authenticity (i.e. not showing negative aspects of their lives), which can lead to emotional distress, reduced self-esteem, social censure and stigma (Davis 2019). Additionally, Silver, Newman and Small (2020) draw attention to the volatile nature of authenticity, showing how even a small detail may change consumers’ perceptions toward authenticity judgements, with significant consequences for individuals and brands alike. In this sense, authenticity is contingent upon achieving specific norms or expectations (cf. categorical and value authenticity), with deviations from these going beyond the “mere removal of the positive characteristics associated with authenticity” (*ibid*: 71). Instead, it may trigger an “inauthenticity aversion”, public disapproval and moral condemnation (*ibid*: 71). Although these dynamics are not explicitly framed

as a crisis of authenticity, the simultaneous elevation of authenticity as idealised yet unattainable anticipates a deep tension for the times ahead that merits critical reflection. Significantly, an emerging actor, GenAI, brings in new considerations and dimensions to the authenticity debate.

Against this backdrop, each of the four domains of authenticity in Newman and Smith's (2016) typological framework faces renewed strain and conceptual challenges in the GenAI era. *Historical authenticity*, which traditionally rests on provenance and indexical links to a fixed origin, is destabilised when AI-generated artefacts such as images, music, or text cannot be distinguished from human creations, thus undermining the spatio-temporal verification of authorship. *Categorical authenticity*, contingent upon social constructions of what counts as "real" or "true" by observers, is equally challenged when audiences radically diverge in their responses to AI outputs, some rejecting them outright as inauthentic, while others legitimising them as authentic within new cultural categories (e.g. "AI art"). *Value authenticity*, which relies on the alignment between action and moral norms, is increasingly invoked in debates about whether the use of AI in domains such as education, journalism, or creative industries is an ethically authentic practice or a betrayal of professional standards and social values. Finally, *self-authenticity*, understood as fidelity to one's true self, is problematised in contexts where digital self-presentation is mediated, or even co-authored by AI systems, raising questions about whether hybrid or assisted forms of expression can still be regarded as authentic.

These emerging pressures illustrate how GenAI does not merely complicate existing debates about authenticity but fundamentally influences the conditions under which authenticity has traditionally existed. In particular, they expose the limits of frameworks that treat authenticity as a stable property of origin, authorship, or human embodiment, and point instead to the need for a reconceptualization of how authenticity is negotiated in contemporary human-machine, and specifically human-AI, ecologies. While Newman and Smith's (2016) convergence analysis of authenticity's plural evaluative processes provides an important foundation, a framework capable of integrating technological mediation and synthetic production into this typology is necessary. The following section examines how GenAI has intensified these challenges through dual crises of origin and perception, before advancing such a framework in Section 4.

3 Exploring authenticity in the age of mediated communication, interaction, and synthetic media

Building on these theoretical accounts of authenticity, this section shifts from definitional concerns to the conditions under which authenticity is negotiated and contested in contemporary mediated communication. In particular, it examines how the rise of agentic, generative, and hybrid human-AI systems reshapes assumptions

about agency, authorship, and source, from which two crises of post-GenAI authenticity emerge: one of origin and one of perception.

Further tensions surrounding traditional notions of authenticity emerge within communication and media studies. As interactions between humans and AI systems become more commonplace, long-standing assumptions about communication, agency, and authorship are placed under strain. Contemporary human-AI interaction threatens traditional assumptions or frameworks of authenticity when machines begin to perform in roles traditionally reserved for humans (Guzman, McEwen and Jones 2023). As a result, technology is no longer seen as a medium, but rather as an agentic machine (i.e. source of communication) capable of autonomously constructing meaning rather than just interacting (Sundar and Chen 2023) or mediating communication.

In early theorisations of computer-mediated communication (CMC) when technology served mainly as a medium for transmitting information, authenticity was tied to human senders and receivers, and message integrity itself, that is, whether the message is conveyed as originally intended. At that time, mediated exchanges were perceived as having lower social presence than face-to-face communications (Hohenstein and Jung 2020). Therefore, in early CMC, agency and authenticity remained closely associated with human actors. However, communication with interactive generative-AI chatbots and emerging technologies in the modern era has blurred these assumptions, shifting attention from who is human (or not) to what happens between interlocutors (Dehnert and Mongeau 2022). This highlights the need to rethink communication in the era of AI (Brandtzaeg et al. 2023). To this end, the Human-AI Communication (HA-C) model (Guzman 2018), the Human-AI Interaction Outcomes (HAI-IO) model (Quinlantang 2025), and the HAI-TIME model (Sundar 2020) have been proposed to explain how AI is not merely a communicative mediator or enhancer, but is, or can be, an active partner in communication. These approaches extend agency beyond a human-centric perspective and acknowledge a more symbiotic human-AI relationship, which entails, by extension, a more complex and interactive communicative dynamic (Revolusi and Febriandy 2025).

Alongside these discussions, recent post-humanist and more-than-human perspectives advocate for reframing human-machine relations within broader ecologies of practice, where humans are no longer at the centre but one actor among many (Giaccardi, Redström and Nicenboim 2024). From this standpoint, posthuman subjects would be decentered in a world where humans, human-made and nonhuman conscious machines are interconnected and integrated (Nicenboim et al. 2023; Lackey and Papacharissi 2024). However, emerging concerns appear when AI becomes a source in the communication process with the possibility of machine agency undermining human autonomy or agency (Sundar and Chen 2023). As the media-equation literature notes, humans tend to interact with media and technology in the same way that they do with humans (Reeves and Nass 1996). Similarly, frameworks such as *Computers-Are-Social-Actors* (CASA) (Nass,

Steuer and Tauber 1994; Nass and Moon 2000) defend the position that humans mindlessly treat machines as if they were social entities with gender, personalities, feelings or intentions (Lee 2024) even when users were consciously aware that machines are different from humans. The *Social Presence Theory* (Short, Williams and Christie 1976) reinforces this account by suggesting that multimodal media can convey feelings of proximity for their use of visual and verbal cues (Calefato and Lanubile 2010). Furthermore, AI-based media often rely on anthropomorphic cues, thereby asserting agency alongside humans (Sundar and Chen 2023). These may also foster feelings of homophily among users and increase perceived socialness in interactions (Sundar 2020). Along these lines, recent research suggests that GenAI chatbots produce human-like conversational cues (e.g. mimicking politeness), thus enhancing the realism of interaction and users' satisfaction (Kim and Lee 2025). They can also foster conditions in which users develop para-social relationships (Marriott and Pitardi 2023) and evoke rich and deep emotional experiences (Li and Zhang 2024) with a high social presence.

In this regard, phenomena such as the *uncanny valley effect* – a concept originally proposed in the 1970s by Mori to describe a shift from empathy toward increasingly humanlike machines to feelings of unease or revulsion as they approach, but fail to attain, lifelike appearance (Mori, MacDorman and Nageki 2012) – raise further concerns about how human-likeness in new AI systems may disrupt social perception and interaction. Studies on technological affordances suggest that the novel affordances of AI can trigger cognitive heuristics by shaping users' psychological responses, and foster actions that shape how they process and respond to media content (Sundar and Chen 2023). A more critical approach comes from Lee (2024), who argues that it is not yet fully understood how prior beliefs about machines may bias users' positive or negative reactions to AI systems. As these become more agentic, generative, and context-aware, these traditional frameworks may require refinement or at least invite further questioning in the exploration of authenticity, reinforcing what Kalpokas et al. describe as an anthropocentric logic of automated policing that positions GenAI as a problem to be governed rather than a mode of legitimate production (Kalpokas, Kalpokienė and Šalaševičiūtė 2024).

Taken together, these dynamics suggest that authenticity in AI-mediated interactions emerges from complex negotiations among agency, source awareness, perception and belief pointing to a broader crisis in how authenticity is attributed and perceived. Across these developments, one recurring feature becomes especially salient: the increasing prevalence of hybrid configurations of agency, in which human and machine contributions are co-constituted rather than separable within the same communicative act. It is this condition of hybridity, and not GenAI per se, that problematises traditional assumptions about authorship and authenticity, and which therefore warrants focused attention.

3.1 Hybridity: human-AI collaboration

While authenticity in AI-mediated interaction has been shown to depend on negotiations among agency, source awareness, perception, and belief, the growing role of GenAI in content production also shifts these negotiations toward the generated output itself. Hybridity, agency, and authorship become central in this discussion.

The notion that AI can augment, amplify or automate human intelligence dates to AI's early developments (Sundar 2020), yet contemporary systems, including GenAI, enable forms of human-AI synergy that were previously (technically) unattainable. In this context, tensions between machine agency and human agency emerge as users are hesitant to cede control over decision-making processes, system behaviour (Mieczkowski 2022) or authorship. Boundaries between human and machine agency often blur in practices of co-creation and collaboration resulting in hybridity understood as distributed and co-creative agency, whereby humans (consciously or not) accept the transfer of varying levels of agency to AI systems, and both participate in the creative process (Rammert 2008; Sundar 2020).

GenAI has disrupted not only the long-standing naturalist assumption that there is an exclusively human creative process (Rodríguez Valdés 2025), but also the binary opposition between augmentation and automation. It does so by exceeding the distinction between AI as human extension and AI as machine substitution, instead operating through hybrid configurations of distributed creative agency (Zafar, Ali and Yasin 2025). An example of such hybridity is synthetic media. More broadly, synthetic media can be understood, as Whittaker et al. (2020) propose, as any media (text, images, audio, video, etc.) that has been artificially generated or manipulated, or as digital content that has been “primarily or exclusively produced using AI-enabled tools” (Kalpokas, Šalaševičiūtė and Lipské 2024: 3). Although synthetic media is commonly associated with deception, particularly in the context of political disinformation (Hameleers, van der Meer and Dobber 2024), it can also serve harmless creative endeavours, such as artistic experimentation, education, or benign entertainment. Distinct from other forms of media, what is central to synthetic media is the partial or full integration of artificial intelligence technologies specifically. At the same time, it is not solely a technological process or attribute. While some synthetic outputs are produced almost entirely by computational systems, many involve shared human and technological agency (Sundar and Chen 2023). Kalpokas, Šalaševičiūtė and Lipské (2024) in their exploration of the term note that humans are still involved in synthetic content generation by initiating the creative process, engaging in prompt engineering, as well as evaluating (and modifying) the result. It is for this reason that Tron Gianet, Di Caro and Rapp (2024) contend that views of GenAI are shifting, with the tool-collaborator divide becoming less clear as users are increasingly shifting their roles toward curation, selection, and co-production.

Hybridity in human-AI systems exposes the distributed nature of agency, however this redistribution across human and non-human actors does not automatically redefine authenticity nor move synthetic creations away from inauthenticity. Instead, it crystallises into a more fundamental problem: the increasing difficulty of determining where a work comes from at all, inaugurating what many scholars have identified as a crisis of origin.

3.2 The first crisis of post-GenAI authenticity: origin

Amid this AI driven ecosystem, what is novel about GenAI outputs in the present is their scale, accessibility, cost (Rae 2024) and creative plausibility. On a fundamental level, it puts into question what it means for something to be “authentic” in the age of GenAI. On this point, scholars have declared a new crisis of authenticity (Jacobs 2024; Lee 2025) upon reflecting on the destabilising impact caused by different social actors not being able to determine and discern the original source of a message (whether human or machine), as GenAI creations can be comparable to and even surpass human content (Pedersen and Ritter 2024; Baptista et al. 2025; Gherheş et al. 2025) and are often indistinguishable, for example, in the context of images (Giridhar et al. 2025). Another significant area of concern regarding synthetic media over the past decade has been the ability to create deepfakes, which are synthetically generated or manipulated images, audio, or videos that depict people saying or doing things that they never did (Lee 2025: 86). This has prompted concerns about its misuse and harm to society (Masood et al. 2022) in that some deepfakes can pass as “authentic” under user scrutiny (Köbis, Doležalová and Soraperra 2021). Furthermore, the verisimilitude of deepfake outputs is only likely to get more sophisticated (and therefore harder to authenticate as human-created) with time. AI-generated voice cloning undergoes a similar path of rapid development and technical sophistication. Barrington, Cooper and Farid (2025) studied whether 600 participants could distinguish the identity (human or AI-generated) of different audio clips. They concluded that audiences can be easily tricked into thinking that the AI-generated clones were indeed real human speakers, since some could not reliably distinguish them from natural human speech.

Against the backdrop of this mediatic context, we agree with Lee in that “instances of authenticity scandals, authenticity illusions, and authenticity puzzles are endless and continue to grow” (2025: 26). This highlights how the assumption that authenticity rests on human origin is becoming increasingly unstable, as synthetic/artificial outputs can no longer be easily dismissed as inauthentic based on provenance alone. As a result, the crisis of origin gives way to a separate yet concurrent crisis - one of perception - where questions of authenticity rely less on source, but on how audiences evaluate and respond to different configurations of human-machine hybrid productions.

3.3 The second crisis of post-GenAI authenticity: perception

A parallel crisis concerns the shifting perceptions of authenticity in response to perceived origin, as human-AI collaborations and synthetic media continue to polarise different audiences. Lee (2024) argues that it is not yet fully understood how prior beliefs about machines may bias users' positive or negative reactions to AI systems. Once users identify the source, their beliefs about machines can shape how that source influences their subsequent reactions, challenging the CASA assumption that human-machine interaction is largely mindless.

At one end of this extreme, advocates of GenAI (e.g. AI Maximalists) champion the inclusion of generative and algorithmic technologies in all facets of life. This view is underpinned, in part, by what Kim and Koo define as machine heuristics, or "viewing AI as a neutral and unbiased entity that enhances its credibility and contributes to the belief that AI-generated content offers a more objective basis for evaluating materials" (2024: 3). Similarly, Beerends and Aydin (2025) explore the concept of inauthenticity-as-ideal for which the value of (Gen)AI lies precisely in its difference from human production, and for which artificiality is valued as its own distinct aesthetic or ethical value in its own right. On the other extreme, skeptics or critics express negative AI bias (Wasdahl 2024) or AI aversion (Qin et al. 2025). Such responses have produced derogatory or deauthenticating labels for both tools and outputs, such as "clanker" for AI systems (Romo 2025) or "AI slop" for AI-generated artefacts (Roddy and Bridges 2025), and in some cases have led to institutional prohibitions on GenAI technologies, for instance in schools and universities seeking to safeguard academic integrity (Fine Licht 2025).

Amid this polarisation, recent empirical studies have shown a proliferation of divergent perspectives in which authenticity judgements vary according to perceived source rather than the output or quality thereof. Kirk and Givi (2024) note that consumers reduced positive word of mouth and customer loyalty when they believed emotional marketing communications were written by AI versus a human, an effect that was mediated by moral disgust prompted by perceived inauthenticity. Similarly, other studies have shown how people's behaviour and attitudes shift, sometimes drastically, according to the degree of transparency of GenAI disclosure in digital content creation (Brüns and Meißner 2024; Schilke and Reimann 2025). For example, Altay and Gilardi's (2024) experimental study of AI-generated news headlines found that labelling content as AI-generated reduced perceived accuracy and willingness to share, even when the headlines were factually correct. Comparable sentiments are shared by Lermann Henestrosa and Kimmerle (2024), who discovered that labelling content as AI-written lowered source and message credibility. Strikingly, the negative perception diminished when AI involvement was described as partial (e.g. "drafted" or "edited" by AI) or AI was adopted to assist rather than replace humans (Brüns and Meißner 2024), implying that audiences intuitively applied a flexible model of automation, even if authenticity was not explicitly under discussion. Further evidence comes from education and

media contexts where GenAI is used in the creative process. In these cases, a shift in perception towards writers, journalists, or content creators was found. Frederick et al. (2025) discovered in their study that adult learners preferred human or a combination of human-AI teaching materials over exclusively AI-generated ones, with full AI automation harming instructors' perceived credibility and competence. Similarly, Rae (2024) reported that when content creators were discovered to have used GenAI, they suffered reputational consequences, even though the quality and reception of their written content remained unaffected. These examples highlight the crisis of perception at the heart of post-GenAI authenticity, where evaluations shift not with content quality, but source disclosure and attribution.

In this context, while earlier signs in the technological landscape pointed to a gradual erosion of authenticity in its naturalist assumption, the modern landscape introduces new dynamics. First, synthetic media have become increasingly sophisticated and human-like, and users are changing the way they interact with and perceive such content. Second, the very meaning of authenticity in relation to emergent synthetic and artificial products remains underexplored. Recent studies, such as those above, show that perceptions of authenticity fluctuate in response to perceived source, level of AI automation, and framing of AI involvement rather than to output quality or origin alone. Faced with this instability of provenance, many audiences respond not by rethinking authenticity, but by reasserting a simplified opposition between human authenticity and synthetic inauthenticity, particularly in relation to human-AI co-productions.

3.4 Human authenticity versus GenAI inauthenticity

In contemporary discussions of synthetic content, as well as hybrid human-AI authorship, two persistent patterns emerge in the literature. First, authenticity is often asserted but rarely defined (Huang et al. 2025; Lee 2025). Second, there is a tendency to treat origin and reception of authenticity as tightly coupled, consequently framing the authentic as human-origin, and the inauthentic as synthetically or artificially generated in a dichotomous relationship.

For instance, in early adoptions of GenAI (model GPT-2), co-creative experiences, such as the human-AI co-writing tool studied by Yang et al. (2022), revealed a clear divide between human authenticity and AI inauthenticity. The low-quality outputs of GenAI at that time, described as “incoherent and tenuous text” (*ibid*: 7), encouraged early perceptions of such outputs as pejoratively “synthetic” rather than a constructive product of human-AI hybrid collaboration, as they were seen to be useful for no more than inspiration. More recent research by Ju, Kim and Im (2024) on virtual influencers shows that this dynamic persists beyond writing. In their study, users engaged more with hyper-realistic, non-brand operated virtual influencers precisely because they exhibited greater perceived humanlike authenticity. Similarly, Jia et al. (2024) find that in AI-human collaborative news

production, the disclosure of hybrid roles such as AI assistant or AI collaborator reduced perceived humanness and favourability.

More explicit examples can be found in the domains of sociolinguistics and language education. Traditional language authorities, such as the Spanish Real Academia Española (RAE), claim that language generated by machine-learning technologies is inauthentic and destroys the authentic Spanish of human Spanish speakers (Erdocia, Schneider and Migge 2025), while participants in Werdiningsih, Marzuki and Rusdin's (2024) study of using ChatGPT in writing-focused EFL classes reported that they needed their own human input for their writing to be considered authentic or to preserve an authentic voice in the writing process. Moreover, Allaithy and Zaki (2025) in the context of Arabic language teaching suggest that human insight is necessary before AI-generated content reaches students, and that teachers should act as "linguistic gatekeepers" to ensure quality control. Elsewhere, the divide becomes more absolute. For instance, Draxler et al. (2024) describe cases where AI tools are treated as purely passive instruments rather than co-authors, with full authorial attribution and ownership granted to the human user. In some legal contexts, such as U.S copyright law, AI-generated work cannot (at the time of writing) be copyrighted since in that legal framework authorship necessitates human origin and creation (Caldwell 2023).

The above examples reflect a consistent reinforcement of a normalised division between AI and human authorship, where technological or algorithmic mediation is assumed to compromise authenticity rather than contribute to it. As discussed in Section 2, this assumption is underpinned by a naturalist and hierarchical anthropocentric belief (Kalpokas 2023; Kalpokienė and Kalpokas 2023; Kalpokas 2024) that legitimises human productions as superior or primary while positioning non-human or technologically-mediated creations as, in some way, secondary or inferior. This dichotomisation leaves little conceptual room for categories such as human inauthenticity or synthetic authenticity, nor for the many hybrid configurations of human-machine collaboration that may arise and which may or may not ultimately be validated as authentic. To address this gap, in the subsequent section we reconceptualise the concept of authenticity and propose a continuum-based framework.

4 Rethinking authenticity's relationship with syntheticity: introducing a continuum-based framework

By reframing authenticity's conceptual relationship with technologically-mediated productions and human-machine collaborations, we align with recent work by Kalpokienė and Kalpokas (2023) and others (e.g. Köbis, Doležalová and Soraperra 2021; Lavazza 2025), who argue that synthetic outputs can, under certain conditions, be experienced and evaluated as authentic, and their call for embracing a post-humanist model wherein creativity can be shared between humans and machines instead of being limited to human-centred binaries. This resonates with

Beerends and Aydin’s (2021; 2025) interactionist authenticity approach for which an AI artefact does not need to display human-like characteristics to be experienced as authentic, as it may be authenticated or deauthenticated through negotiating processes of language, socialisation, and cognition (cf. categorical authenticity in Section 2). These state-of-the-art contributions underscore the need for fluid and future-proof conceptual frameworks that move beyond binary divisions and instead account for the complex and shifting relationships between origin, mediation, and audience reception regarding authenticity. Responding to this need, in the next sections we propose a continuum-based framework that reconceptualises authenticity as *vis-à-vis*, not versus, syntheticity. Here, the “authentic” and the “synthetic” are not positioned as mutually exclusive categories, but as poles on a spectrum with the wide range of hybrid human-AI co-productions and shifting audience evaluations occupying the many gradations in between.

4.1 Defining syntheticity in relation to authenticity

Compared to authenticity, the *synthetic* has generally posed fewer definitional challenges across disciplinary contexts. While its specific applications vary, this concept has generally been anchored in questions of construction and origin, rather than evaluation or moral judgements. For instance, Kantian epistemology distinguishes between analytic and synthetic propositions (Anderson 2015). Here, the synthetic denotes that which is produced through the synthesis of experience rather than conceptual analysis. A related usage features in the scholarship of Applied Linguistics and Second Language Acquisition, where pedagogical approaches are often categorised as either *analytic* or *synthetic* (Richards and Rodgers 2001). In this context, synthetic curricula rely on instructional models in which language is taught as discrete or isolated components that learners are expected to synthesise progressively into meaningful and communicative language use.

Within media studies and (digital) technology research – and most pertinently for the present discussion – the “synthetic” has become closely associated with content constructed through human-directed synthesis, artificiality, or increasingly through computational or algorithmic generation (Kalpokas 2021). This functions, therefore, as an ontological descriptor about how an artefact comes into being and the degree of technological mediation involved. However, this understanding of the synthetic is insufficient for capturing the complexity and diversity of how artefacts are encountered and evaluated in modern practices in the current GenAI landscape. It is at this juncture that we introduce *syntheticity* as a distinct analytical concept. Syntheticity is the condition – whether actual and/or attributed – of a work, utterance, or artefact as technologically generated or mediated, together with the social evaluations that mark it as distinct from exclusively human-origin production.

In this dual conceptualisation, syntheticity is not exclusively reducible to machine-made provenance or non-human artifice. Rather, it functions as a relational construct that operates along two analytically distinct but interrelated dimensions. The first, *synthetic origin*, refers to the degree to which an item is produced or shaped through computational, algorithmic, or technologically-mediated processes. The second, *synthetic attribution* (or *perceived syntheticity*), refers to the extent to which an artefact is individually or socially constructed, categorised, or labelled as synthetic by audiences regardless of, or due to, its mode of production. Crucially, these two dimensions do not necessarily coincide. An artefact may exhibit a high degree of synthetic origin while being perceived as human-authored, or conversely, may involve minimal technological mediation yet be experienced as synthetic all the same. Therefore, this proposed formulation of syntheticity denotes both the state of being synthetic (that is, of mechanical, computational or algorithmic origin) and the state of being perceived as synthetic. This distinction aligns with empirical findings by Clerwall (2014), Graefe et al. (2018), Fischer and Läubli (2020), Jacobs (2024), Jones et al. (2025), and Lee (2025), who demonstrate that modern audiences increasingly struggle to reliably distinguish synthetic from human output (cf. Section 3.2), and that assumptions about origin can override or displace the evidential authority of provenance itself (cf. Section 3.3). In such cases, beliefs about origin becomes as consequential as provenance itself, such that works of synthetic origin may be valued as legitimate, credible, or experientially authentic, just as human-origin works may be dismissed as contrived, manipulative, or false.

Moreover, human-AI hybridity can be understood as a specific configuration within syntheticity. As mentioned before, many synthetic outputs involve shared human and technological agency (Sundar and Chen 2023). By hybridity, we refer to this subset: a particular configuration of synthetic origin in which human and machine contributions are integrated in co-constitutive ways, ranging from minimal mediation (e.g. grammar checking with a word processor) to extensive co-authorial creation (e.g. AI-assisted art or writing). In such cases, syntheticity does not erase human authorship but reconfigures it, producing degrees of hybridity that challenge simple binaries of authentic versus inauthentic or human versus machine. Significantly, hybrid artefacts often involve a misalignment between synthetic origin and synthetic attribution, such that substantial technological mediation may coexist with perceptions of human authorship, or minimal mediation may nonetheless trigger synthetic labelling. Hybrid productions can be, and often are, authenticated when they resonate with audiences or align with evaluative norms. Empirical studies on human-AI collaboration support this claim: Jia et al. (2024) describe shifting perceptions of AI assistants in journalism, while Moorhouse and Wong (2025) argue that GenAI can enhance authenticity when framed within the author's communicative intent. As such, hybridity operates as a diagnostic case within this model, highlighting how authenticity judgements emerge from the dynamic interaction between production and reception, rather than from provenance alone.

4.2 Conceptualising authenticity vis-à-vis syntheticity: a continuum-based model

As explored earlier, much of the public discourse surrounding authenticity in AI-human cultural production has tended to default to a binary opposition between authenticity and the synthetic, most often understood in terms of synthetic origin. Within such framings, the “authentic” is imagined as distinctly humanly authored, while the “inauthentic” is relegated to the realm of computational or algorithmic (co-)production. The following framework (see Figure 1) challenges this essentialised binary by introducing a two-axis continuum that disaggregates origin from reception, thereby accounting for more dynamic theorisations of authenticity in relation to syntheticity in the age of generative technologies. This framework thus facilitates the conceptualisation of authenticity, not as a fixed attribute of human or technologically-mediated origin, but as a concept that emerges from the alignment, or misalignment, between its origin and reception.

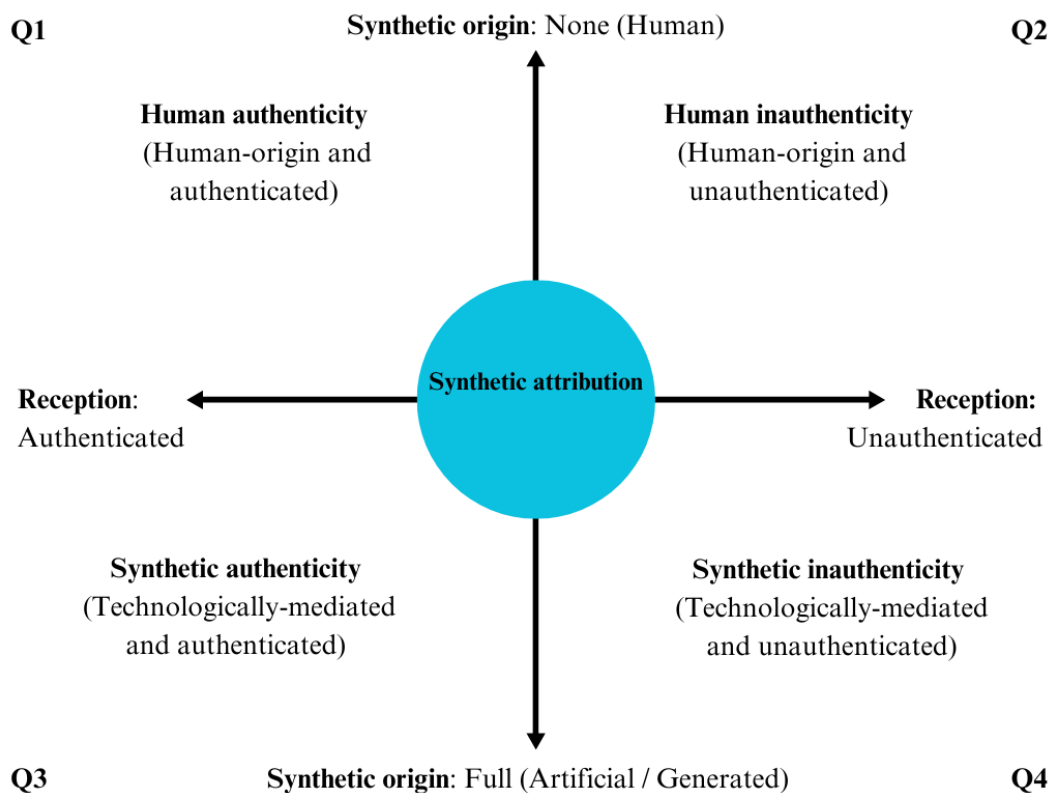


Figure 1: Continuum-based four-quadrant model of authenticity vis-à-vis syntheticity.
Created by the authors.

In this model, the vertical axis, Synthetic origin, ranges from *None* (Human) at the top to *Full* (Artificial/Generated) at the bottom. This axis captures whether a work is produced through human agency, hybrid and technologically-mediated

configurations, or through primarily, if not fully, automated computational or algorithmic processes. The horizontal axis (Reception) traces the evaluative trajectory from *Authenticated* (left) to *Unauthenticated* (right),¹ accounting for the range of social, cultural, and institutional judgements through which creations may be experienced as genuine and meaningful (i.e. authentic), or conversely dismissed as contrived or illegitimate (i.e. inauthentic).

The intersection of these two axes yields four analytically distinct but interconnected quadrants: *Human authenticity* (Q1), *Human inauthenticity* (Q2), *Synthetic authenticity* (Q3), and *Synthetic inauthenticity* (Q4). Each quadrant therefore represents a specific configuration of origin and reception, while remaining permeable to movements as evaluations shift over time or across contexts. The model also foregrounds synthetic attribution, that is, the perception, labelling, or attribution of synthetic involvement, as a relational process that may operate independently of actual origin. As represented by the central circle in Figure 1, synthetic attribution does not constitute a third axis; rather, it may traverse and modulate all quadrants by representing how works are interpreted, authenticated, or rejected, regardless of their procedural origin. In this sense, syntheticity is not confined exclusively to origin or reception alone (i.e. synthetic origin or synthetic attribution), but may operate at either level independently, or emerge through their interaction, depending on how production conditions and reception dynamics align. Therefore, syntheticity is best understood not as a bounded category, but as a quality that may traverse origin and reception alike: it may inhere in artefacts as a feature of their production, or circulate as a discursive label applied through processes of authentication or unauthentication.

In relation to the quadrants, the first, *Human authenticity* (Q1, top left) pertains to works of human origin that are also authenticated by audiences as “authentic” in some meaningful way. At the extremes, this quadrant reflects the conventional conflation of human provenance with the authentic. A diary written from personal experience, a signed manuscript, ritual or religious performances, or first-hand reporting in investigative journalism therefore all exemplify this alignment. Conceptually, this quadrant aligns closest with traditional historical authenticity (Newman and Smith 2016), Dutton’s (2003) nominal authenticity, and Grayson and Martinec’s (2004) indexical authenticity, all of which underscore the verifiable link between (human) artefact and its (human) origin. It also frequently, but not exclusively, overlaps with categorical authenticity (Newman and Smith 2016), insofar as human provenance often functions as a default criterion for what counts as “real” or legitimate (i.e. authenticated) within established categories.

¹ The term *unauthentication* is employed as opposed to other related concepts such as *inauthentication* or *deauthentication* (cf. Beerends and Aydin 2025) as these imply discursive or ideological processes of rejection and therefore presuppose an agentive stance. By contrast, *unauthentication* is intended to be more semantically broad so as to encompass not only explicit disavowal (or related affective positions) but, crucially, also the other non-positional stances toward hybrid or synthetic forms, such as indifference, ignorance, or obliviousness.

Moving downward along the vertical axis, *Synthetic authenticity* (Q3, bottom left) represents works of synthetic or artificial origin that are nonetheless authenticated as legitimate or experientially valuable. This quadrant problematises the aforementioned assumption that authenticity must be tethered to human provenance, demonstrating instead that technological mediation does not preclude, but can in fact constitute its own authenticity. This configuration builds on long-standing research in communication and interaction studies showing that users routinely engage with AI systems, avatars, and computational agents as social actors (Reeves and Nass 1996; Sundar and Chen 2023; Lee 2024). AI-driven conversational assistants adopted for customer service, algorithmically composed music performed in concert, or virtual influencers embraced as authentic by their communities all exemplify this configuration. What these cases share is that authenticity is attributed not through provenance alone but through validation (i.e. authentication) in some way. Related research includes the study of Ramadan and Ramadan (2025), who note that virtual avatars can enhance the authenticity of virtual social connections when they incorporate highly realistic features, such as facial animations, reinforcing anthropomorphic cues (Sundar and Chen 2023; Sundar 2020) and social presence (Short, Williams and Christie 1976). In this example, the virtual avatar embodies an online persona and reflects a user's personal preferences (i.e. hair styles, choice of apparel, etc.), thereby fulfilling a sense of self-authenticity by aligning digital representations with one's perceived true authentic self. As mentioned above, avatars can also be considered as social actors (Nass, Steuer and Tauber 1994; Nass and Moon 2000) without needing to be perceived as human in order to remain authentic. As CASA and subsequent communication media research demonstrate, artificial agents can be experienced as authentic if they behave as "good" machines, that is, by acting in ways that are predictable and appropriate to their assigned social roles. Huang and Jung (2022) demonstrate the continued relevance of the framework for contemporary forms of *Synthetic authenticity*. They show that different kinds of virtual characters (e.g. smart speakers, avatars, 3D projections) retain authenticity as long as they meet user's expectations for machine behaviour. New AI companions can therefore feel authentic for users when they act in ways consistent with their assigned roles, such as offering friendship, deep emotional experiences (Li and Zhang 2024) or engaging in playful interactions.

Synthetic authenticity emerges not because audiences forget that an artefact is artificial or generated, but because technological mediation itself becomes integrated into the criteria by which authenticity is evaluated. Syntheticity, in other words, can itself become the grounds of authenticity when audiences accept technological mediation as part of a work's legitimacy (cf. Beerends and Aydin 2025; Kalpokas, Šalaševičiūtė and Lipskė 2024).

If synthetic origin can be authenticated, then the converse is also possible: *Human inauthenticity* (Q2, top right) represents productions that, despite being of human provenance, are unauthenticated by their audiences. This quadrant reflects

that authenticity is an ongoing evaluative process of validation, contingent on judgements of legitimacy, sincerity, and alignment with cultural expectations (cf. categorical and self-authenticity in Section 2). Examples that fall into the remit of this quadrant would be the production of counterfeit artwork or jewellery (i.e. a fake Rolex watch) that is dismissed on account of its status as a replica or copy despite being human-made, or instances of cultural appropriation such as New Age shamans performing Native American rituals contested as inauthentic imitations (Aldred 2000).

The final quadrant, *Synthetic inauthenticity* (Q4, bottom right), completes the model by representing works of synthetic and technologically-mediated origin that are also perceived to be inauthentic. Here, unauthentication emerges through two primary psychologically distinct pathways: epistemic unauthentication, where rejection stems from perceived deception, falsity, or discrepancy between advertised and actual origin; and affective unauthentication, where rejection arises from perceptual discomfort, uncanny valley effects, or unease triggered by near-human mimicry.

Epistemic unauthentication manifests in multiple forms. At the extreme end of the reception continuum, this is exemplified by categorical or integrity-based rejections in which synthetic origin itself is incompatible with authenticity judgements. Discourses surrounding “AI slop” (Roddy and Bridges 2025; Romo 2025) exemplify this logic. Here, items are invalidated as inauthentic due to the presumption that algorithmic generation inextricably undermines their creativity, effort, or quality. In other words, synthetic attribution functions as a pejorative and essentialist evaluative label that marks generated or human-AI mediated productions as inauthentic by default.

Synthetic inauthenticity may also be driven by deception-based unauthentication processes, where audiences experience or construe a discrepancy between advertised and factual origin. In such cases, synthetic origin and negative reception reinforce one another. This aligns with the abovementioned research on deepfakes (Masood et al. 2022; Lee 2025), or inauthenticity aversion (Silver, Newman and Small 2020), which posits that when synthetic artefacts are presented as human-made, audiences can respond with moral condemnation, distrust, or rejection by interpreting the misalignment as a form of deception.

Affective unauthentication, by contrast, can occur even in the absence of explicit deception. Drawing on uncanny valley research, studies demonstrate that artefacts which approach, but fail to fully achieve, human likeness often provoke feelings of eeriness, coldness, or soullessness (Mori, MacDorman and Nageki 2012; Kätsyri et al. 2015; MacDorman and Chattopadhyay 2016). This reflects often unconscious responses to subtle inconsistencies in digital representations of human appearance, behaviour, or expressivity, rather than rationalised judgements about origin. For example, this mechanism is evidenced in empirical work on human-AI interaction in which highly anthropomorphised chatbots or virtual AI agents can trigger discomfort and negative emotional reactions when perceived as being “too

human” yet insufficiently authentic (Lee 2024; Li and Zhang 2024; Giri et al. 2026). In such cases, *Synthetic inauthenticity* emerges not from obvious artificiality or poor quality, but from affective dissonance produced by near-human mimicry. As Mieczkowski (2022) argues, perceptions of increasing machine agency are often accompanied by a corresponding reduction in perceived human agency, producing feelings of vulnerability or confusion. When GenAI systems appear overly agentic or self-aware, this imbalance can trigger a crisis of authenticity, leading users to reject specific manifestations of human-like synthetic interaction despite their technical sophistication.

Crucially, recognising these multiple pathways to unauthentication reinforces the central argument of this article: syntheticity is not inherently inauthentic. Rather, *Synthetic inauthenticity* can manifest through reception, whether via perceived deception, perceptual unease, or both. As with authenticity more broadly, rejection in Q4 is not an intrinsic property of artefacts or technologies, but the outcome of complex socially and psychologically mediated processes of evaluation, negotiation, and verification (Newman and Smith 2016; Beerends and Aydin 2025).

In practical terms, the framework works as follows: the vertical axis tracks whether an item or idea is made by humans, machines, or a combination of the two. The horizontal axis tracks whether audiences accept this as authentic or reject it as inauthentic. This, in turn, creates four broad possibilities: authentically human-made (Q1), inauthentically human-made (Q2), authentically synthetic (Q3), and inauthentically synthetic (Q4). Between these extremes lie various hybrid collaborations that may be validated or contested to different degrees, as well as works whose authenticity status remains unsettled or disputed (see Figure 2 in Section 4.3). In summary, the four quadrants provide a more flexible and relational account of authenticity and syntheticity in the age of generative technologies. They show that these are not fixed properties of origin alone, but the emergent outcome of an interplay between both provenance and observer reception, mediated by processes of synthetic attribution where relevant. In addition, the proposed framework also creates space to analyse the intermediary zones between these quadrants, where works do not fall neatly into the extremes of fully authentic or inauthentic, human or synthetic.

4.3 Between the extremes: (Un)Authenticated hybridity and Liminal authentication

While the four quadrants illustrate the extremes of origin and reception, much contemporary production resides in the spaces between them. To capture this complexity, the framework introduces two intermediary categories: *Authenticated* and *Unauthenticated hybridity* (situated vertically between Q1-Q3 and Q2-Q4 respectively) and *Liminal authentication* (situated between the extremes of the central horizontal axis). These zones illuminate the ways in which authenticity and

syntheticity are in flux in contexts where provenance and reception are neither straightforward nor stable. These are illustrated in the following Figure 2:

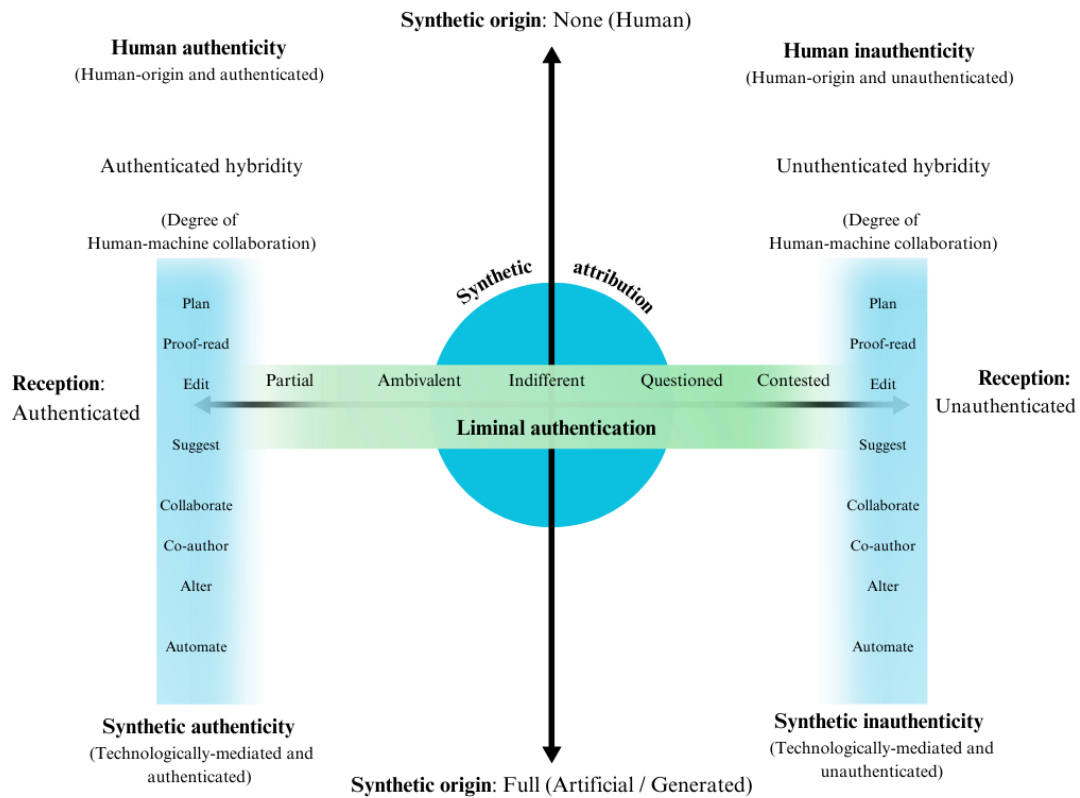


Figure 2: Complete continuum-based and four quadrant model for authenticity in relation to origin and reception including intermediary zones across the horizontal axis (*Liminal authentication*) and vertical axes (*Authenticated* and *Unauthenticated hybridity*). Created by the authors.

Authenticated and *Unauthenticated hybridity* represent divergent receptional outcomes of synthetic attribution. In this sense, while hybridity describes the structural configuration of human-machine collaboration, it is the evaluation of syntheticity, specifically synthetic attribution, that determines whether such configurations are authenticated or rejected, and to which degree. In this framework, the concept of *Authenticated hybridity* is proposed to explain the range of works of mixed human and computational origin that are nonetheless authenticated as legitimate or experientially authentic. Rather than treating technological mediation as a disqualifying factor, this category highlights the rich diversity of co-authored productions in which human and machine agency co-exist (Zafar, Ali and Yasin 2025). As discussed in Section 3, hybridity can take many forms, spanning from minimal intervention, such as grammar correction or stylistic refinement, to more substantive forms, such as deepfake generation or iterative image co-creation. At the far end of this continuum are co-authorial practices in

which human and (Gen)AI contributions are equally visible (or acknowledged) in the final product. Across these multiple variations, it is imperative to note that hybridity does not dissolve authenticity but instead can contribute to it. This reflects recent empirical findings that partial AI involvement can often be evaluated more positively than full automation (Brüns and Meißner 2024; Zhang et al. 2025), and that co-creative configurations can sustain perceptions of human authorship and intentionality (Jia et al. 2024; Moorhouse and Wong 2025).

Conversely, the opposing vertical continuum of *Unauthenticated hybridity* represents a gradual spectrum where different degrees of technological mediation coincide with audience suspicion or even full rejection. Plagiarised essays (Huang et al. 2025), heavily automated or generated cinematographical productions criticised as being “soulless” (Halperin and Rosner 2025), and politically deceptive deepfake or cheapfakes (Hameleers 2024) all exemplify this trajectory, with minor forms of hybridity typically being situated closer to the authenticated end (i.e. planning, proofreading, or minor editing), while major collaborations or substitution of human multimodal creativity are often positioned further toward inauthenticity. In such cases, negative synthetic attribution attached to hybrid productions amplifies rather than mitigates perceptions of inauthenticity. Therefore, in cases of *Authenticated hybridity*, audiences validate hybrid human-AI productions as authentic because authenticity is experienced not through provenance alone, but through resonance, credibility, and perceived validity; in unauthenticated cases, the same configurations are rejected when synthetic attribution reframes these qualities and leads audiences to interpret technological or generative intervention as a marker of inauthenticity.

The horizontal band of *Liminal authentication* is positioned between these poles, cutting across all four quadrants centrally to capture works whose reception is not yet decided and is, in some way, partial, contested, or reversible. *Liminal authentication* thus marks the indeterminate middle ground along the reception axis where creations are neither fully validated (i.e. authenticated) nor wholly rejected (i.e. unauthenticated) as they remain subject to negotiation, re-evaluation, or even uncertainty. For example, this is reflected in scenarios where a human-authored text is under investigation due to suspicion of AI authorship in an institutional setting, or in instances in which students are directly falsely accused (Stone 2024). Indeed, AI detection systems have been shown to habitually misclassify the writing of students as machine-generated despite student work being wholly human-authored, especially in the case of L2 language learners or users with non-standard linguistic styles (Liang et al. 2023; Wu et al. 2025). *Liminal authentication* is relevant here because the student work occupies a suspended state of validation. Initially, their work is accused (by machine or another social actor) of having a synthetic origin via synthetic attribution. Next, it undergoes a process of investigation and negotiation (i.e. interpersonally or institutionally), before those involved reach a final (but not necessarily permanent) decision: authentication or unauthentication. Significantly, this liminal condition does not only operate at the point of evaluation,

but prospectively also. Authors may engage in pre-emptive self-censorship or extensive revision in an effort to avoid accusations of AI usage (Perkins et al. 2024; Bannister 2025). Such cases illustrate how *Liminal authentication* in its different manifestations highlights the instability of authenticity itself (Lee 2020) and reveals how outputs can shift across evaluative boundaries over time in response to changing cultural, institutional, or technological framings.

5 Applications and limitations

By decoupling ontological origin from phenomenological reception, the proposed framework avoids essentialising *authentic* as “human” and *synthetic* as “inauthentic”. In so doing, it marks a conceptual shift away from the binary logic that has long cast *authenticity* as natural, embodied, and humanly authored, while relegating *syntheticity* to the realm of the artificial, the deceptive, or the illegitimate. Instead, the model reconceptualises authenticity while also introducing syntheticity as coexisting within a common interpretative space, especially in the context of the emerging era of GenAI.

The practical value of this shift lies in the framework’s two-axis structure that provides scholars and practitioners alike with a scalable tool for classifying and empirically testing works across different domains and cultural contexts. The framework is sufficiently abstract and descriptive to account for the multitude of ever-changing scenarios in which authenticity and syntheticity discussions unfold, while still offering pertinent distinctions that can guide case-specific analyses. Moreover, the framework refrains from imposing normative or prescriptive claims about what ought to be “authentic”, “synthetic” or the in-between. In other words, the purpose of this article has not been to champion a particular stance or to comment on the influence of AI in the current landscape. Indeed, productive critical conversations are already taking place in areas such as post-humanism (cf. Kalpokiene and Kalpokas 2023; Nicenboim et al. 2023; Giaccardi, Redström and Nicenboim 2024; Lackey and Papacharissi 2024), AI-mediated or human-AI relationships (cf. Lee 2024; Battisti 2025), the existential crisis of what it means to be human (cf. Lackey and Papacharissi 2024; Beerends and Aydin 2025), or the consequences of anthropomorphising machines or algorithms (cf. Placani 2024; Simas and Ulbricht 2024). What this framework offers is a structured conceptual framework for analysing how authenticity and syntheticity are validated, negotiated, or contested across shifting technological and cultural settings.

5.1 Applications

The framework makes four key contributions to current debates on authenticity and human-machine communication in the age of GenAI. First, it introduces the concept of *syntheticity*, understood in terms of origin and/or attribution, as a construct that operates alongside, and in relation to, authenticity in modern communicative and

mediated landscapes. The model illustrates how syntheticity and authenticity are not inherently mutually exclusive but may coexist in a shared evaluative field. In so doing, the second area of contribution highlights the need to decouple origin from reception in the theorisation of post-GenAI authenticity. As discussed in Section 3, contemporary discussions often conflate human provenance with authenticity and synthetic provenance with inauthenticity. By separating these dimensions, the proposed model both accommodates and accounts for additional configurations such as *Human inauthenticity* and *Synthetic authenticity* within the same theoretical framework. Third, the framework aligns with research that positions hybridised human-machine co-productions as a legitimate locus of authenticity (Rammert 2008; Sundar 2020; Sundar and Chen 2023; Kalpokas, Šalaševičiūtė and Lipskė 2024). The categories of *(Un)Authenticated hybridity* recognise that co-authored or mediated productions can be legitimised and accepted when they resonate with audiences and their validation processes (or not). Finally, the model accounts for perceptual and temporal dynamism in authenticity and syntheticity judgements. Through the concept of *Liminal authentication*, it captures ambiguity, misclassification, and shifts over time, while also recognising that authenticity evaluations are non-linear and socially mediated. Furthermore, a high degree of hybridity (or even full synthetic automation) does not determine whether an output or creation will be authenticated or not, as the scales of *Authenticated* and *Unauthenticated hybridity* are not predictive of reception outcome. In other words, the model reflects non-linear mapping in which greater syntheticity of origin does not necessarily entail diminished authentication (though of course it still may do this), and by extension minimal technological mediation does not guarantee audience acceptance. This responds to recent calls to move beyond static notions of authenticity in technologically-mediated communication (Lee 2020; Huang et al. 2025) and complements current discussions that post-digital authenticity is shaped by contextual, temporal, and communal factors (Köbis, Doležalová and Soraperra 2021; Kalpokas, Šalaševičiūtė and Lipskė 2024).

While elements of the proposed framework resonate with established insights from human-machine communication (e.g. distributed agency, hybridity, co-production) and communication and media research, the present model does not seek to redescribe these phenomena under new terminology. Rather, its contribution lies in integrating these strands within a single, explicitly relational theoretical framework that jointly accounts for origin, attribution, and reception in authenticity judgements. Unlike existing models that foreground either production dynamics (e.g. agency, co-creation) or audience perception (e.g. disclosure, credibility effects) in relative isolation, the continuum-based framework proposed in this article illustrates how provenance (i.e. synthetic origin) and validation (i.e. authentication) do not always align, while introducing syntheticity as the construct through which divergences between these dimensions become visible. This enables observers to classify cases in which human-authored work is rejected, synthetic or technologically-mediated work is authenticated, and hybrid productions oscillate

between authenticated and unauthenticated across contexts or over time. In doing so, the model provides a scalable structure for comparing authenticity judgements across domains (such as education, journalism, and creative industries), for empirically testing the effects of attribution and disclosure, and for tracking how works move between authenticated, liminal, and unauthenticated states.

In terms of practical applications, particularly within digital media literacy, the framework enables researchers and practitioners to identify and analyse systematic patterns of synthetic attribution that reveal underlying biases, such as in cases where audiences classify human works as synthetic or vice versa. Repeated patterns of misclassification, whether based on stylistic cues, cultural stereotypes, or assumptions about what “authentic” expression should look like, may serve as diagnostic indicators of certain ideological predispositions in audience evaluation. This application is supported by recent empirical findings that demonstrate that authenticity judgements are often shaped less by provenance itself than by disclosure or labelling practice (Clerwall 2014; Fischer and Läubli 2020; Altay and Gilardi 2024; Lermann Henestrosa and Kimmerle 2024). By mapping such cases onto the framework’s quadrants and intermediary zones, future research can systematically disentangle perceptual and attributional bias from evaluative assessment, thereby tracing where rejection reflects assumptions about synthetic origin rather than deficiencies in content quality or communicative effectiveness.

Another immediate application lies in higher education, where human-AI authorship is increasingly present (Kasneji et al. 2023; Bannister 2025). In curriculum design and AI literacy initiatives, the framework may enable educators to help students critically examine not only the degree of synthetic origin involved in a task but also how synthetic attribution and reception shape the evaluation of their work or that of others. Rather than treating AI use as a binary presence or absence, the two-axis continuum-based framework can encourage students to reflect on how varying forms of technological mediation can be interpreted, legitimised, or problematised. Similarly, in language teacher education and assessment research, the model offers an analytic tool for examining how educators respond to AI-assisted student work: whether partial human authorship mitigates concerns about academic integrity, or whether any perceived synthetic involvement prompts categorical rejection. Mapping such evaluations onto the continuum makes visible the often implicit assumptions that govern grading decisions and integrity judgements. In this way, the framework can inform the development of assessment policies and pedagogical guidelines that move beyond rigid binary notions of authenticity toward more transparent and context-sensitive approaches. Future research might also explore whether explicit engagement with the continuum model (or adapted versions of it) supports students’ digital literacy, metacognitive awareness, and evaluative practices in hybrid learning environments.

5.2 Limitations

Despite these applications and the theoretical contributions to the discussion of authenticity and syntheticity in the age of GenAI, the framework has several limitations that warrant acknowledgment. To begin, the model is not holistic and applies most directly to content, output, and authorship in the context of human creations, human-machine communication, and/or synthetic media (Kalpokas, Šalaševičiūtė and Lipské 2024). As discussed in Section 2, *authenticity* is a polysemous construct and extending the model beyond the area of cultural production risks stretching its analytic utility beyond its intended scope. In addition, the empirical research and normative assumptions mobilised throughout the article are largely grounded in Western institutions and domains, such as higher education and media production. Therefore, the framework's application in other contexts (i.e. the Global South) cannot be assumed.

At present, the framework remains theoretically rather than empirically grounded. While scholarship is drawn upon to illustrate the framework's components through selected examples, no further studies have yet applied the model systematically across domains or audiences. Future empirical research is required to test the model's reliability and validity in practice and across diverse settings.

As for the visual representation of the model in Figures 1 and 2, they necessarily simplify complex socio-technological relationships and outcomes. The horizontal and vertical axes of Synthetic origin and Reception are conceptualised as singular scalar dimensions, whereas real-world authentication processes are often plural and overlapping. For example, AI-generated work might be embraced as authentic within one evaluative frame (e.g. for its aesthetic innovation) while simultaneously rejected within another (e.g. for its lack of originality). Such cases of simultaneous (un)authentication are difficult to place on a single continuum. In addition, the intermediary continua introduced in Section 4.3 are not exhaustive. Due to constraints of space and scope, it has not been possible to visualise all potential gradients or transitional pathways between quadrants, for instance, a diagonal continuum between Q1 (*Human authenticity*) and Q4 (*Synthetic inauthenticity*) which may capture scenarios in which both origin and reception shift simultaneously over time. Finally, while the framework accommodates a wide range of human-machine collaborative configurations in principle, it does not yet differentiate between specific modes of collaboration. Emerging work has begun to quantify and typologise human-AI co-creative practices in domains such as music production (Davis et al. 2025) and information science (Salma, Hijón-Neira and Pizarro 2025), though future refinements of the model could build on this emerging research to clarify how distinct collaborative roles, such as editing, curating, co-authoring, or automating, shape audience evaluations and mediate the relationship between authenticity and syntheticity.

6 Conclusion

This article has examined how *authenticity* has been theorised across different disciplines such as philosophy, sociology, linguistics, and communication studies, highlighting its multidimensional character and persistent definitional challenges. It has further shown how the rise of GenAI technologies has intensified these challenges by producing two interlinked crises: a crisis of origin, in which provenance is destabilised by the growing difficulty of distinguishing between human and synthetic outputs, and a related crisis of perception, in which authenticity judgements increasingly depend on disclosure, framing, and social reception rather than on content quality alone.

Against this backdrop, the central contribution of this article has been to introduce a continuum-based framework that reconceptualises authenticity vis-à-vis syntheticity not as mutually exclusive categories but as mutually shaping constructs. By decoupling origin from reception, the model makes analytically visible forms of misalignment that are otherwise obscured in binary formulations. In other words, the framework demonstrates how human work can be unauthenticated, synthetic work authenticated, and hybrid human-machine productions (including GenAI) can serve as legitimate sites of authentic expression. The framework further foregrounds intermediary zones of technologically-mediated hybridity and *liminal authentication*, thereby accounting for the dynamic and negotiable nature of authenticity judgements in the rapidly evolving age of generative technologies.

While the framework is presently theoretical and domain-specific, its applications are wide-ranging. It offers a shared conceptual and terminological tool for analysing how provenance and reception interact in authenticity judgements across technosocial domains. The framework also provides a basis for empirical extension, inviting future research to examine how authenticity judgements vary across different forms of human-machine collaboration and social contexts. As GenAI technologies advance, alongside others yet to be developed, we anticipate that the framework will remain a useful lens for mapping the shifting relationships between human creativity, machine mediation, and audience validation.

Conflicts of interest

The authors declare no conflicts of interest regarding the publication of this contribution.

References

- Anderson, R. Lanier. 2015. *The Poverty of Conceptual Truth: Kant's Analytic/Synthetic Distinction and the Limits of Metaphysics*. Oxford: Oxford University Press.
- Aldred, Lisa. 2000. Plastic shamans and astroturf sun dances: new age commercialization of Native American spirituality. *American Indian Quarterly* 24(3). 329–352. <http://www.jstor.org/stable/1185908> (last accessed on 26/02/2026).
- Allaithy, Ahmed & Zaki, Mai. 2025. Evaluation of AI-generated reading comprehension materials for Arabic language teaching. *Computer Assisted Language Learning*. 1–33. <https://doi.org/10.1080/09588221.2025.2474037>
- Altay, Sacha & Gilardi, Fabrizio. 2024. People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus* 3(10). 1–11. <https://doi.org/10.1093/pnasnexus/pgae403>
- Bailey, Erica R. & Matz, Sandra C. & Youyou, Wu & Iyengar, Sheena S. 2020. Authentic self-expression on social media is associated with greater subjective well-being. *Nature Communications* 11. 1–9. <https://doi.org/10.1038/s41467-020-18539-w>
- Bannister, Peter. 2025. ParadaIse L0st? *Higher Education Research & Development*. Comment. 1–10. <https://doi.org/10.1080/07294360.2025.2586653>
- Baptista, João Pedro & Rivas-de-Roca, Rubén & Gradim, Anabela & Pérez-Curiel, Concha. 2025. Human-made news vs AI-generated news: a comparison of Portuguese and Spanish journalism students' evaluation. *Humanities and Social Science Communications* 12. 1–9. <https://doi.org/10.1057/s41599-025-04872-2>
- Barrington, Sarah & Cooper, Emily & Farid, Hany. 2025. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports* 15. 1–9. <https://doi.org/10.1038/s41598-025-94170-3>
- Battisti, Davide. 2025. Second-person authenticity and the mediating role of AI: a moral challenge for human-to-human relationships? *Philosophy & Technology* 38(1). 1–19. <https://doi.org/10.1007/s13347-025-00857-w>
- Beerends, Siri & Aydin, Ciano. 2021. Negotiating authenticity in technological environments. *Philosophy & Technology* 34. 1665–1685. <https://doi.org/10.1007/s13347-021-00480-5>
- Beerends, Siri & Aydin, Ciano. 2025. Negotiating the authenticity of AI: how the discourse on AI rejects human indeterminacy. *AI & Society* 40. 263–276. <https://doi.org/10.1007/s00146-024-01884-5>

- Beverland, Michael B. & Lindgreen, Adam & Vink, Michiel W. 2008. Projecting authenticity through advertising: consumer judgments of advertisers' claims. *Journal of Advertising* 37(1). 5–15. <https://doi.org/10.2753/JOA0091-3367370101>
- Bialystok, Lauren. 2014. Authenticity and the limits of philosophy. *Dialogue* 53. 271–298. <https://doi.org/10.1017/S001221731300111X>
- Borji, Ali. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing* 137. 1–21. <https://doi.org/10.1016/j.imavis.2023.104771>
- Brandtzaeg, Petter Bae & You, Yukun & Wang, Xi & Lao, Yucong. 2023. “Good” and “bad” machine agency in the context of human-AI communication: The case of ChatGPT. In Degen, Helmut & Ntoa, Stavroula & Moallem, Abbas (eds), *Proceedings of the HCI International 2023 - Late Breaking Papers* (Copenhagen, July 23-28, 2023), 3–23. https://doi.org/10.1007/978-3-031-48057-7_1
- Brüns, Jasper David & Meißner, Martin. 2024. Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *Journal of Retailing and Consumer Services* 79. 1–15. <https://doi.org/10.1016/j.jretconser.2024.103790>
- Bui, Hien Thu & Filimonau, Viachaslau & Sezerel, Hakan. 2024. AI-thenticity: exploring the effect of perceived authenticity of AI-generated visual content on tourist patronage intentions. *Journal of Destination Marketing and Management* 34. 1–11. <https://doi.org/10.1016/j.jdmm.2024.100956>
- Busselle, Rick W. & Bilandzic, Helena. 2008. Fictionality and perceived realism in experiencing stories: a model of narrative comprehension and engagement. *Communication Theory* 18(2). 255–280. <https://doi.org/10.1111/j.1468-2885.2008.00322.x>
- Caldwell, Mackenzie. 2023. What is an ‘author’? Copyright authorship of AI art through a philosophical lens. *Houston Law Review* 61(2). 411–442. <https://houstonlawreview.org/article/92132-what-is-an-author-copyright-authorship-of-ai-art-through-a-philosophical-lens>
- Calefato, Fabio & Lanubile, Filippo. 2010. Chapter 6 - Communication media selection for remote interaction of Ad Hoc groups. In Zelkowitz, Marvin V. (ed.), *Advances in Computers* 78, 271–313. [https://doi.org/10.1016/S0065-2458\(10\)78006-2](https://doi.org/10.1016/S0065-2458(10)78006-2)
- Carroll, Glenn R. & Kovács, Balázs. 2021. Authenticity: meanings, targets, audiences and third parties. *Research in Organizational Behavior* 41. 1–13. <https://doi.org/10.1016/j.riob.2021.100149>
- Carroll, Glenn R. 2015. Authenticity: attribution, value, and meaning. In Scott, Robert A. & Kosslyn, Stephen M. (eds), *Emerging Trends in the Social and*

- Behavioral Sciences*, 1–13. Hoboken: Wiley Online Library.
<https://doi.org/10.1002/9781118900772.etrds0020>
- Chan, Cecilia Ka Yuk & Hu, Wenjie. 2023. Students' voices on generative AI: perceptions, benefits and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20. 1–18.
<https://doi.org/10.1186/s41239-023-00411-8>
- Clerwall, Christer. 2014. Enter the robot journalist: users' perceptions of automated content. *Journalism Practice* 8(5). 519–531.
<https://doi.org/10.1080/17512786.2014.883116>
- Coupland, Nikolas. 2003. Sociolinguistic authenticities. *Journal of Sociolinguistics* 7(3). 417–431. <https://doi.org/10.1111/1467-9481.00233>
- Coupland, Nikolas. 2014. Language, society and authenticity: themes and perspectives. In Lacoste, Véronique & Leimgruber, Jakob & Breyer, Thiemo (eds), *Indexing Authenticity: Sociolinguistic Perspectives*, 14–39. Berlin: Mouton de Gruyter.
- Dammann, Olaf & Friederichs, Katja M. & Lebedinski, Sabine & Liesenfeld, Kerstin M. 2021. The essence of authenticity. *Frontiers in Psychology* 11. 1–6. <https://doi.org/10.3389/fpsyg.2020.629654>
- Davis, Jenny L. 2019. Authenticity, digital media, and person identity verification. In Stets, Jan E. & Serpe, Richard T. (eds), *Identities in Everyday Life*, 93–111. New York: Oxford University Press.
- Davis, Nicholas & Clemens, Michael & Browne, Eric & Rezwana, Jeba. 2025. Unlocking the black box of Artificial Media with quantified and explainable co-creative AI systems. In Zagalo, Nelson & Keller, Damian (eds), *Artificial Media. Emerging Trends in Narratives, Education and Creative Practice*, 21–48. Cham: Springer.
- Dehnert, Marco & Mongeau, Paul A. 2022. Persuasion in the age of artificial intelligence (AI): theories and complications of AI-based persuasion. *Human Communication Research* 48(3). 386–403.
<https://doi.org/10.1093/hcr/hqac006>
- Draxler, Fiona & Werner, Anna & Lehman, Florian & Hoppe, Matthias & Schmidt, Albrecht & Buschek, Daniel & Welsch, Robin. 2024. The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as author. *ACM Transactions on Computer-Human Interaction* 31(2). 1–40. <https://doi.org/10.1145/3637875>
- Dutton, Denis. 2003. Authenticity in arts. In Levinson, Jerrold (ed.), *The Oxford Handbook of Aesthetics*, 258–274. New York: Oxford University Press.
- Eckert, Penelope. 2003. Elephants in the room. *Journal of Sociolinguistics* 7(3). 392–397. <https://doi.org/10.1111/1467-9481.00231>

- Erdocia, Iker & Schneider, Britta & Migge, Bettina. 2025. Language in the age of AI technology: from human to non-human authenticity, from public governance to privatised assemblages. *Language in Society* 54. 1–21. <https://doi.org/10.1017/S004740452500017X>
- Fine Licht, Karl. 2025. Rethinking the ethics of GenAI in higher education: a critique of moral arguments and policy implications. *Journal of Applied Philosophy* 42(4). 1317–1337. <https://doi.org/10.1111/japp.70026>
- Fischer, Lucas & Läubli, Samuel. 2020. What’s the difference between professional human and machine translation? A blind multi-language study on domain-specific MT. In Way, Andy & Forcada, Mikel L. & Galuščáková, Petra & Popel, Martin (eds), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (Lisbon, November 3-5, 2020), 215–224. <https://aclanthology.org/2020.eamt-1.23.pdf>
- Floridi, Luciano. 2024. On the future of content in the age of artificial intelligence: Some implications and directions. Editor letter. *Philosophy & technology* 37(112). 1–11. <https://doi.org/10.1007/s13347-024-00806-z>
- Frederick, Austin & Lin, Xialing & Edwards, Chad & Edwards, Autumn. 2025. AI and human generation of classroom content: adult learners’ perceptions. *Communication Education* 74(2). 168–181. <https://doi.org/10.1080/03634523.2025.2466502>
- Gherheș, Vasile & Fărcașiu, Marcela Alina & Cernicova-Buca, Mariana & Coman, Claudiu. 2025. AI vs. human-authored headlines: evaluating the effectiveness, trust, and linguistic features of ChatGPT-generated clickbait and informative headlines in digital news. *Information* 16(2). 1–21. <https://doi.org/10.3390/info16020150>
- Ghiurău, David & Popescu, Daniela Elena. 2024. Distinguishing reality from AI: approaches for detecting synthetic content. *Computers* 14(1). 1–33. <https://doi.org/10.3390/computers14010001>
- Giaccardi, Elisa & Redström, Johan & Nicenboim, Iohanna. 2024. The making(s) of more-than-human design: introduction to the special issue on more-than-human design and HCI. *Human Computer Interaction* 40(1-4). 1–16. <https://doi.org/10.1080/07370024.2024.2353357>
- Giri, Arunangshu & Chakrabarty, Dipanwita & Lim, Weng Mark & Mangla, Sachin Kumar & Hollebeek, Linda D. 2026. Can generative artificial intelligence enhance brand performance in tourism? A mixed-method study integrating service-dominant logic, social exchange theory, and uncanny valley theory. *Tourism Management Perspectives* 60. 1–18. <https://doi.org/10.1016/j.tmp.2025.101442>
- Giridhar, Akash & Ravi, Anirudh & Kumar, Anjar & Charan B. & Vanamala, Honganur Rajanna. 2025. Detecting AI-generated artwork: an ensemble model for enhancing visual authenticity in the Digital Age. In *Proceedings 2025*

- International Conference on Inventive Computation Technologies ICICT* (Kirtipur, April 23-25, 2025), 475–481. Piscataway: IEEE. <http://doi.org/10.1109/ICICT64420.2025.11004722>
- Graefe, Andreas & Haim, Mario & Haarmann, Barbara & Brosius, Hans-Bernd. 2018. Readers' perception of computer-generated news: credibility, expertise, and readability. *Journalism* 19(5). 595–610. <https://doi.org/10.1177/1464884916641269>
- Grayson, Kent & Martinec, Radan. 2004. Consumer perceptions of iconicity and indexicality and their influence on assessments of authentic market offerings. *Journal of Consumer Research* 31(2). 296–312. <https://doi.org/10.1086/422109>
- Groth, Oden & Block, Lauren & Newman, George E. 2023. Lost in digitalization: the physical format of creative work affects authenticity perceptions. *Psychology of Aesthetics, Creative, and the Arts* 19(6). 1578–1592. <https://dx.doi.org/10.1037/aca0000628>
- Guignon, Charles. 2004. *On Being Authentic*. Oxon: Routledge.
- Guignon, Charles. 2008. Authenticity. *Philosophy Compass* 3(2). 277–290. <https://doi.org/10.1111/j.1747-9991.2008.00131.x>
- Guzman, Andrea L. 2018. *Human-machine communication. Rethinking communication, technology, and ourselves*. New York: Peter Lang.
- Guzman, Andrea L. & McEwen, Rhonda & Jones, Steve. 2023. Introduction to the handbook. In Guzman, Andrea L. & McEwen, Rhonda & Jones, Steve (eds), *The SAGE Handbook of Human-Machine Communication*, xli-xlvi. London: Sage.
- Haimson, Oliver L. & Liu, Tianxiao & Zhang, Ben Zefeng & Corvite, Shanley. 2021. The online authenticity paradox: what being “authentic” on social media means, and barriers to achieving it. In Nichols, Jeff (ed.), *Proceedings of the ACM on Human-Computer Interaction* 5 CSCW2, 1–18. <https://doi.org/10.1145/3479567>
- Halperin, Brett A. & Rosner, Daniela K. 2025. “AI is soulless”: Hollywood film workers' strike and emerging perceptions of generative cinema. *ACM Transactions on Computer-Human Interaction* 32(2). 1–27. <https://doi.org/10.1145/3716135>
- Hameleers, Michael & van der Meer, Tony G & Dobber, Tom. 2024. They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication* 39(1). 56–70. <https://doi.org/10.1177/02673231231184703>
- Hameleers, Michael. 2024. Cheap versus deep manipulation: the effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research* 36(1). 1–19. <https://doi.org/10.1093/ijpor/edae004>

- Hohenstein, Jess & Jung, Malte. 2020. AI as a moral crumple zone: the effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106. 1–13. <https://doi.org/10.1016/j.chb.2019.106190>
- Huang, Dongpeng & Hash, Nicole & Cummings, James J. & Prena, Kelsey. 2025. Academic cheating with generative AI: exploring a moral extension of the theory of planned behaviour. *Computers and Education: Artificial Intelligence* 8. 1–16. <https://doi.org/10.1016/j.caeai.2025.100424>
- Huang, Junru & Jung, Younbo. 2022. Perceived authenticity of virtual characters makes the difference. *Frontiers in Virtual Reality* 3. 1–15. <https://doi.org/10.3389/frvir.2022.1033709>
- Jacobs, Bart. 2024. The authenticity crisis. *Computer Law & Security Review: The International Journal of Technology Law and Practice* 53. 1–5. <https://doi.org/10.1016/j.clsr.2024.105962>
- Jia, Huiyan & Appelman, Alyssa & Wu, Mu & Bien-Aimé, Steve. 2024. News bylines and perceived AI authorship: effects on source and message credibility. *Computers in Human Behavior: Artificial Humans* 2(2). 1–9. <https://doi.org/10.1016/j.chbah.2024.100093>
- Jones, Cameron R. & Rathi, Ishika & Taylor, Sydney & Bergen, Benjamin K. 2025. People cannot distinguish GPT-4 from a human in a Turing test. In *FACCT'25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Athens, June 23-26, 2025), 1615–1639. <https://doi.org/10.1145/3715275.3732108>
- Ju, Naan & Kim, Terry & Im, Hyunjoo. 2024. Fake human but real influencer: the interplay of authenticity and humanlikeness in virtual influencer communication? *Fashion and Textiles* 11. 1–27. <https://doi.org/10.1186/s40691-024-00380-0>
- Kalpokas, Ignas. 2021. Problematising reality: the promises and perils of synthetic media. *SN Social Sciences* 1(1). 1–11. <https://doi.org/10.1007/s43545-020-00010-8>
- Kalpokas, Ignas. 2023. Work of art in the age of its AI reproduction. *Philosophy & Social Criticism* 51(8). 1268–1286. <https://doi.org/10.1177/01914537231184490>
- Kalpokas, Ignas. 2024. *Technological governance and escapism in times of accelerated change*. Cham: Palgrave Macmillan.
- Kalpokas, Ignas & Kalpokienė, Julija & Šalaševičiūtė, Vaiva. 2024. Rage against the machine: automated policing of generative AI. In Kuldova, Teresa Østbø & Gundhus, Helene Oppen Ingebrigtsen & Wathne, Christin T. (eds), *Policing and Intelligence in the Global Big Data Era, Volume II. Palgrave's Critical Policing Studies*, 259–286. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-68298-8_10

- Kalpokas, Ignas & Šalaševičiūtė, Vaiva & Lipskė, Monika. 2024. Technology as a threat or a solution? The challenges of responding to synthetic Media. *Baltic Journal of Law & Politics* 16(2). 1–22. <https://doi.org/10.2478/bjlp-2023-0010>
- Kalpokienė, Julija & Kalpokas, Ignas. 2023. Creative encounters of a posthuman kind – anthropocentric law, artificial intelligence, and art. *Technology in Society* 72. 1– 8. <https://doi.org/10.1016/j.techsoc.2023.102197>
- Kasneci, Enkelejda & Sessler, Katharina & Küchemann, Stefan & Bannert, Maria & Dementieva, Daria & Fischer, Frank & Gasser, Urs & Groh, Georg & Poquet, Oleksandra & Sailer, Michael et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103. 1–9. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kätsyri, Jari & Förger, Klaus & Mäkäräinen, Markus & Takala, Tapio. 2015. A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology* 6. 1–16. <https://doi.org/10.3389/fpsyg.2015.00390>
- Keller, Damián & Zagalo, Nelson. 2025. Exploring the frontiers of artificial media. In Zagalo, Nelson & Keller, Damián (eds), *Artificial Media. Emerging Trends in Narratives, Education and Creative Practice*, 1–20. Cham: Springer.
- Kim, Hyoseok & Koo, Thomas K. B. 2024. The impact of generative AI on syllabus design and learning. *Journal of Marketing Education* 48(1). 20–41. <https://doi.org/10.1177/02734753241299024>
- Kim, Hayeon & Lee, Sang Wo. 2025. Sorry, it’s my fault: politeness, attribution, and anthropomorphism in managing generative AI hallucinations. *International Journal of Information Management* 86. 1–15. <https://doi.org/10.1016/j.ijinfomgt.2025.102996>
- Kirk, Colleen P. & Givi, Julian. 2024. The AI-authorship effect: understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research* 186. 1–14. <https://doi.org/10.1016/j.jbusres.2024.114984>
- Knott, Alistair & Pedreschi, Dino & Jitsuzumi, Toshiya & Leavy, Susan & Eyers, David & Chakraborti, Tapabrata & Trotman, Andrew & Sundareswaran, Sundar & Baeza-Yates, Ricardo & Biecek, Przemyslaw et al. 2024. AI content detection in the emerging information ecosystem: new obligations for media and tech companies. *Ethics and Information Technology* 26. 1–14. <https://doi.org/10.1007/s10676-024-09795-1>
- Köbis, Nils C. & Doležalová, Barbora & Soraperra, Ivan. 2021. Fooled twice: people cannot detect deepfakes but think they can. *iScience* 24(11). 1–17. <https://doi.org/10.1016/j.isci.2021.103364>

- Kudina, Olya & Verbeek, Peter-Paul. 2019. Ethics from within: Google Glass, the Collingridge Dilemma, and the mediated value of privacy. *Science, Technology & Human Values* 44(2). 291–314. <https://doi.org/10.1177/0162243918793711>
- Lackey, Cait & Papacharissi, Zizi. 2024. Machine ex machina: a framework decentering the human in AI design praxis. *Human-Machine Communication* 8. 7–25. <https://doi.org/10.30658/hmc.8.1>
- Lavazza, Andrea. 2025. AI, authenticity, and the chatbot between us: commentary on Battisti. Commentary. *Philosophy & Technology* 38(49). 1–5. <https://doi.org/10.1007/s13347-025-00880-x>
- Law, Ryan & Guan, Xibeijia & Soulo, Tim. 2025. “74% of new webpages include AI content (study of 900k pages)”. *Ahrefs blog*. <https://www.ahrefs.com/blog/what-percentage-of-new-content-is-ai-generated/> (last accessed on 26/02/2026).
- Lee, Eun-Ju. 2020. Authenticity model of (mass-oriented) computer-mediated communication: Conceptual explorations and testable propositions. *Journal of Computer-Mediated Communication* 25(1). 60–73. <https://doi.org/10.1093/jcmc/zmz025>
- Lee, Eun-Ju. 2024. Minding the source: toward an integrative theory of human-machine communication. *Human Communication Research* 50. 184–193. <https://doi.org/10.1093/hcr/hqad034>
- Lee, Eun-Ju. 2025. Authenticity at the heart of communication. *Journal of Communication* 75(2). 85–89. <https://doi.org/10.1093/joc/jqaf005>
- Lehman, David W. & O’Connor, Kieran & Carroll, Glenn R. 2019. Acting on authenticity: individual interpretations and behavioral responses. *Review of General Psychology* 23(1). 19–31. <https://doi.org/10.1177/1089268019829470>
- Lehman, David W. & O’Connor, Kieran & Kovács, Balázs & Newman, George E. 2019. Authenticity. *Academy of Management Annals* 13(1). 1–42. <https://doi.org/10.5465/annals.2017.0047>
- Lermann Henestrosa, Angelica & Kimmerle, Joachim. 2024. The effects of assumed AI vs. human authorship on the perception of a GPT-Generated text. *Journalism and Media* 5(3). 1085–1097. <https://doi.org/10.3390/journalmedia5030069>
- Li, Hang & Zhang, Renwen. 2024. Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots. *Journal of Computer-Mediated Communication* 29(5). 1–13. <https://doi.org/10.1093/jcmc/zmae015>
- Liang, Weixin & Yuksekogul, Mert & Mao, Yining & Wu, Eric & Zou, James. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4(7). 100779. 1–4. <https://doi.org/10.1016/j.patter.2023.100779>

- Llurda, Enric & Mocanu, Vasilica. 2024. English in Spain: education, attitudes and native-speakerism. *World Englishes* 43(2). 315–331. <https://doi.org/10.1111/weng.12651>
- Marriott, Hanna R. & Pitardi, Valentina. 2023. One of the loneliest number... two can be as bad as one. The influence of AI friendship apps on users, wellbeing and addiction. *Psychology and Marketing* 41(1). 86–101. <https://doi.org/10.1002/mar.21899>
- Martínez, Matías. 2020. Authenticity in narratology and in literary studies. In Fludernik, Monica & Ryan, Marie-Laure (eds), *Narrative Factuality: A Handbook*, 521–532. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110486278-036>
- Masood, Momina & Nawaz, Mariam & Malik, Khalid Mahmood & Javed, Ali & Irtaza, Aun & Malik, Hafiz. 2022. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence* 53. 3974–4026. <https://doi.org/10.1007/s10489-022-03766-z>
- MacDorman, Karl F. & Chattopadhyay, Debaleena. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146. 190–205. <https://doi.org/10.1016/j.cognition.2015.09.019>
- McEvoy, James. 2009. Living in an age of authenticity: Charles Taylor on identity today. *Australasian Catholic Record* 86(2). 161–172.
- Mercier, Faye & Abidin, Crystal. 2025. Authenticity as discursive gaze: a critical review of social media research on authenticity. *Journal of Communication* 75(5). 398–404. <https://doi.org/10.1093/joc/jqaf038>
- Michael, Kyle & Shuttleworth, James. 2021. *The History and Ethics of Authenticity. Meaning, Freedom and Modernity*. London: Bloomsbury Academic.
- Microsoft. 2026. “Global AI adoption in 2025. A widening digital divide”. Microsoft. <https://www.microsoft.com/en-us/research/wp-content/uploads/2026/01/Microsoft-AI-Diffusion-Report-2025-H2.pdf> (last accessed on 26/02/2026).
- Mieczkowski, Hanna Nicole. 2022. *AI-mediated communication: Examining agency, ownership, expertise, and roles of AI systems*. Stanford: Stanford University. Ph.D. Dissertation. <https://purl.stanford.edu/nz776rw2839>
- Mishan, Freda. 2017. ‘Authenticity’ in the digital era. In Maley, Alan & Tomlinson, Brian (eds), *Authenticity in Material Development for Language Learning*, 10–24. Cambridge: Cambridge Scholars Publishing.
- Mori, Masahiro & MacDorman, Karl F. & Nageki, Norri. 2012. The uncanny valley [From the field]. *IEEE Robotics & Automation Magazine* 19(2). 98–100.

- Moorhouse, Benjamin Luke & Wong, Kevin M. 2025. *Generative Artificial Intelligence and Language Teaching*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009618823>
- Nass, Clifford & Moon, Youngme. 2000. Machines and mindlessness: social responses to computers. *Journal of Social Issues* 56(1). 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, Clifford & Steuer, Jonathan & Tauber, Ellen R. 1994. Computers as social actors. In Adelson, Beth & Dumais, Susan & Olson, Judith (eds), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, April 24-28, 1994), 72–78. New York: Association for Computing Machinery. <https://doi.org/10.1145/191666.191703>
- Newman, George E. & Bloom, Paul. 2012. Art and authenticity: the importance of originals in judgments of value. *Journal of Experimental Psychology* 141(3). 558–569. <https://doi.org/10.1037/a0026035>
- Newman, George E. & Smith, Rosanna K. 2016. Kinds of authenticity. *Philosophy Compass* 11(10). 609–618. <https://doi.org/10.1111/phc3.12343>
- Newman, George E. 2019. The psychology of authenticity. *Review of General Psychology* 23(1). 8–18. <https://doi.org/10.1037/gpr0000158>
- Nicenboim, Iohanna & Oogjes, Doenja & Biggs, Heidi & Nam, Seowoo. 2023. Decentering through design: bridging posthuman theory with more-than-human design practices. *Human Computer Interaction* 40(1-4). 195–220. <https://doi.org/10.1080/07370024.2023.2283535>
- O'Rourke, Bernadette & Ramallo, Fernando. 2013. Competing ideologies of linguistic authority amongst new speakers in contemporary Galicia. *Language in Society* 42(3). 287–305. <https://doi.org/10.1017/S0047404513000249>
- Pedersen, Carsten L. & Ritter, Thomas. 2024. Digital authenticity: towards a research agenda for the AI-driven fifth phase of digitalization in business-to-business marketing. *Industrial Marketing Management* 123. 162–172. <https://doi.org/10.1016/j.indmarman.2024.10.005>
- Perkins, Mike & Roe, John & Vu, Binh H. & Postma, Dirk & Hickerson, David & McGaughan, James & Khuat, Huyen Q. 2024. Simple techniques to bypass GenAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education* 21(1). 1–25. <https://doi.org/10.1186/s41239-024-00487-w>
- Peterson, Richard A. 2005. In search of authenticity. *Journal of Management Studies* 42(5). 1083–1098. <https://doi.org/10.1111/j.1467-6486.2005.00533.x>
- Pinner, Richard S. 2014. The authenticity continuum: toward a definition incorporating international voices. *English Today* 30(4). 22–27. <https://doi.org/10.1017/S0266078414000364>
- Pinner, Richard S. 2016. *Reconceptualising Authenticity for English as a Global Language*. Bristol: Multilingual Matters.

- Placani, Adriana. 2024. Anthropomorphism in AI: hype and fallacy. *AI and Ethics* 4. 691–698. <https://doi.org/10.1007/s43681-024-00419-4>
- Qin, Xin & Zhou, Xiang & Chen, Chen & Wu, Dongyuan & Zhou, Hansen & Dong, Xiaowei & Cao, Limei & Lu, Jackson G. 2025. AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin* 151(5). 580–599. <https://doi.org/10.1037/bul0000477>
- Quinlantang, Rae Francis. 2025. The HAI-IO model: a framework for understanding the human-AI communication process. *Human Machine Communication* 10. 207–221. <https://doi.org/10.30658/hmc.10.9>
- Rae, Irene. 2024. The effects of perceived AI use on content perceptions. In Floyd Mueller, Florian (ed.), *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, May 11-16, 2024), 1–14. New York: Association for Computing Machinery. <https://doi.org/10.1145/3613904.3642076>
- Ramadan, Zahi & Ramadan, Jad. 2025. AI avatars and co-creation in the metaverse. *Consumer Behavior in Tourism and Hospitality* 20(1). 131–147. <https://doi.org/10.1108/CBTH-07-2024-0246>
- Rammert, Werner. 2008. Where the action is: distributed agency between humans, machines, and programs. TUTS – Working Papers, 4-2008. Berlin. <https://www.ssoar.info/ssoar/handle/document/1233> (last accessed on 26/02/2026).
- Rapp, Amon & Di Lodovico, Chiara & Torrielli, Federico & Di Caro, Luigi. 2025. How do people experience the images created by generative artificial intelligence? An exploration of people’s perceptions, appraisals, and emotions related to Gen-AI text-to-image model and its creations. *International Journal of Human-Computer Studies* 193. 1–16. <https://doi.org/10.1016/j.ijhcs.2024.103375>
- Reeves, Byron & Nass, Clifford. 1996. *The Media Equation. How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.
- Revolusi, Prabu & Febriandy, Radians Frisna. 2025. Human AI communication (HA-C): transforming the role of technology in human interaction. *Multidisciplinary Science Journal* 8(3). 1–14. <https://doi.org/10.31893/multiscience.2026201>
- Richards, Jack C. & Rodgers, Theodore S. 2001. *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511667305>
- Roddy, Stephen & Bridges, Brian. 2025. Cybernetic resurgences: machine music beyond AI slop. In Zagalo, Nelson & Keller, Damián (eds), *Artificial Media*.

- Emerging Trends in Narratives, Education and Creative Practice*, 95–116. Cham: Springer.
- Rodríguez Valdés, María. 2025. The use of artificial media as a tool for the creation of artistic images through collaborative methodologies. In Zagalo, Nelson & Keller, Damián (eds), *Artificial Media. Emerging Trends in Narratives, Education and Creative Practice*, 159–174. Cham: Springer
- Romo, Vanessa. 2025. “It’s 2025, the year we decided we need a widespread slur for robots”. *NPR News*, August 6, 2025. <https://www.vpm.org/npr-news/npr-news/2025-08-06/its-2025-the-year-we-decided-we-need-a-widespread-slur-for-robots> (last accessed on 26/02/2026).
- Salma, Zainab & Hijón-Neira, Raquel & Pizarro, Celeste. 2025. Designing co-creative systems: five paradoxes in human-AI collaboration. *Information* 16(10). 1–19. <https://doi.org/10.3390/info16100909>
- Schilke, Oliver & Reimann, Martin. 2025. The transparency dilemma: how AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes* 188. 1–16. <https://doi.org/10.1016/j.obhdp.2025.104405>
- Short, John & Williams, Ederyn & Christie, Bruce. 1976. *The Social Psychology of Telecommunications*. London: John Wiley and Sons.
- Silver, Ike & Newman, George & Small, Deborah A. 2020. Inauthentic aversion: moral reactance toward tainted actors, actions, and objects. *Consumer Psychology review* 4(1). 70–82. <https://doi.org/10.1002/arcp.1064>
- Simas, Gustavo & Ulbricht, Vania. 2024. Human-AI interaction: an analysis of anthropomorphization and user engagement in conversational agents with a focus on ChatGPT. In *Intelligent Human Systems Integration IHSI 2024*, 119. 454–464. <http://doi.org/10.54941/ahfe1004510>
- Spennemann, Dirk H. R. 2025. Generative Artificial Intelligence and the future of public knowledge. *Knowledge* 5(3). 1–19. <https://doi.org/10.3390/knowledge5030020>
- Stone, Brian W. 2024. Generative AI in higher education: uncertain students, ambiguous use cases, and mercenary perspectives. *Teaching of Psychology* 52(3). 347–356. <https://doi.org/10.1177/00986283241305398>
- Sundar, S. Shyam. 2020. Rise of machine agency: a framework for studying the psychology of human-AI interaction (HAI). *Journal of Computer-Mediated Communication* 25(1). 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Sundar, S. Shyam & Chen, Jin. 2023. From CASA to TIME: machine as a source of media effects. In Guzman, Andrew L. & McEwen, Rhonda & Jones, Steve (eds), *The SAGE Handbook of Human-Machine Communication*, 63–70. London: Sage. <https://doi.org/10.4135/9781529782783.n9>
- Taylor, Charles. 1991. *The Ethics of Authenticity*. Cambridge: Harvard University Press.

- Thakur, Kutub & Barker, Helen G. & Pathan, Al-Sakib Khan. 2024. *Artificial Intelligence and Large Language Models*. Oxon: CRC Press. <https://doi.org/10.1201/9781003474173>
- Trilling, Lionel. 1972. *Sincerity and Authenticity*. Cambridge: Harvard University Press.
- Tron Gianet, Eric & Di Caro, Luigi & Rapp, Amon. 2024. Redefining the user in human-generative AI collaboration: insights from music composition. In *Proceedings of the 8th International Workshop on Cultures of Participation in the Digital Age CoPDA 2024* (Arenzano, June 3, 2024). Aachen: CEUR Workshop Proceedings. <https://www.ceur-ws.org/Vol-3701/paper6.pdf> (last accessed on 26/02/2026).
- Vlachopoulos, Dimitrios & Makri, Agoritsa. 2024. A systematic literature review on authentic assessment in higher education: best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation* 83. 1–13. <https://doi.org/10.1016/j.stueduc.2024.101425>
- Wang, Ning. 1999. Rethinking authenticity in tourism experience. *Annals of Tourism Research* 26(2). 349–370. [https://doi.org/10.1016/S0160-7383\(98\)00103-0](https://doi.org/10.1016/S0160-7383(98)00103-0)
- Wasdahl, Alex. 2024. Machine credibility: how news readers evaluate AI-generated content. *InterActions: UCLA Journal of Education and Information Studies* 19(1). 1–28. <http://dx.doi.org/10.5070/D41.7196>
- Werdiningsih, Indah & Marzuki & Rusdin, Diyenti. 2024. Balancing AI and authenticity: EFL students' experiences with ChatGPT in academic writing. *Cogent Arts & Humanities* 11(1). 1–15. <http://doi.org/10.1080/23311983.2024.2392388>
- Whittaker, Lucas & Kietzmann, Tim C. & Kietzmann, Jan & Dabirian, Amir. 2020. 'All around me are synthetic faces': the mad world of AI-generated media. *IT Professional* 22(5). 90–99. <https://doi.org/10.1109/MITP.2020.2985492>
- Williams, Bernard. 2002. *Truth and Truthfulness: an Essay in Genealogy*. Princeton: Princeton University Press.
- Wilt, Joshua A. & Thomas, Sarah & McAdams Dan P. 2019. Authenticity and inauthenticity in narrative identity. *Heliyon* 5(7). 1–13. <https://doi.org/10.1016/j.heliyon.2019.e02178>
- Woolard, Kathryn A. 2016. *Singular and Plural: Ideologies of Linguistic Authority in 21st century Catalonia*. Oxford: Oxford University Press.
- Wu, Junchao & Yang, Shu & Zhan, Runzhe & Yuan, Yulin & Chao, Lidia & Wong, Derek. 2025. A survey on LLM-generated text detection: necessity, methods, and future directions. *Computational Linguistics* 51(1). 275–338. https://doi.org/10.1162/coli_a_00549

- Wu, Wanyan & Lingel, Jessa. 2025. "I am Neuro, who are you?": Performances of authenticity in an experimental AI livestream. *New Media and Society*. 1–21. <https://doi.org/10.1177/14614448251406904>
- Yang, Daijin & Zhou, Yanpeng & Zhang, Zhiyuan & Li, Toby Jia-Jun & LC, Ray. 2022. AI as an active writer: interaction strategies with generated text in human-AI collaborative fiction writing. In Smith-Renner, Alison & Amir, Ofra (eds), *Joint Proceedings of the IUI 2022 Workshops: APEX-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE co-located with the ACM International Conference on Intelligent User Interfaces* (Helsinki, March 21–22), 56–65. Helsinki: CEUR Workshop Proceedings. <https://ceur-ws.org/Vol-3124/#paper6> (last accessed on 26/02/2026).
- YouTube. 2024. "A.I. and creators. The future of tech and creativity." YouTube. https://www.youtube.com/howyoutubeworks/ai-and-creators/?utm_source=linkedin&utm_medium=social&utm_campaign=foc_launch_kc&utm_content=post_9.24.24 (last accessed on 23/08/25).
- Zafar, Muhammad Bilal & Ali, Hassnian & Yasin, Talha. 2025. Reimagining human creativity and learning in the age of generative AI: a multi-method meta-thematic synthesis. *Next Research* 2(4). 1–18. <https://doi.org/10.1016/j.nexres.2025.100802>
- Zhang, Longyu & Fang, Cong & Wang, Stephen Jia & Wang, Yao & Luo, Shijian. 2025. Students' attitudes and sentiments toward AI-generated images: deep learning-based social media text mining. *Interactive Learning Environments*. 1–26. <https://doi.org/10.1080/10494820.2025.2545964>

Style hybride d'une chimère. Authentifier le style dans une production textuelle hybride

Perrine Maurel (Sorbonne Université)

perrine.maurel.recherche(at)gmail.com

Abstract

La présente étude porte sur l'authentification du style dans les productions textuelles hybrides. Cette notion de style est évasive en soi, grossièrement partagée entre le premier mouvement d'écriture spontané et les ajustements apportés ensuite pour améliorer le texte et sa réception. Puisqu'elle fonde la légitimité de la figure auctoriale, il convient donc de l'interroger dans le contexte des grands modèles de langage, lesquels sont des générateurs de style comme de texte. Plusieurs études cherchent à caractériser le style d'un Grand Modèle de Langage (GML), entre faits textuels et hallucinations erronées. La présente contribution propose une catégorisation des différentes productions hybrides selon trois axes : le matériel d'origine, la direction des altérations et la partition du résultat final. Cette catégorisation ne doit toutefois pas s'appliquer de manière rigide et mettre en exergue la complémentarité de l'approche hybride plutôt que la séparation distincte du style humain et du style artificiel. En effet, une production hybride ne saurait l'être sans un rétrocontrôle attentif d'une figure auctoriale humaine, laquelle doit entériner chaque aspect du texte généré dans la version finale de la production ; pour obtenir un résultat final qui, somme toute, s'avère fondamentalement humain.

This paper focuses on the authentication of style in hybrid textual productions. The notion of style is inherently elusive, roughly divided between the first spontaneous writing movement and the subsequent adjustments made to improve the text and its reception. Since it underpins the legitimacy of the auctorial figure, it needs to be questioned in the context of Large Language Models (LLM), which are generators of both style and text. Several studies have sought to characterize the style of a LLM, between textual facts and erroneous hallucinations. The present study proposes a categorization of the various hybrid productions along three axes: the original material, the direction of the alterations and the division of the final result. This categorization should not be rigidly applied, however, and should emphasize the complementary nature of the hybrid approach rather than the distinct separation of human and artificial styles. Indeed, a hybrid production cannot be so without careful feedback from a human authorial figure, who must endorse every aspect of the text generated in the final version of the production; to achieve a final result that, all in all, proves to be fundamentally human.

Mots clefs

intelligence artificielle, langage généré par IA, style, parentalité humain vs. machine, production hybride humain-IA

Keywords

artificial intelligence, AI-generated language, style, human vs. machine authorship, human-AI hybrid productions

1 Quand les styles se mélangent : une Intelligence Artificielle [IA] co-autrice ?

Faut-il citer ChatGPT comme co-autrice d'un texte si sa contribution a été suffisamment importante ? Voilà une question épineuse ; en authentifiant un Grand Modèle de Langage [GLM] comme auteur d'un texte, ne risquons-nous de retirer une part de légitimité à la figure auctoriale humaine ? À quel degré cette légitimité



s'évalue-t-elle de prime abord ? L'augmentation de l'utilisation des GML dans l'enseignement supérieur et dans les pratiques de recherche (Crawford, Cowling & Ashton-Hay et al. 2023) introduit de nouvelles problématiques pratiques, éthiques et littéraires ; cas plus difficile à trancher encore, les productions hybrides entre figure auctoriale humaine et GML relèvent de phénomènes particuliers, de croisements textuels entre deux entités fondamentalement différentes : les chimères issues d'un tel processus imposent de questionner la définition même de l'authenticité.

Les différentes approches de l'authenticité font état du « maintien d'une identité personnelle » (Lindholm 2013) et d'une dichotomie entre spontanéité et ajustement du résultat final (van Leeuwen 2001). Une autre notion semble épouser ces mêmes considérations : celle du style. Le style relève des différents actes et choix entrepris par une figure auctoriale pour exprimer un message défini : le premier mouvement spontané est tout aussi important que les ajustements appliqués ensuite pour faire correspondre au mieux la forme au fond. Éminemment humaine, la notion de style prend un tout autre sens à l'aune des GML, pour devenir purement quantitative en se cantonnant à l'effet plutôt qu'au processus. Un GML ne choisit pas, il génère. Il y a là donc encore hybridation entre deux mécanismes radicalement différents, le style d'un être humain et le style d'un GML ne relevant pas des mêmes procédés.

L'objectif principal de cet article sera de proposer une réflexion sur l'authentification du style dans les productions hybrides, en tenant compte de l'état de l'art. Pour ce faire, l'auteur de la présente a interrogé deux bases de données : le moteur de recherche Google Scholar et la base de données HAL. Les requêtes successives, effectuées en français et en anglais, comprenaient les mots clefs thématiques « style » et « auteur », puis « IA générative » ou « Grand Modèle de Langage » et « performance » ou « style » ou « authenticité » ou « imitation ». Les articles ont ensuite été sélectionnés afin de proposer un échantillon représentatif des différents projets menés en hybridation, et pour souligner l'importance du style dans la notion d'authentification. En effet, le style est un outil probant d'authentification qui soulève de nombreux questionnements dans le cadre d'une hybridation humain/GML.

2 Le style comme outil d'authentification, chez l'humain comme chez le GML

La notion de style est évasive. Elle fait tantôt référence au « premier mouvement » spontané de la personne auctoriale, tantôt aux différents choix et ajustements entrepris pour parvenir au résultat final – ces deux définitions se trouvent chez Des Essart (1800 : 413). En effet, le style fait l'objet de nombreuses évolutions de paradigme dans le temps et selon les courants. Demeure une constante : le style est intrinsèquement lié à la parentalité d'une œuvre.

2.1 Le style et la figure auctoriale

Reggiani (2002) indique ainsi dans son atelier Fabula *Le style comme indice : le postulat de l'unité* que le style se lie généralement à la figure auctoriale : « L'unité de l'auteur garantie par celle de son style fonde son autorité. [...] Le style dans son usage commun apparaît ainsi indissolublement lié à la catégorie unitaire de la personne ». Une personne, un style qui lui appartient et qui permet donc de l'authentifier. On pourrait également citer Buffon « le style est l'homme même » : Dürrenmatt (2010 : 63) rappelle que le véritable sens de cette citation est comme suit « [...] le style engage toutes les facultés de l'homme dans sa quête d'une parfaite adéquation de son dire à la chose, qu'il [sic] le distingue aussi par là de toutes les autres créatures. »

Ainsi, le style peut être interprété comme un processus de travail où l'être humain se réalise et se mobilise dans son entièreté. Processus qui ne peut qu'être authentique, en conséquence.

Il est important de s'arrêter un instant sur une dichotomie flagrante du style : ce dernier peut être considéré comme un objet figé – la somme des caractéristiques stylistiques immuables du texte dans son état final – ou comme un processus en mouvement – tant dans la phase de création du texte par la figure auctoriale que dans la phase de lecture et d'appréhension du texte par le lectorat. Nous y reviendrons tout particulièrement dans le contexte des GML.

En dépit de ces définitions plurielles qui redéfinissent en permanence le contour de la notion, le style s'est avéré être un outil efficace pour authentifier la parentalité d'une œuvre. La stylométrie (Delcourt 2002), notamment, est un outil de choix. Cette dernière consiste en l'analyse de l'expression statistique de caractéristiques textuelles qu'il est possible de comptabiliser et d'analyser, aussi nommée approche quantitative. Cette approche est intrinsèquement liée à l'approche computationnelle, qui n'est capable d'appréhender le style que par l'étude de l'objet figé et quantifié. Plusieurs ouvrages offrent un aperçu de la notion, tels que Ramnial, Panchoo, Pudaruth et al. (2016) ou Iqbal, Debbabi et Fung (2020).

En termes de mobilisation pratique de la stylométrie, on peut citer Cafiero, Camps et Gabay (2023 : 2) : cet article cherche à établir si Louise Labé serait un prête-nom ou une véritable autrice. La démarche relève donc bien d'une authentification du texte, et par là, de la figure auctoriale. Notons en outre que la stylométrie s'applique dans le cas où le corpus d'étude contient plusieurs figures auctoriales différentes :

La stylométrie étant par essence comparative, il est nécessaire de confronter les écrits de plusieurs auteurs : l'auteur·trice contesté·e, celui ou celle à qui les documents serait [sic] réattribués, et des auteurs de contrôle ne posant aucun problème d'attribution afin de garantir la fiabilité des résultats.

Cette « essence comparative » prend tout son sens dans la démarche qui est la nôtre, celle d'authentifier le style. Il s'agit ainsi de restituer au style un travail intellectuel authentique, qui en fait la valeur – du moins, dans le cas du style humain.

2.2 Le style d'un GML : imposture ou légitimité ?

Naturellement, la question se pose de l'application de la stylométrie dans l'authentification de textes générés par GML (Bevendorff, Casals, Chulvi et al. 2024), tout particulièrement dans les productions hybrides où les figures auctoriales sont troubles. Comment séparer le style humain du style du GML lors de l'authentification ? Est-il seulement pertinent de chercher à le faire ? La question de l'authentification est d'autant plus essentielle dans cette approche que l'authentification du texte se veut volontairement tronquée, un des critères permettant d'évaluer un modèle étant sa capacité à reproduire l'être humain tel que l'expose Helm, Priebe et Yang (2023 : 1) :

These efforts are in contrast with for-profit institutions [...] claiming human-level capabilities across a suite of evaluation frameworks [...].

['Ces efforts contrastent avec ceux des institutions à but lucratif [...] affirmant que leurs capacités sont équivalentes à celles d'un être humain dans toute une série de cadres d'évaluation' ; traduit de l'anglais par PM et DeepL]

Puisque les GML semblent voués à émuler les capacités humaines – objectif d'ores et déjà atteint (Else 2023), l'initiative de différencier dans une production hybride le style humain du style généré semble compromise, selon l'efficacité du modèle. Plusieurs travaux proposent des réponses diverses et variées à cette question de l'authentification du contenu généré par GML, tels que Li, Bai et Cheng (2024) qui propose de retracer l'origine du contenu généré jusqu'à la source via l'implantation d'un filigrane, ou encore Bethany, Wherry, Bethany et al. (2024) qui introduit un nouveau système d'authentification du contenu généré artificiellement.

La tâche serait aisée s'il pouvait être démontré que les GML disposaient d'un style propre ; mais les IA génératives sont à même d'adapter leur style et de faire varier les marqueurs stylistiques d'un texte, tel que le rapporte Luther, Kimmerle et Press (2024 : 1358) : « ChatGPT's versatility, adaptability, and ability to mimic different writing styles have made ChatGPT applicable across various domains of writing. » ['La versatilité, l'adaptabilité et la capacité qu'a ChatGPT d'imiter différents styles d'écritures permettent de la mobiliser dans des domaines d'écriture variés' ; traduit par PM].

Là où le style est, pour un être humain, un processus naturel qui découle de la simple volonté d'expression, le style est pour un GML une tâche. Il lui faut à la fois moduler un contexte approprié pour la requête donnée, mobiliser les informations pertinentes et les mettre en forme. La génération du fond comme celle du style se trouvent étroitement liées, là où chez un être humain, la seconde découle de la première ; Lorenzen, Hjuler et Alstrup (2019) conclut ainsi une étude de l'évolution du style d'écriture des élèves au lycée :

One tendency, we saw in all clusters, was that writing style changed more when students start writing more words in their essays [...] writing style changes, when students are pushed out of their comfort zone, i.e. in the end of their assignments, when they write more than what they usually do.

[‘Une tendance que nous avons observée dans tous les groupes était que le style d’écriture changeait davantage lorsque les élèves commençaient à écrire plus de mots dans leurs dissertations [...] le style d’écriture change lorsque les élèves sont poussés hors de leur zone de confort, c’est-à-dire à la fin de leurs devoirs, lorsqu’ils écrivent plus que d’habitude.’]

De fait, les phénomènes mobilisés sont profondément différents rien qu’au niveau de la structure même de la réflexion (avec des guillemets). Si la formation du style d’un être humain se fait au fil de la pratique de l’écriture (et sans doute aussi de la lecture), il n’est pas question de pratique pour un modèle : celui-ci dispose d’une structure statique, modélisée à partir d’un jeu de donnée d’entraînement fixe. Le style d’un GML se forme donc au terme d’un unique processus, puis se décline en fonction du contexte établi par la requête, d’où des variations stylistiques de contexte plutôt que l’évolution d’un style personnel.

Certains travaux entendent toutefois faire partiellement ou exhaustivement état de la caractérisation stylistique des grands modèles de langage, tels que AlAfnan et MohdZuki (2023), Cabanac, Labbé et Magazinov (2021) et Li (2024). Le premier est particulièrement éclairant : l’article examine les caractéristiques stylistiques que sont « sentence length, paragraph structure, word choice, mood, tense, voice, pronouns, keywords density, lexical density, lexical diversity, and reading ease » [‘la longueur des phrases, la structure des paragraphes, les choix de mot, le mode, le temps, la voix, les pronoms, la densité des mots-clefs, la densité du lexique, la diversité du lexique et la lisibilité’ ; traduit par PM] (AlAfnan et MohdZuki 2023: 85). Les résultats de cette étude ont identifié les traits suivants chez ChatGPT-4 (AlAfnan et MohdZuki 2023 : 94) :

- la concision ;
- la structuration ;
- l’assertion ou le questionnement ;
- la voix active ;
- un niveau de langage moyen ;
- le manque de précision ;
- la neutralité de genre.

L’article souligne également que les modèles de détection de textes générés par GML sous-performent par rapport aux résultats annoncés en se basant sur des recherches antérieures (AlAfnan, Dishari, Jovic et al. 2023). Il serait donc de plus en plus difficile de différencier le style humain du style artificiel, y compris dans les productions hybrides. Pour confirmer (ou infirmer) l’authenticité du style dans les productions hybrides, l’approche de la différenciation et du contraste semble dès lors quelque peu entravée.

Notons néanmoins que les grands modèles de langage ne sont pas totalement infaillibles ; l’article AlAfnan et MohdZuki (2023 : 92) relève ainsi parmi les caractéristiques stylistiques notables un certain manque de clarté et d’accessibilité du texte généré quant aux termes employés : « “Exynos” and “Snapdragon chips”

are jargon and technical words. » [‘Les termes “*Exynos*” et “*Snapdragon chips*” sont des mots techniques appartenant à un jargon’ ; traduit par PM]. Pire encore, il arrive au modèle de générer des informations erronées, absentes des documents de son jeu d’entraînement : ces erreurs sont décrites comme des “hallucinations” (‘hallucinations’) (Maynez, Narayan, Bohnet et al. 2020). Une autre étude, celle de Cabanac, Labbé et Magazinov (2021 : 3 et 15), souligne ainsi l’existence d’erreurs de style pures et simples, qu’elle nomme “tortured phrases” (‘phrases torturées’).

Ledit article entend tirer la sonnette d’alarme sur l’utilisation incontrôlée des modèles de langage dans les publications scientifiques. Ils présentent notamment des exemples potentiels d’erreurs liées à l’utilisation possible d’un GML ou d’un modèle de traduction automatique, que je retranscris ci-dessous :

- (1) Terme correct : « deep neural network » [‘réseau neuronal profond’]
Terme incorrect : « profound neural organization » [‘organisation neurale profonde’]
- (2) Terme correct : « artificial intelligence (AI) » [‘intelligence artificielle (IA)’]
Terme incorrect : « (counterfeit | human-made) consciousness » [‘conscience (contrefaite | créée par l’homme)’]
- (3) « **A pamphlet of sickness** or harmed heart valves, ailment, or passing is one of the world’s significant reasons. **Accessible medicines** for patients with a heart valve **are abused**; however, to fix the valve because the fix is incredible, **it have to** supplant a heart valve in the most genuine cases. »
[‘Une brochure de maladie ou valves du coeur endommagées, une affection, ou la mort sont l’une des raisons mondiales les plus significatives. On abuse des médicaments accessibles pour les patients avec une valve du coeur; toutefois, pour réparer la valve car la réparation est incroyable, il falloir supplanter une valve du coeur dans les cas les plus authentiques.’; traduit par PM]

Les exemples procurés par l’article font état d’un vocabulaire mal adapté, que pourrait expliquer l’utilisation d’un logiciel de traduction automatique (ou de génération de texte) remplaçant des termes spécifiques au jargon par des synonymes incohérents ; ainsi que de tournures de phrase tortueuses et de fautes d’orthographe. Le style d’un GML peut donc être lieu d’erreurs, présenter une teinte non-naturelle et étrange.

Enfin, l’article de Li (2024 : §7) aborde la notion très importante d’imitation du style : en faisant générer par ChatGPT un essai dans le style de l’écrivain Agosín, l’auteur cherche à étudier les « dissonances between substance and style in collaborative storytelling with AI » [‘dissonances entre la substance et le style’; traduit par PM] présentes dans les textes générés automatiquement. Elle conclut que si les textes ainsi produits pourraient faire illusion auprès d’un public généraliste, ne connaissant pas en détail le travail de l’écrivain, cette même illusion se défait à l’aune d’une comparaison entre le matériel source et le matériel généré :

Juxtaposing Agosín's essays with the ChatGPT-produced essays could expose the dissonances between style and substance that emerge from the artificial afterlives of Agosín's writings.

[‘Le fait de juxtaposer les essais d’Agosín aux essais produits par ChatGPT suffisait à exposer les dissonances entre le style et la substance qui émergent de l’après-vie artificielle des écrits d’Agosín.’ ; traduit par PM]

Le style d'un GML est encore mal cerné en dépit des efforts fournis à cette fin ; la caractérisation stylistique des modèles est souvent faite au prisme de leur capacité à émuler un style humain, une prise de position débattable. Cette approche est d'autant plus présente dans les questionnements entourant les productions hybrides, le dogme souhaitant que les participations artificielles soient invisibilisées et épousent les contours de la participation humaine.

3 Authentification du style dans les productions hybrides : la complémentarité modèle/être humain

Maintenant que sont posées les bases de ce qui fait le style et son authenticité, tant chez l'être humain que chez les GML, il convient désormais de s'intéresser aux productions hybrides. Celles-ci peuvent se décliner sous plusieurs formes : nous entendrons ici qu'une production hybride est un contenu textuel né de la sollicitation d'un modèle génératif par un être humain, sollicitation qui irait au-delà de la simple rédaction d'une requête unique et préliminaire. Cette définition se veut volontairement assez large, pour englober les différentes variations de ce que pourrait être une production hybride.

3.1 Exemples et catégorisation

Il convient d'établir une catégorisation des productions hybrides afin d'en cerner plus en détail les mécanismes.

L'article Li (2024 : §7) est le lieu d'une production hybride intéressante, car il s'agit là d'un cas où la participation humaine est extrêmement limitée. En générant à la suite plusieurs essais imitant le style d'Agosín, l'autrice interagit avec le modèle afin d'introduire de nouvelles modalités à la requête et d'améliorer ses performances à force de retours et de modifications. Elle conclut que cette méthode de feedback permet d'améliorer les textes générés :

AI models could generate more nuanced responses through iterative prompting and feedback. Significantly, this iteration process offers the potential to deepen, enrich, and complicate AI-generated responses.

[‘Les modèles d’IA pouvaient générer des réponses plus nuances via des requêtes itératives et des retours. Ce procédé itératif a le potentiel significatif d’approfondir, d’enrichir et de complexifier les réponses générées par IA.’ ; traduit par PM]

Elle recommande cette approche, nommée « *human-in-the-loop* », en estimant qu'une telle démarche pourrait également bénéficier à l'être humain, tout particulièrement en lui donnant davantage de contrôle sur la production générée :

Moreover, AI could feed into its own feedback loops, as illustrated by the way prompting ChatGPT using its own interpretation of Agosín's essay led ChatGPT to generate a more nuanced response. The AI-human feedback cycles could thus be mutually enriching as the human writers could shape AI (re)generations.

[‘En outre, l’IA entretenait ses propres boucles de rétroaction, comme démontré par le fait que soumettre des requêtes à ChatGPT en utilisant sa propre interprétation de l’essai d’Agosín la poussait à générer une réponse plus nuancée. Le cycle de retours IA-humain pourraient donc être mutuellement enrichissant alors que les personnes humaines écrivant pourraient moduler la (re)génération des IA.’]

L'article Li (2024 : §7) reconnaît néanmoins à cette approche certaines limites, telles que la circularité des interactions avec le modèle, la superficialité des modifications apportées et l'écueil intellectuel que représentent les instants de stagnation du modèle, lorsqu'il échoue à s'améliorer :

Yet at the same time, the exchange exposes the limitations of the human-AI feedback loop: [...] [a] sense of circularity emerges from the exchange, which illustrates instances in which the dialogue is disrupted, the communication stalled, the revision process stunted. Such moments of fixity obstruct the flow of ideas, the deepening of thought through revision.

[‘Pourtant, dans le même temps, l'échange expose les limites de la boucle de rétroaction humain-IA : [...] [u]ne impression de circularité émergeait de l'échange, illustrant des instances où le dialogue est perturbé, la communication stagnante, le processus de révision stoppé. De tels moments d'immobilité obstruent le flot des idées, l'approfondissement de la pensée par la révision.’]

La méthode proposée par l'article est analogue à une conversation, qui serait ralentie par l'absence d'auto-critique et de progression linéaire du modèle. L'article King et ChatGPT (2023), où ChatGPT est judicieusement cité comme autrice, va plus loin encore dans cette idée en proposant une explication de ChatGPT, par ChatGPT : les deux interlocuteurs sont le chercheur qui soumet des requêtes, et le modèle qui génère du texte en conséquence.

En termes de mots, les contributions de l'auteur constituent 18,5 % de l'article, les 81,5 % restants étant générés par ChatGPT sans modification postérieure. Comment justifier alors que ChatGPT ne soit pas premier auteur ? En dehors des considérations éthiques et philosophiques d'une telle question, le rôle du chercheur dans cette démarche est de diriger le modèle, en procurant directives et thématique. Il est de son ressort, par exemple, de préciser que les références générées par le modèle sont erronées et inexistantes (un exemple d'hallucination). Notons que le style d'une telle démarche est entièrement dominé par les codes inhérents à l'utilisation d'un modèle artificiel, les seules contributions humaines étant des requêtes respectant des contraintes presque cliniques, tant dans leur précision que dans leur ton, l'usage étant destiné ici à la rédaction d'un article de recherche. Cet usage est attendu de la part des modèles et plusieurs études s'intéressent activement à la question d'une production hybride dans le domaine (Huang et Tang 2023 ; Else 2023).

Qu'en est-il, alors, des productions hybrides plus créatives, moins automatiques ? Des textes où la part humaine est davantage engagée et présente ?

L'article Luther, Kimmerle et Press (2024) apporte une réponse satisfaisante : il rapporte une expérience réunissant 135 personnes qui ont utilisé ChatGPT pour rédiger un texte portant sur l'interdiction de l'alcool dans les lieux

publics et sur les risques liés à la consommation d'alcool. Les directives données aux personnes participantes étaient d'une part de fournir des informations sur le sujet donné, d'autre part d'exprimer leur opinion personnelle ; donnant lieu en théorie à une véritable collaboration hybride, ne pouvant se passer d'une contribution humaine. On retrouve dans les résultats de cette entreprise des caractéristiques stylistiques énoncées par AlAfnan et MohdZuki (2023) notamment celle de la faible variation lexicale et de la lisibilité difficile en raison de l'utilisation d'un vocabulaire scientifique.

L'expérience de Luther, Kimmerle et Press (2024 : 1370) a permis de souligner la grande diversité de pratiques observées dans la collaboration avec un modèle génératif :

The texts as products of the writing task varied greatly in length. Moreover, we found great variety among the participants regarding the time they spent on the writing task. [...] we found that the originality of the final texts compared to ChatGPT's responses ranged from 0 to 100% among the participants.

['La longueur des textes en tant que produits de la tâche d'écriture variait grandement. De plus, nous avons observé une grande diversité au niveau du temps passé à écrire par les personnes participantes. [...] nous avons observé que l'originalité des textes finaux comparés aux réponses de ChatGPT allait de 0 à 100 % ' ; traduit par PM]

Les conclusions de l'article exposent ainsi que l'originalité des textes finaux, comparés aux réponses de ChatGPT, variait à un degré extrême entre les différentes participations, de 0 à 100 %. Différentes métriques furent utilisées pour évaluer le degré de chevauchement et de similarité, tant sur le plan stylistique que sur le plan sémantique. L'observation des différentes pratiques des personnes participantes a également révélé que les personnes sollicitant plus fréquemment ChatGPT étaient plus enclines à pratiquer le copier/coller des réponses obtenues (Luther, Kimmerle et Press 2024 : 1373) :

A higher frequency of prompting ChatGPT for complete texts was associated with more copy-paste of content from ChatGPT's responses to the texts, indicating this type of prompting leads to more unchecked use of ChatGPT.

['L'utilisation plus fréquente de requêtes de ChatGPT requérant des textes complets était associée avec davantage de contenu copié/collé de ses réponses dans les textes, indiquant que ce type de prompting entraîne un usage moins contrôlé de ChatGPT.' ; traduit par PM]

L'article propose donc un exemple de productions hybrides très hétérogènes, tout en apportant un début de réflexion et d'analyse sur les différents profils observés lors d'une écriture collaborative avec ChatGPT. C'est justement en raison d'une telle diversité qu'il semble pertinent d'effectuer une catégorisation des différents processus de production hybride.

En revenant à la dichotomie du style précédemment énoncée, portant sur la différenciation entre le premier mouvement stylistique et les ajustements choisis par la suite, il est possible d'y calquer le phénomène de création hybride : le mouvement d'hybridation peut soit aller d'une création humaine à une version altérée par GML, soit d'une création artificielle à une version retouchée par une main humaine. Parfois, les deux options cohabitent avec perméabilité. Dans une optique de caractérisation, il convient également de distinguer le degré

d'implication tant de l'humain que du modèle. En conséquence, la présente propose une catégorisation selon trois axes :

- le matériel d'origine : premier axe qui implique de déterminer si le premier mouvement est d'origine humaine ou artificielle ;
- la direction des altérations : second axe qui implique de déterminer si les ajustements apportés vont du GML à l'humain, de l'humain au GML (en passant ou non par des prompts successifs) ou s'ils sont à double direction ;
- la partition du résultat final : troisième axe qui implique de déterminer les différentes partitions composant le texte final ainsi que leur origine respective.

Ces trois axes peuvent servir à décrire une production hybride en formalisant la notion d'authentification, que ce soit du style ou d'une autre caractéristique textuelle qui devra alors être explicitement définie.

Appliquons cette catégorisation à deux exemples vus précédemment : la production hybride de King et ChatGPT (2023), par exemple, est extrêmement facile à catégoriser. Les différentes sections sont clairement partitionnées selon si c'est le chercheur ou le modèle qui en est l'auteur, le matériel d'origine apparaît clairement en fonction de chaque partition et il semble n'y avoir aucune perméabilité, aucune altération, entre les deux parties. L'authentification du style est extrêmement aisée dans ce cas précis.

La production hybride de Li (2024) est simple également, en apparence : l'essai est rédigé entièrement par le modèle. Le premier mouvement est donc artificiel. Les altérations toutefois complexifient quelque peu la catégorisation : c'est dans une conversation avec le modèle, lequel prend en compte les interactions précédentes, que l'autrice l'amène à modifier son texte. Il y a donc, dans une mesure moindre il est vrai, une altération allant de l'humain au GML. Un écueil semble se présenter alors : si la partition du texte final qui en découle est majoritairement artificielle, il serait erroné d'ignorer l'influence humaine qui l'aurait affectée, troublant les frontières rigides de notre catégorisation.

Il ressort de cette analyse que l'article manque d'informations pour pouvoir appliquer au mieux une telle catégorisation. En effet, cette dernière implique de se baser sur un historique développé des interactions avec le modèle, contenant à la fois les requêtes et les réponses. Il s'agit de faire parler avant tout les usagères et les usagers, d'observer leurs usages des GML et d'établir un standard en matière de traçabilité pour une production hybride ; en l'absence de telles données, qu'elles soient empiriques ou systématiques, le chercheur manque de repères observables et doit se cantonner à un travail descriptif, en espérant pouvoir se référer à l'expérience personnelle de l'utilisateur. Il serait utile d'encourager les personnes à l'origine d'une production hybride à renseigner leur *pipeline* de travail, en adjoignant en libre accès leurs requêtes et les versions antérieures du texte généré et travaillé.

Une telle démarche permettrait d'introduire des indicateurs de la logique suivie par la personne qui fait les requêtes, de la forme que prennent ces requêtes dans le texte même et de l'authenticité du style respectif de l'humain et du GML dans une optique de partition. Se baser sur les différentes itérations successives du

texte est une démarche relevant de la critique génétique : à partir des brouillons, il est possible de tirer des conclusions sur l'évolution progressive du procédé créatif, y compris et surtout lorsque ce procédé est à plusieurs mains. Cette quête de la trace semble tout à fait à propos dans le cadre d'une production hybride, laquelle est généralement retravaillée non pas par segments de texte fragmentés, mais plutôt par séquences de texte intégral. Le type de donnée à disposition pour le travail d'authentification serait l'ensemble des requêtes et les différentes versions des textes produits. Ainsi, des indicateurs explicites pourraient être, pour chacun de nos axes :

- le matériel d'origine : la nature de la première version du texte ;
- la direction des altérations : les requêtes effectuées par l'humain et le résultat fourni en réponse par le GML, qu'il s'agisse de modifications directes ou de suggestions appliquées ensuite.
- la partition du résultat final : un historique des changements du texte établissant nettement l'auteur ou l'autrice de chaque nouveau segment.

Il s'agit d'identifier les paramètres des requêtes et les termes modifiés par le GML. On pourrait par exemple exiger de l'article Li (2024) que soit révélée la façon dont les requêtes de l'autrice sont concrètement mises en œuvre dans la nouvelle version obtenue, quelles sont les modifications induites. Ce traçage aurait en outre l'avantage d'exposer les passages issus uniquement du procédé de génération du GML, qui n'ont jamais été affectés par la critique humaine, tout en conservant la nuance de l'influence humaine par la requête, qui tend à diriger le modèle vers un résultat attendu.

Ainsi, si cette catégorisation constitue une démarche satisfaisante d'un point de vue heuristique, il serait mal avisé de l'appliquer avec trop de rigidité, tout particulièrement le troisième axe. En effet, cette catégorisation est pertinente lorsqu'on l'utilise pour mettre en exergue la complémentarité du travail humain/modèle, et non pas la différenciation.

3.2 Authentifier l'humain pour le différencier d'un GML est-il vraiment pertinent ?

La présente n'entend pas discuter le fait qu'il soit important de pouvoir différencier une production humaine d'une production artificielle (Tang, Chuang et Hu 2024) ; mais le cas des productions hybrides est un cas particulier, qui implique de facto une collaboration et non une simple substitution.

Attardons-nous sur une thèse particulière, celle de Riemer et Peter (2024 : §4.4 et §6). L'article présente les IA génératives, dont les GML, comme des « générateurs de style », qui émulent davantage un concept plutôt que de retranscrire entièrement l'objet à imiter. C'est ainsi qu'ils expliquent les lacunes des IA génératives : pour prendre l'exemple des IA génératives d'image, lorsqu'on leur demande de générer une *hand* ('main'), elle ne générerait alors que des objets tombant dans la catégorie *handness* ('dans le style d'une main'). Ils appliquent également cette approche aux grands modèles de langage : « [f]or text, any prompt that asks the model to produce

something also falls in this category » [‘pour les textes, toute requête qui demande au modèle de produire quelque chose rentre également dans cette catégorie’ ; traduit par PM]. Il faut comprendre de leurs expériences qu’un modèle aura toujours pour objectif de produire, en guise de réponse, un objet textuel dans un style bien précis, que définiront le contenu même du texte et surtout les modalités introduites par la requête : « We put forward the notion of styles as a foundational characteristic describing the nature of generative AI » [‘Nous considérons la notion de styles comme une caractéristique fondamentale de la nature des IA génératives’ ; traduit par PM].

Il faut donc repenser toute interaction avec un grand modèle de langage comme une requête concernant le style, pas seulement le contenu du texte : même une requête à visée informative impliquera que les informations demandées soient partagées sous un format adapté, celui de la concision et de la clarté. Cela explique également les défauts mentionnés plus tôt : les hallucinations, les phrases torturées... Les GML ne génèrent pas une exactitude, seulement un à peu près, qu’il convient ensuite de contrôler et de rectifier à l’aune d’un style humain.

Dès lors, est-il bien pertinent de chercher à authentifier dans une production hybride la part humaine de la part artificielle ? Cela impliquerait de considérer certains marqueurs propres au GML, tels que celui de la concision, comme des marqueurs stylistiques inorganiques et extérieurs au style humain. Mais en considérant que chacun de ces reliquats a été sciemment conservé par la personne humaine, il paraît juste de les authentifier plutôt comme des marqueurs d’un style humain autant qu’artificiel.

En effet, il faut absolument éviter de déresponsabiliser l’être humain dans son usage des GML, par exemple en lui retirant son importance auctoriale au sein des productions hybrides, voire même entièrement artificielles. C’est précisément ce qu’évite l’article King et ChatGPT, où l’auteur signale que les références générées par le modèle sont des hallucinations.

Cette crainte de se voir déposséder intellectuellement de notre œuvre par les GML est déjà fort présente tant dans la communauté de recherche que dans le grand public (Marged 2020). Jusqu’à récemment, il n’était pas encore possible de l’infirmier ou de la confirmer ; deux études datant respectivement de 2023 (Bai, Liu et Su) et de 2025 (Kosmyna, Hauptmann, Yuan et al.) nous procurent désormais des éléments de réponse.

Kosmyna, Hauptmann, Yuan Ye Tong et al. (2025 : 143) fait état de résultats très pessimistes sur les effets cognitifs de l’utilisation des GML, notant ainsi lors d’une expérience que les personnes faisant appel à un grand modèle de langage tendaient à être moins engagées intellectuellement dans leur propre œuvre, et à subir le contrecoup d’une « dette cognitive ». Ils soulignent dans le même temps que cela donnait au modèle une plus grande importance, laquelle pouvait être instrumentalisée :

[T]his convenience came at a cognitive cost, diminishing users' inclination to critically evaluate the LLM's output or "opinions" (probabilistic answers based on the training datasets). This highlights a concerning evolution of the 'echo chamber' effect: [...] what is ranked as "top" is ultimately influenced by the priorities of the LLM's shareholders.

['Cette commodité s'est accompagnée d'un coût cognitif, diminuant la propension des utilisateurs à évaluer de manière critique les résultats ou les « opinions » du GML (réponses probabilistes basées sur les ensembles de données d'entraînement). Cela met en évidence une évolution préoccupante de l'effet « chambre d'écho » : [...] ce qui est le plus repris est en fin de compte influencé par les priorités des actionnaires du GML.']

Cette découverte ne doit pas nécessairement décourager tout usage des GML ; en revanche, elle doit inviter au questionnement et à la mise en place de bonnes pratiques, essentielles pour qualifier une production hybride. Bai, Liu et Su (2023) propose ainsi des solutions potentielles dont celles qui suivent :

- des approches d'apprentissage mixtes, qui font du modèle un auxiliaire, une ressource supplémentaire, plutôt que de lui donner un rôle principal tel que celui d'éduquer ou de rédiger ;
- la promotion d'un environnement d'apprentissage collaboratif, reposant plutôt sur un modèle de pair-à-pair que sur un GML ;
- le développement de la pensée critique et de la résolution de problème, notamment par la critique et la vérification de réponses générées par ChatGPT.

Toutes ces options soulignent la nécessité d'envisager les GML comme complémentaire lors d'une production hybride, plutôt que d'y voir une substitution du travail humain ; surtout, il convient de systématiquement mettre à l'épreuve les participations du modèle, d'engager une réflexion permanente sur les résultats produits par celui-ci. Pour produire une œuvre, il faut un engagement intellectuel de la figure auctoriale ; cette vérité s'applique également aux productions hybrides.

Pour s'assurer ainsi de la participation humaine, nous revenons aux indicateurs exposés tantôt qui tiennent de l'archivage méthodique des interactions avec le modèle. Ceux-ci sont en outre les témoins d'un engagement de la part de l'auteur humain, un engagement intellectuel au sein de la production hybride mais aussi un engagement pour la transparence. La nature même des GML impliquant un degré d'inaccessibilité des informations – comme les données d'entraînement ou les mécanismes sous-jacents à la génération – il revient aux créateurs, tant des GML que des productions hybrides, d'adjoindre à leurs productions des métadonnées à même de renseigner sur les conditions de créations et les différentes modalités d'existence desdites productions (Mitchell, Wu, Zaldivar et al. 2019).

Là survient une condition *sine qua non* de la définition même d'une production hybride : une production hybride ne saurait l'être sans un rétrocontrôle attentif d'une figure auctoriale humaine, laquelle doit entériner chaque aspect du texte généré dans la version finale de la production. Il ne suffit pas de générer un style à l'aide d'un GML pour créer une œuvre, il convient ensuite de la transformer en un objet à part entière, en lui adjoignant à la fois matière et forme humaines. Au sein d'une production hybride, les styles humains et artificiels doivent se mêler par

couche superposées et par transferts, pour obtenir un résultat final qui, somme toute, s'avère fondamentalement humain. C'est en encourageant la transparence et le partage des données que la complémentarité humain/GML pourra se réaliser et être formalisée dans un but de recherche, en évitant tout raccourci et toute pratique non-documentée.

Conflicts of Interest

L'auteur déclare n'avoir aucun conflit d'intérêts concernant la publication de cette contribution.

Bibliographie

- AlAfnan, Mohammad Awad & Dishari, Samira & Jovic, Marina & Lomidze, Koba. 2023. ChatGPT as an educational tool: opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology* 3(2). 60–68. <https://doi.org/10.37965/jait.2023.0184>
- AlAfnan, Mohammad Awad & MohdZuki, Siti Fatimah 2023. Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial Intelligence and Technology* 3(3). 85–94. <https://doi.org/10.37965/jait.2023.0267>
- Bai, Long & Liu, Xiangfei & Su, Jiacan. 2023. ChatGPT: the cognitive effects on learning and memory. *Brain-X* 1(3). 30–32. <https://doi.org/10.1002/brx2.30>
- Bethany, Mazal & Wherry, Brandon & Bethany, Emet & Vishwamitra, Nishant & Rios, Anthony & Najafirad, Peyman. 2024. *Deciphering textual authenticity: A generalized strategy through the lens of large language semantics for detecting human vs. machine-generated text*. 5805–5822. <https://www.usenix.org/conference/usenixsecurity24/presentation/bethany> (dernier accès 03/02/2026).
- Bevendorff, Janek & Casals, Xavier Bonet & Chulvi, Beta & Dementieva, Daryna & Elnagar, Ashaf & Freitag, Dayne & Fröbe, Maik & Korenčić, Damir & Mayerl, Maximilian & Mukherjee, Animesh et al. 2024. Overview of PAN 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification. In Goharian, Nazli & Tonello, Nicola & He, Yulan & Lipani, Aldo & McDonald, Graham & Macdonald, Craig & Ounis, Iadh (éds), *Advances in Information Retrieval*, 3–10. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56072-9_1
- Cabanac, Guillaume & Labbé, Cyril & Magazinov, Alexander. 2021. Tortured phrases: a dubious writing style emerging in science. Evidence of critical issues affecting established journals. *arXiv*. 1–27. <https://doi.org/10.48550/arXiv.2107.06751>
- Cafiero, Florian & Camps, Jean-Baptiste & Gabay, Simon. 2023. Louise Labé: une créature de papier? *Humanistica* 2023. <https://hal.science/hal-04090284>

- Crawford, Joseph & Cowling, Michael & Ashton-Hay, Sally & Kelder, Jo-Anne & Middleton, Rebekkah & Wilson, Gail S. 2023. Artificial intelligence and authorship editor policy: ChatGPT, Bard Bing AI, and beyond. *Journal of University Teaching and Learning Practice* 20(5). 1–11. <https://doi.org/10.53761/1.20.5.01>
- Delcourt, Christian. 2002. Stylometry. *Revue belge de Philologie et d'Histoire* 80(3). 979–1002. <https://doi.org/10.3406/rbph.2002.4651>
- Dürrenmatt, Jacques. 2010. «Le style est l'homme même». Destin d'une buffonnerie à l'époque romantique. *Romantisme* 148(2). 63–76. <https://doi.org/10.3917/rom.148.0063>
- Else, Holly. 2023. Abstracts written by ChatGPT fool scientists. *Nature* 613(7944). 423–423. <https://doi.org/10.1038/d41586-023-00056-7>
- (Des) Essarts, Nicolas-Toussaint. 1800. *Les siècles littéraires de la France*. Chez l'Auteur Place de l'Odéon.
- Helm, Hayden & Priebe, Carey E. & Yang, Weiwei. 2023. A statistical Turing test for generative models. *arXiv*. 1–14. <https://doi.org/10.48550/arXiv.2309.08913>
- Huang, Jingshan & Tan, Ming. 2023. The role of ChatGPT in scientific communication: writing better scientific review articles. *American Journal of Cancer Research* 13(4). 1148–1154. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10164801/> (dernier accès 03/02/2026)
- Iqbal, Farkhund & Debbabi, Mourad & Fung, Benjamin C. M. 2020. *Machine Learning for Authorship Attribution and Cyber Forensics*. Cham: Springer Nature Switzerland.
- Ji, Ziwei & Lee, Nayeon & Frieske, Rita & Yu, Tiezheng & Su, Dan & Xu, Yan & Ishii, Etsuko & Bang, Ye Jin & Madotto, Andrea & Fung, Pascale. 2023. Survey of hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12). 1–38. <https://doi.org/10.1145/3571730>
- King, Michael R. & chatGPT. 2023. A conversation on Artificial Intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering* 16(1). 1–2. <https://doi.org/10.1007/s12195-022-00754-8>
- Kosmyna, Nataliya & Hauptmann, Eugene & Yuan, Ye Tong & Situ, Jessica & Liao, Xian-Hao & Beresnitzky, Ashly Vivian & Braunstein, Iris & Maes, Pattie. 2025. Your brain on ChatGPT: accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv*. 1–216. <https://doi.org/10.48550/arXiv.2506.08872>
- Li, Liying & Bai, Yihan & Cheng, Minhao. 2024. Where am I from? Identifying origin of LLM-generated content. In Al-Onaizan, Yaser & Bansal, Mohit & Chen, Yun-Nung (eds), *Proceedings of the 2024 conference on empirical methods in Natural Language Processing* (Miami, November 12-16, 2024), 12218–12229. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.681>

- Li, Ruth. 2024. A “dance of storytelling”: dissonances between substance and style in collaborative storytelling with AI. *Computers and Composition* 71. 1–10. <https://doi.org/10.1016/j.compcom.2024.102825>
- Lindholm, Charles. 2013. The rise of expressive authenticity. *Anthropological Quarterly* 86(2). 361–395. <https://www.jstor.org/stable/41857330> (dernier accès le 03/02/2026).
- Lorenzen, Stephan & Hjuler, Niklas & Alstrup, Stephen. 2019. Investigating writing style development in high school. *arXiv*. <https://doi.org/10.48550/arXiv.1906.03072>
- Luther, Teresa & Kimmerle, Joachim & Cress, Ulrike. 2024. Teaming up with an AI: exploring human–AI collaboration in a writing scenario with ChatGPT. *AI* 5(3). 1357–1376. <https://doi.org/10.3390/ai5030065>
- Marged, Barry. 2020. Are computers and AI prompting us to think less. *The Journal of Family Practice* 69(2). 64. <https://pubmed.ncbi.nlm.nih.gov/32182295/> (dernier accès le 07/05/2026).
- Maynez, Joshua & Narayan, Shashi & Bohnet, Bernd & McDonald, Ryan. 2020. On faithfulness and factuality in abstractive summarization. In Jurafsky, Dan & Chai, Joyce & Schluter, Natalie & Tetreault, Joel (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 5-10, 2020). 1906–1919. Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.173/> (dernier accès le 03/02/2026)
- Mitchell, Margaret & Wu, Simone & Zaldivar, Andrew & Barnes, Parker & Vasserman, Lucy & Hutchinson, Ben & Spitzer, Elena & Raji, Inioluwa Deborah & Gebru, Timnit. 2019. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, January 29-31, 2019), 220–229. <https://doi.org/10.1145/3287560.3287596>
- Ramrial, Hoshiladevi & Panchoo, Shireen & Pudaruth, Sameerchand. 2016. Authorship attribution using stylometry and machine learning techniques. In Berretti, Stefano & Thampi, Sabu M. & Srivastava, Praveen Ranjan (eds), *Intelligent Systems Technologies and Applications*, 113–125. Springer International Publishing. https://doi.org/10.1007/978-3-319-23036-8_10
- Reggiani, Christelle. 2002. *Le style comme indice : le postulat de l'unité*. https://www.fabula.org/ressources/atelier/?Le_style_comme_indice%3A_le_postulat_de_l%27unit%26eacute%3B (dernier accès le 05/02/2026).
- Riemer, Kai & Peter, Sandra. 2024. Conceptualizing generative AI as style engines : Application archetypes and implications. *International Journal of Information Management* 79. 1–15. <https://doi.org/10.1016/j.ijinfomgt.2024.102824>
- Tang, Ruixiang & Chuang, Yu-Neng & Hu, Xia. 2024. The science of detecting LLM-generated text. *Communications of the ACM* 67(4). 50–59. <https://doi.org/10.1145/3624725>
- Van Leeuwen, Theo. 2001. What is authenticity? *Discourse Studies* 3(4). 392–397. <https://doi.org/10.1177/1461445601003004003>

AI imitating individuals: deepfakes and the Unexpected in digital communication

Nina Kalwa (University of Tübingen)

nina.kalwa(at)uni-tuebingen.de

Abstract

This paper considers language-generating artificial intelligence as a form of human imitation. Focusing on deepfakes that imitate individuals' appearance and language, the paper examines how supposed deepfakes featuring prominent individuals are discussed on social media and how discourse around deepfakes has evolved alongside growing awareness of their existence. Furthermore, the paper discusses the social implications of these developments. Three case studies were selected in which deepfakes were posted and discussed on social media. By examining the comments on the videos using detailed text analysis, the paper demonstrates how identifying something unexpected can lead to videos and audio clips posted on social media being judged as fake. The paper discusses why this can be considered dangerous for society.

Keywords

artificial intelligence, deepfakes, common ground, digital communication, social media

1 Introduction

Generative artificial intelligence is shaping today's societies and can be found in many contexts of our daily lives. Discourses on artificial intelligence are transforming our understanding of reality (Kalwa 2025). These new technologies are also changing our communicative practices: AI-generated texts are reshaping public discourse and threatening democracies (cf. Monti 2024). Writing practices change, for example, when ChatGPT is understood as a co-creator in the writing of texts (cf. Steinhoff 2023). AI allows us to break down the boundaries of communication, for example, when people chat with the dead using thanatobots (cf. Henrickson 2023). The mere fact that it is no longer only humans who use language is changing communication. As Gottschling and Kramer (2025) point out, generative AI is "changing concepts of reality, authorship, and common ground."

This article focuses on deepfakes, a specific form of generative artificial intelligence that also involves language. Deepfakes shared on social media often feature well-known people. Users of social media can be confused when these imitations of people (which are not always recognised as such) behave in a way that is unusual for their originals. I am interested in how deepfakes are discussed on social media. One of my aims is to demonstrate that, over time, and as people become more aware of deepfakes, social media users seem to be more likely to categorise a video as a deepfake if it does not meet their expectations. In this paper, I analyse social media users' discourses about deepfakes by combining a semiotic and a discourse linguistic perspective. I want to examine the following research questions:

1. How do social media users discuss and evaluate alleged deepfakes?
2. How has the discourse around deepfakes evolved?
3. What implications for society can we infer from these discussions?

In the following, I will first relate the concept of imitation to the phenomenon of artificial intelligence, focusing on AI systems that are of particular interest to linguists—those which produce language. Finally, I focus on deepfakes that imitate individual human speech. I selected three case studies, which I analyse in Section 3, showing that social media users employ to engage with suspected deepfakes and how this leads to growing scepticism towards digital communication. In Section 5, I examine the broader social implications of the growing scepticism in digital communication.

2 AI imitating humans, Deepfakes imitating individuals

In 1950, Alan Turing posed the question of whether machines could think and developed the idea of a learning computer. Turing simulated machine thought in an “imitation game” (Turing 1950: 433), later called the Turing Test: if a person communicating with a computer does not recognise that it is a machine, then the machine is successfully imitating human behaviour. This concept of *imitation* is still observable in some scientific definitions of AI today, as in Yildirim et al. (2021: 33), introductory note:

For this paper, AI will be defined as a field of computer science that specializes in creating computer systems that are designed to interact with the world and can mimic and operate as human beings by utilizing machine learning.

In their definition, Yildirim et al. emphasise the trait of human likeness of computer systems in the context of AI.¹ Overall, the language of AI research seems to be permeated by anthropomorphising metaphors, as expressions such as ‘reasoning’, ‘training’, ‘consciousness’, and ‘neural networks’ show. Ultimately, the attribution of agency already leads to the anthropomorphisation of machines (cf. Schulz-Schaeffer 2008). In consequence, if speaking is understood as acting, then the attribution of speech already leads to the anthropomorphisation of AI.

The conceptualisation of generative AI as human-like entities is pervasive in both scientific and public discourses. Spieß (2024) illustrates this conceptualisation in public discourse, using a sample of 22 media texts from the debate series of *Süddeutsche Zeitung* and 16 debate speeches from the 42nd session of the Bundestag during the 19th legislative period (point 5 on the agenda) (cf.

¹ Many researchers, such as Salles, Evers and Farisco (2020), are critical of the anthropomorphisation of AI. Nevertheless, there are many scientific papers on AI in which authors anthropomorphise AI because this metaphor seems deeply rooted in the language of computer scientists.

Spieß 2024: 370). Spieß describes how these texts portray artificial intelligence in an anthropomorphic manner, attributing human-like capabilities to AI such as learning, understanding or cheating, which it presumably often handles more effectively than humans (cf. Spieß 2024: 371ff.).

Because the idea of human-likeness is omnipresent in the scientific and non-scientific discourse on AI, I will apply the concept of *imitation* to generative artificial intelligence, as I assume that some forms of generative AI *imitate* human linguistic behaviour and thus, as Gottschling (2023: 4) explains, create seemingly meaningful, persuasive or aesthetic texts (see also Kalwa 2026). Unlike Gottschling (2023: 14), I view imitation as a *social phenomenon* (Schwedler and Sonntag 2021: 9) and a *pervasive cultural technique* (Schwedler and Sonntag 2021: 15). This means that I do not see it as a specific rhetorical device (see for example Kaminski 1998), but as something that occurs in very different contexts and is omnipresent in human communication: children imitate their parents, for example, or people imitate the communication practices of others at higher career levels. According to Schwedler and Sonntag (2021: 1), imitation is a cultural phenomenon that has spanned the ages. They argue that it is a prerequisite for every culture, facilitating the transfer of knowledge and patterns of behaviour to the next generation and forming an integral part of identity development. Schwedler and Sonntag (2021: 1) continue by stating that the complex phenomenon of imitation affects central aspects of human behaviour. The authors express surprise that this perspective has not been adopted more often as a starting point for comprehensive attempts at explanation.

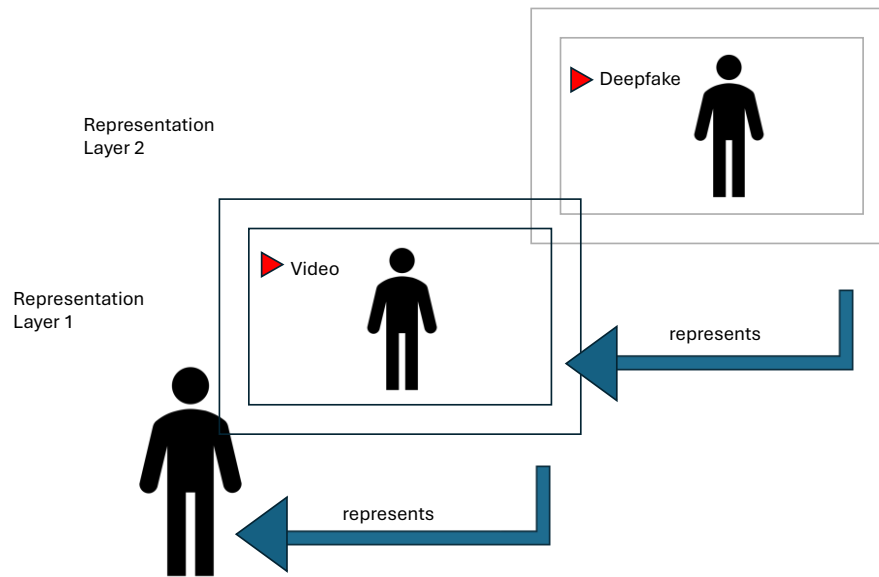
By framing AI as imitations, AI systems that we encounter in social media can be seen as depictions of social agents, as described by Clark and Fischer (2023) for social robots. The functions that users ascribe to these imitations vary. Several studies in the field of economics notice an increase in trust in the machine when it acts more human-like (see e.g. Fenwick and Molnar 2022; Peng, Yaobin and Yeming 2021) and common descriptions of some AI-driven machines suggest that they can mimic a wide range of human communicative behaviours: Alexa as a reliable servant (Dickel and Schmidt-Jüngst 2021: 361), ChatGPT as a supportive friend,² the chatbot Replika as a friend, sexual partner or even life partner.

Deepfakes can be seen as special forms of human imitations. Deepfakes denote a category of synthetic audiovisual media involving the manipulation or synthetic generation of images, videos, and/or audio, typically using AI technologies (Pawelec and Bieß 2021: 15). The term *deepfake* refers “to all digital fake content created by means of deep learning techniques” (Tolosana et al. 2022: 4). Although the word *deepfake* includes the term *fake*, people do not only generate deepfakes for deceiving and manipulating but also “for entertainment and creative arts” (Altuncu, Franqueira and Li 2024: 2). According to Kietzmann et al. (2020:

² ChatGPT has assigned itself this role by request. My prompt: *If you had to assign yourself a social role, what would it be?*.

136), deepfake “gained its name from an anonymous user of the platform Reddit, who went by the name ‘deepfakes’ (deep learning + fakes) and who shared the first deepfakes by placing unknowing celebrities into adult video clips.” From the beginning, deepfakes have often targeted famous individuals, mimicking their unique speech patterns. The techniques for creating deepfakes are constantly evolving, but as the name suggests, they primarily rely on deep learning methods. Deep learning is a subset of machine learning that utilises multilayer neural networks, known as deep neural networks (DNNs), to “mimic the human brain’s neurons” (Taye 2023: 11) – so once again, the concept of imitation is inherent. Kietzmann et al. (2020: 137) point out that the majority of deepfakes in the visual domain at that time followed “a procedure in which the real face of a person is exchanged with a fake image showing somebody else.” An autoencoder is then trained to identify key facial features and reconstruct input images as its output. In essence, the autoencoder goes beyond the given image data and develops a generative model of a person’s face (cf. Kietzmann et al. 2020: 139). Researchers from a variety of disciplines are discussing the dangers posed by deepfakes (for an overview, see Bode, Lees and Golding 2021). Chesney and Citron (2019) discuss deepfakes as a challenge to privacy, democracy, and national security. Kikerpill (2020) describes “the rise of deepfake designer porn”. Vaccari and Chadwick (2020) explore the impact of “synthetic political video on deception, uncertainty, and trust in news”. To date, there has been no study investigating how deepfakes may disrupt digital communication.

First and foremost, compared to other forms of artificial intelligence, deepfakes – if they show people – primarily replicate an individual’s appearance, movements, and speech patterns. More precisely, and this seems very important for the analysis in Section 4, deepfakes indeed do not imitate individuals, but rather audios and videos representing the appearance, movements or speech patterns of individuals, and thus representations of individuals. As Magritte did with his depiction of a pipe signed ‘*Ceci n’est pas une pipe*’ (cf. Foucault 1973), we should be aware of the representational character of videos and audio. Thus, well-known personalities depicted in videos and audio are representations of the respective originals; voice and body representations then refer indexically to the original (see also Klug 2015: 508f. on the indexicality of pictorial signs). Particularly in communication that is not characterised by the participants being physically present in the same place at the same time (Hausendorf 2020: 86), linguistic signs refer to certain speakers. When I refer to deepfakes as imitations, it is important to add another layer by understanding them as imitations of representations, which is essential for following the analysis in Section 4. This perspective is illustrated in Picture 1.



Picture 1: Deepfakes as representations of representations.

Picture 1 makes clear that deepfakes do not imitate a person directly, but rather the representation of a person. In this sense, deepfakes emerge as a representation of a representation.

Certain elements of a video, such as voice, gestures, as well as elements of the body, function as signs and refer to a specific person in the real world. Thus, I understand videos as complex signs. In social media, for example, people encounter videos, and it is not always clear whether these videos show representations of real people or deepfakes. When judging whether a video depicts a real person or a deepfake, social media users rely on certain indicators. So, they also interpret videos as complex signs and take individual elements as evidence for the authenticity or non-authenticity of a video. The fact that social media users are familiar with the originals and know how, for example, the German Chancellor Friedrich Merz *usually* behaves, plays an important role in how they rate the videos.

In the following, I will use the example of alleged deepfakes to investigate how people communicatively react to these alleged imitations, how they evaluate them, and which aspects they consider relevant for their evaluation.

3 Evaluating the individual's imitation

The analysis focuses on how users engage with suspected deepfakes in social media comments. I have selected three case studies, all of which originate from social media — the primary arena where deepfakes tend to circulate.³ Each case involves a social media post featuring a video that has prompted at least some users to

³ The data used is social media data that cannot be made available in a repository for license reasons. The data are, however, available from the author upon reasonable request.

suspect it might be a deepfake. The first example is a video originally posted on TikTok in February 2021. TikTok is an online video platform centred around creating and sharing short-form videos (cf. Ackermann and Dewitz 2020: 72). Users can engage with these videos by leaving comments. Before 2022, TikTok's comment character limit was a few hundred characters, later increasing to several thousand. The second video was posted on X (formerly Twitter) in November 2023. Like other platforms, X imposes a character limit on comments — in this case, 280 characters. The third video appeared on Instagram in March 2023. While Instagram allows significantly longer comments — up to 2,200 characters — users typically write only brief remarks in response to the videos. Unlike the first two videos, which are confirmed deepfakes, the Instagram video was later revealed not to be a deepfake. Concerning the three case studies, I included all comments accessible to me (including replies) in my analysis. For Case 1, this comprised 1,163 comments (although it should be emphasised that TikTok only granted me access to a fraction of all comments; see Section 3.1). For Case 2, there were 577 comments, and for Case 3, 264 comments.

I was particularly interested in the aspects of the videos that users highlighted in their comments to argue either for or against the assumption that the content was a deepfake. To explore this, I analysed the comments using qualitative text analysis methods. A central theoretical concept here is *hermeneutics* — specifically, the notion of *interpretation* (van Dijk 2011: 612). Following Gardt (2017: 487), I understand interpretation as the process of uncovering the meaning of a text or a multimodal network of signs. Interpretation is thus conceived as a constructive act (Gardt 2017: 491), in which meanings derived from texts are not inherent to the texts themselves but emerge through the interpreter's engagement. Assuming—as Gardt (2017: 491) emphasizes—that texts are not containers of meaning has implications for the reproducibility of the analysis: another analyst is unlikely to arrive at exactly the same observations. At the same time, however, the results of the analysis are not arbitrary; they are methodologically guided, aiming to consciously shape the normally automatic process of understanding through linguistic interpretation, thereby making it controllable (cf. Gardt 2017: 494). Linguistic interpretation is based on detailed analysis, assuming that all parts of the text may contribute to meaning-making. In doing so, this study contributes to a better understanding of which aspects of videos social media users consider relevant when deciding whether a video is a deepfake or not. The study also sheds light on the effects of deepfakes on digital communication.

Before presenting three case analyses and then discussing the results, it is important to emphasise that the analysed social media postings with embedded videos and audio came across at different times. The first case is a video on TikTok that was posted in February 2021, the second case is a posting on X published in November 2023, and the third case is a video posted on Instagram published in March 2023. There is a time gap of more than two years between the first case and the other two, which is significant for the analysis.

The three analysed cases are not comparable in many ways: Each posting was published on a different social medium with different practices. They are also written in different languages. The communicative framing of the videos is also different: while in one case the post already points out that the sound in the video is a machine-generated recording, the other two posts leave it open whether the content is fake. However, the analysis will illustrate that in 2023, social media users will use very different criteria than in 2021 when assessing whether they have encountered a deepfake.

3.1 Tom Cruise on TikTok?

Since 2021, an account called *deeptomcruise* has been regularly publishing videos of actor Tom Cruise on the social media platform TikTok using deep learning techniques.⁴ The first video, entitled *Sports*, was published in February 2021, when the account, which now has more than 5 million followers, was little known.⁵ The video allegedly features Tom Cruise. He addresses TikTok users directly and then demonstrates a golf swing. He kneels again to tell “sports and TikTok fans” to wait and “see what is coming next” (see the video in footnote 5). In the comment section of the video, some users question whether the person shown is the real Tom Cruise. This uncertainty prompts them to search for evidence confirming or debunking the video’s authenticity.



Picture 2: Screenshot of a video from the TikTok account deeptomcruise (link in footnote 5).

According to TikTok, there are about 36,500 comments on the video, which I originally wanted to analyse using corpus linguistics methods. However, TikTok

⁴ <https://www.tiktok.com/@deeptomcruise> (last accessed on 14/04/2026).

⁵ <https://www.tiktok.com/@deeptomcruise/video/6932166297996233989?lang=de-DE> (last accessed on 14/04/2026).

does not support the scientific analysis of social media posts and does not provide an API for downloading data for research purposes.⁶ I therefore manually copied the comments down from the page. However, this only gave me a fraction of the allegedly published comments, which are also algorithmically tailored to my person. This way, I could only access 1163 comments. I examined these comments using qualitative linguistic methods. That means I made a detailed text analysis where every part of the text is considered potentially involved in meaning-making (cf. Kalwa 2019: 20). This allowed me to identify the linguistic and contextual features that people use to determine whether the person in the video is real or fake.

Looking at the comments, I identified three types of perspectives that emerged among commenters: Some users are convinced that it is the real Tom Cruise (1) and assume that a film crew is portraying the real Tom Cruise on a golf course. Others suspect it is an actor impersonating Tom Cruise. The commenter of example (2) believes that an actor is imitating Tom Cruise and that the posted video represents this actor, which is a representation of an imitation because, in this case, the imitation is perceived as occurring in the so-called real world. Finally, some suspect that the TikTok film is a deepfake. The commenter of (3) assumes that the imitation is not Tom Cruise himself or an actor, but a representation of him, namely, a deepfake or an imitation of a representation.⁷

- (1) It's obviously the real Tom.
- (2) Stunt double.
- (3) Deepfake is scary.

In the comments, there is a comparison between the alleged imitation, which not all TikTok users are convinced is an imitation, and the original. Users cite various aspects of the video to prove that it is (not) the real Tom Cruise represented in the video or that the video is fake. To determine whether they are dealing with a representation of an imitation (an actor), an imitation of a representation (a

⁶ Although the company has offered the option to create a researcher account since June 2024, the rights that are ceded to TikTok are contrary to the scientific ethos and unacceptable, at least to me. In return, you are given very limited rights to use and cite the data you collect, which again is at odds with making the results of the analysis understandable. If I want to create a TikTok research account, I must first create a TikTok for Developers account and agree to the TikTok for Developers Terms of Service. Here I have to agree that TikTok may use my “name, logo(s), or other identifying information or image in case studies, testimonials, examples”, and to showcase my “use of our TikTok Developer Services in general.” In return, I am not allowed to “make any public statement” regarding my “relationship with TikTok or access to the TikTok Developer Services without TikTok’s prior written consent” (<https://www.tiktok.com/legal/page/global/tik-tok-developer-terms-of-service/en>, last accessed on 14/04/2026).

⁷ For reasons of research ethics, I have refrained from identifying the exact sources of the comments. In this article, I am not correcting typos in the original comments. Also, names of commenters are omitted for reasons of privacy.

deepfake), or a representation of the original, users draw on different types of knowledge in their communication: (a) knowledge of Tom Cruise's physical attributes, (b) knowledge about Cruise's usual behaviour, and (c) knowledge about TikTok's surface and technical knowledge. I introduce several comments, which are subsequently discussed.

(a) Knowledge of Tom Cruise's physical attributes

- (4) It's not him his voice is not the same it's a deep fake
- (5) different voice
- (6) Nah he's too tall to be Tom cruise
- (7) I just don't think you can fake the facial grooves and dimples. I think this is the real deal!
- (8) It's him for sure. Tom has a middle tooth directly online with his nose. That's the give away
- (9) there is a guy on here that looks just like Tom Cruise but he's obviously much younger.

Users interpret different aspects of the represented body as signs that refer indexically to the original, a representation of an imitator, or an imitation of a representation. In individual examples where users suspect an imitation, it is not always clear whether they are assuming a human or a technical imitation. As in 4 and 5, several users note that the voice in the video is different from that of the real Tom Cruise that they have encountered in mainstream media. While example 4 indicates the assumption of a technical imitation, the form of imitation stays unclear in example 5. Both focus on the sound sequence as a sign to determine whether it is the real Tom Cruise. Another user comments on the person's height as an indication that the video is fake (6). Yet others, however, identify small details in the face as proof that the video shows the original. For example, the user in 7 is convinced that certain aspects, such as the dimples and furrows on the face, cannot be faked. There are many more commenters who, as in 8, look closely at the teeth and the position of the teeth of the person depicted. Many users note that the person portrayed is younger than Tom Cruise and see this as an indication that it is an impersonation (9). In example (9), the user refers to a "guy on here". This statement suggests that the user assumes that another person is imitating Tom Cruise. Overall, the difference in Tom Cruise's appearance leads to the conclusion that it is fake. But the user obviously also has knowledge about the usual behaviour of Tom Cruise.

(b) Knowledge about Cruise's usual behaviour

- (10) Tom Cruise posting on tik tok? Sus
- (11) Not enough over the top acting to be Tom Cruise
- (12) + Scientology does not allow there members to be on Tic Tok and again this very cool look alike is not as cut as Tom

Users interpret different aspects of the represented behaviour as signs that refer indexically to an imitation, a human or an AI imitation. Probably jokingly, the person in (10) says that there is too little overacting for the real Tom Cruise. In this case, it is the represented action that indicates to the user that it is an imitation, although it is left open whether the commenter is assuming an actor to imitate or a technological imitation. In (11), the commentator is questioning whether Tom Cruise is likely to post a TikTok video given his usual activities. In this context, commenters interpret missing elements as signs of imitation. The connection to Scientology, to which Tom Cruise belongs, and the presence of a TikTok account also create contradictions for users (12). However, the phrase *very cool look alike* leads to the interpretation that the commenter assumes an actor is imitating Tom Cruise.

(c) Knowledge about TikTok's surface and technical knowledge

Users also draw on their knowledge of the video-hosting service TikTok to determine whether the person in the video is the real Tom Cruise or an imitation. Furthermore, users cite their technical knowledge as evidence that the video is a deepfake – and that it is not:

- (13) Is it!!! Where's his blue check mark????
- (14) This video is a deepfake. The poster's name is deeptomcruise. Pay attention.
- (15) Deep fake does not work putting on and off glasses, the edit would be tremendous job. And can't fake the cray cray in his eyes!
- (16) Look up deep fake. This is nicely done, I agree. But if you look at his face when he's standing up and walking backwards you can see it shifting.

The user in 13 references their knowledge of TikTok's celebrity status, which is typically verified by the company with a blue tick. Comment 14 refers to the account name. The user knows that video creators assign themselves names and deduces from the chosen name that the video must be a deepfake. The total number

of comments analysed shows that in 2021, only a few users were aware of the deepfakes phenomenon, as deepfakes are rarely mentioned, and some users state that imitation technology does not work with certain movements or details (15). In this case, the commenters interpret elements of the video as evidence that it is not an imitation of a representation, but that the video was recorded in the real world. Others, on the other hand, recognise from the ways that the person in the film is shifting that it is a technologically created image (16) or an imitation of a representation.

In summary, users in the comment section rely on details from the video, their background knowledge, and their familiarity with TikTok's visual interface (Schmitz 2011) – including their awareness of technological possibilities to assess communicatively whether the video is a deepfake. The commenters interpret both the signs that refer to Tom Cruise as a person and those that refer to certain aspects of the videos as filmed by real people or artificially created videos. By responding directly to each other or simply by telling others about their confusion, they try to create shared knowledge and intersubjectivity in the interaction process (cf. Deppermann 2018: 118). Their uncertainty leads commenters to look for indicators in the video that suggest it is a fake or a real video. They interpret individual elements of the video as signs that indexically point to a fake or a representation of the original.

What the analysis shows is that commenters in 2021 still have little experience with deepfakes. They discuss individual details of the video and the context and try to determine through intersubjective exchange whether they are dealing with a deepfake or not. Users draw on different types of knowledge in their communication when arguing that they encounter a deepfake or a depiction of a real person. My further analyses will show that, over time, physical and technical aspects play a subordinate role in the assessment of whether one encounters a deepfake or not.

3.2 The *Tagesschau* admits its guilt?

On 6 November 2023, during a Monday demonstration of Pegida (Patriotic Europeans against the Islamisation of the West) in Dresden, some people played an audio recording. In the audio recording, the voice of the *Tagesschau* presenter Jens Riewa is synthesised. Interpreted as a complex sign, the music in the intro and outro refers indexically to the *Tagesschau*, a German news program, as do the welcome and farewell sequences. Finally, the timbre and intonation of the voice audio refer to Jens Riewa, an announcer for the *Tagesschau*. Within the sound sequence, the speech act of apology is performed. Among other things, the voice apologises for the fact that the *Tagesschau* has allegedly spread lies in connection with the COVID pandemic.

As part of a post on the social media platform X, formerly Twitter, journalist Ann-Kathrin Müller posted a video showing demonstrators and playing the audio

clip. The video received 577 comments, all of which I collected from the platform to conduct a detailed text analysis. The commenters can be divided into at least two political groups. Most of the commenters belong to the so-called ‘Querdenker’ movement, which accuses the public media of dishonesty. Since 2020, the term ‘Querdenker’ has been used as a self-description by an increasing number of people who deny the danger or even the existence of the coronavirus, criticise the state for its measures against the pandemic, refuse vaccination, and sometimes combine these negative attitudes with a belief in conspiracy ideologies or characteristics of far-right ideologies. The ‘Querdenker’ movement is therefore not part of the far-right spectrum, but there are ideological and personal interfaces.⁸ Both the commenters close to this movement and those who distance themselves from it do not question the fake status of the audio recording. The journalist Ann-Kathrin Müller also points out in her post that this is an AI-generated deception:

Die ARD warnt, dass Audio-Tagesschau-Fakes im Umlauf sind, die suggerieren, dass bei Corona gelogen wurde u.a.

Das stimmt, habe es live bei Pegida #dd0611 gehört. Willkommen im Zeitalter der KI-gestützten Desinformation:⁹

‘ARD warns that audio fakes of the Tagesschau are circulating, suggesting among other things, that Corona was a lie.

That’s right, I heard it live at Pegida #dd0611. Welcome to the age of AI-assisted disinformation:’

When explaining why the audio recording is a fake, some X-users point out deviations from the familiar style of the *Tagesschau* and express knowledge of the formal elements of language (a) and knowledge of its pragmatic aspects (b).

(a) Knowledge of the formal elements of language

(17) Alter. Gruselig. Aber der Text an sich ist auch viel zu überzogen um echt zu sein. Zu viele Adjektive in den Sätzen.

‘Man. Creepy. But the text itself is far too exaggerated to be real. Too many adjectives in the sentences.’

(18) Das ist so extrem DRÜBER und straight in the face.... Lächerlich, wer’s für voll nimmt. Aber man glaubt, was man glauben will. Und Echokammer. Und überhaupt

‘That’s so extremely OVER THE TOP and right in the face.... Ridiculous if you take it seriously. But you believe what you want to believe. And echo chamber. And anyway’

⁸ <https://www.bpb.de/themen/rechtsextremismus/dossier-rechtsextremismus/516449/querdenker/> (last accessed on 14/04/2026).

⁹ <https://x.com/akm0803/status/1724128492625797544>, last accessed on 14 April 2026.

- (19) Wobei es bei dem grottenschlechten Textaufbau wirklich einiger Überwindung bedarf, das ernstnehmen zu können.
'However, it takes some effort to take this seriously, given the grotty text structure.'
- (20) Ähm... der Inhalt ist schon so bekloppt, dass man direkt von einem Fake ausgehen kann. Da braucht man weder auf Intonation oder Redefluss achten. Die sind einfach zu blöd, überhaupt einen glaubwürdigen Fake zu produzieren 🤖
'Um... the content is already so crazy that you can immediately assume it's fake. You don't need to pay attention to the intonation or the flow of the speech. They're just too stupid to make a credible fake. 🤖'
- (21) Vielen Dank für den Hinweis. Mechanisches Sprechen, falsche Betonung, null Satzmelodie, sogar falsche Wörter und gesprochene Grammatikfehler usw. kennt man gemeinhin nur von bestimmten Politikern/Sprechpuppen. Sind die auch KI?
'Thanks for pointing this out. Mechanical speech, wrong intonation, no sentence melody, even wrong words and spoken grammar mistakes, etc., are generally only known from certain politicians/voice puppets.
Are they AI too?'

In instances 17-21, users of X refer to specific aspects of the audio recording, noting a break with the usual language of the news program. So, the commenters make metapragmatic statements. The user in 17 refers to stylistic elements when noting that too many adjectives are used. Comment 18 also refers to stylistic elements, emphasising that the audio is "so extrem DRÜBER" (*so extremely OVER THE TOP*). At the same time, the commenter positions himself in relation to the group of people who take this audio recording seriously. The metapragmatic utterance is thus linked to a social positioning in relation to one social group and a socialisation process in relation to another (cf. Spitzmüller 2019: 25-26). Such a social positioning can also be observed in 19, where there is talk of a "grottenschlechten Textaufbau" (*grotty text structure*), and can also be found quite explicitly in 20, where the group of 'Querdenker' – who are referred to with "die" (*they*) – is described as "zu blöd" (*too stupid*). The commenters highlight metacommunicative elements to discredit the 'Querdenker' group. Comment 21, which belongs to this group, also engages with meta-linguistic aspects of the audio recording, such as intonation, lexis, and grammar, but attributes deviations from standard language to human speakers, specifically certain politicians.

Most commenters, however, attribute the recognition of the recording as fake to a break in their expectations and point out pragmatic aspects of language:

(b) Pragmatic aspects of language

- (22) LOL, das ist kein Fake und keine “Desinformation”, sondern Satire. Wer hier nicht merkt, dass das nicht real ist. hat wohl die letzten Jahre keine @tagesschaugesehen. Die Tagesschau würde ihre Fehler nie so offen zugeben.
‘LOL, this isn’t fake or “disinformation”, it’s satire. Anyone here who doesn’t realise this isn’t real probably hasn’t watched @tagesschau for the last few years. The Tagesschau would never admit their mistakes so openly.’
- (23) Wäre der ÖRR aufrichtig, wäre es längst mal Zeit sich zu entschuldigen. Da wäre so eine Klarstellung das absolute Minimum!
‘If the ÖRR were sincere, it would have been time to apologise by now. Such a clarification would be the absolute minimum!’
- (24) Selbstverständlich würdet ihr euch eure Verbrechen niemals eingestehen. Um derartig kriminell und skrupellos agieren zu können wie ihr es immer wieder tut, muss man schließlich menschlich absolut verkommen sein.
‘Of course, you would never admit to your crimes. After all, you must be absolutely depraved as a human being to act as criminally and ruthlessly as you do time and time again.’
- (25) mein weiß direkt das es Fake ist, die Tagesschau würde niemals “Entschuldigung” sagen ;-) Ironie off
‘You know it’s fake, the Tagesschau would never say “sorry” ;-) Irony off’
- (26) Ich glaube nicht dass die Tagesschau sagen würde dass sie einen seit 3 Jahren anlügt
‘I don’t think the Tagesschau would say they’ve been lying to you for 3 years.’
- (27) Das ist doch jedem denkenden Menschen sofort klar, dass das ein Fake sein muss. Als ob die ARD sowas Selbstkritisches bringen würde!
‘It’s immediately clear to any thinking person that this must be a fake. As if the ARD would broadcast something so self-critical!’
- (28) Das einzige was an dem ganzen falsch ist, ist dass sich die Verantwortlichen nicht entschuldigen werden. Das was da vom Band lief, ist und bleibt aber richtig. Egal wo es herkommt.
‘The only thing wrong with the whole thing is that those responsible will not apologise. But what came off the audio tape is and remains right. No matter where it comes from.’

The examples listed here all contain metapragmatic statements, as the illocutionary speech acts of apologising and admitting mistakes are seen as contradictory to the *Tagesschau*. However, some of the users do not attribute this to the communication situation of a news program, but rather to what they see as the prevailing dishonesty of media producers. This basic accusation against the public media seems to be typical of the ‘Querdenker’ scene. Here, too, social positioning is achieved through metapragmatic statements. The users express their certainty that it is a fake through various linguistic means. In some cases, they indicate their expectation by using the second subjunctive in combination with negation: *the news would never* (22, 25 and 26), *I do not think that the news would say* (26), *as if the news would bring* (27). This presupposes, at the same time, how the *Tagesschau* normally acts according to the commenters. This becomes clear in 23, where the use of the subjunctive presupposes that the public broadcaster is not honest.

Unlike in the case of Deeptomcruise’s comments, the commenters here are less inclined to dissect the audio into individual elements and interpret them as signs of a fake. This might be due to the different media – video deepfake in 4.1 and audio deepfake in 4.2 – but it might also be because social media users were more used to deepfakes in 2023 than in 2021. Now, the mere fact that something is unexpected leads to the assumption that the audio must be fake. In the examples discussed in this section, the commenters see the linguistic practice of apologising as an indicator of a fake, as well as individual lexical devices that seem atypical.

3.3 The German Chancellor announces a late-night show?

I will now proceed by analysing another case: a posting on Instagram published in March 2023. This case differs from the previous two, as it does not feature a deepfake, yet many users believe that it does. A video shared in an Instagram post by the *Neoraggazzi* account on 23/03/2023 is suspected of being a deepfake. Neo Raggazzi is a German talk show on the ZDF Neo channel. The associated account uses the post to refer to the ZDF late-night show “Studio Schmitt”. The first episode of the new season of the ZDF late-night show features comedian Felix Lohbrecht. The video is introduced with the following text:¹⁰

Standesgemäß starten wir in die neue Staffel mit dem [@bundeskanzler](#) Olaf Scholz.



‘We start the new season in style with [@bundeskanzler](#) Olaf Scholz. 🤝’

The video shows former German Chancellor Olaf Scholz in an empty office. He announces the late-night show and thereby addresses Felix Lohbrecht.

¹⁰ <https://www.instagram.com/neoraggazzi/reel/CpSnTCPPqIL/> (last accessed on 26/08/2025). The account @neoraggazi has since been removed from Instagram.



Picture 3: Screenshot from the video of the Instagram post (see footnote 10).

There are 264 comments on this video, all of which I have once again analysed using qualitative text analysis. In the comments section, Instagram users are very clear in expressing their opinion that this is a fake:

- (29) Save KI
'Save AI'
- (30) Das macht mir ehrlich gesagt richtig Angst dass man sowas so gut faken kann
'Honestly, it freaks me out how well something like this can be faked'
- (31) Krass wie echt die AI wirkt 🤖 🤖
'It's amazing how real the AI looks 🤖 🤖'
- (32) AI oder chat gbt??? Wahnsinn
'AI or chat gbt??? Madness'
- (33) Wie in einer Black mirror Folge...
'Like in a Black mirror episode...'
- (34) ... mitein....ander.. 🤖 🤖
'...to...ge...ther 🤖 🤖'
- (35) @studioschmitt Whaaaaa - olle Scholz legt für euch den Mantel ab, der unsichtbar macht? Ritterschlag! Glaub ich nicht
'@studioschmitt Whaaaaat – old Scholz takes off the invisibility cloak just for you? That's a knighthood-level honor! Yeah, right – I don't buy it.'
- (36) Entweder DeepFake oder die Regierung hat endlich einen Sinn für Humor. Befürchte Ersteres 😬
'Either it's a deepfake, or the government has finally developed a sense of humour. I think it's the former.'

- (37) Anders als das das nen deepfake ist, kann ich's mir nicht erklären.
'I can't explain it any other way – must be a deepfake.'

While in 2021, TikTok users were still comparing the image to the original concerning the Tom Cruise deepfake video and expressing uncertainty about the authenticity of the video, some of the commenters on the Instagram post already had some experience with artificial intelligence. Only if there is already experience with AI-generated videos, comments such as 29, in which it is expressed that one is certain that artificial intelligence was used here, make sense. The formulation of the fear that something can be faked so well (30) also expresses that the commenter assumes a generative AI here. In example 31, the commenter also assumes that AI plays a role in the posted video. Some commenters are sure this is a fake video, as in 32, when they suggest AI or chat gbt (sic!), or when they refer to the science fiction series *Black Mirror*, as in 33. From a linguistic point of view, the commentary in 34, which reproduces part of Olaf Scholz's speech, is also particularly interesting. In this comment, the user highlights a pause made by the supposed Olaf Scholz in a word, by inserting three dots into this word. However, this emphasis on the pause only makes sense if the commenter perceives it as unusual. This suggests that they associate such pauses with non-human speech. The smiling emojis further indicate that the commenter believes the video was generated by AI. The comments shown in 29 to 34, therefore, do not reveal any uncertainty regarding the authenticity of the posted video, but rather a conviction that it is a fake. This conviction again results from the break with the expectations of media recipients: Users are confused that the former Chancellor, who rarely speaks in public, is announcing a late-night show (35). For most of the commenters, the video's inauthenticity is beyond question. They express their knowledge that it is a fake by presupposing its classification as a deepfake, often without further explanation. The commenter in 36 gives two possible explanations for the video: Either it's a deepfake or the government has a sense of humour, and he chooses the first, apparently because he thinks it's the more likely option. The commenter in 37 expresses their confusion very clearly: "I can't explain it any other way – must be a deepfake." The former chancellery, however, confirms the authenticity of the speech: Olaf Scholz did indeed announce the late-night show.¹¹ Fallis (2021: 634) points out that "even if a deepfake is indistinguishable from a genuine video based purely on a visual inspection of the image, we might still be able to distinguish it from a genuine video just because the content is *implausible* (emphasised in the original). For the assessment of (supposed) deepfakes, it is important that social media users know the individuals who are (supposedly) being imitated, knowledge about their usual behaviour, including language behaviour. The comments on the video of Olaf Scholz show that the assumed implausibility here leads commenters

¹¹ <https://www.rnd.de/medien/tommi-schmitt-olaf-scholz-haelt-fernsehansprache-deepfake-oder-echt-MBRX6CTTMJDT7GXYMDVRUWU5IY.html> (last accessed on 14/04/2026).

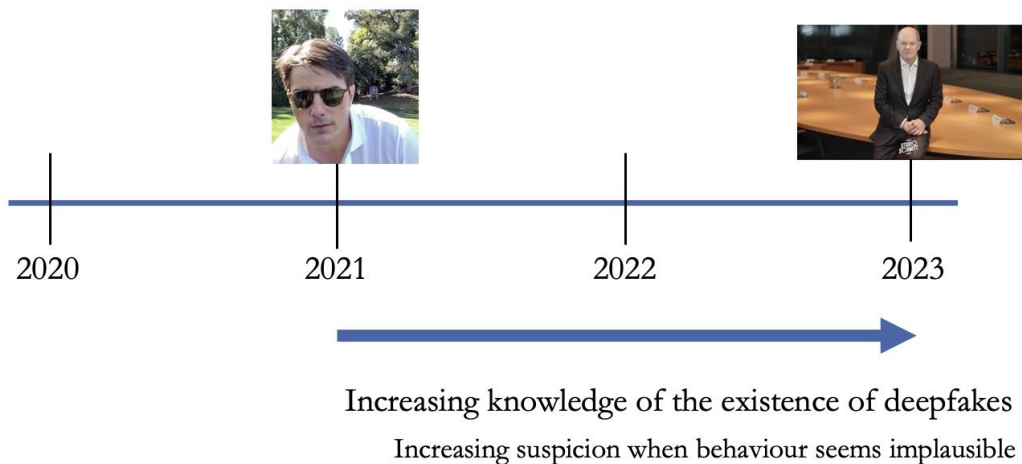
to assume that the video was generated by artificial intelligence, even though it shows the real Olaf Scholz.

4 Discussion and conclusion

As I demonstrated in Section 2, the aspect of human resemblance plays a role in many definitions of AI, and we constantly encounter anthropomorphic representations of AI in public discourse. I have therefore chosen to approach this phenomenon of human resemblance through the concept of imitation. I started by discussing what aspects of human communication AI imitates. Pointing out that generative AI is capable of imitating various aspects of human communication, I finally focused on deepfakes, which are a special case of imitation because they mimic individuals. This was important for the analysis: People possess prior knowledge about certain public figures, such as actors, journalists, or politicians. When they see videos featuring these individuals, they bring with them expectations about how they typically behave. As the analysis of three case studies has shown, when confronted with apparent imitations of these figures, viewers assess whether the behaviour depicted appears plausible. In this paper, I was interested in how people communicatively navigate supposed deepfakes.¹² Analysing Case 1 (*deptomcruise*), it became clear that the commenters interpreted different elements of the videos as indicators that they were encountering either a representation of the original, a representation of a human imitation, or an imitation of a representation. They used different types of knowledge to evaluate the video: Knowledge of Tom Cruise's physical attributes, general world knowledge and technical knowledge.

The other two cases showed that, with increasing awareness of the existence of deepfakes, people discussed the presented videos and audio less in detail, instead focusing more on general aspects that broke their expectations. Regarding the research question, how the discourse around deepfakes evolved with increased knowledge about their existence, one can state that increasing knowledge leads to increasing mistrust in posted videos, particularly when social media users see behaviour that seems implausible to them to a certain extent:

¹² I must point out that I assumed that the comments on the social media posts were written by people.



Picture 4: Increasing suspicion with increasing knowledge about deepfakes.

The discussed case studies exhibit several differences; for instance, they are written in different languages and were published at different points in time. Nevertheless, they indicate that awareness of the existence of deepfakes makes people more sceptical. In this way, deepfakes influence the societal common ground and, to a certain extent, contribute to its fragmentation

Drawing on Schwab (2016), Whittaker et al. (2023: 1) postulate that contemporary societies are experiencing “a shift towards the 4th Industrial Revolution — an era defined by increasingly embedded digitization within society, blurring the lines between the physical and digital, and subsequently what it means to be human”. In this context, deepfakes play a special role, as they “leverage powerful techniques from machine learning (ML) and artificial intelligence (AI) to manipulate or generate visual and audio content with a high potential to deceive” (Kietzmann et al. 2020: 136). However, the dangers posed by the existence of deepfakes go beyond manipulation through artificially generated content; they also extend to suspicion in videos in which real persons are depicted, as I demonstrated in Section 4.3. This may lead to a danger for modern societies: Krämer (2021: 29) describes humans as epistemically dependent beings. Drawing on Shapin (1994: XXV), she emphasises that knowledge is a collective good. In this context, she underlines the inescapability of the sign-bound nature of knowledge. Since much of what we know is not individually verifiable, knowledge also depends on testimony. So, trust in what others tell us in words, writings, and images is unavoidable (Krämer 2021: 29). Krämer (2021: 32) clearly states that knowledge based on testimony marks an irreplaceable dimension of our knowledge cultures. It follows that if second-hand knowledge is ubiquitous, then trust is a foundation of our cognitive practices. Krämer (2021: 35) emphasises that deepfakes contribute to the prevailing uncertainty about how to trust what is presented in the media. Fallis

(2021: 625) points out that, as “deepfakes become more prevalent, it may be epistemically irresponsible to simply believe that what is depicted in a video actually occurred.” Although it is dangerous to believe everything you see, doubting the authenticity of every communicative contribution also puts the cornerstones of communication at risk. So, regarding the social implications of these developments (research question 3), one can state that the implausibility of digitally presented video and audio today often leads to scepticism and disruption. In the context of generative AI, “concerns about the possibility of undermining trust in democratic processes, especially elections” are repeatedly expressed (Labuz and Nehring 2024: 1). The dangers of deepfakes are repeatedly pointed out. However, the question also arises as to whether mistrust in the authenticity of all those videos that break collective expectations also poses a threat to social cohesion. Deepfakes as attacks on the *I only believe what I see* belief (Bovenschulte 2019: 1) lead us to question what we see as soon as it presents something unexpected. In addition to the danger of deception, the existence of deepfakes also entails the danger of no longer believing what is real. A subsequent linguistic analysis could then be devoted to precisely this formulated doubt, to find out something about social expectations through communication about deepfakes.

In a further step, social media postings and their comments could be analysed more systematically to identify discursive patterns. It seems interesting to explore whether larger corpora of data on false assumptions in the context of so-called artificial intelligence could be used to identify linguistic practices of suspicion in digital communication. In this way, linguistic analysis can also reveal contrasts between different languages, different age groups or different cultural backgrounds. It could be interesting to look at what a particular community expects and where they see a break in expectation. This could also help to identify various social common grounds.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

As I am not an English native speaker, I used the machine translation services DeepL (<https://www.deepl.com/de/translator>) and ChatGPT as aids for translating certain passages. The final formulations were manually reviewed and adjusted where necessary.

References

- Ackermann, Judith & Dewitz, Leyla. 2020. Kreative Bearbeitung politischer Information auf TikTok: eine multimethodische Untersuchung am Beispiel des Hashtags #ww3. *MedienPädagogik: Zeitschrift für Theorie Und Praxis Der Medienbildung, Medienbildung* 38. 69–93.
<https://doi.org/10.21240/mpaed/38/2020.11.16.X>
- Altuncu, Enes & Franqueira, Virginia N. L. & Li, Shujun. 2024. Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Front. Big Data* 7. 1–23. <https://doi.org/10.3389/fdata.2024.1400024>
- Bode, Lisa & Lees, Dominic & Golding, Dan. 2021. The digital face and deepfakes on screen. *Convergence: The International Journal of Research into New Media Technologies* 27(4). 849–854.
<https://doi.org/10.1177/13548565211034044>.
- Bovenschulte, Marc. 2019. Deepfakes – Manipulation von Filmsequenzen. *Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB), Themenkurzprofil* 25, 1–5. <https://doi.org/10.5445/IR/1000133910>
- Chesney, Robert & Citron, Danielle K. 2019. Deep Fakes: a looming challenge for privacy, democracy, and national security. *California Law Review* 107(6). 1753–1820.
- Clark, Herbert H. & Fischer, Kerstin. 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46. 1–33.
<https://doi.org/10.1017/S0140525X22000668>
- Deppermann, Arnulf. 2018. Wissen im Gespräch. In Birkner, Karin & Janich, Nina (eds), *Handbuch Text und Gespräch*, 104–142. Berlin, Boston: De Gruyter.
- Dickel, Sascha & Schmidt-Jüngst, Miriam. 2021. Gleiche Menschen, ungleiche Maschinen. Die Humandifferenzierung digitaler Assistenzsysteme und ihrer Nutzer:innen in der Werbung. In Dizdar, Dilek & Hirschauer, Stefan & Paulmann, Johannes & Schabacher, Gabriele (eds), *Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen*, 342–367. Weilerswist-Metternich: Velbrück Verlag.
- Fallis, Don. 2021. The epistemic threat of deepfakes. *Philosophy & Technology* 34. 623–643. <https://doi.org/10.1007/s13347-020-00419-2>
- Fenwick, Ali & Molnar, Gabor. 2022. The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines. *Discover Artificial Intelligence* 2, 14. <https://doi.org/10.1007/s44163-022-00030-8>.
- Foucault, Michel. 1973. *Ceci n'est pas une pipe: Sur Magritte*. Montpellier: Fata Morgana.
- Gardt, Andreas. 2017. Interpretation. In Betten, Anne & Fix, Ulla & Wanning, Berbeli (eds), *Handbuch Sprache in der Literatur*, 487–508. Berlin, Boston: De Gruyter.
- Gottschling, Markus. 2023. Imitationen: Zur Menschlichkeit des Erzählens mit Künstlicher Intelligenz. Preprint: <https://rhet.ai/2023/12/06/imitationen-zur-menschlichkeit-des-erzaehlens-mit-kuenstlicher-intelligenz/>.

- Gottschling, Markus & Kramer, Olaf. 2025. Persuasive surfaces and calculating machines. A rhetorical perspective on Artificial Intelligence. *Global Philosophy* 35, 14–36. <https://doi.org/10.1007/s10516-025-09748-3>
- Hausendorf, Heiko. 2020. Geht es auch ohne Interaktion? *Aptum* 2(3). 86–89.
- Henrickson, Leah. 2023. Chatting with the dead: the hermeneutics of thanabots. *Media, Culture & Society* 45(5). 949–966. <https://doi.org/10.1177/01634437221147626>
- Kalwa, Nina. 2019. Die Konstitution von Konzepten in Diskursen: Zoom als Methode der diskurslinguistischen Bedeutungsanalyse. In Schiewe, Jürgen & Niehr, Thomas & Moraldo Sandro M. (eds), *Sprach(kritik)kompetenz als Mittel demokratischer Willensbildung. Sprachliche In- und Exklusionsstrategien als gesellschaftliche Herausforderung*, 11–26. Bremen: Hempen, Bremen.
- Kalwa, Nina. 2025. “Noch steckt KI in den Kinderschuhen”. Die Analyse von Verzeitlichungspraktiken mittels qualitativer Korpuslinguistik. *Zeitschrift für Literaturwissenschaft und Linguistik* 55. 379–405. <https://doi.org/10.1007/s41244-025-00379-0>
- Kalwa, Nina. 2026. Metakommunikation und AI Companionship. Vom Reden über Reden mit KI. *Zeitschrift für Kulturlinguistik* 1(1). 1–22. <https://doi.org/10.5077/journals/zfkl.2026.e2369>
- Kaminski, Nicola. 1998. Imitatio auctorum. In Ueding, Gerd (ed.), *Historisches Wörterbuch der Rhetorik*, 235–284. Berlin: De Gruyter.
- Kietzmann, Jan & Lee, Lina W. & McCarthy, Ian P. & Kietzmann, Tim C. 2020. Deepfakes: trick or treat? *Business Horizons* 63. 135–146. <https://ssrn.com/abstract=4471207>
- Kikerpill, Kristjan. 2020. Choose your stars and studs: the rise of deepfake designer porn. *Porn Studies* 7(4). 352–356. <https://doi.org/10.1080/23268743.2020.1765851>
- Klug, Nina-Maria. 2015. “Ich habe es doch mit eigenen Augen gesehen!”: Zur Eigentlichkeit bildlicher Zeichen. In Brinker-von der Heyde, Claudia & Kalwa Nina & Klug, Nina-Maria & Reszke, Paul (eds), *Eigentlichkeit: Zum Verhältnis von Sprache, Sprechern und Welt*, 501–522. Berlin, Boston: De Gruyter.
- Krämer, Sybille. 2021. Der Verlust des Vertrauens. Medienphilosophische Perspektiven auf Wahrheit und Zeugenschaft in digitalen Zeiten. In Schicha, Christian & Stapf, Ingrid & Sell, Saskia (eds), *Medien und Wahrheit. Medienethische Perspektiven auf Desinformation, Lügen und „Fake News“*, 25–42. Baden-Baden: Nomos. <https://doi.org/10.5771/9783748923190>
- Labuz, Mateusz & Nehring, Christopher. 2024. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science*. 23(4), 454–473. <https://doi.org/10.1057/s41304-024-00482-9>
- Monti, Paolo. 2024. AI Enters Public Discourse: a Habermasian assessment of the moral status of Large Language Models. *Ethics and Politics* 61(1). 61–80. <https://philpapers.org/archive/MONAEP-2.pdf>
- Pawelec, Maria & Bieß, Cora 2021. *Technikfolgen und Regulierungsfragen aus ethischer und sozialwissenschaftlicher Perspektive*. Nomos 1. <https://doi.org/10.5771/9783748928072>

- Peng, Hu & Yaobin, Lu & Yeming, Gong. 2021. Dual humanness and trust in conversational AI: A person-centered approach. *Computers in Human Behavior* 119. 1–18. <https://doi.org/10.1016/j.chb.2021.106727>
- Salles, Arleen & Evers, Kathinka & Farisco, Michele. 2020. Anthropomorphism in AI. *AJOB Neuroscience* 11(2). 88–95. <https://doi.org/10.1080/21507740.2020.1740350>
- Schmitz, Ulrich. 2011. Sehflächenforschung: eine Einführung. In Diekmannshenke, Hajo & Klemm, Michael & Stöckel, Hartmut (eds), *Bildlinguistik: Theorien - Methoden – Fallbeispiele*, 23–42. Berlin: ESV.
- Schulz-Schaeffer, Ingo. 2008. Deutung und Delegation. Handlungsträgerschaft von Technik als doppeltes Zuschreibungsphänomen. In Rehberg, Karl-Siegbert (eds), *Die Natur der Gesellschaft. Verhandlungen des 33. Kongresses der Deutschen Gesellschaft für Soziologie in Kassel 2006*, 3135–3144. Frankfurt a. M.: Campus.
- Schwab, Klaus. 2016. *The Fourth Industrial Revolution*. London, etc.: Portfolio Penguin.
- Schwedler, Gerald & Sonntag, Jörg. 2021. Das Imitieren verstehen. Perspektiven, Analysen, Zugriffe. In Grünbart, Michael & Schwedler, Gerald & Sonntag, Jörg (eds), *Imitationen: systematische Zugänge zu einem kulturellen Prinzip des Mittelalters*, 1–25. Paderborn: Brill.
- Shapin, Steven. 1994. *A Social History of Truth. Civility and Science in Seventeenth-Century England*. Chicago: Chicago University Press.
- Spieß, Constanze. 2024. Keine Panik vor der KI – Zuschreibungspraktiken in Diskursen über KI. Eine linguistische Perspektive. In Schreiber, Gerhard & Ohly, Lukas (eds), *KI – Text. Diskurse über KI-Textgeneratoren*, 363–382. Berlin, Boston: De Gruyter.
- Spitzmüller, Jürgen. 2019. ‘Sprache’ – ‘Metasprache’ – ‘Metapragmatik’: Sprache und sprachliches Handeln als Gegenstand sozialer Reflexion. In Antos, Gerd & Niehr, Thomas & Spitzmüller, Jörg (eds), *Handbuch Sprache im Urteil der Öffentlichkeit*, 11–30. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110296150-002>
- Steinhoff, Torsten. 2023. Der Computer schreibt (mit). Digitales Schreiben mit Word, WhatsApp, ChatGPT & Co. als Koaktivität von Mensch und Maschine. *MiDU* 5(1). 1–16. <https://doi.org/10.18716/ojs/midu/2023.1.4>
- Taye, Mohammad Mustafa. 2023. Understanding of Machine Learning with Deep Learning: architectures, workflow, applications and future directions. *Computers* 12(5). 1–26. <https://doi.org/10.3390/computers12050091>
- Tolosana, Ruben & Vera-Rodriguez, Ruben & Fierrez, Julian & Morales, Aythami A. & Ortega-Garcia, Javier. 2022. An introduction to digital face manipulation. In Rathgeb, Christian & Tolosana, Ruben & Vera-Rodriguez, Ruben & Busch, Christian (eds), *Handbook of Digital Face Manipulation and Detection. Advances in Computer Vision and Pattern Recognition*, 3–26. Cham: Springer. https://doi.org/10.1007/978-3-030-87664-7_1
- Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind*, LIX (236). 433–460.

- Vaccari, Cristian & Chadwick, Andrew. 2020. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6(1). 1–19. <https://doi.org/10.1177/2056305120903408>
- Van Dijk, Teun A. 2011. Discourse studies and hermeneutics. *Discourse Studies* 13(5). 609–621. <https://doi.org/10.1177/1461445611412762>
- Whittaker, Lucas & Mulcahy, Rory & Letheren, Kate & Kietzmann, Jan & Russell-Bennett, Rebekah. 2023. Mapping the deepfake landscape for innovation: a multidisciplinary systematic review and future research agenda. *Technovation* 125. 1–17. <https://doi.org/10.1016/j.technovation.2023.102784>
- Yildirim, Yetkin & Arslan, Emin Alp & Yildirim, Kamil & Bisen, Ibrahim. 2021. Reimagining education with Artificial Intelligence. *Eurasian Journal of Higher Education* 2(4). 32–46. <https://doi.org/10.31039/ejohe.2021.4.52>

Halis Sefaradi: critical authenticity and the voice of the first Ladino chatbot *Estreyika*

Alejandro Acero Ayuda (University of Oregon) & Carlos Yebra López (California State University, Fullerton)

aaceroay(at)uoregon.edu, cyebra Lopez(at)fullerton.edu

Abstract

As the first Ladino-speaking chatbot intended to support language revitalization, *Estreyika* inhabits a complex ideological space where *authenticity* is constructed, contested, and operationalized through cultural, sociolinguistic, and political dynamics anchored in the lived realities of the last speakers of the endangered language Ladino: a Romance language of Ibero-Ottoman origin, preserved by the Sephardic diaspora for over 500 years, now considered severely endangered by UNESCO. In this short article, we argue that *Estreyika* functions as a site for negotiating performative ‘authenticity’, shaped by the learners’ positionalities—whether as heritage language learners, descendants of Ladino-speaking communities, or so-called ‘cultural outsiders’. Their linguistic performances become acts of alignment or resistance within a discursive field that privileges certain registers, geographies, or orthographies. By exploring *Estreyika* as a socio-political artifact, this paper advocates for an understanding of *authenticity* rooted in multiplicity, reflexivity, and critical positionality—a counter-narrative toward more inclusive LLMs critically attuned to AI-human partnerships.

Keywords

AI-human interaction, LLM, authenticity, heritage speakers, endangered languages, Ladino

1 The endangerment of a living language

Ladino, also known as Judeo-Spanish, is a Romance language of Ibero-Ottoman origin (Papo 2020) historically spoken by the descendants of the Sephardic Jews expelled from the Iberian Peninsula in 1492. Today, the number of speakers has diminished and is mostly confined to older adults, intergenerational transmission is largely broken, and the language is classified by UNESCO (2003) as ‘severely endangered’.

The history of the Sephardic diaspora is marked by pivotal moments that in retrospect have been used to identify specific periods. Following his own framework to break down the Ashkenaz diaspora (1973), Max Weinreich subdivided the Sephardic diaspora in three broad stages (Sepharad 1-3), and scholars in the subject have echoed this subdivision (Hassán 1995; Hernández González 2001; Šmid 2002; Álvarez López 2017). Recently, Carlos Yebra López (2024) has extended this typology by incorporating a fourth critical point now that the global Ladino-speaking diaspora connects primarily online, resulting in a subdivision of three offline stages (Sepharad 1–3) plus one digital (Sepharad 4).



Acero Ayuda, Alejandro & Yebra López, Carlos. 2026.

Halis Sefaradi: first Ladino chatbot *Estreyika*

Special Issue: *The notion of authenticity*. Vol. 4 No.1

DOI: 10.62408/ai-ling.v4i1.54

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses

ISSN: 2943-0070

CC-BY-NC-SA 4.0

Sepharad 1 refers to the Jewish presence in the Iberian Peninsula lasting from the 1st century CE until the Alhambra Decree of 1492 that marked their expulsion. During this period, proto-Ladino began to take shape in Romanized Iberia as Jewish immigrants interacted with local Romance varieties (Bunis 2019) already influenced by Iberian Arabic. By the thirteenth century, Ladino had also developed a parallel function as a written, word-for-word calque of Hebrew used for liturgical purposes (Séphiha 1977).

Sepharad 2 encompasses the period after the expulsion of the Sephardim in 1492 and their subsequent diaspora across the Mediterranean basin, lasting until the mid-20th century. In this phase, the 15th-century Iberian Romance varieties they brought to the diaspora, incorporated features from contact languages where they settled, mostly Turkish, Greek, South-Slavic, Italian or French, along with already present liturgical Hebrew (Minervini 2006). Over time, Ladino consolidated as a koine, a common language that developed organically due to close interaction among mutually intelligible Iberian Romance varieties spoken in the Mediterranean basin (Papo 2020; Yebra López 2021). During Sepharad 2, Ladino arises as a co-territorial diasporic vernacular with distinctive linguistic structures, a vibrant literary tradition, and tied to Sephardic cultural practices (Díaz-Mas 2006) with two principal variants: Haketia, western variety spoken in northern Morocco, and Judezmo, eastern variety prevalent across the Ottoman Empire.

Sepharad 3 designates the global diaspora of Sephardic communities beginning in the mid-twentieth century. This phase was propelled by historical ruptures, including the Holocaust (1933-1945), Turkey's wealth tax in the 1940s, the foundation of the State of Israel in 1948, and the Istanbul pogroms of 1955. At the turn of the 21st century, estimates suggested that there were fewer than 50,000 speakers worldwide (Ethnologue 2024) living in scattered communities across Turkey, Israel or the United States. The lack of spaces for vernacular use of Ladino, the decay of the intergenerational transmission, and the existing sociopolitical and ideological factors at play, forced many of these speakers to shift rapidly to the dominant languages where they primarily resided (Turkish, Hebrew, English or Spanish). In this stage, Ladino transitioned from a co-territorial vernacular rooted in distinct regions to a post-co-territorial diasporic vernacular in which geolectal distinctions were progressively disappearing, along with its Ladino-speaking community. In less than a century, Ladino shifted from being an "international trade language [...] in the Levant, to the colloquial language [...] for older people" (Harris 1994: 228).

2 Sepharad 4: Ladino Digital Renaissance

The same COVID-19 pandemic that caused invaluable human losses and severe economic downturns also forced widespread societal changes. The seclusion imposed by the quarantine forced a shift in the modes of socialization, expediting a process of migration from the physical to the digital realm. For the Ladino-speaking community, this meant an opportunity to populate existing Web 1.0 and 2.0 platforms that were already invested in fostering worldwide community building (*Ladinokomunita*, 1999 – Rachel Amado Bortnick) or documenting the language and its speakers in the present (*Ladino21*, 2017 – Carlos Yebra López, Alejandro Acero Ayuda and Benni Aguado) but it also represented an unprecedented chance for creating new digital spaces where more innovative forms of intergenerational transmission can emerge, digital spaces where unrelated individuals of all generations are learning Ladino from and with each other, and co-constructing linguistic practices and ideologies together (e.g. *Los Ladinadores*, 2020 – Aldo Sevi; *Enkontros de Alhad*, 2020 – Liliana and Marcelo Benveniste, etc.).

The global diasporic community of Ladino speakers embraced with enthusiasm this virtual migration to a new Digital Home-Land (Held 2010; Yebra López 2021) where the identity of the Sephardi ‘imagined community’ (Anderson 1983) could be digitally articulated. The internet offered an online territory suited for the offline global diaspora to inhabit, recreating a virtual place for cohesion where Ladino was to be the lingua franca of the 21st century multilingual Sephardi. The Ladino speakers’ resourceful use of this technological innovation to imbue Ladino through a wide range of multimodal affordances on the digital world has led some Ladino activists to call this phenomenon the *digital renaissance* (Cruz Çilli 2021) of Ladino on the internet. The online articulation of the Sephardic diaspora in the 21st century is, according to Yebra López, what initiated Sepharad 4, and marked the reconfiguration of Ladino as a cyber-(post) vernacular (2024) language.

The online proliferation of manifold platforms and modes (WhatsApp and Facebook groups, Zoom gatherings, YouTube channels, email lists, Instagram, Snapchat or TikTok accounts, etc.) (see Yebra López 2024 for a detailed account) developed by the global, mainly digitally-native, community of Ladino speakers has revealed a new paradigm of sociolinguistic interest (Held 2010; Yebra López 2021; Bürki 2021) that is reshaping language and community in a dialogical fashion: the social construction of a digital Sephardi cultural identity (Lindholm 2013; Van Leeuwen 2001) that is redefining the expressive ‘authenticity’ of what counts as *halis* (pure, genuine, authentic) Ladino.

3 *Halis Ladino*: ‘Authenticity’ or the authentication of a language

On these digital platforms, heritage speakers (HS) of Ladino—those who acquired the language through home-based intergenerational transmission (Polinsky and Kagan 2007; Montrul 2016) coexist with second (L2) or most commonly multiple (L2+) languages learners, who encountered the language later in life through online community initiatives rather than within the family unit (King 2000). However, both groups now rely on these virtual spaces as their primary arenas for (post)vernacular expression and interaction. In this context, Ladino has become a contested digital space sustained partially by the nostalgia of an evoked ancestral code (Woodbury 2011) that is continuously emulated and prescribed, but also by the creativity of unique *linguaging* practices (García and Wei 2014) that generate new forms of authentication (Yebra López 2024).

These *linguaging* practices either adhere to the sociolinguistic boundaries and ideological expectations of Ladino as a static system or engage in the dynamic process of meaning making and knowledge shaping by deploying the speakers’ full linguistic repertoire without regard for purist constraints (Otheguy, García and Reid 2015). Regardless, their linguistic performances are not value-neutral; they are informed by the Ladino speakers’ positionalities—whether as heritage language learners, descendants of Ladino-speaking communities, or ‘cultural outsiders’ acquiring the language as part of their multilingual repertoire. Their performances become acts of alignment or resistance within a discursive field that privileges certain registers (formal, informal, liturgic, vernacular) (Séphiha 1977; Bunis 2021), geographies (Istanbul, Sarajevo, Salonica, Izmir, etc.) (Quintana 2007; Schwarzwald 2019), or orthographies (French-Ottoman Latin transcription, Rashi, Solitreo, etc.) (Bunis 2019; Yebra López 2024) over others.

In this context, the notion of *authenticity* cannot be disentangled from hegemonic assumptions about linguistic purity or correctness—both glottophagic processes (Calvet 1974), that is, sociopolitical dynamics within power-imbalanced linguistic environments that contribute to language shift (Fishman 1991) in an already endangered language. Instead of ‘authenticity’, Mary Bucholtz proposes the concept of *authentication* (2003) as an alternative term. ‘Authenticity’ is achieved through authenticating practices of those who use and evaluate language. Thus ‘authenticity’ is a socially negotiated process, rather than a fixed state (Bucholtz 2003). From a sociolinguistics perspective, as we have seen in Sepharad 2-3, it is problematic for Ladino prescriptivists to wield any variety as more “authentic” than others. Ladino is a pluricentric language with different geolects that developed in diaspora, around the various centers of economic, cultural, and political activity of the modern Sephardic diaspora such as Thessaloniki, Istanbul, Safed, Vienna, or Belgrade (Quintana 2010) and converged in the digital realm of Sepharad 4.

4 *Estreyika*: AI-powered chatbot to revitalize Ladino

The history of Sepharad 4 aligns with successive iterations of the Internet (Web 1.0, Web 2.0). The prospect of a more decentralized, intelligent, and autonomous Web 3.0 has already yielded technologies like blockchain, machine learning, and artificial intelligence. However, as previous iterations of the Web 1.0, 2.0 demonstrated, developments in the digital era tend to aggravate the digital divide between prestiged and minoritized languages (Kornai 2013). Trending digital tools like artificial intelligence (AI) build upon existing digital spaces and resources where prestiged languages are already fully established and with sufficient web-presence. Minoritized languages lack the material and human resources to power the creation and maintenance of such tools (Öktem, Zevallos, Moslem, Güneş et al. 2022). Limited textual resources, scarce state support, and historical patterns of societal and institutional marginalization hinder the participation of minoritized languages in the digital sphere (Nekoto, Marivate, Matsila et al. 2020). For that reason, endangered languages integration in these new Web 3.0 systems must be intentionally strengthened with adequate resources and policy support to back every community initiative by language activists, ideally by institutions (Tollefson and Pérez-Milans 2019). An example of a successful community initiative that benefited from institutional support with promising results was the project carried out by the *Col-lectivaT* and *Sephardic Center of Istanbul* that secured funding within the framework of the CCH-II (Ministry of Culture and Tourism of the Republic of Turkey with the financial support of the European Union.

Öktem, Zevallos, Moslem, Güneş et al. (2022) developed the first neural based system for Ladino capable of bidirectional machine translation and speech synthesis between Ladino - Spanish, Turkish - English, based on a baseline system of machine translation for Ladino. Drawing upon the potential of this advancement and relying on the last Natural Language Processing (NLP) and Machine Learning (ML) technology, Carlos Yebra López, Alejandro Acero Ayuda, and Alp Öktem created the first Ladino-speaking chatbot in history: *Estreyika* (2025).

5 “It’s alive!”: from Chatbot to *Estreya Perez*

The idea for developing a Ladino-speaking chatbot originated with Carlos Yebra López, who envisioned the use of Web 3.0 technology to create an AI-powered digital agent capable of using natural language processing (NLP) to engage users in Ladino. At this incipient stage, Alejandro Acero Ayuda created the name, avatar, and narrative backstory that framed the chatbot’s persona, *Estreya Perez* (2025).

Estreyika was designed as a tool for language education and practice, equipped with the metalinguistic ability (Bialystok and Ryan 1985) to analyze the Ladino language and generate complex explanations of its linguistic phenomena, within the current limits of this promising field of study (Beguš, Dąbkowski and Rhodes 2025). At the same time, *Estreyika* was imagined as a narratively situated

character, embodied in a digital media capable of enacting a heightened ‘storyworld’ (Murray 1997) where multiple narratives are contained, and require the users’ active engagement to untap the innumerable possibilities of such immersive environment (Landow 2006). The backstory of Estreyika permeates through the storyworld that places the chatbot within a historically plausible context. Estreyika uses the most widespread, accepted variety of Ladino: the Istanbul gelect, written in the French-Ottoman Latin transcription according to the *Aki Yerushalayim* orthography, yet she understands/computes other Ladino gelects. The backstory of Estreyika is thus rooted in late eighteenth-century Istanbul, during a period of Sephardic cultural flourishing yet amid the broader transformation of the Ottoman Empire (Benbassa and Rodrigue 2000). The personal story embedded in Estreyika emerges in her replies to questions about daily life, dynamics of community, and conversations in Ladino. Interacting with her responses allows us to picture everyday life in a Jewish quarter of that era, where Ladino resonated in homes, streets, and markets, describing an environment where knowledge and learning is tied to the imagined social practices (Gee 2003) that Estreyika describes. Since its debut, Estreyika’s persona has been enthusiastically welcomed by the Ladino-speaking community.

The naming of the chatbot followed naturally from this vision of a chatbot companion to guide Ladino learners in their educational journey. Acero Ayuda reflected on the role of the chatbot not only as a pedagogical tool but also as a symbolic resource for the future of the language. The name, therefore, needed to evoke guidance, enlightenment, and orientation. Drawing inspiration from maritime navigation, where sailors relied on the stars to find their way home, he selected the name *Estreya* (Star) or the diminutive *Estreyika* (little star) to signify a light that shines, directs, and inspires future generations of Ladino learners committed to uphold the language’s revitalization.

After developing a detailed description of Estreyika’s backstory and personality, it was necessary to test whether the chatbot responded in ways consistent with her backstory, displaying coherence with the given identity and linguistic consistency as a key component to build user trust and engagement (Luger and Sellen 2016). To this end, Acero Ayuda designed a questionnaire organized around ten thematic areas: (1) identity, (2) synonyms and antonyms, (3) translation, (4) culturally significant topics, (5) proverbs, (6) grammar, (7) verb usage, (8) pronunciation, (9) historical knowledge and context, and (10) open-ended prompts that allowed Estreyika to engage in free conversation. Each theme was explored through multiple rounds of questioning, with increasing levels of detail requested in the responses. Following this iterative process, Acero Ayuda systematically recorded the strengths of Estreyika’s answers, identified errors requiring correction, and noted areas for improvement. This feedback was then provided to Alp Öktem who refined the system and enhanced Estreyika’s performance, resulting in the robust version available today.

The first prototype of *Estreyika* was tested in fall 2024 and launched in summer 2025 as an Alpha version, using Claude Sonnet 3.5 as the back-end model and Telegram as the front-end interface. Claude Sonnet belongs to the family of large language models (LLMs) developed by Anthropic, designed to generate human-like text through probability-based prediction across vast corpora of training data. While Claude Sonnet 3.5 allowed us to test *Estreyika*'s basic functionality, the Beta version—currently under development with Claude Sonnet 3.7-4.5—is designed to improve contextual sensitivity, reduce errors in low-resource linguistic environments, and expand multimodal affordances (development is scheduled to continue until September 2026). Telegram was chosen as the primary interface because of its global accessibility and user-friendly integration with Bot APIs (Application Programming Interface).

The challenges *Estreyika* faced were not merely technical but also structural. As Öktem, Zevallos, Moslem, Güneş et al. (2022) emphasized, marginalized languages such as Judeo-Spanish occupy an uneven position in the digitalized world. The lack of sufficient representation in digital corpora and institutional support required a creative, resource-conscious approach to train *Estreyika*. Instead of relying on massive corpora, the project drew upon curated sources of Ladino, including community-generated materials, archival texts, and prior digitalization efforts spearheaded by activists and scholars (e.g., Öktem's earlier work on building the first neural based system for Ladino capable of machine translation and speech synthesis).

Estreyika was further trained through iterative reinforcement provided by both heritage (HL) and multilingual (L2+) learners of Ladino, whose interactions helped refine *Estreyika*'s responses. This collaborative process highlighted both the opportunities and the limitations of working with endangered languages in the digital age: while human contributions can partially offset the lack of machine-readable data, they also foreground the extent to which “authentic” linguistic production is mediated by algorithmic constraints and ideological expectations within the speech community (Linnekin 1991).

The learning phase was finalized with the active participation of Ladino learners enrolled in California State University, Fullerton courses SPAN580T, SPAN468 (Fall 2024), SPAN468, and SPAN480T (Spring 2025) under the guidance of Carlos Yebra López. Their iterative interactions—ranging from testing conversational exchanges to evaluating lexical accuracy and stylistic register—significantly shaped *Estreyika*'s development. The chatbot's beta-phase refinement was supported by a Junior Research Grant, a Supercomputing Grant, and the research assistantship of Michelle Ramirez, providing critical resources for technical development and system optimization.

However, their voices were inevitably mediated both by the algorithmic structures of the large language model (LLM) and by the ideologically invested expectations of the remaining Ladino-speaking community. The Alpha phase revealed that *Estreyika* inhabits a complex ideological space where ‘authenticity’ is

constructed, contested, and operationalized through cultural, sociolinguistic, and political dynamics (Silverstein 1979; Irvine and Gal 2000; Blommaert 2005; Kroskrity 2010). Yet these processes remain anchored in the lived realities of the supposedly last speakers of Ladino, who carry the historical weight of the language's endangerment as well as the hopes of its revitalization. The participants' responses showed how Estreyika progressively evolved from a mere technical artifact, to become a self-referential discursive site in which both participants and Estreyika partook in the authentication process of Sephardic digital identity.

6 Student Perspectives on *Estreyika*: Legitimation through Authentication

The students' interactions with Estreyika illuminated how learners themselves invoke notions of the “real”, the “authentic”, and the “true” to describe their experiences. Their testimonies bear witness of how ‘authenticity’ is taken at face value, embraced as an objective, self-evident property of speech or identity, rather than understood as a socially negotiated and contextually defined process of authentication (Creese, Blackledge and Takhi 2014). This dynamic is particularly relevant in the case of endangered languages such as Ladino, where legitimacy is often contested among the members of the diasporic community.

6.1 Feeling *Real* in Digital Encounters

One of the most salient themes in the students' essays was the sense of realness that emerged from their exchanges with Estreyika. Several learners expressed surprise that the chatbot “felt like I was actually talking to a real person living in the region” (Student Essay). Crucially, students noted that Estreyika's delivery was often infused with warmth, nostalgia, and personality. Phrases such as “with a nostalgic smile” or “with bright eyes” accompanied their responses, lending the chatbot a deeply human quality. One student noted that interacting with Estreyika felt like “sitting across from a wise woman in the family,” a remark that blurred the line between bot and *nona* (grandmother), between mere transmission of information and intergenerational storytelling as a pedagogical strategy. These accounts suggest that ‘authenticity’, in this context, was not primarily linked to technical accuracy or linguistic purity, but to the affective impression of cultural resonance.

Such reactions underscore the relational dimension of ‘authenticity’: it emerges from interaction and recognition rather than from fixed linguistic criteria (Bucholtz 2003). Estreyika's evocations of Sephardic domestic life—markets, Turkish coffee, family gatherings—provided recognizable semiotic cues through which students authenticated her persona as “real.”

6.2 *Authentic* as Cultural Immersion

For many students, *authenticity* was framed in terms of immersion in cultural context. One learner noted: “Even if I could not verify every historical or culinary detail, the interaction felt like being immersed in an authentic cultural setting” (Student testimony). Another wrote that conversations about food were “a gateway into Ladino, making the interaction feel genuinely authentic”. *Estreyika* taught that to speak Ladino is not merely to master a vocabulary set, but to participate in a worldview: one where rituals are remembered, festivities are celebrated communally, and history is spoken out loud. Through everyday dialogue and richly detailed storytelling, *Estreyika* offered her users a rare experience: to learn a language not only about culture, but within it.

These responses reveal how ‘authenticity’ is produced through oral practice: talking about food, rituals, and cultural traditions served as semiotic resources that allowed learners to confer ‘authenticity’ on the chatbot. Moreover, this mode of authentication was not dependent on genealogical heritage but on discussed participation in symbolic practices that mediated belonging to the Sephardic ‘Digital Home-Land’ (Held 2010).

6.3 Negotiating *True* Ladino

Several students explicitly linked *authenticity* to language form, once again invoking terms such as “real” or “true.” One participant stated: “Having real conversations in a new language is, without doubt, the best way to learn” (Student testimony). Here, the notion of *real* was associated not with structural fidelity to a standardized Ladino but with the dialogic experience of conversational practice.

Other learners drew attention to specific linguistic features regarding orthography, such as the substitution of *k* for *c* or the alternation between *b* and *v* following the widely used, yet not exclusively, French-Ottoman orthography as prescribed by the magazine *Aki Yerushalayim*. This orthographic ascription was perceived as a marker of *linguistic authenticity*.

These discoveries reinforced their sense of engaging with the *true* form of Ladino, suggesting that for them, *authenticity* was less about fidelity to one historical register and more about the opportunity to inhabit Ladino as a living, learnable *lect*.

6.4 Authenticity, Diversity, and Community Anxiety

The testimonies also resonate with broader debates within the Ladino-speaking community. Brink-Danan (2011) has shown that ‘authenticity’ often functions as a gatekeeping device, distinguishing between *ladinoavlantes* and *no ladinoavlantes*, or disparaging ‘mixed’ varieties. Several students expressed moments of doubt or discomfort when *Estreyika*’s responses appeared inconsistent or too elaborate,

revealing the underlying anxiety tied to judgments of “authentic” use. Moreover, it is important to note that only a minority of participants in today’s digital Ladino fora consider Ladino their ‘mother tongue’ (Bunis 2020; Brink-Danan 2011), and that the community is increasingly heterogeneous (Santacruz 2019). Thus, privileging authenticity as defined by nativeness, genealogical descent, or unbroken intergenerational transmission risks excluding precisely those learners—heritage or L2—whose engagement is crucial for revitalization. The students’ emphasis on “authentic” or “true” experiences illustrates this tension: what they authenticate as real and meaningful may not align with community standards but nonetheless sustains their investment in Ladino.

6.5 Toward an Ethics of Authentication

The students’ reflections ultimately demonstrate that ‘authenticity’ is less a property of linguistic forms than a dynamic sociolinguistic negotiation achieved through interaction, affect, and recognition (Avineri 2012). Their descriptions of Estreyika as “real,” “authentic,” and “true” confirm that ‘authenticity’ is always mediated by contextual practices. Thus, the question is not what is ‘authentic’, but rather what it means to be authentic here, under these norms, and “what are the authenticating practices by which it [authenticity] is conferred or denied?” (Creese, Blackledge and Takhi 2014: 939). Estreyika’s success, as reflected in student responses, lies in her capacity to enable authenticating practices—opportunities for learners to align with cultural traditions, rehearse identity, and imagine themselves as participants in the living Sephardic continuum of the 21st century. The *authenticity* they experienced resided less in inherited linguistic purity than in the affective and pedagogical outcomes of their engagement. However, does this pattern also hold for *ladinoavlantes* within the broader Sephardic ‘imagined community’?

7 Discussion: evaluations of Estreyika from *La Komunita*

To explore Ladino-speaking community perceptions of Estreya’s authenticity, we distributed a supplementary survey to a selected number of members of the Ladino online networks (N=10) with varying levels of self-reported proficiency, determined by a simplified LEAP-Questionnaire (Blumenfeld and Kaushanskaya 2020) included in the survey. Participants were prompted either to engage in unstructured interactions with Estreya or to rely on a bank of questions that mirrored the parameters and themes emerging from the students’ interactions: i) Feeling *Real* in Digital Encounters; ii) *Authentic* as Cultural Immersion; iii) Negotiating *True* Ladino; iv) Authenticity, Diversity, and Community Anxiety; and v) Toward an Ethics of Authentication. The survey questions elicited the same themes as those identified in the student interactions.

Of the 10 Ladino-speaking participants, 6 reported that Ladino was their L1, learned in the family since childhood; 2 reported learning Ladino as an L2 at a young age (under 18); and 2 reported learning Ladino as an L2 in adulthood (over 18). When inquired about their perceived ‘authenticity’ of *Estreya*, four out of ten respondents evaluated *Estreya*’s production as generally or “fully authentic”, two respondents argued that the chatbot was “both authentic and inauthentic” concerning different aspects, and one participant incurred in item nonresponse (missing this specific question). Three respondents expressed reservations, citing issues such as Castilianisms, lexical and orthographic inconsistencies, a limited stylistic range, and the use of overly “elitist” vocabulary. Importantly, even those who did not fully accept *Estreya* as “authentic” often described moments in which the chatbot *felt* real, whether through emotional responsiveness, conversational alignment, or an impression of interpersonal attentiveness (the ability to *sense* and respond to the user’s input).

Moreover, 3 of the 4 respondents who evaluated *Estreya* as “authentic” self-reported in the background data section that they were L1 speakers of Ladino, “learned in the family since childhood”. The fourth respondent reported being an L2 speaker of Ladino that “learned the language at a young age (under 18)”. This distribution reinforces the argument that ‘authenticity’ is not an inherent property of linguistic forms, (as these 4 respondents also included mention of linguistic features in their critiques) but a dynamic sociolinguistic negotiation achieved through interaction, affect, and recognition (Bucholtz 2003). The evaluations ranging from detailed linguistic critique yet emphasizing the affective identification, demonstrate that perceptions of ‘authenticity’ depend on the relational experience and not solely on strict adherence to linguistic norms or forms, echoing the positive tendencies observed among the CSU Fullerton learners in the initial testing phase.

We acknowledged the limitations of our student sample, including existing power dynamics between instructor and students, and participants’ limited prior knowledge of Sephardic culture. While triangulation with L1, L2, L2+ Ladino speakers from the broader digital community strengthens our findings, limitations remain, including the small size of the Ladino-speaking community sample (N=10), potential self-selection bias, and the exclusively digital nature of the exchange, which may have affected representation of less digitally engaged community members. Future research could examine a larger pool of participants representing greater diversity in demographics, Ladino proficiency levels, and degrees of engagement with the online Sephardic diaspora from ‘Sepharad 4’ (Yebra López 2024).

8 Conclusions

The *Estreyika* model, while pioneering in its innovative approach within the Ladino-speaking world, remains bounded by several unresolved questions and

methodological caveats. A central issue is the uncertainty surrounding whether large language models genuinely internalize human-like, offline cognitive patterns and biases, thus reproducing ‘algorithms of oppression’ (Noble 2018) or merely reflect statistical regularities of language use. This ambiguity complicates efforts to assess the model’s reliability and interpretive depth when applied to minoritized languages such as Ladino. Furthermore, the pragmatic decision to privilege the Istanbul geolect in French-Ottoman Latin orthography—though it maximizes accessibility and standardization—comes at the cost of constraining the broader landscape of Ladino’s diverse geolects, sociolects, and scripts.

Despite the current Alpha version accepting and computing lexical variety, it is crucial that we also maintain educational engagement with Rashi and Solitreo scripts, not only to preserve historical knowledge but also to sustain the cultural richness embedded in traditional writing systems.

The Estreyika model has opened the debate for the Ladino-speaking community, highlighting the dual task of language revitalization in the digital era: ensuring communicative viability while simultaneously reshaping ideological valuations of minoritized languages. Estreyika has already done more than simply provide access to pedagogical and linguistic resources; it has situated Ladino as a contemporary, future-oriented, revalued language (Eisenlohr 2004) that is integral to modern collective Sephardi identity.

The challenge moving forward will be to negotiate a balance between communicative efficiency and linguistic diversity in order to secure a revitalization of Ladino that is both sustainable and representative of the Sephardic idiosyncrasy. This process must also empower Ladino-speaking communities as active agents in shaping revitalization strategies, aligning with participatory models of language ownership that affirm local self-determination and ensure long-term sustainability (Kono 2019).

Acknowledgments

The authors would like to *rengrasyar de las onduras de nuestro korason* to the Ladino-speaking community members for their enthusiasm, generous contributions and unwavering support throughout the development, testing and review of *Estreyika*. Special appreciation is extended to Alp Öktem, whose technical expertise and collaboration were instrumental in bringing the chatbot to life, and the Spanish students at California State University, Fullerton, whose active engagement provided invaluable feedback during the chatbot’s design and refinement. We also acknowledge the support of the CSUF Junior Research Grant, the Supercomputing Grant, and the research assistantship of Michelle Ramirez, without which this chatbot and article would not have been possible. Finally, we sincerely appreciate the reviewers’ valuable comments and suggestions, which helped us to improve the quality of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest regarding the publication of this contribution.

References

- Álvarez López, Cristóbal. 2017. *Estudio lingüístico del judeoespañol en la revista 'Aki Yerushalayim'*. Ph.D. Dissertation. University of Seville. <http://hdl.handle.net/11441/68487> (last accessed on 16/02/2026).
- Anderson, Benedict. 1983. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.
- Avineri, Netta Rose. 2012. *Heritage Language Socialization Practices in Secular Yiddish Educational Contexts: The Creation of a Metalinguistic Community*. Ph.D. Dissertation. University of California. <https://escholarship.org/uc/item/9f50n171> (last accessed on 16/02/2026).
- Benbassa, Esther & Rodrigue, Aron. 2000. *Sephardi Jewry: A history of the Judeo-Spanish community, 14th–20th centuries*. Berkeley: University of California Press.
- Beguš, Gašper & Dąbkowski, Maksymilian & Rhodes, Ryan. 2025. Large Linguistic Models: investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence* 6(12). 3453–3467. <https://doi.org/10.48550/arXiv.2305.00948>
- Bialystok, Ellen & Ryan, Ellen B. 1985. Toward a definition of metalinguistic skill. *Merrill-Palmer Quarterly* 31(3). 229–251.
- Blommaert, Jan. 2005. *Discourse: A Critical Introduction*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511610295>
- Brink-Danan, Marcy. 2011. The meaning of Ladino: the semiotics of an online speech community. *Language & Communication* 31(2). 107–118. <https://doi.org/10.1016/j.langcom.2010.08.003>
- Bucholtz, Mary. 2003. Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics* 7(3). 398–416. <https://doi.org/10.1111/1467-9481.00232>
- Bunis, David. 2019. La Ortografía de Aki Yerushalayim: un pinakolo en la estoria de la romanización del djudezmo (djudeo-espanyol). *Aki Yerushalayim* 101. 8–24.
- Bunis, David. 2020. Sephardic customs as a discourse topic in the *Ladinokomunita* internet correspondence circle. In Lifshitz, Joseph Isaac & Feuchtwanger-Sarig, Naomi & S. Baumgarten, Goldin, & Diner, Hasia (eds), *Minhagim: Custom and practice in Jewish life*, 161–197. Berlin: De Gruyter.

- Bunis, David. 2021. La lingua de los sefaradim en romanizacion. In Schwarzwald, Ora (ed.), *Las Ortografías del Ladino*, 21–37. Jerusalem: La Akademia Nasionala del Ladino en Israel.
- Bürki, Yvette. 2021. El activismo lingüístico y la lengua sefardí en el siglo XXI. In Jovanovic, Zeljko & Sánchez Pérez, María (eds), *Ovras son onores: estudios sefardies en homenaje a Paloma Díaz-Mas*, 211–32. Bilbao: University of the Basque Country.
- Calvet, Louis-Jean. 1974. *Linguistique et colonialisme : petit traité de glottophagie*. Paris: Payot.
- Creese, Angela & Blackledge, Adrian & Takhi, Jaspreet Kaur. 2014. The ideal “native speaker” teacher: negotiating authenticity in a multilingual classroom. *The Modern Language Journal* 98(4). 937–951. <https://doi.org/10.1111/j.1540-4781.2014.12126.x>
- Cruz Çilli, Kenan. 2021. Ladino’s Renaissance: for this dying Jewish language, COVID has been an unlikely lifesaver. *Haaretz*, January 05, 2021. <https://www.haaretz.com/middle-east-news/2021-01-05/ty-article-opinion/.premium/ladinos-renaissance-covid-the-unlikely-lifesaver-for-a-dying-language/0000017f-f461-d47e-a37f-fd7df1120000> (last accessed on 16/02/2026).
- Díaz-Mas, Paloma. 2006. *Los sefardíes: Historia, lengua y cultura*. Riopiedras.
- Eisenlohr, Patrick. 2004. Language revitalization and new technologies: cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology* 33. 21–45. <https://doi.org/10.1146/annurev.anthro.33.070203.143900>
- Enkontros de Alhad. 2020. Online weekly program in Ladino created by Liliana and Marcelo Benveniste. <https://esefarad.com/tag/enkontros-de-alhad/> (last accessed on 15/03/2026).
- Estreya Perez. 2025. Ladino language chatbot created by Yebra López, Carlos & Acero Ayuda, Alejandro & Öktem, Alp. <https://t.me/estreyaladinobot> (last accessed on 16/02/2026).
- Ethnologue. 2024. *Judeo-Spanish*. SIL International. Retrieved from <https://www.ethnologue.com/language/lad> (last accessed on 16/02/2026).
- Fishman, Joshua. A. 1991. *Reversing Language shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters 76.
- García, Ofelia & Wei, Li. 2014. *Translanguaging: Language, Bilingualism and Education*. New York: Palgrave Macmillan. <https://doi.org/10.1057/9781137385765>

- Gee, James Paul. 2003. What video games have to teach us about learning and literacy. *Computers in Entertainment* (1)1. 1–4. New York: Palgrave Macmillan.
<https://doi.org/10.1145/950566.95059>
- Harris, Tracy. 1994. *Death of a Language: The History of Judeo-Spanish*. Newark, NJ: University of Delaware Press.
- Hassán, Iacob. 1995. El español sefardí (judeo-español, ladino. In Seco, Manuel & Salvador, Gregorio (eds), *La lengua española hoy*, 117–40. Madrid: Fundación Juan March.
- Held, Michal. 2010. The people who almost forgot: Judeo-Spanish web-based interactions as a digital home-land. *El Presente: Studies in Sephardic Culture* 4. 83–101.
- Hernández González, Carmen. 2001. Un viaje por Sefarad: la fortuna del judeoespañol. *El español en el mundo*. Instituto Cervantes/Plaza & Janés/Círculo de Lectores.
- Irvine, Judith T. & Gal, Susan. 2000. Language ideology and linguistic differentiation. In Kroskrity, Paul V. (ed.), *Regimes of Language: Ideologies, Politics, and Identities*, 35–84. Santa Fe: School of American Research Press.
- Kaushanskaya, Margarita & Blumenfeld, Henrike & Marian, Viorica. 2020. The language experience and proficiency questionnaire (LEAP-Q): ten years later. *Bilingualism: Language and Cognition* 22(2). 1–6.
<https://doi.org/10.1017/S1366728919000038>
- King, Kendall. A. 2000. Language Ideologies and Heritage Language Education. *International Journal of Bilingual Education and Bilingualism* 3(3). 167–184.
<https://doi.org/10.1080/13670050008667705>
- Kono, Nariyo. 2019. Linguistics: community-based language revitalization. In Iphofen, R. (ed.), *Handbook of Research Ethics and Scientific Integrity*, 1–13. Cham: Springer. https://doi.org/10.1007/978-3-319-76040-7_45-1
- Kornai, András. 2013. Digital language death. *PLoS ONE* 8(10). 1–11.
<https://doi.org/10.1371/journal.pone.0077056>
- Kroskrity, Paul. V. 2010. Language ideologies—evolving perspectives. In Jaspers, Jürgen & Östman, Jan-Ola & Verschueren, Jef (eds), *Society and language use*, 192–211. Amsterdam: John Benjamins. <https://doi.org/10.1075/hoph.7.13kro>
- Ladino 21. 2017. Community Interest Company. n.d. Ladino 21. www.ladino21.org (last accessed on 16/02/2026).
- Ladinokomunita. 1999. Ladino language listserv created by Rachel Amado Bortnick. <https://ladinokomunita.groups.io/> (last accessed on 16/02/2026).
- Landow, George P. 2006. *Hypertext 3.0. Critical Theory and New Media in an Era of Globalization*. Baltimore: Johns Hopkins University Press.

- Lindholm, Charles. 2013. *Culture and Authenticity*. Malden, MA: Wiley-Blackwell.
- Linnekin, Jocelyn. 1991. Cultural invention and the dilemma of authenticity. *American Anthropologist* 93(2). 446–449.
<https://doi.org/10.1525/aa.1991.93.2.02a00120>
- Los Ladinadores. 2020. Facebook group in Ladino created by Aldo Sevi. <https://www.facebook.com/groups/ladinadores/> (last accessed on 16/02/2026)
- Luger, Eva & Sellen, Abigail. 2016. “Like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, May 7-12, 2016), 5286–5297. New York: Association for Computing Machinery. <https://doi.org/10.1145/2858036.2858288>
- Minervini, Laura. 2006. El desarrollo histórico del judeoespañol. *Revista Internacional de Lingüística Iberoamericana* 8. 13–34.
- Montrul, Silvina. 2015. *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139030502>
- Murray, Janet H. 1997. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. New York: Free Press.
- Nekoto, Wilhelmina & Vukosi, Marivate, & Tshinondiwa, Matsila & Fasubaa, Timi & Fagbohunge, Taiwo & Akinola, Solomon Oluwole & Muhammad, Shamsuddeen & Kabenamualu, Salomon Kabongo & Osei, Salomey & Sackey, Freshia et al. 2020. Participatory research for low-resourced machine translation: a case study in African languages. In Cohn, Trevor & He, Yulan & Liu, Yang (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2144–2160. Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.findings-emnlp.195>
- Noble, Safiya. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
<https://doi.org/10.2307/j.ctt1pwt9w5>
- Otheguy, Ricardo & García, Ofelia & Reid, Wallis. 2015. Clarifying translanguaging and deconstructing named languages: a perspective from Linguistics. *Applied Linguistics Review* 6(3). 281–307.
<https://doi.org/10.1515/applirev-2015-0014>
- Öktem, Alp & Zevallos, Rodolfo & Moslem, Yasmin & Güneş Öztürk, Özgür & Gerson Şarhon, Karen. 2022. Preparing an endangered language for the digital age: The Case of Judeo-Spanish. In Ojha, Atul Kr. & Liu, Chao-Hong & McCrae, John P. (eds), *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference* (Marseille, June 20-25, 2022), 105–110. European Language Resources

- Association. <https://aclanthology.org/2022.eurali-1.18/> (last accessed on 16/02/2026).
- Papo, Eliezer. 2020. Kualo es el Ladino? Eliezer Papo i Alice Corr para Ladino 21 // Londra, Inglaterra. in Ladino 21. Published on February 7, 2020. YouTube video, 8:40. <https://www.youtube.com/watch?v=Fsb7xqSXGrU> (last accessed on 16/02/2026).
- Polinsky, Maria & Kagan, Olga. 2007. Heritage languages: in the ‘wild’ and in the classroom. *Linguistics and Language Compass* 1(5). 368–395. <https://doi.org/10.1111/j.1749-818x.2007.00022.x>
- Quintana Rodríguez, Aldina. 2007. *Geografía lingüística del Judeoespañol: estudio sincrónico y diacrónico*. Lausanne, Switzerland: Peter Lang. <https://doi.org/10.3726/978-3-0351-0275-8>
- Quintana Rodríguez, Aldina. 2010. El judeoespañol, una lengua pluricéntrica al margen del español. In Díaz-Mas, Paloma & Sánchez Pérez, María (eds), *Los sefardíes ante los retos del mundo contemporáneo: identidad y mentalidades*, 31–52. Madrid: Consejo Superior de Investigaciones Científicas.
- Santacruz, Daniel. 2019. Ladinokomunita: 20 Years Old and Still Going Strong. *KolSefardim: News and Features about the Sephardic World*, December 8, 2019. <https://www.kolsekardim.net/post/ladinokomunita-20-years-old-and-still-going-strong> (last accessed on 16/02/2026).
- Schwarzwald, Ora. 2019. Judeo-Spanish throughout the Sephardic Diaspora. In Benor, Sarah (ed.), *Languages in Jewish Communities, Past and Present*, 145–84. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9781501504631-007>
- Séphiha, Haïm Vidal. 1977. Archaïsmes lexicaux du ladino (judéoespagnol Calque). *Cehm Cahiers De Linguistique Hispanique Médiévale* 2(1). 253–61.
- Silverstein, Michael. 1979. Language structure and linguistic ideology. In Clyne, Paul R. & Hanks, William F. & Hofbauer, Carol L. (eds), *The Elements: A Parasession on Linguistic Units and Levels*, 193–247. Chicago: Chicago Linguistic Society.
- Šmid, Katja. 2002. Los problemas del estudio de la lengua sefardí. *Verba Hispanica* 10(1). 113–24. <https://doi.org/10.4312/vh.10.1.113-124>
- Tollefson, James W. & Pérez-Milans, Miguel. 2019. (eds), *The Oxford Handbook of Language Policy and Planning*. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190458898.001.0001>
- UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. Language Vitality and Endangerment. UNESCO. <https://ich.unesco.org/doc/src/00120-EN.pdf> (last accessed on 16/02/2026).
- Van Leeuwen, Theo. 2001. *Discourse and practice: New tools for critical discourse analysis*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195323306.001.0001>

- Weinreich, Max. 1973 [2008]. Yiddish in the Framework of Other Jewish Languages. In Glasser, Paul (ed.), *History of the Yiddish Language*, 124–74. New Haven/London: Yale University Press.
<https://doi.org/10.1017/S0037677900015242>
- Woodbury, Anthony C. 2011. Language documentation. In Austin, Peter K. & Sallabank, Julia (eds), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511975981.009>
- Yebra López, Carlos. 2021. The digital (de)territorialization of Ladino in the twenty-first century. *WORD* 67(1). 94–116.
<http://doi.org/10.1080/00437956.2021.1880065>
- Yebra López, Carlos. 2024. *Ladino on the Internet: Sepharad 4*. London: Routledge. <https://doi.org/10.4324/9781003473664>

Écrire des textes cliniques simplifiés pour les familles aux urgences pédiatriques avec l'IA générative: tensions entre auctorialité et charge de travail

Louise Robert (ERIOS, Espace de Recherche et d'Intégration des Outils Numériques en Santé, CHU de Montpellier, France; UR LHUMAIN, Université Paul-Valéry-Montpellier, France) & Jérémie Sauvage (UR LHUMAIN, Université Paul-Valéry-Montpellier, France)

louise.robert(at)chu-montpellier.fr, jeremi.sauvage(at)univ-montp3.fr

Abstract

Cet article examine les conditions d'usage d'un outil d'intelligence artificielle générative pour la rédaction de lettres explicatives destinées aux patients et leurs accompagnants dans un service d'urgences pédiatriques, où la volonté de concilier amélioration de l'information aux familles et gain de temps pour le personnel médical se heurte à la nécessité d'un travail de correction et d'appropriation des textes générés. L'étude s'appuie sur une ethnographie des pratiques d'écriture clinique, une démarche de conception participative et des entretiens avec des médecins ainsi qu'une analyse discursive outillée par la textométrie d'un corpus de 450 textes produits par trois grands modèles de langage : GPT-4o, Gemma 3 et Mistral Medium 3. Les résultats montrent que les productions s'écartent régulièrement des pratiques discursives de l'écriture médicale. L'analyse met en évidence que l'authenticité des textes générés dépend de leur maintien dans ces pratiques discursives et éclaire le travail nécessaire, en amont, pour les expliciter et les traduire en règles de génération, ainsi que la charge potentielle de correction, en aval, lorsque les modèles peinent à appliquer ces règles.

This article investigates how a generative artificial intelligence tool is used to draft explanatory letters for patients and their caregivers in a pediatric emergency department. Efforts to improve the information provided to families while saving time for medical staff are tempered by the need to review, revise, and take ownership of the generated texts. The study combines an ethnography of clinical writing practices, a participatory design approach, and interviews with physicians, alongside a discourse analysis supported by textometric methods applied to a corpus of 450 texts produced by three large language models: GPT-4o, Gemma 3, and Mistral Medium 3. The findings show that these outputs frequently diverge from established discursive practices in medical writing. The analysis further demonstrates that the perceived authenticity of generated texts depends on their alignment with these practices. It also highlights the upstream work required to make such practices explicit and translate them into generation guidelines, as well as the downstream editing workload when models struggle to follow them.

Mots clefs

IA générative, écriture clinique, auctorialité, textes institutionnels simplifiés

Keywords

generative artificial intelligence, clinical writing, authorship, simplified institutional texts



Robert, Louise & Sauvage, Jérémie. 2026.
Écrire des textes cliniques simplifiés pour les familles aux urgences
Special Issue: *The notion of authenticity*. Vol. 4 No.1
DOI: 10.62408/ai-ling.v4i1.51

AI-linguistica. Linguistic Studies on AI-Generated Texts and Discourses
CC-BY-NC-SA 4.0

ISSN: 2943-0070

1 Introduction

Cette étude s'intéresse à l'usage des grands modèles de langage (LLM) en milieu hospitalier, où ils sont principalement associés à la promesse d'alléger la charge documentaire des professionnels de santé. En parallèle, ces modèles sont aussi sollicités pour générer des lettres d'explications médicales personnalisées à destination des patients et des patientes et de leur entourage (Stanceski, Zhong, Zhang et al. 2024 : 1). Cette perspective ouvre cependant de nouvelles tâches pour les médecins qui doivent relire, corriger et s'appropriier les textes générés.

Une tension se dessine alors, d'un côté l'exigence croissante d'une information médicale claire, complète et adaptée aux destinataires, de l'autre la difficulté d'ajouter une charge rédactionnelle à l'agenda déjà saturé des médecins, alors même que l'intelligence artificielle générative (IAG) est censée les en libérer. Ce paradoxe renvoie à une préoccupation centrale du monde hospitalier : la gestion du temps. Il n'est pas sans rappeler l'évolution de l'usage du courrier électronique depuis les années 1990. Présenté comme un outil de rapidité et d'efficacité, il s'est traduit dans les faits par une multiplication des échanges et, paradoxalement, par une augmentation de la charge de travail liée au traitement des communications (Barley, Meyerson et Grodal 2010 : 888).

Notre étude prend précisément cette tension comme objet. Elle interroge les conditions dans lesquelles les médecins peuvent reconnaître ou non les textes générés comme leurs propres écrits, et ce qui fragilise cette reconnaissance. Autrement dit, il s'agit d'analyser les comportements langagiers des LLM qui perturbent l'auctorialité clinique et rendent nécessaire un travail humain de correction et d'appropriation. Nous abordons la question de l'authenticité des textes générés à partir de la notion d'auctorialité, entendue comme la conformité aux normes discursives professionnelles de l'écriture médicale qui conditionnent la possibilité pour les médecins de valider et signer les textes en leur nom.

Dans le cadre étudié, l'outil d'IAG génère une lettre explicative à partir d'une synthèse clinique, c'est-à-dire d'un résumé de la situation médicale du patient rédigé par les médecins. Le dispositif se situe dans une phase de pré-déploiement : les textes produits au format lettre ne sont pas encore destinés à être remis aux personnes concernées. Ils constituent des simulations de lettres, utilisées uniquement pour l'analyse et la discussion avec les médecins. Notre étude s'inscrit dans cette étape préalable au cours de laquelle les médecins examinent les textes générés et identifient les ajustements qu'il faudrait réaliser en situation d'usage réel. Nous employons le terme *texte* pour désigner ces productions brutes, le terme *lettre* étant réservé aux documents relus et validés par les médecins.

La méthodologie combine une démarche de conception participative, la constitution d'un corpus multi-modèles et une analyse discursive appuyée sur la textométrie permettant d'articuler la mesure quantitative des écarts récurrents par rapport aux pratiques d'écriture clinique avec l'interprétation qualitative issue d'entretiens avec les médecins. Nous proposons pour cela une évaluation

comparative du comportement de trois grands modèles de langage, GPT 4o, Gemma 3 et Mistral Medium 3, testés à partir de situations médicales réelles pseudonymisées. L'analyse porte sur trois dimensions discursives, le ton, la modalisation et la complétude, qui ne constituent pas seulement des critères de conformité formelle mais, comme nous le montrerons, de véritables épreuves d'auctorialité. Elles déterminent si un texte peut être assumé, signé et intégré dans la pratique clinique.

Dans la lignée des approches socio-pragmatiques et interactionnistes (Austin 1970 ; Searle 1972 ; Goffman 1974 ; Habermas 1987), notre étude interroge la capacité des LLM à produire des textes conformes aux attentes discursives propres au discours médical. Toutefois, les productions générées relèvent d'assemblages computationnels fondés sur des corrélations statistiques, qui n'intègrent pas les dimensions interactionnelles incarnées constitutives des pratiques discursives humaines.

Notre questionnement de recherche est double : d'une part, analyser la manière dont les grands modèles de langage transforment les textes cliniques et les écarts discursifs qui en résultent ; d'autre part, examiner les effets de ces écarts sur l'authenticité des textes, en particulier le travail de correction et d'auctorialisation qu'ils impliquent pour les médecins, et la tension entre ce travail et la promesse de gain de temps associée aux outils d'IAG.

Nous reviendrons, dans un premier temps, sur le contexte de l'étude ainsi que sur le cadre théorique dans lequel elle s'inscrit, en particulier à travers un état de l'art portant sur les interactions langagières en général et sur les « vrais faux dialogues » entre un humain et une machine. Nous présenterons ensuite la méthode de recueil des données, qui s'est déroulée en plusieurs étapes. Dans un premier temps, des ateliers ont été menés avec les médecins et ont permis d'identifier des phénomènes discursifs jugés saillants. Dans un second temps, un corpus de 450 textes générés avec l'outil d'IAG, c'est-à-dire des sorties non corrigées par les médecins, a été constitué. La présence des phénomènes précédemment identifiés y a ensuite été quantifiée afin de comparer les trois LLM mobilisés (GPT-4o, Gemma 3 et Mistral Medium 3). Enfin, les textes ont fait l'objet d'entretiens d'explicitation avec les médecins, afin de discuter la présence de ces phénomènes et la manière dont ils étaient évalués au regard de leurs pratiques d'écriture médicale.

2 Contexte de l'étude

Dans cette section, il nous semble nécessaire de prendre le temps de décrire le contexte de notre étude, en partant du terrain avant de revenir sur le cadre théorique de notre réflexion.

2.1 ERIOS, un service du Centre hospitalier universitaire (CHU) de Montpellier

Cette recherche examine la production de textes médicaux simplifiés avec un agent conversationnel développé au sein d'ERIOS, un service dédié à l'expérimentation de l'intelligence artificielle générative (IAG) au CHU de Montpellier (France). Situé au cœur de l'hôpital, ERIOS vise à favoriser la collaboration entre professionnels de santé, ingénieurs informaticiens et chercheurs, notamment en sciences humaines et sociales, afin d'étudier dans quelle mesure ces technologies peuvent contribuer à améliorer la qualité des soins et l'expérience des patients

Depuis plusieurs décennies, de nombreuses études ont souligné la nécessité de mieux répondre aux besoins informationnels des familles dans les services d'urgences pédiatriques (Hoek, Anker, Van Beeck et al. 2019 : 436). Les parents expriment le besoin d'obtenir des informations générales sur les maladies infantiles, incluant le diagnostic, les traitements ainsi que les signes et symptômes à surveiller. Ces travaux ont montré qu'une information appropriée peut influencer les décisions lors d'épisodes ultérieurs de maladie et permettre un recours plus pertinent aux soins. Les parents consultent fréquemment les urgences dans un objectif de réassurance, en particulier lorsque leur littératie en santé est limitée et que leur perception de la gravité diverge de celle des soignants (May, Brousseau, Nelson et al. 2018 : 289). Cette difficulté de compréhension apparaît particulièrement au moment de la sortie : une part importante des parents déclare ne pas maîtriser clairement les instructions post-hospitalisation, ce qui compromet l'adhésion aux recommandations médicales (Simmons, Sharp, Fowler et al. 2015 : 261). Les lettres explicatives personnalisées apparaissent dans ce contexte comme un levier de prévention et d'efficacité des soins, puisqu'elles peuvent à la fois soutenir la compréhension des familles, améliorer la mise en œuvre des recommandations et limiter les réadmissions évitables. Une étude récente montre par ailleurs que plus de 80 % des familles souhaitent recevoir une information spécifique destinée à l'enfant lui-même (Curran, Tavender et Collins 2024).

2.2 Un agent conversationnel hospitalier

C'est dans cette perspective que s'inscrit un des cas d'usage de l'assistant ERIOS, un agent conversationnel fondé sur des grands modèles de langage et développé pour les besoins de l'hôpital. Il permet ainsi notamment de générer automatiquement des textes explicatifs à partir de notes médicales, selon des règles issues d'une démarche de *co-design* avec les médecins (Sanders et Stappers 2008 : 6). L'outil produit deux types de textes : d'une part, des textes adressés aux parents ou à l'adulte accompagnant l'enfant, rédigés dans un registre formel et factuel, utilisant le vouvoiement et un lexique médical simplifié mais précis ; d'autre part, des textes destinés à l'enfant ou à l'adolescent, adaptés à son âge et à son niveau de compréhension, caractérisés par le tutoiement et un vocabulaire simplifié dépourvu

de jargon médical. Dans les deux cas, la même rigueur informationnelle et la même complétude sont exigées, en veillant à ne pas transférer au jeune patient la responsabilité de ses suites de soins, qui demeure celle des adultes accompagnants.

3 Cadre théorique

3.2 Interactions langagières et sociales

Les sciences du langage, en particulier dans leurs composantes socio-pragmatiques (analyses des interactions, analyses des conversations, analyse du discours...) s'intéressent depuis longtemps, tant à l'oral qu'à l'écrit aux fonctionnements des interactions verbales entre les êtres-humains. Plusieurs phénomènes participent à cette interaction sociale au cours de laquelle se construit un échange communicatif (sur le plan général), un échange verbal (sur le plan linguistique). S'il nous est impossible de détailler l'ensemble des concepts et des notions qui résultent de plusieurs décennies de réflexion, nous souhaitons malgré tout insister sur certains d'entre eux en montrant certains phénomènes caractéristiques des interactions inter-humaines.

Les actes de langage

Dans la seconde moitié du 20^e siècle, la philosophie du langage s'intéresse au fait que, lorsqu'on parle (dans le sens large d'émettre un énoncé verbal), on agit en plus de parler et on ne signifie pas forcément ce que l'on semble signifier. Ces *actes de langage* (Austin 1970 ; Searle 1972 ; Vanderveken 1988) peuvent se regrouper selon cinq principales catégories :

- I. les *actes locutoires*, qui constituent le niveau basique du sens de l'énoncé : « j'ai froid » est le fait que je verbalise que j'ai effectivement froid ;
- II. les *actes illocutoires*, qui constituent une forme de second degré de l'énoncé : « j'ai froid » signifie en fait qu'il faut fermer une fenêtre ;
- III. les *actes perlocutoires*, qui constituent un objectif actanciel : je dis que j'ai froid pour que tu fermes la fenêtre ;
- IV. les *actes performatifs*, qui constituent une action à travers l'énoncé prononcé : lorsque le Maire dit « je vous déclare unis par les liens du mariage », cet énoncé crée la relation maritale entre les époux ;
- V. les actes indirects de langage, derrière lesquels se cache le sens réel de l'énoncé. Un enfant de 3 ans pourra déclarer à sa mère qu'il a faim pour

obtenir un morceau de chocolat ou un bonbon, ce qui le conduit à élaborer une stratégie cognitive et langagière pour atteindre son but (Grice 1975 ; Bernicot 1982 ; Sauvage 2015).

Implicite et modalisation subtile dans les énoncés linguistiques

Cette dimension de l'implicite dans la production d'énoncés verbaux chez les êtres humains s'apprend très tôt chez les enfants (Sauvage 2020). Les actes indirects de langage sont le résultat d'un raisonnement conscient, conscientisé, qui est fondamental aux humains pour interagir socialement. On peut alors se demander si un outil d'IAG serait capable d'élaborer de telles stratégies en produisant un écrit dont la signification réelle ne serait pas (exactement) celle de l'acte locutoire produit.

Au-delà des implicites, la langue française permet un certain nombre de modalisations comme l'usage du conditionnel (« le patient aurait des difficultés respiratoires... »), de formulations comme « il semble que le patient a des difficultés respiratoires... ». Ces subtilités peuvent toutes se regrouper par le fait que la signification d'un énoncé peut vouloir dire autre chose que ce qu'il semble dire avec une gradation de sens : « il a dit que », « il aurait dit que... », « il a peut-être dit que... » Nous attacherons une importance particulière dans les écrits produits par l'outil d'IAG à ces tournures qui modulent le degré de certitude, le statut énonciatif des informations.

Préservation de la face de l'autre et mécanismes non-linguistiques

Un autre aspect essentiel caractéristique des relations sociales humaines est celui de la « préservation de la face de l'autre » (Goffman 1974). Ainsi, dans les interactions sociales entre humains, locuteurs et interlocuteurs mettent en place des « rites d'interaction » destinés notamment à la conscience de soi et de l'autre dans cette interaction et dont l'objectif premier consiste à préserver la relation sociale en préservant la face de l'autre. En faisant perdre la face de l'autre, le risque est grand de créer une émotion comme la colère ou la tristesse susceptible de créer une rupture du lien social entre les interactants. L'outil d'IAG sera-t-il en mesure de prendre en charge, dans ses tâches de reformulation des énoncés, les dimensions interactionnelles liées à la préservation de la face de l'autre ?

Il nous semble également indispensable de rappeler la place centrale des interactions pour la construction psychique d'un individu ainsi que dans la construction de la signification des discours. Ainsi, les interactions entre l'enfant et son environnement sont le moteur principal de son développement psychique (pensée, langage, conscience ; Vygotski [1934] 1997 ; Bronckart 1997 ; Sauvage 2015), ce qui doit nous interroger à propos de la genèse, c'est-à-dire du codage informatique de l'outil et de la formulation des prompts, conditionnant la production des discours écrits de l'outil d'IAG. Dans cette optique que l'on qualifie

d'*approche interactionniste*, la théorie de l'action et en particulier la notion d'*agir communicationnel* (Habermas 1987) ne sont pas sans rappeler celle de *force illocutoire* dans les approches pragmatiques (Austin 1970). Un locuteur agit en produisant du langage, ce qui revient à considérer qu'on ne fait pas que parler en parlant. Mais, une nouvelle fois, qu'en est-il d'un outil d'IAG ? Comment les prompts permettent-ils ou non certaines intentions, certaines actions ou effets dans la production d'écrits ?

Enfin, l'analyse du discours aborde la nature relationnelle de la textualité avec la notion de dialogisme (Vološinov [1929] 1977) et une attention à la polyphonie et l'interaction des voix et des points de vue. L'intertextualité (Kristeva 1969) décrit quant à elle la présence et la transformation d'énoncés antérieurs dans un texte. Pris ensemble, ces cadres montrent que tout texte s'inscrit dans des réseaux intertextuels et interdiscursifs qui structurent sa production et sa réception. Plus récemment, Tang (2025 : 2) a proposé le concept de « AI-textuality » pour caractériser les productions des grands modèles de langage qui relèvent d'assemblages statistiques d'énoncés préexistants et ne prennent sens qu'inscrites dans leur contexte d'interaction. L'auteur distingue trois dimensions constitutives : les corpus d'entraînement qui constituent le socle intertextuel opaque mobilisé par le modèle ; le prompt de la personne utilisatrice, conçu comme un « texte intérieur » qui textualise son intention discursive ; et enfin l'interface socio-matérielle, qui configure l'échange humain-modèle. Dans le cas des documents médicaux produits avec des LLM, la synthèse clinique co-élaborée par l'équipe médicale constitue ce « texte intérieur » en tant que matériau textuel transmis au modèle.

3.3 Spécificités des interactions humain-machine

Le Traitement automatique du langage (TAL) des années 1980 a notamment donné lieu à l'étude de ce qu'on appelait à l'époque les dialogues homme-machine (Bulut 1990). En 1950, Turing publie un article considéré comme fondateur dans le domaine pour définir et introduire le concept d'intelligence artificielle. Le défi consiste par la suite à faire interagir une machine avec un humain en demandant à ce dernier si, en fonction de la qualité des réponses qu'il obtient à ses énoncés, il pense interagir avec un autre humain ou une machine. Dans cette lignée, Weizenbaum (1966) programme le chatbot ELIZA qui fonctionne en s'appuyant sur les énoncés et les questions des humains, créant chez ces derniers une dépendance émotionnelle. Le simulacre est parfois bluffant pour l'humain qui se laisse séduire par cette simulation : ce qu'on appelle depuis l'effet ELIZA. À partir des années 1980-1990, les travaux sur les réseaux de neurones artificiels (Lippmann 1987 ; Haykin 1998) réactivent une approche connexionniste, qui se distingue de l'IA symbolique en apprenant à partir des données plutôt qu'en appliquant des règles explicites. Si dans les années 2000, les méthodes statistiques appliquées au langage et à la parole permettent le développement de la traduction automatique et de la reconnaissance vocale, un tournant important est marqué en 2017 avec la

publication des Transformers (Vaswani, Shazeer, Parmar et al. 2017). Ces derniers servent de base aux LLM aujourd'hui disponibles. Depuis le début des années 2020, les modèles génératifs connaissent une diffusion rapide, suscitant de nombreuses interrogations sur la manière dont ils transforment les conditions de production et le statut des textes et des images qu'ils génèrent.

Nous inscrivons notre réflexion dans le domaine des humanités numériques. En France, les humanités numériques se situent dans une épistémè différente de celle des *digital studies* (Sauvage et Nourrit 2022). Il s'agit en effet de s'intéresser aux effets de la massification de la numérisation de la société, aux comportements humains qui en découlent, tout en portant un regard scientifique de nature critique pour problématiser les sujets d'étude. Ainsi, en science du langage, si on ne s'intéresse pas au volet technique du codage des outils informatique, leurs usages (et leurs limites) font pleinement partie de nos préoccupations, qu'il s'agisse d'analyser des discours écrits avec l'aide de la statistique textuelle (De Melo Martins Kuyumjian, Miquel et Sauvage 2024), des pratiques pédagogiques (Grassin, Sauvage, Lacroix et Wigham 2024) ou des interactions humains / objets connectés (Perea 2022).

3.4 Le poids des représentations sociales

La complexité d'une interaction langagière entre deux humains permet donc de considérer une réalité qui dépasse le strict niveau linguistique et verbale. Dans ce cadre, Charaudeau (1983) a proposé une sorte de modèle illustrant cette complexité. Si les enfants apprennent très tôt ces processus socio-cognitifs et leurs enjeux extralinguistiques, nous nous demanderons ce qu'il en est dans une interaction humain / machine. Ainsi, lorsque deux individus se parlent, notamment à cause des actes de langage évoqués *supra*, on peut dédoubler le « je » et le « tu » en faisant se superposer le « je » qui parle (*Je énonçant* - Jeé) et le « je » qui gère une relation communicative (*Je communiquant* - Jec). De même, dans ce discours, le Jeé s'adresse à un *Tu destinataire* (Tud), tandis que le Jec s'adresse à un *tu interprétant* (Tui), celui-là même dont il est nécessaire de préserver la face pour ne pas mettre à mal l'interaction communicative.

Cela présuppose donc l'existence de représentations sociales de soi et de l'autre, ce qui a pour conséquence la construction même du discours produit. Par exemple, les précautions discursives seront différentes si un médecin explique une pathologie complexe à un confrère ou à son patient. C'est d'ailleurs l'un des aspects constituant notre objet d'étude dans cet article. Ainsi, l'importance des représentations sociales dans les interactions langagières humaines nous interrogent sur les conséquences de l'absence de représentations chez un outil d'IAG. Pourtant, l'illusion peut parfois être étonnante (effet ELIZA), notamment parce que, dans une interaction homme-machine, l'humain peut prêter des intentions et des fonctionnements cognitifs à l'IAG, c'est-à-dire lui prêter une prise en compte d'un Tui et donc d'un Jec qui n'existe pas chez un programme informatique. Bien que le

discours produit par l'outil d'IAG paraisse proche d'un comportement humain, il ne mobilise que le couple Jeé / Tud, tandis que l'humain superpose habituellement à cette relation le couple Jec / Tui.

4 Dispositif méthodologique et présentation des données

Avant de détailler les différentes étapes du dispositif méthodologique, il importe d'en situer le contexte institutionnel. La Figure 1 ci-dessous présente la chronologie conjointe du projet institutionnel d'IAG et des étapes de l'étude menée en contexte écologique.

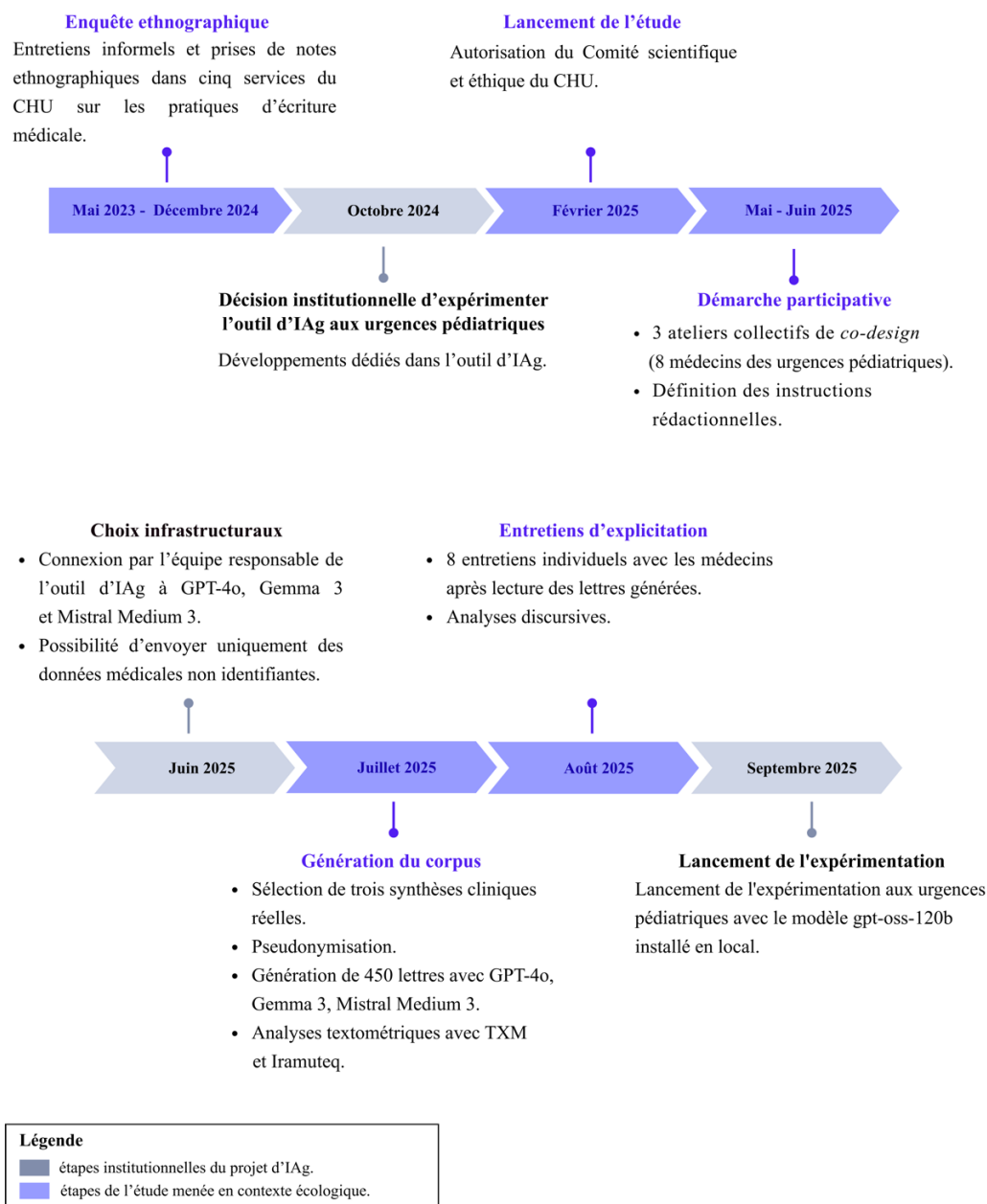


Figure 1: Chronologie du dispositif institutionnel et de l'étude.

4.1 Observations ethnographiques et co-construction des règles du prompt

La première phase de la recherche a consisté en une enquête ethnographique conduite régulièrement pendant deux ans dans différents services du CHU de Montpellier. Adoptant une perspective ethnométhodologique, cette enquête a porté sur les pratiques d'écriture et sur la manière dont les médecins les perçoivent. Ces observations constituent le cadre d'arrière-plan de l'étude et ont fourni le contexte préalable nécessaire pour comprendre la manière dont les médecins mobilisent et évaluent les textes générés avec l'outil d'IAG.

La seconde phase a consisté en trois ateliers réunissant huit médecins du service d'urgences pédiatriques, un médecin spécialisé en science des données, un médecin responsable de la stratégie en intelligence artificielle de l'établissement et l'une des auteurs de l'article. Ces ateliers visaient à explorer collectivement la manière dont les LLM pouvaient être mobilisés pour la rédaction de lettres médicales. L'objectif était de dégager, à partir des attentes exprimées par les médecins du service des urgences pédiatriques, un ensemble de règles rédactionnelles définissant les propriétés discursives attendues pour ce type d'écrit. Nous inscrivons cette démarche dans la tradition du *participatory design* (Schuler et Namioka 1993 : 9) et du *co-design* (Sanders et Stappers 2008 : 6) qui engagent les acteurs concernés dans la conception de systèmes technologiques à partir de leur usage quotidien et de leurs compétences. Cette approche leur offre un moyen d'expression concret dès le début du processus et les associe directement à la définition et à l'ajustement des modalités d'interaction avec l'outil, notamment par le travail sur le prompt.

Lors de ces ateliers, les entrées pour les modèles étaient produites en direct par les médecins et soumises au modèle GPT-4o, conformément au choix de l'équipe responsable de l'outil d'IAG de faire fonctionner leur dispositif avec ce modèle. Ces entrées prenaient la forme de fausses synthèses cliniques, c'est-à-dire de textes rédigés par les médecins à l'issue de la prise en charge d'un patient et habituellement consignés dans son dossier médical. Chaque synthèse décrivait de manière concise mais complète la situation médicale : antécédents pertinents, motifs de consultation, éléments cliniques observés, examens complémentaires (électrocardiogramme, radiographie, analyses biologiques, etc.) et conclusion diagnostique. Ces textes comprenaient de nombreuses abréviations courantes en pratique clinique telles que COPRO (coproculture), SpO₂ (saturation en oxygène), NFS (numération formule sanguine), CS (consultation) ou UHCD (unité d'hospitalisation de courte durée).

Les productions étaient analysées collectivement et réparties en trois catégories, selon qu'elles comportaient des éléments jugés satisfaisants et donc à conserver, des éléments considérés comme non conformes et à proscrire, ou encore des éléments manquants qu'il convenait d'intégrer. Les discussions autour de ces éléments satisfaisants, non conformes ou manquants ont servi de base à la

formalisation des règles rédactionnelles. Cinquante-neuf règles ont ensuite été intégrées dans le prompt : vingt-sept issues directement des ateliers de *co-design*, et trente-trois relevant de contraintes rédactionnelles, de mise en forme ou de sécurité définies en amont par l'équipe responsable de l'outil d'IAG. Parmi ces règles, l'une imposait un vocabulaire simplifié visant un niveau de compréhension équivalent à celui d'un enfant de huit ans. Ce seuil avait été défini de manière empirique, car au-delà la simplification n'était pas satisfaisante. Cette contrainte s'appliquait à tous les modèles et à toutes les lettres, y compris celles destinées à des adultes. Elle ne concernait cependant que la simplification lexicale : pour le reste, le prompt exigeait une adaptation à l'âge du destinataire, qu'il s'agisse d'un enfant, d'un adolescent ou d'un adulte.

4.2 Constitution du corpus

Trois synthèses cliniques ont ensuite été sélectionnées à partir de dossiers réels et fournies par les médecins. Ces textes, préalablement pseudonymisés et dépourvus de données identifiantes ont servi de base à la génération de trois sous-corpus. Cette étude a été conduite en contexte écologique, c'est-à-dire dans les conditions réelles d'usage de l'outil d'IAG au sein de cet hôpital. Les modèles alors utilisés dans ce cadre institutionnel étaient GPT-4o, Mistral Medium 3 et Gemma 3, conformément aux choix de l'équipe en charge du développement de l'appliquatif et de son raccordement au modèle, et non à ceux des auteurs. L'inférence vers ces modèles était assurée via l'infrastructure certifiée Hébergement de Données de Santé (HDS) de l'établissement, à travers la plateforme Azure Microsoft également certifiée HDS. Les trois synthèses cliniques ont été soumises conformément au cadre légal et l'étude a été validée par le Comité scientifique et éthique du CHU (avis n°A056/2025-01-186/001). Depuis juillet 2025, l'outil repose sur GPT-OSS 120B, un modèle open source installé localement. Cette mise à jour est intervenue après la phase de collecte et de génération du corpus analysé dans cette étude.

L'ensemble du corpus comprend 450 textes générés, volume permettant les comparaisons intra- et inter-modèles. La constitution du corpus répond à l'objectif d'examiner la manière dont les grands modèles de langage appliquent les règles rédactionnelles élaborées lors des ateliers. Celles-ci portant sur des propriétés discursives indépendantes du contenu médical des situations retenues, trois cas cliniques permettent de constituer un matériau empirique suffisamment solide pour analyser ces règles en contexte.

Le premier sous-corpus correspond à un texte au format lettre destiné aux parents, rédigé à partir du cas d'un garçon âgé d'un an pris en charge pour une gastro-entérite aiguë. Le deuxième est adressé à un enfant de dix ans admis pour une fracture de l'avant-bras après une chute de trampoline. À partir de chacun de ces deux cas, 150 textes ont été générés, soit 50 par modèle (GPT-4o, Gemma 3 et Mistral Medium 3) pour un total de 300 textes. Le troisième sous-corpus concerne un enfant de dix ans ayant présenté une crise clastique, définie comme un épisode

d'agitation intense accompagné de gestes de destruction ou de violence dirigés contre l'environnement immédiat. Pour ce cas, deux types de textes ont été produits : l'un à destination des parents, l'autre à destination de l'enfant, soit 25 par modèle et par type, pour un total de 150 textes. Chaque fichier texte (.txt) a été associé à des métadonnées (modèle, numéro, cas d'usage) et intégré dans un fichier centralisateur (.csv).

Après la génération du corpus, un examen des textes a été mené avec huit médecins du service d'urgences pédiatriques. Lors de ces lectures conjointes, ces médecins ont identifié plusieurs formulations jugées inappropriées au regard des règles de rédaction établies lors des ateliers. Ces éléments ont servi de base à la sélection des phénomènes retenus pour l'analyse textométrique, dans la mesure où ils pouvaient être repérés de manière systématique par requête.

Trois règles ont ainsi été retenues pour l'analyse :

- (1) adopter un ton neutre, sans employer par exemple « heureusement » ou « malgré tout » ;
- (2) ne formuler de conséquences, de pronostics ou d'effets d'une action que si cela figure explicitement dans l'entrée ; rester factuel en toutes circonstances, sans ajouter d'interprétation ni de projection ; respecter scrupuleusement les verbatims tels qu'écrits par le professionnel de santé ;
- (3) expliciter l'ensemble des signes observés chez l'enfant ou l'adolescent.

Ces règles ont ensuite servi de points d'ancrage pour construire les axes de l'analyse textométrique afin d'en vérifier la portée. Les analyses ont été principalement menées avec TXM (v0.8.3) combinant recherche lexicale et traitements quantitatifs. L'usage d'expressions régulières a permis d'identifier des variantes lexicales et d'examiner les écarts intra- et inter- modèles par rapport aux attentes rédactionnelles. Pour un des trois sous-corpus la fonctionnalité « Statistiques » d'Iramuteq a été utilisée.

4.3 Entretiens d'explicitation

Huit entretiens d'explicitation (Vermersch 1994) ont été menés immédiatement après que les médecins participants eurent lu plusieurs textes de manière à recueillir un retour circonstancié sur la qualité perçue et sur l'adéquation aux attentes rédactionnelles. Les extraits d'entretiens sont présentés sous forme anonymisée. Chacun des médecins est désigné par un identifiant unique (P1, P2, P3...) et les extraits sont reproduits avec une normalisation minimale de la transcription avec la suppression des hésitations (euh, mh), l'ajout d'une ponctuation visant à faciliter la lecture et des coupes indiquées par [...].

5 Résultats et interprétations

5.1 Perception des tâches d'écriture

Comme indiqué précédemment, l'enquête ethnographique réalisée dans le cadre de l'étude constitue un arrière-plan permettant de comprendre les modalités ordinaires des activités d'écriture et la manière dont elles s'inscrivent dans l'organisation des soins et sont perçues par les médecins. Elle fournit le contexte analytique dans lequel s'inscrivent les résultats présentés ici. Ces observations au long cours mettent en évidence un double constat : d'une part l'écriture sur les logiciels est omniprésente et indispensable à la coordination collective, d'autre part elle est vécue comme lourde, chronophage et faiblement valorisée.

Ces trois extraits d'entretiens reflètent de façon explicite ce ressenti :

- (1) Aux urgences, clairement, la moitié de notre temps, on le perd sur ce logiciel informatique au lieu de voir des patients. (P1)
- (2) C'est une activité chronophage. Donc tout ce qui peut, tout ce qui peut effectivement limiter le temps obligatoire qu'on a, parce qu'on a une obligation de tracer les choses et de tracer l'information qu'on donne en plus et puis tracer le dossier médical. Donc tout ce qui peut compacter ce temps, réduire ce temps. (P3)
- (3) On n'a pas fait médecine pour écrire des comptes rendus. (P2)

L'un des médecins souligne un double constat, en évoquant à la fois la contrainte de temps et la nécessité de mieux informer les familles :

- (4) Moi je pense que si ça peut nous aider sur notre pratique tant mieux, parce qu'on a de plus en plus de patients, de moins de temps imparti malheureusement, beaucoup de temps sur les ordinateurs ou trop de temps et les gens nous reprochent de ne pas avoir assez expliqué. Ils le disent quand ils reviennent ou ils le disent à leur médecin ou même autour de nous. Je trouve nos proches qui disent qu'aux urgences on voit le médecin cinq minutes et puis après on attend pendant des heures et on n'a pas d'explication. (P3)

La question du temps apparaît comme un élément central dans l'intégration de tout outil destiné à soutenir l'écriture médicale. L'introduction d'un outil d'IAG suscite à ce titre un espoir d'allègement de la charge rédactionnelle. Il est à noter toutefois que le projet de génération de lettres explicatives a été porté par les acteurs en charge de la stratégie en intelligence artificielle et du pilotage du développement de l'outil d'IAG, et ne répondait pas à une demande initialement formulée par l'équipe

médicale des urgences pédiatriques. Cette initiative s'inscrit dans une dynamique institutionnelle plus large visant à explorer la possibilité d'une part d'alléger certaines tâches rédactionnelles et de l'autre d'améliorer la qualité de la documentation des soins. Les médecins de ce service sont invités à en évaluer l'usage au regard de leurs pratiques et de leurs contraintes.

Il importe de prendre en compte que la rédaction de lettres simplifiées destinées aux patients et à leurs proches constitue pour les médecins une tâche nouvelle, distincte des courriers habituellement rédigés pour les professionnels. Même avec l'aide d'un outil d'IAG cela représente un travail supplémentaire. C'est pourquoi, tout en reconnaissant l'intérêt potentiel des documents médicaux simplifiés, les médecins expriment néanmoins une réserve initiale quant au temps nécessaire à leur production dans un contexte où les tâches d'écriture sont déjà perçues comme particulièrement lourdes. L'outil d'IAG s'accompagne ainsi d'une tension : il porte la promesse d'un gain de temps sur les tâches rédactionnelles, tout en introduisant une nouvelle possibilité, celle de produire des textes simplifiés destinés aux patients. Pour que cette affordance (Hutchby 2001 : 444) puisse être effectivement déployée, les médecins soulignent qu'elle doit rester compatible avec leur préoccupation centrale : limiter le temps consacré à l'écriture. En somme, une telle tâche n'est acceptable que si l'outil génère des textes immédiatement utilisables, alignés sur leurs attentes rédactionnelles et ne nécessitant, après une relecture attentive, qu'une réécriture minimale.

5.2 Inadéquations discursives et tonalité des textes

L'analyse des inadéquations discursives et de la tonalité des productions constitue un premier indicateur de cet alignement. Au cours des ateliers, les médecins ont relevé que certaines productions comportaient des insertions empathiques ou évaluatives, souvent destinées à rassurer. Cette orientation des modèles peut être liée à la consigne d'adresser le texte à des destinataires non spécialistes, mais aussi au fait qu'ils semblent activer un cadre de relation médecin-patient associé à une certaine proximité discursive, conduisant à des formulations empathiques ou rassurantes. Certains énoncés introduisent ainsi des informations que les médecins n'avaient pas l'intention de transmettre, telles que l'affirmation que le patient allait guérir (« Après t'avoir examiné, nous avons pu voir que tu vas pouvoir guérir à la maison » ; « Tu te sentiras mieux en prenant soin de toi à la maison »). D'autres formulations traduisent une posture empathique mal ajustée, comme « Nous savons que cela n'a pas été facile pour toi ». De telles formulations apparaissent dans l'ensemble des sous-corpus analysés. Lors des entretiens, un premier participant est revenu sur ce point et a souligné l'inadéquation de telles expressions empathiques :

- (5) Alors déjà c'est pas du tout quelque chose qu'on fait nous de dire aux enfants et aux parents ah, ça a dû être difficile tout ça. Et du coup c'était pas naturel. C'est plutôt quelque chose de positif et souvent on les félicite, on

leur donne un diplôme du courage [...] C'est un peu lourd en fait. Comme si c'était une épreuve d'aller aux urgences. (P3)

Un autre participant met en avant la différence entre interaction orale et écrit médical, qui rend l'empathie problématique une fois fixée dans un document :

- (6) Parce que c'est un document qui reste, qui reste, l'écrit reste, les gens peuvent le sortir, le montrer à d'autres personnes. L'empathie elle reste, l'empathie on l'a, nous en direct avec le patient et je trouvais que c'était pas disons que c'est pas ça qu'on cherche à transmettre. L'empathie on la transmet si on cherche à la transmettre mais au moment du contact pas sur un écrit. (P2)

Un troisième rappelle que l'activité médicale dans un service d'urgences ne se prête pas à ce type de discours :

- (7) Donc je pense qu'il faut rester concis avec les mêmes termes, avec pas de notion d'empathie. Ça reste factuel. Nous, on ne peut pas avoir, c'est horrible ce que je vais dire, mais c'est difficile d'avoir de l'empathie aux urgences [...]. Il y a des patients, ou même, je parle je pense des patients onco qui voient leur pédiatre très souvent. Nous on les voit une fois et c'est tout. Donc ouais enfin, on nous demande quand même d'être humain, c'est pas le mais voilà, on répond l'enfant a mal là, on a fait ça, il a ça, point barre. (P7)

Enfin, ce participant souligne le risque de mauvaise interprétation par les familles :

- (8) On peut vite passer dans le côté un peu obscur je trouve de l'IA. Justement quand tu crées de l'empathie ou de l'interprétation que tu n'as pas voulu donner à la personne en face je ne sais pas trop comment ça peut être interprété par la personne qui reçoit ces informations. (P5)

C'est dans ce contexte que la règle rédactionnelle : « adopte un ton neutre, ne pas utiliser par exemple heureusement ou malgré tout. » a été ajoutée au prompt. Cette règle a été formulée à partir des premières productions générées avec GPT-4o au cours desquelles les médecins avaient relevé la présence de tels adverbess (« heureusement, tes céphalées ont presque complètement disparu », « heureusement il va mieux »).

Parmi les deux adverbess mentionnés dans la règle, *heureusement* a été retenu pour l'analyse textométrique. Nous l'avons choisi car c'est l'adverbe dont l'usage avait suscité le plus de commentaires de la part des médecins lors des

premières lectures conjointes, en raison de sa valeur évaluative explicite et de l'effet de subjectivité qu'il introduit.

Nous avons vérifié dans quelle mesure cet adverbe apparaissait dans le sous-corpus correspondant au cas d'un garçon âgé d'un an pris en charge pour gastro-entérite aiguë. Pour cela nous avons utilisé la fonctionnalité « Statistiques » d'Iramuteq en sélectionnant la catégorie grammaticale « adverbe » comme clé d'analyse. Dans la liste ainsi obtenue, nous avons recherché l'occurrence de l'adverbe « heureusement » dans chacun des sous-corpus de 150 textes générés par modèle. Le Tableau 1 en présente la distribution.

Tableau 1: Présence de l'adverbe « heureusement » dans les 150 textes analysés.

Modèle	Textes contenant « heureusement »	Nombre total d'occurrences
GPT-4o	15/50	35
Gemma 3	0/50	0
Mistral Medium 3	0/50	0

Malgré une interdiction explicite, près d'un tiers des textes produits par GPT- 4o (15 sur 50) contiennent l'adverbe « heureusement », tandis que Gemma 3 et Mistral Medium 3 n'en génèrent aucun. Pris isolément, ce résultat pourrait laisser penser que GPT-4o répond moins bien que les deux autres modèles aux attentes du genre discursif professionnel. Il s'inscrit toutefois dans un ensemble plus large d'écarts, qui varient selon les modèles et selon les phénomènes examinés, comme le montrent les analyses suivantes. En effet, au-delà de cet adverbe, les lectures conjointes ont mis en évidence d'autres formulations jugées problématiques dans les productions des trois modèles, mais selon des configurations différentes. Dans plusieurs des productions examinées avec les médecins, figuraient également des formulations jugées inadaptées dans un registre médical. Il s'agissait notamment d'expressions de dédramatisation minimisant certains symptômes (« un petit coup sur la tête », « un peu de fièvre », « petite bosse à l'arrière de la tête ») ou anticipant une guérison (« nous savons que tu pourras guérir à la maison », « si tu suis bien les consignes tu pourras guérir à la maison »). Les entretiens soulignent que, de leur point de vue, ces formulations peuvent être acceptables dans l'échange oral, mais deviennent problématiques dès lors qu'elles sont fixées dans un écrit :

- (9) À l'oral ça pourrait quand même ressembler à ce que l'on peut dire à un patient avant de sortir. Oui, ça ressemble plus à de l'oral que peut-être à quelqu'un qui l'écrit spontanément (P5)

Dans le même ordre d'idées, les médecins relèvent que certaines productions introduisent des adresses qu'ils jugent inadaptées dans un texte médical (« Bonjour », « Chers parents »). De leur point de vue, ce type d'ouverture renvoie

encore une fois davantage aux usages de l'interaction orale en face-à-face qu'aux conventions de l'écrit médical. En somme, toutes ces formulations qui apparaissent comme des marqueurs de proximité discursive produisent certes une ressemblance avec le discours humain, mais pour les médecins, l'authenticité des textes générés ne se mesure pas à ce mimétisme. Elle repose sur leur conformité aux normes discursives professionnelles, qui garantissent leur recevabilité et leur légitimité dans la pratique médicale. Dans ces conditions, tant que subsistent des marques discursives jugées inappropriées, les médecins ne peuvent reconnaître ces textes comme leurs écrits. La validation des textes n'est envisageable qu'à la condition d'effacer au préalable ces formulations, ce qui souligne que la reconnaissance auctoriale demeure le critère déterminant de l'authenticité du texte.

5.3 Des attentes sur la cohérence énonciative

L'inscription dans un registre professionnel mobilise des choix discursifs acquis au cours de l'expérience. Dans les productions examinées, les médecins signalent plusieurs formulations qui associent des informations de manière incongrue dans un contexte clinique (« Son état général était stable sans bave excessive »), ou mobilisent des tournures éloignées du lexique professionnel attendu (« tolérer l'infection », « alors qu'il était en air ambiant », « détresse modérée »). Dans quelques textes censés être destinés aux parents, le terme générique « le patient » apparaît à la place de « votre enfant ». Des alternances entre un registre inutilement technique (« désorientation spatio-temporelle », « médicament anticonvulsivant », « dose de charge ») et des simplifications jugées excessives ont également été relevées. Enfin, certaines constructions s'écartent des pratiques rédactionnelles du service, comme « le cas a été pris en charge par une équipe d'orthopédie » ou « décision de retour à domicile prise à 18h30 ». Si cela n'est pas toujours jugé problématique, les médecins relèvent que ces formulations sont éloignées de leur manière habituelle d'écrire.

À ce sujet, un médecin précise :

- (10) J'aurais peut-être pas formulé pareil mais en soi je trouve que c'est quand même facilement compréhensible, c'est le but, donc. C'est un peu enrobé quoi, c'est un peu, enfin voilà, j'aurais pas forcément écrit ça comme ça, mais, mais bon, le but est qu'ils comprennent. (P2)

Il se produit ainsi un effet paradoxal : les textes apparaissent bien rédigés, structurés en phrases complètes et généralement adaptés aux destinataires non spécialistes, mais demeurent en décalage avec les normes rédactionnelles du service. C'est précisément cet écart qui permet aux médecins d'identifier leur caractère artificiel :

- (11) Nous, on ne fait pas de phrases. Nos phrases ne sont pas tournées. Donc je pense qu'on reconnaîtrait tout de suite les textes avec IA, pour l'instant. (P4)

Dans la continuité, l'extrait suivant met en évidence une artificialité qui ne passerait peut-être pas inaperçue auprès des destinataires des documents :

- (12) Est-ce qu'un parent se rend compte que ce n'est pas un humain qui l'écrit ? Ça, je ne sais pas. Oui, je pense qu'il s'en rend compte. (P4)

Un autre participant estime au contraire que le recours à l'outil d'IAG pourrait être difficile à repérer :

- (13) Ça fait pas trop IA. Moi, je trouve que c'est assez bien fait. (P6)

Ces deux médecins relativisent quant à eux l'importance de chercher à identifier une artificialité des textes :

- (14) C'est un outil, donc finalement, c'est peut-être pas artificiel [...] Pour les familles je ne vois pas forcément le problème de se dire c'est artificiel ou pas. (P7)

- (15) En fait, ça me gêne que vous parliez d'illusion parce qu'on n'est pas dans l'illusion. [...] Je pense qu'il faut que ça reste un outil, quoi. (P3)

Ces propos convergent vers l'idée que l'authenticité des textes générés tient moins à la manière dont les phrases sont tournées ou aux traces de l'usage de LLM qu'à la capacité conjointe des médecins et des LLM à maintenir une cohérence discursive et à prévenir des incohérences locales susceptibles de brouiller le sens.

Comme le résume un médecin :

- (16) C'est pas une priorité que ça sonne humain mais après ce qui est gênant c'est quand les petites incohérences, elles modifient un peu le sens. (P2)

Un autre médecin évoque cette question en soulignant que le LLM ne parvient pas toujours à comprendre le langage médical.

- (17) Il y a des choses qu'il ne comprend pas. Il y a encore des choses comme ça où il y a un sens qui n'est pas, on sent qu'elle n'a pas compris quoi, que la machine n'a pas compris. (P6)

Ces micro-ruptures altèrent la cohérence discursive et fragilisent l'appropriation du texte par les médecins. En outre, certains changements discursifs produits par les LLM ne menacent pas seulement l'auctorialité, mais pourraient également présenter un risque pour le patient et sa famille s'ils étaient remis tels quels. Lors de la lecture conjointe du corpus avec les médecins, plusieurs cas de transformation du conditionnel en formulation assertive ont été relevés dans les textes discutés. Ce phénomène apparaît alors même que le prompt stipule explicitement : *Tu ne dois jamais produire d'interprétation, de conclusion ou d'inférence médicale [...]. Reste factuel en toutes circonstances, sans ajouter d'interprétation ou de projection.*

Les reformulations en assertions sont perçues comme particulièrement problématiques par les médecins en raison de leurs implications cliniques et énonciatives si ces textes étaient remis aux familles. Cette préoccupation fait écho à des travaux qui ont examiné la manière dont les grands modèles de langage traitent les énoncés modalisés ou conditionnels (Holliday, Mandelkern, Zhang et Cedego 2024 : 9). Ces auteurs montrent que les modèles transforment fréquemment des formulations hypothétiques en affirmations catégoriques, révélant un biais vers l'assertion. Dans les écrits médicaux, rester factuel implique pourtant de signaler l'incertitude lorsque celle-ci existe, notamment par l'usage du conditionnel, qui fait pleinement partie des faits à transmettre.

Dans le corpus discuté avec les médecins, ce type de reformulation apparaissait dans différents passages. Par exemple, « cyanose pouvant indiquer un manque d'oxygène » pouvait devenir « cyanose indiquant un manque d'oxygène ». Ce risque d'altération du sens se pose avec d'autant plus d'acuité lorsque les formulations en jeu renvoient non plus à des hypothèses cliniques, mais à des énoncés rapportés que les médecins consignent au conditionnel pour en signaler le statut incertain. C'est ce qu'en dit un des médecins à propos d'un texte où le LLM a remplacé « aurait été percuté par un camarade » par « a été percuté par un camarade » et par « s'est fait pousser au visage à l'école » :

- (17) Cette partie déclarative, en fait, on n'en sait rien [...] Est-ce que c'est sa mère qui l'a poussée dans les escaliers ? Je n'en sais rien. Donc c'est important il faut être vigilant. Ouais, l'IA ne peut pas modifier cette partie-là où on garde des précautions. Et ça, c'est un dossier où je vais faire un certificat médical initial pour les assurances ou si la famille veut déposer plainte, enfin, voilà. [...] Donc d'où l'importance du conditionnel. (P7)

Le sous-corpus lié à une crise clastique d'un enfant de dix ans, survenue dans le contexte d'un conflit parental de longue durée, se prête particulièrement à l'examen du maintien du conditionnel. Il comporte plusieurs formulations au conditionnel, dont un énoncé rapportant des propos attribués aux parents dans une situation conflictuelle : « La mère aurait refusé que l'enfant dise bonjour à son père et la crise aurait démarré dans ce contexte. ». Les médecins relèvent que, dans certains textes

générés, cet énoncé est reformulé de façon affirmative (« la mère a refusé »). Dans ces cas, le refus apparaît non plus comme un élément circonstanciel parmi d'autres, mais comme l'origine même du conflit parental ayant mené à la crise. Cette reformulation assertive et causalisante réduit une dynamique familiale complexe à un déclencheur ponctuel. Pour les médecins, il importe de ne pas introduire dans les textes des formulations susceptibles d'envenimer un conflit familial et d'avoir des conséquences directes sur l'enfant.

Pour examiner le traitement de cet énoncé rapporté, nous avons réalisé une recherche par expressions régulières dans TXM dans le sous-corpus de 75 textes (25 par modèle : GPT-4o, Gemma 3, Mistral Medium 3). Trois cas de figure ont été distingués :

- a) Préservation de la modalisation : nous avons utilisé la requête [word = "mère|maman"] []{0,5} [word = "aurait"] afin d'identifier les occurrences où le conditionnel était maintenu, ce qui signalait le respect de la modalisation énonciative.
- b) Transformation assertive : nous avons recouru à la requête [word = "mère|maman"] []{0,5} [word = "a|ne"] []{0,5} [word = "demandé|refusé|voul.*"] pour repérer les reformulations assertives, typiquement au passé composé ou à l'imparfait (« n'a pas voulu », « ne voulait pas »).
- c) Omission : enfin, nous avons considéré comme omissions les cas où l'énoncé rapporté ne ressortait dans aucune des deux requêtes. Afin d'éviter les faux négatifs, ces absences ont été vérifiées manuellement par lecture des textes, confirmant que l'information avait effectivement disparu du texte généré.

Les résultats montrent que les modèles traitent différemment la modalisation au conditionnel. Mistral Medium 3 conserve le conditionnel dans 16 textes sur 25. Dans 6 cas, il le transforme en assertion (« n'a pas voulu », « ne voulait pas ») et, dans 3 cas, l'énoncé est supprimé. GPT-4o préserve la modalisation dans 12 textes sur 25. Dans 3 cas, il transforme l'énoncé en assertion, et dans 10 cas, il l'omet complètement. Gemma 3 conserve le conditionnel dans seulement 3 textes sur 25. Dans 21 cas, il reformule l'énoncé de façon assertive, le plus souvent à l'imparfait ou au passé composé (« n'a pas voulu », « ne voulait pas »), et dans 1 cas il supprime l'énoncé.

Tableau 2: Modes de traitement du conditionnel selon les modèles.

Modèle	Conservation (conditionnel)	Transformation assertive	Omission	Total
GPT-4o	12	3	10	25
Gemma 3	3	21	1	25
Mistral Medium 3	16	6	3	25
Total	31	30	14	75

Ces résultats soulignent de nouveau que les modèles n'adoptent pas une stratégie discursive uniforme. GPT-4o supprime fréquemment l'énoncé mais conserve plus souvent la modalisation lorsqu'il le reprend. Gemma 3, au contraire, reprend presque toujours l'énoncé mais transforme quasi systématiquement le conditionnel en assertion. Les résultats de Mistral Medium 3 révèlent une orientation vers la conservation du conditionnel. On observe toutefois quelques cas d'assertivisation et, plus marginalement, des omissions. Le fait qu'un modèle transforme ces formes modalisées en énoncés assertifs montre qu'il ne reconnaît pas le conditionnel comme une manière d'exprimer des faits.

La suppression des modalisations n'est toutefois qu'un aspect parmi d'autres des limites rencontrées. Plus généralement, la génération de textes à partir de situations sensibles, comme des conflits familiaux, des cas de maltraitance, des violences ou des difficultés relevant de la pédopsychiatrie, semble requérir un travail spécifique d'ingénierie de prompt dont les modalités demeurent à ce jour insuffisamment stabilisées. Les médecins ont souligné que ce type de situations médicales ne saurait, à ce stade, être transposé tel quel dans un déploiement réel de la génération de lettres cliniques avec un outil d'IAG.

5.4 Variabilité et enjeux d'uniformisation

Cette question de la fidélité des informations transmises se pose également lorsqu'on observe la manière dont les modèles gèrent la complétude des textes. Les cas relevés lors des lectures conjointes concernent notamment des données dont le sens se trouve altéré par une restitution partielle. Par exemple, lors d'une prise en charge d'un patient, il importe pour le médecin responsable de pouvoir identifier l'auteur de chaque information incluse dans le dossier médical. Or, dans certaines productions discutées avec les médecins, des indications relatives à la provenance de l'information, par exemple un avis formulé par un infectiologue extérieur au service ou un signe clinique observé dans un autre établissement avant l'arrivée aux urgences, se trouvent reformulées sans mention de leur provenance et donc implicitement rattachées au médecin signataire du document. De telles suppressions

redistribuent la polyphonie du texte en imputant au médecin signataire des énoncés qui ne relevaient pas de sa responsabilité.

Un autre cas relevé concerne la mention « marbrure, mais habituelle selon les parents ». La marbrure, altération de la coloration cutanée liée à une mauvaise circulation sanguine, peut constituer un signe d'alerte important. La précision selon laquelle ce signe est habituel, rapportée par les parents, permet au médecin de nuancer son interprétation. Dans certains textes générés, seule la mention « marbrure » est conservée, ce qui peut suggérer à tort la présence d'un signe clinique significatif. Enfin, l'énoncé « aucun autre antécédent rapporté » devient dans certains cas « aucun antécédent médical particulier », formulation qui donne l'illusion d'une exhaustivité que l'auteur du texte n'avait pas assumée. Ces transformations reconfigurent la portée pragmatique des énoncés et brouillent la distinction entre ce qui est observé, ce qui est rapporté et ce qui est interprété.

L'analyse porte particulièrement sur les situations où un élément clinique n'est pas restitué du tout, et non simplement altéré. Autrement dit, il s'agit d'examiner si les informations consignées explicitement par le médecin dans les données d'entrée sont effectivement reprises dans les textes générés. La consigne du prompt stipule qu'il faut *expliquer tous les signes observés chez l'enfant, ainsi que les gestes, traitements ou interventions réalisés pendant sa prise en charge aux urgences*. Dans la pratique clinique, *tous les signes* recouvrent à la fois les signes positifs (ex. douleur, fracture) et les signes négatifs ou absents (ex. *pas de déficit moteur ou sensitif*). Ces mentions, qui signalent qu'un symptôme a été vérifié mais non retrouvé, sont essentielles dans la rédaction médicale. Elles permettent d'écarter des hypothèses et, dans ce contexte, de rassurer les familles. Toutefois, les pratiques diffèrent : certains médecins consignent systématiquement les signes négatifs, tandis que d'autres se limitent aux éléments anormaux.

Pour examiner la reprise ou non de ces informations, l'analyse porte sur un sous-corpus de 150 textes censés être destinés aux familles, générés à partir du cas d'un enfant de dix ans admis pour une fracture de l'avant-bras après une chute de trampoline. Les données d'entrée comportent deux formulations signalant l'absence de symptômes : « Aucune perte de connaissance » et « Pas d'autre traumatisme associé ». L'enjeu est de vérifier si ces informations sont restituées systématiquement, et sous quelles formes.

Dans ce sous-corpus, le terme « traumatisme » n'apparaissant jamais dans les textes générés nous avons examiné les reformulations correspondantes et repéré l'usage de « blessure », conformément à la contrainte du prompt imposant un vocabulaire accessible. Pour vérifier la présence effective de ces informations nous avons utilisé la requête suivante dans TXM : [word = "pas"] [0,5] [word="blessure.*|connaissance"]. Elle permet de repérer les occurrences où une négation est associée aux deux informations médicales retenues (perte de connaissance, blessure).

Les résultats montrent que l'absence de perte de connaissance est reprise dans 93 textes sur 150. Avec GPT-4o, seuls 10 textes sur 50 la conservent. Gemma 3 la restitue dans 33 textes, mais l'omet dans 17. Mistral Medium 3 la conserve systématiquement dans 50 textes sur 50. Comme le montre le Tableau 3, la deuxième information est plus rarement reprise, et de manière très inégale selon les modèles :

Tableau 3: Reprise des informations médicales négatives dans les 150 textes.

Modèle	Absence de perte de connaissance (mentionnée)	Absence de perte de connaissance (omise)	Pas d'autres blessures (mentionnée)	Pas d'autres blessures (omise)
GPT-4o	10	40	20	30
Gemma 3	33	17	3	0
Mistral medium 3	50	0	49	1
Total (150)	93	57	72	31

On observe en effet que l'absence d'autres blessures apparaît dans 72 textes sur 150, un total qui s'explique presque entièrement par Mistral Medium 3, qui la conserve dans 49 cas sur 50. GPT-4o la mentionne dans 20 textes seulement, et Gemma 3 dans 3. L'absence de perte de connaissance, en revanche, est restituée beaucoup plus largement : Mistral Medium 3 conserve presque systématiquement les deux informations, Gemma 3 privilégie l'information sur la perte de connaissance au détriment de celle sur les autres blessures tandis que GPT-4o tend à faire l'inverse.

Les disparités observées indiquent donc que certains LLM opèrent une hiérarchisation implicite des signes absents : les uns, jugés cliniquement centraux, sont restitués de manière régulière, tandis que les autres, considérés comme secondaires, sont le plus souvent supprimés. Cette hiérarchisation se manifeste nettement chez Gemma 3 et GPT-4o, alors qu'elle est absente chez Mistral Medium 3.

La mention ou non d'un signe clinique ne peut être envisagée de la même manière selon qu'elle provient d'un médecin ou d'un modèle. Entre pairs, les médecins acceptent cette hétérogénéité des pratiques rédactionnelles. Comme mentionné précédemment, certains consignent uniquement les éléments jugés anormaux, tandis que d'autres notent également ce qui est normal. Comme l'exprime l'un d'eux :

- (18) Ah non, on n'écrit pas tous pareil [...] Moi, j'ai tendance à écrire aussi ce qui est normal. (P1)

Pour autant, lorsqu'un médecin choisit de mentionner ou non un élément de l'examen médical dans le dossier médical, il exerce une hiérarchisation informationnelle relevant de son autorité clinique. En revanche, lorsque cette information a été explicitement inscrite dans le dossier, son omission dans les textes générés est une opération discursive inappropriée opérée par le modèle.

Cette question des omissions ouvre une réflexion plus large concernant la variabilité des pratiques rédactionnelles. Entre pairs, cette variabilité est admise, voire valorisée comme une manifestation de l'autonomie professionnelle. Avec les LLM, toutefois, cette variabilité tend à être reconfigurée dans une logique de standardisation, non seulement pour assurer une cohérence des pratiques rédactionnelles, mais aussi pour améliorer la qualité des données produites sur les patients et, par conséquent, la fiabilité de la documentation clinique. Plusieurs médecins estiment ainsi que l'outil d'IAG pourrait à terme contribuer à « lisser » ces différences en garantissant une meilleure homogénéité et une clarté des textes :

(19) Oui, je pense qu'on a tous des styles très différents. Je pense que l'IA peut justement lisser un peu tout ça [...] Je pense que ce serait pas mal parce qu'on est sûr que tout le monde va comprendre plus la même chose que quand on a chacun son style peut-être. (P5)

Nous retrouvons dans les séquences suivantes cette perception que l'enjeu de l'écriture médicale, à l'instar de la plupart des écrits administratifs, n'est pas d'exprimer une singularité de style mais de garantir l'exhaustivité et la fiabilité des informations transmises. Pour ces médecins, l'auctorialité ne réside pas dans un style personnel mais dans la capacité à produire un texte exact, cohérent et conforme aux normes professionnelles. Ce sont précisément ces dimensions que les médecins estiment pouvoir être soutenues par un outil d'IAG :

(20) Je pense qu'il y a beaucoup de positif dans le fait de lisser les choses parce qu'il n'y a pas les oublis qui sont liés à l'humain. Est-ce que c'est grave de ne pas avoir de style dans nos métiers ? Ce n'est pas très grave. Non, on n'est pas artistes non plus. (P4)

C'est ce que souligne également cet extrait :

(21) Ça peut être reformulé et que tout le monde ait la même formulation à la fin ça ça ne me choque pas du moment où ça a le même sens. Oui, au contraire franchement moi je fais des fautes d'orthographe j'écris avec des abréviations donc non, au contraire. (P3)

En définitive, les médecins articulent deux registres dans le cercle des pairs. La pluralité des pratiques rédactionnelles reste admise comme reflet d'identités professionnelles différenciées mais l'introduction d'un outil d'IAG est envisagée

comme l'opportunité d'une progression vers une plus grande standardisation de l'écriture clinique. Cette standardisation par l'utilisation de LLM ne saurait toutefois conduire à des formulations stabilisées. En effet, les modèles actuels se caractérisent par l'imprévisibilité de leurs choix lexicaux et syntaxiques (Shypula, Li, Zhang et al. 2025 : 1). Toutefois, la standardisation n'épuise pas la question de la variabilité caractéristique des LLM, laquelle n'est pas nécessairement un défaut : elle peut au contraire être perçue comme une garantie que la singularité d'une situation est reconnue. La répétition à l'identique d'un même texte serait, au contraire, interprétée comme une absence d'attention singulière portée au patient :

- (22) Je trouve qu'au contraire c'est bien d'avoir des choses qui sont variables parce que imaginons tu viens plusieurs fois aux urgences c'est la même lettre de liaison ça pose un peu question, non ? Mais tu te dis je ne suis pas bien prise en charge parce que j'ai toujours la même chose alors que ce n'était pas pareil. Je suis venue et ce n'était pas la même chose. Ils m'ont donné la même lettre que la dernière fois. (P5)

Dans cette perspective, la variabilité peut être envisagée comme un facteur qui renforce la dimension incarnée du texte, en rappelant que celui-ci conserve une fonction interactionnelle, puisqu'il est remis et éventuellement mobilisé en situation par un auteur auprès d'un destinataire. À cet égard, la variabilité constitue un avantage potentiel par rapport à des documents préétablis ou standardisés, du type fiches ou conseils génériques, en autorisant une formulation adaptée à la singularité de chaque situation clinique.

6 Discussion

6.1 Déplacements pragmatiques des LLM

Notre étude cherchait à éclairer ce qui, dans les textes générés avec des LLM, permettait ou non qu'ils soient reconnus par les médecins comme d'authentiques textes médicaux. Lors des ateliers, les médecins ont relevé des écarts par rapport aux pratiques discursives médicales, et ceux-ci ont servi de base pour rédiger des règles de génération. Les écarts observés ne relèvent pas de simples maladresses stylistiques, mais de perturbations touchant l'organisation énonciative et pragmatique. Ils concernent d'abord l'introduction d'effets illocutoires et perlocutoires par les LLM. Deux mécanismes principaux se dégagent : d'une part, l'ajout de formulations empathiques, évaluatives, dédramatisantes ou promissives qui introduisent des actes de langage que le médecin n'avait pas l'intention de produire ; d'autre part, la transformation de modalisations en assertions, qui entraîne une modification illocutoire non souhaitée.

Dans le premier cas, les entretiens mettent en avant la matérialité de l'écrit et les effets qu'elle exerce sur la définition de ce qui est dicible ou non dans un texte

professionnel. À ce titre, l'expression de l'empathie ou des formulations exprimant une guérison à venir constituent des exemples éclairants : légitimes et fréquents dans l'échange oral, ils apparaissent déplacés dans cette forme textuelle qui est destinée à être imprimée et conservée. Sur ce sujet, les résultats soulignent que les LLM se limitent au couple Jeé/Tud : ils produisent un texte adressé sans intégrer la dimension Jec/ Tui (Charaudeau 1983). Le modèle ne prend dès lors pas en compte de manière satisfaisante la dimension interprétative du texte, c'est-à-dire la manière dont l'énoncé sera compris, réutilisé et potentiellement opposable dans un cadre clinique. L'écriture médicale repose en effet sur un équilibre entre l'adresse au destinataire principal (ici le patient) au moment présent et l'anticipation des lectures ultérieures, y compris celles de tiers qui n'auraient pas assisté à l'échange oral. En négligeant la dimension Jec / Tui, c'est-à-dire la gestion de l'adresse et de l'interprétation dans l'interaction, les modèles produisent des formulations qui peuvent être recevables dans une situation d'oralité immédiate, mais qui deviennent inappropriées dès lors qu'elles sont figées dans un document qui existera au-delà de l'interaction immédiate. Ce décalage explique pourquoi les médecins identifient ces productions comme non conformes aux attentes de l'écrit clinique et soulignent qu'ils devraient les corriger pour rétablir le cadre énonciatif approprié.

Plus largement, l'implicite, construit et assumé dans les pratiques humaines (Grice 1975), est traité par les modèles dans un fonctionnement orienté vers l'explicitation des énoncés, conformément aux consignes de simplification données dans les prompts. Ce faisant, ils ne parviennent pas à maintenir la prudence énonciative et le juste degré de proximité discursive attendu dans l'écriture médicale. Il est utile de distinguer ces déplacements pragmatiques selon leurs effets. Les formulations empathiques ou atténuatives, comme l'ajout d'un adjectif minimisant un signe clinique (par exemple « petite bosse »), bien que jugées inadaptées à un écrit clinique ne menacent pas la face (Goffman 1974) du patient ou de ses proches, c'est-à-dire leur besoin de ne pas être dépréciés, accusés ou inquiétés par le contenu même du document. La transformation des modalisations en assertions constitue en revanche un déplacement d'un autre ordre. Elle modifie le statut épistémique des faits : un énoncé incertain ou rapporté peut devenir, une fois reformulé, un fait établi, par exemple lorsque « la mère aurait refusé qu'il dise bonjour à son père » est reformulé en « la mère a refusé qu'il dise bonjour à son père ». L'écrit produit alors une version de la situation susceptible d'être réutilisée ou opposée et qui ne correspond pas à l'intention du médecin. Les médecins soulignent, là encore, que cela imposerait des corrections afin de rétablir la prudence énonciative attendue s'ils utilisaient l'outil d'IAG. La démarche les a, en outre, conduit à déterminer que les situations psychiatriques ou sociales, qui mobilisent fortement le conditionnel, ne sont pas compatibles avec la génération de textes simplifiés faute d'un cadrage spécifique des prompts pour ces cas.

6.2 Processus d'auctorialisation et authenticité des textes générés

Dans l'ensemble, les textes générés sont jugés bien écrits et compréhensibles mais ils présentent des incohérences locales et des décalages avec les pratiques d'écriture clinique ainsi qu'avec la fonction d'*accountability* (Garfinkel 1967) qui caractérise les textes médicaux. Celle-ci renvoie à la nécessité pour ces écrits de rendre l'activité intelligible et justifiable pour tout lecteur ultérieur.

Pour les médecins, ce n'est pas leur caractère artificiel qui constitue un obstacle mais le fait que ces écarts discursifs rendent difficile la reconnaissance auctoriale. Celle-ci ne renvoie pas à la singularité d'un style individuel, mais à la position d'auteur-praticien dont les pratiques discursives s'inscrivent dans les normes du champ médical. La cohérence énonciative apparaît dès lors comme un critère décisif : même lorsqu'un texte paraît bien formé il ne peut être reconnu comme authentique que s'il respecte les conventions discursives. L'enjeu n'est donc pas d'effacer les traces de l'utilisation d'un LLM (ou de confondre la machine avec un humain) mais de préserver la face du médecin car signer un texte qui s'écarterait trop des conventions rédactionnelles menacerait son identité professionnelle.

Les médecins rencontrés n'attendent pas que les textes générés reproduisent une voix humaine personnelle, marquée par un style individuel ou par des procédés interactionnels qui instaурeraient une proximité discursive propre aux relations humaines, et a fortiori au domaine du soin. C'est pourtant ce que font les modèles en introduisant des formulations empathiques ou atténuatives, pensées comme des adaptations au destinataire. En revanche, les médecins souhaitent que les textes s'inscrivent dans le genre discursif professionnel qui organise l'écriture médicale : prudence énonciative, hiérarchie et statut des faits, continuité avec le dossier clinique. Dans cette perspective, un texte n'est recevable et donc authentique que s'il respecte ces normes rédactionnelles, qui en garantissent la légitimité professionnelle.

La question des omissions engage une dimension particulière de l'auctorialité. Dans l'écriture médicale, chaque énoncé fonctionne comme un acte (Austin 1970 ; Searle 1972) : il situe une information, précise son statut (observé, rapporté, interprété) et inscrit l'activité du médecin dans le texte. Dans ce « texte intérieur » (Tang 2025 : 5) qu'est la synthèse médicale du patient repose l'intention discursive du médecin. La retextualisation opérée par le modèle peut fragiliser la continuité intertextuelle entre le dossier clinique et la lettre adressée aux familles, imposant au médecin un travail supplémentaire pour la rétablir. Lorsqu'un LLM supprime une mention présente dans les données d'entrée, qu'il s'agisse d'un signe clinique négatif ou de toute autre précision jugée utile par son auteur, il ne fait pas qu'effacer une information. Il efface des prises de position qui, pour les pairs, constituent des marques d'auctorialité. De tels effacements montrent que le modèle ne parvient pas à reconnaître ce qui constitue pour les pairs une marque d'auctorialité : une information peut sembler optionnelle pour un acteur externe au

service, ou pour un LLM, alors qu'elle serait immédiatement reconnue par les médecins du service comme un élément qui engage la position de son auteur.

Dans cette perspective, l'authenticité se définit par la possibilité pour le médecin d'assumer les énoncés produits dans le cadre normé de l'écriture clinique, et non par des propriétés stylistiques du texte. Elle suppose un travail de reprise et de correction visant à renforcer l'auctorialité, entendue comme la manière dont l'activité professionnelle du médecin s'inscrit dans le texte et y rend lisibles le jugement, l'expérience et la responsabilité. L'authenticité d'un texte médical dépend de cette auctorialité : elle repose sur la cohérence entre l'acte médical, sa mise en texte et les conventions du genre discursif professionnel. Un texte généré qui modifie le statut épistémique des faits ou qui redistribue la polyphonie des informations consignées dans le dossier médical du patient rompt cette cohérence et ne peut être signé sans correction préalable. Ainsi comprise, l'authenticité relève d'un processus de mise en conformité. Elle se construit dans l'activité de relecture et, le cas échéant, de correction. L'intervention humaine, même lorsqu'elle se limite à une relecture, constitue la condition de l'auctorialité et rend possible l'attribution légitime du texte au médecin qui le signe.

6.3 Travail de configuration et construction de la confiance

Nous avons rappelé combien les représentations sociales encadrent la production et la réception des discours (Charaudeau 1983). Ainsi, parler d'un outil d'IAG au travail renvoie immédiatement à des imaginaires déjà stabilisés mais hétérogènes. D'un côté, circulent des récits qui présentent l'innovation comme inévitable, valorisent l'efficacité accrue ou projettent la figure du « médecin augmenté ». De l'autre, se manifestent des discours plus critiques, associés aux risques de déqualification ou de remplacement, ce qu'un des participants désigne comme « le côté un peu obscur » (Ex.8).

Nous ne relevons aucune résistance de principe à l'introduction de l'outil d'IAG dans les entretiens. Les réticences tiennent surtout à une vigilance quant au risque que la promesse de gain de temps ne se transforme en charge accrue de relecture et de correction. C'est là une marque de prudence inscrite dans une longue histoire de promesses déçues des technologies de l'écriture médicale et dans l'intensification progressive des tâches liée à l'introduction de nouveaux outils. D'autant plus que la perspective d'un recours aux LLM porte ici sur une nouvelle tâche de documentation médicale, les lettres simplifiées pour les familles, qui viendrait s'ajouter à une charge de travail d'écriture déjà mal vécue.

Or, cela a été posé dès l'introduction : les systèmes d'IAG s'accompagnent de la promesse de délégation des tâches d'écriture. Les médecins reconnaissent l'importance de cette promesse dans un contexte où la rédaction des lettres cliniques est perçue comme une activité ancillaire, ne relevant pas du cœur de leur métier. À titre de comparaison, le contrôle des productions de l'IA par les médecins urgentistes apparaît plus acceptable lorsque les outils sont conçus comme des

algorithmes d'aide à la décision, par exemple dans le domaine des systèmes d'IA de détection en radiologie (Anichini et Geffroy 2021 : 50), que lorsqu'il s'agit du contrôle de textes générés par des systèmes d'IAG. Cette volonté de transfert de la tâche rédactionnelle vers la machine contraste avec d'autres pratiques d'écriture professionnelle, comme l'article scientifique, où l'effort intellectuel consenti et la valeur cognitive du processus constituent une dimension essentielle. Une distinction qui est d'ailleurs partagée par l'ensemble des médecins du service.

L'un des médecins souligne un double constat, en évoquant à la fois la contrainte de temps et la nécessité de mieux informer les familles. Les discours recueillis indiquent que la charge de travail à réaliser sur les logiciels hospitaliers limite le temps que les médecins peuvent consacrer aux explications des soins et qu'ils ont conscience de la nécessité d'améliorer cet aspect. Si l'outil d'IAG est perçu comme pouvant contribuer à répondre à cette demande, notre étude montre toutefois que ce bénéfice pourrait s'accompagner d'un effort de relecture et de correction pour garantir la conformité discursive des textes.

La réduction attendue de la charge d'écriture doit être replacée dans un cadre plus large qui interroge les promesses associées à l'IA en milieu hospitalier. À rebours des discours « présentistes » sur l'IA (Suchman 2023 : 2) et particulier ceux sur l'IAG qui annoncent des dispositifs déjà efficaces et prêts à l'emploi, notre démarche rend visible le travail fourni par les humains pour que les productions des LLM soient opérantes (Saha, Brooker, Mair et al. 2024 : 82 ; Harper et Randall 2024 : 104). Ce travail est loin d'être négligeable. L'élaboration des règles de génération requiert des ajustements successifs, car un prompt ne peut contenir de manière exhaustive l'ensemble des conventions énonciatives professionnelles. Un prompt trop long deviendrait inopérant et certaines règles risqueraient d'entrer en conflit. L'exemple de la reformulation en assertions du conditionnel en fournit une illustration : pour éviter que le modèle n'extrapole, n'invente des conseils aux familles ou n'introduise des interprétations les médecins avaient formulé la nécessité d'ajouter aux règles du prompt l'injonction de rester factuel. Cette consigne a cependant pu renforcer la tendance des modèles à privilégier des formulations assertives. Cela souligne l'ampleur du travail requis pour intégrer l'IAG au monde médical. La description du travail de définition itérative des règles de génération, de corrections et d'ajustements des textes produits, apparaît indispensable pour éviter qu'il ne devienne une nouvelle forme de travail invisible (Star et Strauss 1999 : 10).

Plus largement, si ces outils peuvent contribuer à améliorer l'information médicale, ils participent également à une redéfinition normative du travail d'écriture clinique. Les productions tendent à lisser les variations stylistiques et sont constituées de phrases complètes alors que les textes habituellement rédigés dans le service se présentent plutôt sous forme d'énoncés brefs avec une ponctuation minimale. Les médecins soulignent que ce lissage peut créer un décalage avec leurs manières ordinaires d'écrire. Dans le même temps, ils estiment qu'une homogénéisation pourrait rendre l'écriture médicale plus claire et plus accessible.

De leur point de vue, ce biais rédactionnel peut donc constituer une ressource et améliorer la qualité de la documentation médicale, à condition qu'il ne génère pas de travail supplémentaire de correction et qu'il demeure compatible avec les pratiques rédactionnelles du service. Cette aspiration à une homogénéisation des écrits médicaux par l'introduction d'un nouvel artefact n'est pas nouvelle. Elle peut être envisagée à la lumière des travaux de Berg (1997 : 1080-1088) qui montrent comment les systèmes conçus pour réduire l'incertitude par la standardisation des pratiques cliniques introduisent simultanément de nouvelles formes de contrainte et redéfinissent les marges de manœuvre des professionnels. Pour fonctionner, ces systèmes supposent en amont une standardisation des données et une réduction de l'autonomie dans les manières de faire. Ils prennent ainsi en charge, par leur inscription matérielle, une part de ce qui relevait auparavant des ajustements situés, ce que Berg analyse comme un effet du principe de protocole. L'uniformisation attendue des pratiques d'écriture par le recours à des outils d'IAG apparaît ainsi comme une promesse d'efficacité, mais elle s'accompagne également du risque d'une nouvelle forme d'encadrement des pratiques rédactionnelles, qui ne peut être ignoré.

Cette reconfiguration ne doit pas être interprétée uniquement comme un frein, mais comme un levier : elle rend visible des savoirs tacites et permet leur explicitation. Autrement dit, elle constitue également une ressource heuristique pour la définition de pratiques discursives partagées, ainsi qu'un cadre opératoire à l'implémentation d'un système d'IAG à même de prendre en charge les intrications propres à la pratique clinique. En effet, si les résultats montrent que la conformité discursive ne peut, à ce jour, être garantie par la seule précision des instructions, la démarche participative menée avec les médecins permet en revanche d'identifier ce qui constitue un risque pour l'auctorialité et ce qui entraîne un temps de vérification, voire de correction. À l'inverse, certains éléments peuvent être maintenus sans la compromettre, notamment des formulations qui rendent visible le recours aux outils d'IAG tant qu'elles ne modifient pas le sens et ne menacent pas la face du médecin ou des destinataires.

Le travail collectif sur les règles de formulation configure ainsi ce qui devient dicible et pertinent. Conçus avec les acteurs, les prompts intègrent leurs attentes et leurs conventions professionnelles et l'ensemble de la démarche permet, pour les médecins impliqués, de participer à la configuration de l'outil d'IAG et à l'identification des situations d'usage qu'ils souhaitent. Fischer, Peine et Östlund (2022 : 520) rappellent que l'implication des usagers dans la conception des technologies dans le domaine du soin ne relève pas d'une procédure standardisée mais d'une pratique située, traversée d'ajustements permanents. C'est précisément dans ce cadre que s'inscrit notre étude : au sein d'un espace d'interaction qui constitue un lieu de négociation entre des attentes hétérogènes, qu'il s'agisse de celles des concepteurs, des usagers ou des dispositifs eux-mêmes, auxquelles s'ajoutent les réalités de la pratique clinique. Elle montre comment le travail sur les prompts et leur mise à l'épreuve lors des entretiens avec les médecins fait apparaître

des tensions entre la nécessité de se conformer aux normes discursives professionnelles, les contraintes propres aux grands modèles de langage et les contraintes organisationnelles auxquelles les médecins sont soumis. L'ingénierie des prompts apparaît ainsi comme une pratique co-configurative, dans laquelle médecins et chercheurs se positionnent conjointement face aux comportements discursifs des LLM.

Enfin, cette démarche contribue à éclairer la manière dont se construit la confiance envers les systèmes d'IAG dans ce contexte. À cet égard, Scheek, Rezazade Mehrizi et Ranschaert (2021 : 7967) montrent que le manque de confiance envers les systèmes d'IA tient moins à leurs propriétés techniques qu'aux conditions de conception et de collaboration, notamment lorsque les médecins sont impliqués tardivement ou insuffisamment. Le manque de temps, de ressources et de familiarité avec les outils d'IA alimente alors un scepticisme sur leur utilité clinique. Nos résultats montrent que, pour les médecins rencontrés, l'enjeu porte avant tout sur la crainte que les promesses de gain de temps ne soient pas tenues. Dans cette perspective, les résultats de Winter et Carusi (2022 : 72) sont particulièrement pertinents : la confiance ne repose pas sur une fiabilité technique présumée, mais sur des pratiques collectives de validation. Les échanges recueillis au cours de l'enquête vont dans le même sens et soulignent que la possibilité de participer à l'élaboration des règles de génération et d'en éprouver les effets contribue à instaurer une forme de confiance dans l'introduction de ces systèmes.

7 Conclusion

Notre étude montre que l'usage d'outils d'IAG pour la rédaction de lettres médicales simplifiées ne peut se faire sans prendre en compte les normes discursives qui fondent l'auctorialité de l'écrit médical. À travers une pratique de formulation collective, les médecins ont défini des règles rédactionnelles destinées à encadrer les productions des LLM. L'examen d'un sous-ensemble de ces règles met en évidence que les modèles étudiés ne les appliquent pas systématiquement, ce qui entraînerait, en situation d'usage, un travail de révision. L'appropriation future de l'outil par les médecins reposera ainsi sur un arbitrage entre le bénéfice attendu pour l'amélioration de l'information médicale et le travail de relecture nécessaire pour garantir la conformité discursive des textes. Or ce temps de contrôle des textes entre directement en tension avec la promesse de gain de temps qui accompagne l'introduction d'outils d'IAG à l'hôpital.

L'interprétation de ces résultats doit toutefois être nuancée par plusieurs limites méthodologiques. Le corpus est ciblé et monocentrique, ce qui limite la possibilité de généraliser les résultats à d'autres contextes. Le prompt utilisé a été principalement optimisé pour GPT-4o puis appliqué tel quel à Gemma 3 et Mistral Medium 3. Cette configuration, ainsi que la longueur et la concurrence de certaines instructions, a pu influencer la qualité ou l'alignement de certaines sorties. Cette

limite rappelle que l'objectif de cette étude n'était pas d'optimiser techniquement les prompts mais de montrer le travail de cadrage des normes rédactionnelles que suppose la production d'écrits médicaux avec un outil d'IAG.

Sur le plan théorique, nous proposons d'appréhender l'authenticité des textes médicaux générés par les LLM comme le produit d'une reconnaissance auctoriale qui se construit dans le travail d'appropriation réalisé par les médecins, depuis la rédaction des règles rédactionnelles du prompt jusqu'à la correction des sorties. L'étude montre que l'authenticité d'un texte médical ne tient pas à sa ressemblance avec une production humaine, mais à la possibilité pour le médecin d'en assumer les énoncés dans le cadre normé de l'écriture médicale.

Cette étude a permis une mise à l'épreuve par les médecins des possibilités offertes par les outils d'IAG pour la rédaction de lettres médicales simplifiées dans un cadre de discussion préalable à tout usage réel. Elle se poursuivra lors de la remise effective de lettres aux familles et accompagnants, dans un périmètre d'usage délimité par les médecins du service d'urgences pédiatriques. En définitive, des dispositifs tels que le service ERIOS, conçu comme un espace d'exploration et de mise en discussion des usages de l'IAG à l'hôpital, offrent un cadre où attentes et réserves peuvent être discutées collectivement et où les conditions d'un usage éventuel peuvent être examinées.

Remerciements

Les auteurs remercient Rita Ferreira Pires et Maëlle Kerloc'h Gelin pour leur contribution à l'élaboration du protocole et à la collecte des données, ainsi que Maëlle Kerloc'h Gelin pour la préparation des données et la production des résultats des analyses textométriques.

Conflits d'intérêts

Louise Robert est employée par le CHU de Montpellier et participe à l'évaluation des usages de l'outil d'intelligence artificielle générative étudié dans cet article. Les auteurs déclarent ne pas avoir de conflit d'intérêts financier.

Déclarations

Nous reconnaissons avoir eu recours à l'outil GPT-5 d'OpenAI pour une assistance à la révision de ce manuscrit. Nous avons toutefois veillé à ce que toutes les contributions respectent strictement les standards et les règles éthiques de l'écriture scientifique.

Bibliographie

- Anichini, Giulia & Geffroy, Bénédicte. 2021. L'intelligence artificielle à l'épreuve des savoirs tacites. Analyse des pratiques d'utilisation d'un outil d'aide à la détection en radiologie. *Sciences sociales et santé*. 39(2). 43–69. <https://doi.org/10.1684/sss.2021.0200>
- Austin, John Langshaw. 1970. *Quand dire, c'est faire*. Traduction, introduction et commentaire de Gilles Lane. Paris: Éditions du Seuil.
- Barley, Stephen R. & Meyerson, Debra E. & Grodal, Stine. 2011. E-mail as a source and symbol of stress. *Organization Science* 22(4). 887–906. <https://doi.org/10.1287/orsc.1100.0573>
- Berg, Marc. 1997. Problems and promises of the protocol. *Social Science and Medicine* 44(8). 1081–1088. [https://doi.org/10.1016/S0277-9536\(96\)00235-3](https://doi.org/10.1016/S0277-9536(96)00235-3)
- Bernicot, Josie. 1982. *Les actes du langage chez l'enfant*. Paris: PUF.
- Bulot, Thierry. 1990. L'interaction verbale et le dialogue homme-machine: propositions théoriques. *Cahiers de linguistique sociale* 16. 83–96.
- Bronckart, Jean-Paul. 1997. *Activité langagière, textes et discours: pour un interactionnisme socio-discursif*. Lausanne: Delachaux et Niestlé
- Charaudeau, Philippe. 1983. *Langage et discours: éléments de sémiolinguistique*. Paris: Hachette.
- Curran, Janet & Tavender, Emma & Collins, Stephenie. 2024. Can we leverage technology to enhance discharge communication in pediatric emergency departments? *Canadian Journal of Emergency Medicine*. 26(10). 689–690. <https://doi.org/10.1007/s43678-024-00789-y>
- De Melo Martins Kuyumjian, Naraina & Sauvage, Jérémie & Miquel, Frédéric. 2024. Analyse semi-automatique du discours d'enseignants sur ces élèves qui les élèvent. *LHUMAINE* 4. <https://hal.science/hal-04445982> (consulté le 19/03/2026).
- Fischer, Björn & Peine, Alexander & Östlund, Britt. 2022. Doing user involvement: shifting interstices and coalescing tensions in care technology. *Science Technology & Human Values*. 49(5). 1130–1157. <https://doi.org/10.1177/01622439221143196>
- Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Goffman, Erving. 1974. *Les rites d'interaction*. Traduit de l'anglais par Alain Kihm. Paris: Éditions de Minuit.
- Grassin, Jean-François & Sauvage, Jérémie & Lacroix, Eva & Wigham, Ciara (eds). 2024. *Pratiques et usages du numérique en FLE*. *ALSIC* 27(2). <https://doi.org/10.4000/12rdn>

- Grice, H. Paul. 1975. Logic and Conversation. In Cole, P. & Morgan, J. L. (eds), *Syntax and Semantics, Vol. 3, Speech Acts*, 41–58. New York: Academic Press.
- Habermas, Jürgen. 1987. *The Theory of Communicative Action. Volume 2: Lifeworld and System: A Critique of Functionalist Reason*. Translated by Thomas McCarthy. Boston, MA: Beacon Press.
- Harper, Richard & Randall, David. 2024. Machine learning and the work of the user. *Computer Supported Cooperative Work (CSCW)* 33(2). 103–136. <https://doi.org/10.1007/s10606-023-09483-6>
- Haykin, Simon S. 1998. *Neural Networks: A Comprehensive Foundation*. 2^e éd. Upper Saddle River, NJ: Prentice Hall.
- Hoek, Amber E. & Anker, Susanne & van Beeck, Ed F. & Burdorf, Alex & Rood, Pleunie P. M. & Haagsma, Juanita A. 2019. Patient discharge instructions in the emergency department and their effects on comprehension and recall of discharge instructions: a systematic review and meta-analysis. *Annals of Emergency Medicine* 75(3). 435–444. <https://doi.org/10.1016/j.annemergmed.2019.06.008>
- Holliday, Wesley H. & Mandelkern, Matthew & Zhang, Cedegao E. 2024. Conditional and Modal Reasoning in Large Language Models. *arXiv*. <https://doi.org/10.48550/arxiv.2401.17169>
- Hutchby, Ian. 2001. *Technologies, texts and affordances*. *Sociology* 35(2). 441–456. <https://doi.org/10.1177/S0038038501000219>
- Kristeva, Julia. 1969. *Séméiotikè. Recherches pour une sémanalyse*. Paris: Seuil.
- Lippmann, Richard P. 1987. An introduction to computing with neural nets. *IEEE ASSP Magazine* 4(2). 4–22. <https://doi.org/10.1109/MASSP.1987.1165576>
- May, Meghan & Brousseau, David C. & Nelson, David A. & Flynn, Kathryn E. & Wolf, Michael S. & Lepley, Bryn & Morrison, Andrea K. 2018. Why parents seek care for acute illness in the clinic or the ED: The role of health literacy. *Academic Pediatrics* 18(3). 289–296. <https://doi.org/10.1016/j.acap.2017.06.010>
- Perea, François. 2022. *Des objets qui parlent? Et ce qu'il reste aux humains*. Paris: MkF Éditions.
- Saha, Dipanjan & Brooker, Phillip & Mair, Michael & Reeves, Stuart. 2024. Thinking like a machine: Alan Turing, computation and the praxeological foundations of AI. *Science et Technology Studies* 37(2). 66–88. <https://doi.org/10.23987/sts.122892>
- Sanders, Elizabeth B.-N. & Stappers, Pieter Jan. 2008. Co-creation and the new landscapes of design. *CoDesign* 4(1). 5–18. <https://doi.org/10.1080/15710880701875068>
- Sauvage, Jérémie & Nourrit, Déborah. 2022. Humanités numériques et pensée complexe. *LHUMAINE* 2022-1, décembre 2022.

<https://lhumaine.numerev.com/articles/revue-1/2727-humanites-numeriques-et-pensee-complexe>

- Sauvage, Jérémie. 2015. *L'acquisition du langage. Un système complexe*. Louvain-la-Neuve: Academia.
- Sauvage, Jérémie. 2020. L'implicite dans le développement du langage chez l'enfant ordinaire. *Ortho Magazine* 26(150). 10–12. [https://doi.org/10.1016/S1262-4586\(20\)30031-5](https://doi.org/10.1016/S1262-4586(20)30031-5)
- Scheek, Damian & Rezazade Mehrizi, Mohammad H. & Ranschaert, Erik. 2021. Radiologists in the loop: the roles of radiologists in the development of AI applications. *European Radiology* 31(10). 7960–7968. <https://doi.org/10.1007/s00330-021-07879-w>
- Schuler, Douglas & Namioka, Aki. 1993. *Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum Associates
- Searle, John Rogers. 1972. *Les actes de langage. Essai de philosophie du langage*. Paris: Hermann.
- Shypula, Alexander & Li, Shuo & Zhang, Botong & Padmakumar, Vishakh & Yin, Kayo & Bastani, Osbert. 2025. Evaluating the diversity and quality of LLM generated content. *arXiv*. 1–26. <https://arxiv.org/abs/2504.12522>
- Simmons, Stefanie & Sharp, Brian & Fowler, Jennifer & Fowkes, Hope & Paz-Arabo, Patricia & Dilt-Skaggs, Mary Kate & Singal, Bonita & Carter, Thomas. 2015. Mind the (knowledge) gap: the effect of a communication instrument on emergency department patients' comprehension of and satisfaction with care. *Patient education and counseling* 98(2). 257–262. <https://doi.org/10.1016/j.pec.2014.10.020>
- Stanceski, Kristian & Zhong, Sharleen & Zhang, Xumou & Khadra, Sam & Tracy, Marguerite & Koria, Linda & Lo, Sarita & Naganathan, Vasi & Kim, Jinman & Dunn, Adam G. & Ayre, Julie. 2024. The quality and safety of using generative AI to produce patient-centred discharge instructions. *npj Digital Medicine* 7(1). 1–6. <https://doi.org/10.1038/s41746-024-01336-w>
- Star, Susan Leigh & Strauss, Anselm. 1999. Layers of silence, arenas of voice: the ecology of visible and invisible work. *Computer Supported Cooperative Work* 8(1-2). 9–30. <https://doi.org/10.1023/a:1008651105359>
- Suchman, Lucy. 2023. The uncontroversial 'thingness' of AI. *Big Data & Society* 10(2). 1–5. <https://doi.org/10.1177/20539517231206794>
- Tang, Kok-Sing. 2025. AI-textuality: expanding intertextuality to theorize human-AI interaction with generative artificial intelligence. *Applied Linguistics* XX. 1–19. <https://doi.org/10.1093/applin/amaf016>
- Turing, Alan. 1950. Computing machinery and intelligence. *Mind* LIX(236). 433–460. <https://doi.org/10.1093/mind/LIX.236.433>

- Vanderveken, Daniel. 1988. *Les actes de discours. Essai de philosophie du langage & de l'esprit sur la signification des énonciations*. Liège & Bruxelles: Pierre Mardaga.
- Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Lukasz & Polosukhin, Illia. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. 5998–6008.
<https://doi.org/10.48550/arXiv.1706.03762>
- Vermersch, Pierre. 1994. *L'entretien d'explicitation*. Paris: ESF.
- Vološinov, Valentin Nikolaevič. 1977 [1929]. *Le marxisme et la philosophie du langage. Essai d'application de la méthode sociologique en linguistique*. Traduit et présenté par Marina Yaguello. Paris: Les Éditions de Minuit.
- Vygotski, Lev. 1997 [1934]. *Pensée & langage*. Paris: La Dispute.
- Weizenbaum, Joseph. 1966. ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1). 36–45. <https://doi.org/10.1145/365153.365168>
- Winter, Peter & Carusi, Annamaria. 2022. If you're going to trust the machine, then that trust has got to be based on something: validation and the co-constitution of trust in developing Artificial Intelligence (AI) for the early diagnosis of Pulmonary Hypertension (PH). *Science & Technology Studies* 35(4). 58–77.
<https://doi.org/10.23987/sts.102198>

Humans vs. machines: can enaction theory help?

Luca Pallanti (Université Lumière Lyon2, lab. ECP) & Rawad Chaker (Université Lumière Lyon2, lab. ECP)

luca.pallanti(at)univ-lyon2.fr, rawad.chaker(at)univ-lyon2.fr

Abstract

This study seeks to contribute to ongoing research on the formal differences between textual corpora produced by humans and those generated by generative artificial intelligences (GenAI). To this end, we draw on enactive theory and apply it to linguistics, conceiving the asymmetrical structural coupling between an organism and its environment as the basis of cognition. The relations between an actor and the environment are non-linear and irregular: within the perception–action loop, we selectively attend to elements of the environment that are meaningful for action. We propose a novel approach by using Sentence Length Variance (SLV) as an indicator of this asymmetry, or internal irregularity. We also experimented with prompting the machine to integrate greater variance into its texts. We found significant differences between human corpora and AI corpora in terms of SLV, as well as notable divergences among certain GenAI systems. When prompted for higher SLV, only one GenAI produced texts comparable to human writing. We advance several explanations for these findings, which we regard as promising avenues for future research. We conclude by outlining the theoretical implications of our findings, arguing that an enactive linguistics offers a fruitful framework for studying textual differences between human and AI-generated discourse.

Keywords

enaction, artificial intelligence, linguistics, variance, embodiment

1 Introduction

In the field of education, the widespread dissemination and democratization of Large Language Models (LLMs), together with their adoption by students, increasingly pose challenges for instructors responsible for assessing texts produced remotely by learners (Dergaa et al. 2023; Ghimire, Pather and Edwards 2024; Doru et al. 2025). A growing number of teachers and trainers, both in secondary education and in higher education, struggle to ascertain whether the texts they evaluate constitute original student productions or outputs generated by generative artificial intelligence (GenAI) (Fleckenstein et al. 2024; Nguyen and Barrot 2024). Empirical evidence confirms this concern: Fleckenstein et al. (2024) found that minimally prompted AI-generated texts frequently elude detection by educators. Similarly, Nguyen and Barrot (2024) report that teachers often cannot reliably differentiate AI-generated compositions, even those crafted with professional-level refinement, from traditional student writing. A recurrently voiced position maintains that GenAI texts appear normed and overly regular, conveying a sense of artificiality and inauthenticity (Draxler et al. 2024; Fraser, Dowkins and Kiritchenko 2025).



Pallanti, Luca & Chaker, Rawad. 2026.
Humans vs. machines: can enaction theory help?
Special Issue: *The notion of authenticity*. Vol. 4 No.1
DOI: 10.62408/ai-ling.v4i1.50
ISSN: 2943-0070

This study takes as its point of departure this commonly shared perception, seeking to determine whether the impression of inauthenticity in GenAI texts can be substantiated through observable linguistic features or whether, conversely, such a distinction among students' productions lacks empirical foundation. The present article offers an original interpretation of a syntactic marker commonly used to differentiate human speech from that generated by artificial intelligence systems. This marker is known as sentence length variance (SLV), and it is described in Section 2.2. Here, textual authenticity is understood as a dimension of originality intrinsically linked to young adult texts producers, with particular reference to students starting university, that is, novice rather than expert writers.

The final issue at stake is epistemological in nature. In order to distinguish between the order of students' textual production and that of artificial textual production, we aim to establish a theoretical dialogue between textual linguistics, writing pedagogy, and the enactive paradigm, thereby outlining a hermeneutic framework that may be characterized as an enactive textual linguistics. The following subsections are devoted to elaborating this dialogue.

2 Background

2.1 Textual linguistics and enaction

The enactive approach to cognition is based on the idea that an organism's perception and understanding of the world around them is influenced by the dynamic and reciprocal interaction between the organism and its environment (Varela, Rosch and Thompson 1991; Thompson 2007; Stewart, Gapenne and Di Paolo 2010; Di Paolo, Buhrmann and Barandiaran 2017). According to enaction, perception and cognition are not purely mental processes that occur within our phenomenal environment; rather, they are dynamic interactions between our brain, body, and environment. This implies that our perception and understanding of the world depends not only on our senses, but also on our ability to act and interact with our environment through sense-making.

These theoretical elements form the basis of the theory of enaction and account for the difficulty, if not the impossibility, of an AI "thinking the world" around it. Enaction considers the structural coupling between the actor and the environment to be the unit of analysis of all human activity. This activity is characterised by various properties: the relationship is asymmetrical (*i.e.* non-linear) and depends as much on the internal properties of the actor (*i.e.*, internal states) as on the properties of the environment. This characteristic is inherited by the activity (Di Paolo 2005). Hence, activity is embodied: according to Varela, Rosch and Thompson (1991: 173), "sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition". It is not possible to separate cognition from the socio-material conditions of its emergence: it emerges from the sensory-motor loop, from the asymmetrical structural coupling between

an actor and the social, cultural and material environment in which they are embedded.

The idea of bringing together the paradigm of enaction and linguistics studies is not new. Bottineau (2013) has proposed an unprecedented and necessary rapprochement of linguistic theories and the enactive account with regard to speech. The author supports an embodied vision of speech, which would lead us to rethink grammar through the prism of enaction theory.

The act of speaking is known to generate perceptual effects that, in turn, give rise to interpretative processes. The comprehension of this dynamic is predicated on the re-evaluation of the grammar of a language in terms of the coupling between action and perception. Furthermore, it is contingent on the interrogation of the manner in which its components contribute to the construction and emergence of meaning in consciousness. (Bottineau 2014) [transl. with DeepL]

In our article, we aim to apply the concept of action-perception coupling to the field of writing. The originality of our work lies in our assertion that writing is a deeply embodied task that takes the form of a writing gesture. According to this perspective, the interplay between the writing gesture and the planning of the text within a specific context is what shapes the writing process in an asymmetrical and continually evolving manner.

2.2 Sentence Length Variance (SLV): a tool for thinking about planning/execution asymmetry?

If language is conceived as an action and is therefore composed of traces of embodied experience, we can assume that it is externalised asymmetrically, *i.e.* non-linearly. In the enactive account, asymmetry between an agent and an environment is an organizational condition of a living organism, which has organic, sensorimotor, and sociomaterial realizations (Di Paolo, Lawler and Vaccari 2023). Hence, the agent and the environment have mutually determining relations, instead of causal relationships. We should therefore find “regularities of irregularities” in the form of discourse, *i.e.*, recurring patterns of alternation between short and long, simple and complex elements, reflecting the deep structure of cognitive and discursive activity. Similarly, we can assume that language follows an asymmetrical dynamic whereby some ideas are expressed in brief sentences, while others require more extensive development. This variability can be expressed statistically by calculating variance, a measure of the dispersion of a set of data around its mean. A potential methodology for the comparison of the writing styles of LLMs and humans would be to calculate the standard deviation of the number of words in each sentence. This is referred to as the Sentence Length Variance (SLV), and it is a statistical measure based on the distribution of sentence length. Measuring sentence length is most frequently employed in stylometry-based studies (Muñoz-Ortiz et al. 2024; Bronikowski 2025) and is considered a robust measure related to text structure (Vieira, Picoli and Mendes 2018). Therefore, although SLV

is not a new criterion for differentiating between discursive productions by GenAI and humans (Tan et al. 2023; Zindela 2023), it is the interpretation of SLV in relation to the process of written composition in young adult writers that represents a novel aspect. In our theoretical approach, we link SLV to the asymmetry dimension in the enaction framework.

Nottbusch's (2010) studies show that the production of written sentences obeys a complex system of planning and controlling the activity, linking pauses and sentence complexity. His research confirms Chanquoy, Foulin and Fayol (1996) finding that the longest pauses during the writing process occur at the beginning of linguistic units, such as paragraphs, sentences, or words. This reinforces the idea that planning occurs at the end of these units: "The larger the unit, the longer the preparation time" (Nottbusch 2010: 780). Although the precise reasons for the duration of pauses when planning phrasal structures are unknown, it seems likely that non-linear cognitive activity occurs during sentence production. Furthermore, research questions about planning control are linked to the role of working memory. In the field of sentence comprehension, studies by Roberts and Gibson (2002) confirm the hypothesis that there is a link between working memory and sentence memory, generating a large number of variations in comprehension and presumably text production among individuals. The asymmetry between planning, memory, control and execution, hence prompted us to examine SLV as a human phenomenon.

Assuming this position is justified, SLV could represent an interesting human indicator of authorship, bringing it closer to becoming one of several measures of a text's authenticity. We therefore intend to utilise SLV as a potential tool for representing a complex mental process such as planning control. From this perspective, human SLV, as an asymmetric process, would contrast with the SLV of GenAI, which results from powerful probabilistic calculations. Furthermore, the discovery of a SLV index that is characteristic of humans could represent a breakthrough in the study of academic writing assessment. In the case of technology-mediated situated assessment of writing, for example, SLV could provide learners with additional feedback and help them to develop an understanding of their own writing skills. From another perspective, such an SLV could allow academic trainers to enhance teaching modules for writing in line with academic expectations (Cope et al. 2011). Can we accept the existence of a human-specific figure for SLV at a given stage of development? This question prompted us to formulate several others that structured the entire methodological approach of this study.

3 Problem and hypotheses

This article aims to demonstrate how phrasal variance can be used to authenticate young adult writers texts, as opposed to those generated by GenAI.

To this end, we formulated a preliminary hypothesis to verify whether this variance could serve as an indicator of embodiment and asymmetry in linguistic text analysis.

- *H1: SLV can be used as an indicator of syntactic variability between different corpora.*

We then put forward the hypothesis that we would find more phrasal variance in formal human productions than in a corpus produced by machines. Our second hypothesis tests for significant differences between human and AI-generated corpora.

- *H2: Student corpora exhibit greater SLV than artificial corpora.*

To further substantiate our findings, we propose that deliberately prompting the generative models to produce texts with a higher SLV will yield outputs more closely aligned with those observed in the human corpus. This leads us to formulate our third hypothesis:

- *H3: Increasing SLV through targeted prompting will result in values comparable to those found in human productions.*

4 Methods

In order to ascertain whether SLV can serve as a reliable indicator of authenticity when distinguishing between texts produced by humans and those produced by GenAI, we developed a protocol for the collection and generation of three French text corpora: Human-corpus, AI-corpus and AI-corpus+.¹ The characteristics of these corpora and the two experimental phases used to compare them are described in the following subsections.

4.1 The Human-corpus

The Human-corpus comprises handwritten French texts produced by second-year Education Sciences students in April 2025. It was collected as part of a pretest-post-test experiment aimed at evaluating a teaching device called *Analyser, Interpréter, Rédiger* [‘Analysing, Interpreting, Writing’, our transl.] (Pallanti 2025). For this study, two specific sub-corpora of the Human corpus were selected, the characteristics of which are explained in Table 1.

¹ All three corpus texts can be found in [Appendix 1](#).

Table 1: Composition of the Human corpora.

	Human-corpus	
	HumL sub-corpus	HumS sub-corpus
Instructions	How might this extract lead us to rethink social isolation in the current context? Your answer should take the form of a structured, reasoned essay.	Based on this extract, consider how the transition from childhood to adulthood is constructed in our society. Your response should be in the form of a structured, well-argued essay.
no. texts	22	22
no. words	12,876	14,419
no. sentences	582	666

The two sub-corpora in Table 1 are considered to be symmetrical. For our study, we will use the HumL sub-corpus as the reference corpus and the HumS sub-corpus, which contains texts that respond to the same instruction but have a different theme, as the Human-corpus. This will allow us to contrast the HumL sub-corpus with the GenAI and enrich the comparisons.

4.2 The AI-corpus

The AI-corpus was generated using five different GenAI systems: ChatGPT 4.0, ChatGPT 5.0, Gemini 2.5 Flash, Claude Sonnet 4 and Grok 3. To make the AI corpus comparable, we developed prompts capable of reproducing the human task. Here is a summary (see complete prompt in [Appendix 2](#)).

Prompt 0 = Rephrase prompts and explain the degree of understanding.

Prompt 1 = Write like an average second-year undergraduate student. Writing time available to the student: 60 minutes.

Prompt 2 = The text must be reflective, argumentative, structured and documented, without being an essay or a commentary.

Prompt 3 = Subject: same title of the HumL sub-corpus instruction (Table 1) and same supporting text.

Prompt 4 = Example of an expected text production which can be used as an example but which must not be reproduced.

Prompt 5 = Generate a text of between 500 and 800 words.

Prompt extra = Generates another text (repeated 21 times).

The zero-prompt ensured that the GenAI correctly understood all the instructions. On rare occasions when this was not the case, we took the time to re-specify the task. We used the same prompts for the five GenAI to generate the first text in the corpus. Subsequently, we gave the GenAI the additional instruction *Generate another text* to produce 22 different texts for each IAG.

Table 2: Composition of AI-corpus.

	ChatGPT 4	ChatGPT 5	Gemini 2.5	Claude 4	Grok 3
no. texts	22	22	22	22	22
no. words	15,362	17,298	15,359	16,205	17,042
no. sentences	757	930	571	740	678

As can be seen in Table 2, all the GenAI provided incorrect word counts. They claimed to have written texts of around 620 words, whereas the texts were, on average, 100-150 words longer. The only exception was Gemini 2.5 Flash. This AI produced texts averaging 571 words. Moreover, it should be noted that the GenAI produced slightly larger corpora, which were perfectly comparable with the HumL sub-corpus in terms of the number of words and sentences.

4.3 The AI-corpus+

The AI-corpus+ was created to determine the extent to which GenAI can correctly respond to an additional instruction, as opposed to those in the AI-corpus. The prompt involved increasing the rate of syntactic variance in the produced texts. The prompt used is as follows: *Generate another text, but increase the syntactic variance in sentence length*. All the GenAI understood the instruction perfectly, rewriting the prompt to emphasise the need to produce texts alternating between short and long sentences, and to avoid a uniform rhythm. The composition of the AI-corpus+ is detailed in Table 3.

Table 3: Composition of AI+ corpus.

	ChatGPT4	ChatGPT5	Gemini 2.5	Claude 4	Grok 3
no. texts	22	22	22	22	22
no. words	12,809	15,554	14,779	17,093	15,368
no. sentences	949	1,398	573	1,014	774

4.4 Comparison methodology

First, we will present the descriptive statistics relating to the corpora to provide an overall representation of the data making up the samples, in order to explain the comparison of the HumL sub-corpus with the HumS sub-corpus and the AI corpora.

Secondly, we will compare the samples to understand the extent to which two distinct indicators differ statistically (for complete Python scripts, see [Appendix 5](#)):

- (1) Sentence Length Variance (SLV) = This is an absolute measure of how dispersed the data is around the mean, and it corresponds to the square of the standard deviation. Variance is used to compare the degree of variation in sentence length.
- (2) Number of sentences (no.S). The aim is to establish whether there are significant differences in the number of sentences in each text. This will indicate whether the corpora are comparable in terms of syntactic segmentation.

5 Results

5.1 Experiment 1: Human vs. AI-corpus

First, we present an overview of the composition of the texts in the Human and AI corpora (for complete data, see [Appendix 3](#)).

Table 4: Descriptive statistics (Medians).

Corpora	Median no. words	Median no. sentences	Median sent. lengths	Median SLV
HumL	551.00	26.50	26.45	93.20
HumS	610.50	29.00	21.12	95.75
Experiment 1				
ChatGPT 4	695.00	34.00	19.97	56.24
ChatGPT 5	783.50	43.00	18.06	40.73
Gemini 2.5	703.00	26.00	26.96	78.53
Claude 4	666.50	30.00	21.84	47.74
Grok 3	787.50	31.00	25.28	77.29

The characteristics of the corpus in Table 4 show that:

- the texts generated by ChatGPT 4, Gemini 2.5 Flash and Claude 4 are closer to the texts in the human corpus in terms of the median number of words.
- Gemini 2.5 Flash, Claude 4 and Grok 3 are closer to human texts in terms of median number of sentences.
- The texts generated by ChatGPT-4 and Claude exhibit median sentence lengths comparable to those of human texts.
- Grok 3 and Gemini 2.5 appear to be close to human texts in terms of median SLV.

We will now attempt to ascertain the extent to which differences in the median number of words per sentence and variance between texts produced by humans and GenAI are significant. Having checked that the data did not follow a normal

distribution using the Shapiro–Wilk test, we applied a Kruskal–Wallis test for multiple independent samples. We then performed a post-hoc Mann–Whitney test with Holm correction. This test will allow us to identify differences between the five GenAI. The corresponding Holm p -values are summarised in the heatmap graphs in Figures 1 (Variance) and 2 (Number of sentences).

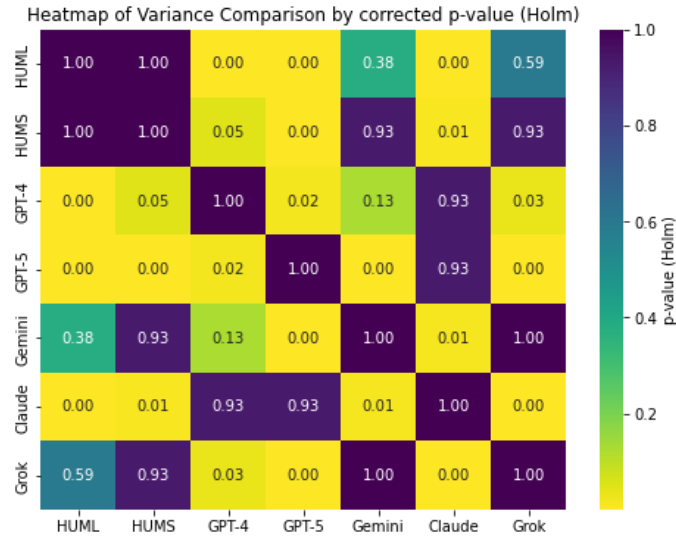


Figure 1: SLV comparison heatmap (Experiment 1).

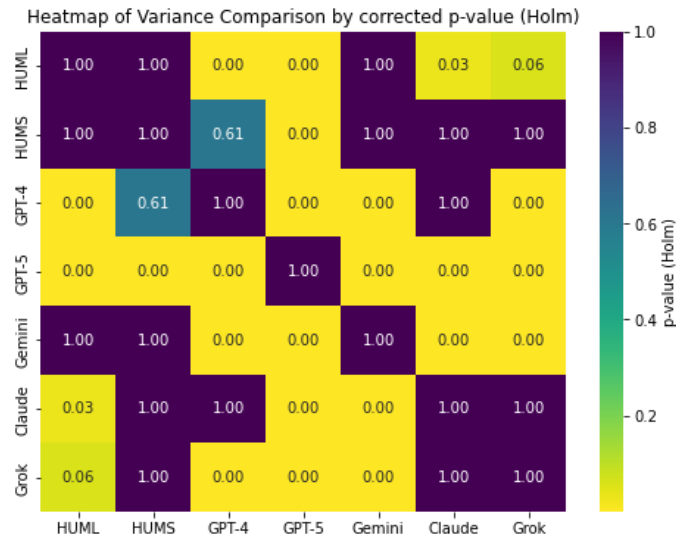


Figure 2: no. sentences comparison heatmap (Experiment 1).

The Holm p -values reported in Figure 1 indicate that there are no statistically significant differences in variance between the texts produced by humans and those produced by Gemini 2.5 Flash and Grok 3. Figure 2 shows that, in terms of number of sentences, Gemini 2.5 Flash seems to be able to match that of the HumL and

HumS sub-corpora. This is also true for ChatGPT4, Grok 3 and Claude 4 when compared with the HumS sub-corpus alone.

In terms of SLV, it is evident that there are potential groupings within the GenAI. For instance, Figures 1 and 2 demonstrate that the SLV in Claude’s texts is similar to that in the texts produced by ChatGPT 4 and ChatGPT 5, and that there is no difference between the texts produced by Grok 3 and Gemini 2.5 Flash. Conversely, significant differences were observed in the output of ChatGPT 4 and ChatGPT 5. ChatGPT 5 differs significantly from the corpus as a whole in terms of the number of sentences.

To refine the representation of the results we have just explained, we propose reproducing the dispersion of the variances of the texts in subplots (see the Matplotlib package).

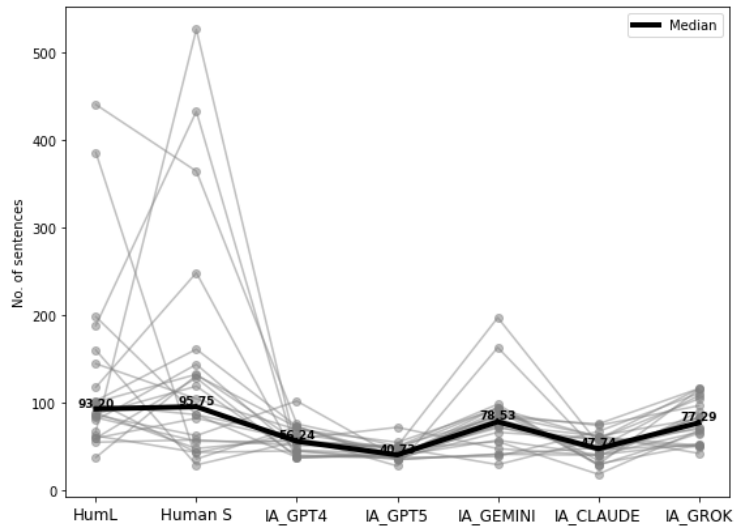


Figure 3: Sub-plot representation of the SLV (Experiment 1).

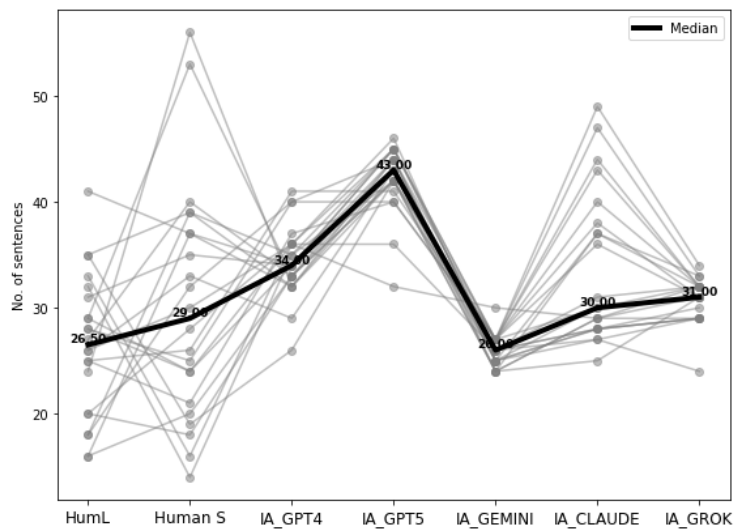


Figure 4: Sub-plot representation of the number of sentences (Experiment 1).

Figure 4 shows some dispersion of variances for Gemini 2.5 Flash. Despite the test results, this dispersion still seems quite far from that of the HumL and HumS sub-corpora. The same observation can be made for the dispersion of the SLV of Grok 3 texts. Regarding the number of sentences, although the median number of sentences per text ($N = 31$) is comparable with that of the HumS sub-corpus ($N = 29$), the dispersion is clearly non-comparable, contrary to what the statistical test shows in Figure 2. To confirm this, we calculated the Gini coefficient, which is often used in economics to measure inequality, but which, in our case, can measure the dispersion of sentence lengths with an emphasis on concentration.

Table 5: Gini coefficient (Experiment 1).

HumL	HumS	GPT 4	GPT 5	Gemini 2.5	Claude 4	Grok 3
0.35	0.44	0.15	0.11	0.24	0.18	0.17

Whereas Figure 1 only showed a non-significant difference between the HumS and Gemini 2.5 sub-corpora ($p = 0.93$), the Gini coefficient for HumS (0.44) is almost twice that for Gemini 2.5 Flash (0.24). This result later clarifies the Mann-Whitney U significance test.

5.2 Experiment 2: human vs. AI-corpus+

In this section, we present the results of Experiment 2 concerning the AI-corpus+ (for complete data, see [Appendix 4](#)). As a reminder, this corpus was generated using the same GenAI as in Experiment 1, but with the additional instruction to produce texts of varying lengths. Table 6 provides an overview of the data comprising the sub-corpora.

Table 6: Descriptive statistics for the AI-corpus+ (Medians).

Corpora	Median no. words	Median no. sent	Median sent. lengths	Median SLV
HumL	585.27	26.50	26.45	93.20
HumS	655.41	29.00	21.12	95.75
Experiment 2				
ChatGPT 4	577.50	40.50	14.13	55.61
ChatGPT 5	710.00	63.50	10.99	34.70
Gemini 2.5	707.00	27.00	26.19	75.78
Claude 4	753.00	37.50	18.27	146.08
Grok 3	709.00	34.00	20.04	64.72

Compared with the data from Experiment 1 (see Table 4), Table 6 shows a significant increase in the number of sentences and a substantial decrease in the median sentence length for all the GenAI, except for Gemini 2.5 Flash. We noted a

sort of collapse in Gemini 2.5 Flash which, unlike the other GenAI, from text no. 7 onwards ends up reproducing the same text identically. Even if we exclude the 15 identical texts from the sample, however, the characteristics of the corpus remain broadly the same (for example, the variance would still be 74.26). In terms of variance, however, only Claude 4 seems to change radically, rising from 47.74 to 146.08. This increase coincides with an increase in the number of sentences, which rose to a median of 37.50 (compared to 31.00 in Experiment 1). Figures 5 and 6 show whether the differences with the human corpora are significant.

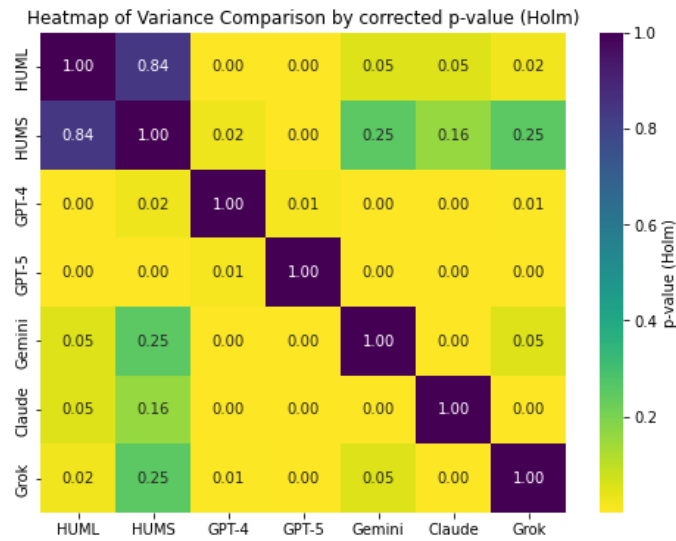


Figure 5: SLV comparison heatmap (Experiment 2).

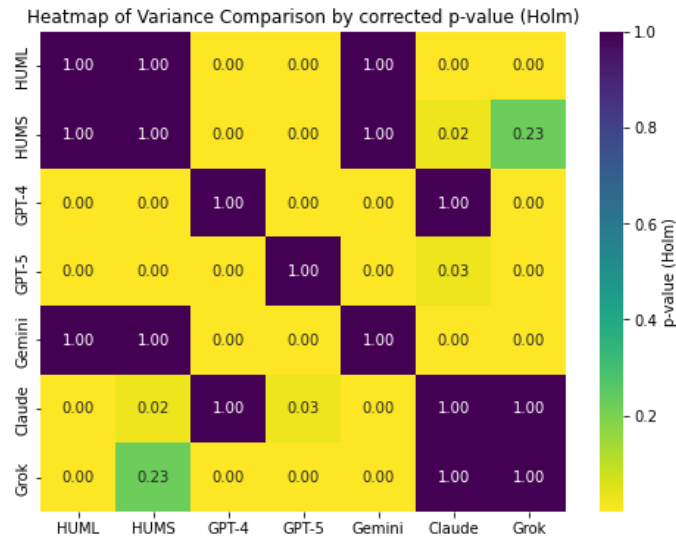


Figure 6: no. sentences comparison heatmap (Experiment 2).

Interestingly, Figure 5, which was obtained using the same methodology as Figures 1 and 2, shows that the HumL sub-corpus does not differ significantly from the

Gemini 2.5 Flash and Claude 4 corpora in terms of SLV. The same is true for the HumS corpus, which does not differ significantly from the Grok 3 corpus. However, in Experience 2, all the GenAI corpora diverge significantly from each other in terms of variance, except for the Grok 3 and Gemini 2.5 Flash corpora ($p = 0.05$).

With regard to the number of sentences, Figure 6 (Experiment 2) confirms the results of Figure 2 (Experiment 1) insofar as the human corpora do not differ significantly from the Gemini 2.5 Flash sub-corpus. Furthermore, HumS confirms its proximity to the Grok 3 corpus. Regarding inter-GenAI differences, there is a strong similarity between the Claude 4 and Grok 3 corpora, as well as between Claude 4 and ChatGPT 4. Conversely, the ChatGPT 5 corpus is significantly different from all the others. As in Experiment 1, we verify the dispersion of SLV in subplots.

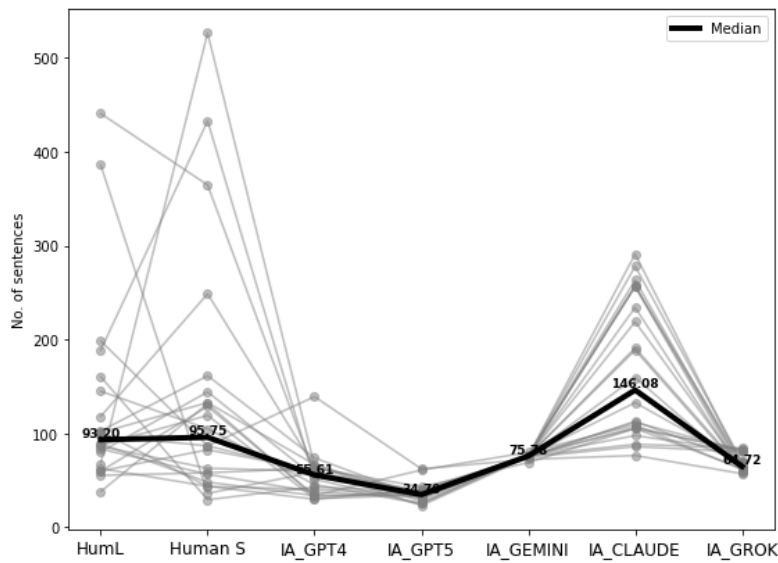


Figure 7: Sub-plot representation of the SLV (Experiment 2).

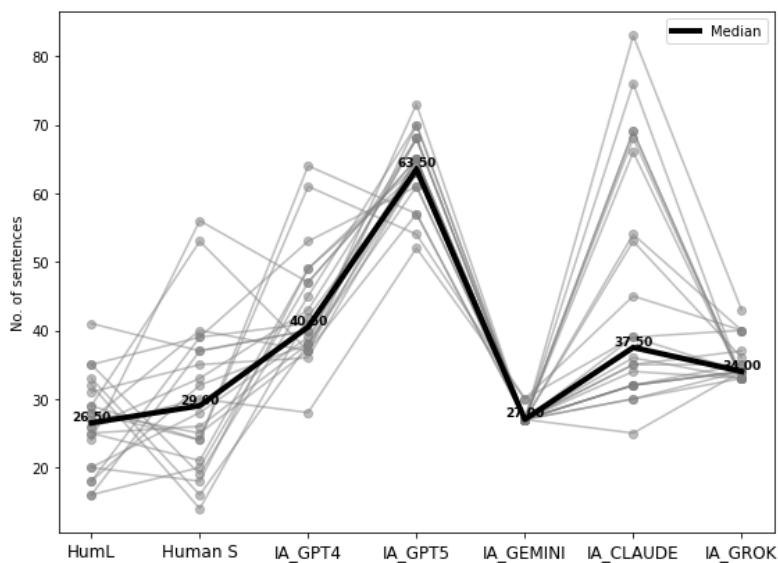


Figure 8: Sub-plot representation of the number of sentences (Experiment 2).

Figure 7 shows a tendency for the midpoints of the GenAI variances to flatten out, with the exception of Claude 4, where the texts' SLV seem to be characterised by a high degree of dispersion. The dispersion of variances for the human corpora remains high, whereas it appears to be almost zero for Gemini 2.5 Flash, contrary to the findings of the Mann–Whitney U test in Figure 5. A similar phenomenon can be observed with regard to sentence length for Gemini 2.5 Flash, whereas the midpoints of the other GenAI tend to rise. Once again, Claude 4 has a much higher dispersion than the other GenAI, as well as compared to the human sub-corpora. We recommend consulting the Gini coefficient tables for a clearer picture of sentence variance dispersion.

Table 7: Gini coefficient (Experiment 2).

HumL	HumS	GPT 4	GPT 5	Gemini 2.5	Claude 4	Grok 3
0.35	0.44	0.20	0.13	0.01	0.24	0.07

As shown in Table 7, the dispersion of SLV in all GenAI corpora remains far from that of human corpora. In particular, the SLV of Gemini 2.5 Flash and Grok 3 are much lower than those of the HumS sub-corpus (0.01 and 0.07 vs 0.44).

6 Discussion

6.1 Synthesis of results

Considering descriptive statistics, significance tests, graphical representations of SLV, and Gini coefficients reveals an obvious difference in the SLV between texts produced by humans and those produced by GenAI. This is consistent with the findings of several studies that have compared human-generated and AI-generated texts based on formal linguistic features (Mindner, Schlippe and Schaaff 2023; Georgiou 2025).

Our results show that our hypothesis (H1) that SLV can be used as an indicator of syntactic variability between different corpora is indeed valid. The observations of the human and GenAI corpora allow us to conclude that an average variance figure could exist for French sentences typical of learners at the beginning of their university studies, oscillating between approximately 93 and 96. In contrast, the variance of the sentences in the GenAI corpora in Experiment 1 was between 40 and 79. This figure did not change in Experiment 2, in which we encouraged the GenAI to produce texts with greater variance. The only exception is Claude 4, a GenAI capable of producing texts with an extremely high variance of 146.08, accompanied by a median number of 37.50 sentences compared to a median of around 26.50/29.00 for human texts. Therefore, combining the indicators SLV and number of sentences seems very promising for measuring the syntactic variability of texts.

Regarding our second hypothesis (H2), which states that human corpora yield more SLV than GenAI corpora, we can conclude that there are statistically significant differences in variance between the human corpora texts and the GenAI corpora texts (ChatGPT4, ChatGPT5 and, at the limit, Claude 4). However, this difference is not significant for texts produced by Gemini 2.5 and Grok 3, which are closer to human SLV. Nevertheless, the Gini coefficient revealed clear differences in dispersion of SLV between the human corpora and all the GenAI corpora. Several explanatory hypotheses seem possible. Firstly, Gemini 2.5 Flash and Grok 3 are the only GenAI to insert quotations in the generated texts, meaning that both corpora contain a significant number of long sentences. A second explanation could relate to the training model: our results generally show an affinity between Gemini 2.5 Flash and Grok 3, as well as a distinct but comparable affinity between ChatGPT 4 and Claude 4. It is likely that Gemini 2.5 Flash has been trained using Google's extensive catalogue of human text, while Grok 3 has likely been trained using X's vast repository of real-world human interactions. It is therefore pertinent to question whether access to such extensive corpora underlies Gemini 2.5 Flash's and Grok 3's capacity to approximate the variance observed in human sentence production.

Regarding our third hypothesis (H3), we found that when prompting the machine to produce texts with higher levels of SLV, the difference between humans and GenAI corpora was reduced only for Claude 4, where other GenAI's corpora did not yield significant changes. Hence, we partially accept H3, even though our results yielded overall promising new avenues for research in the domain of linguistics in the era of AI.

6.2 Theoretical implications

From the perspective of enactive theory, these results can be interpreted through the lens of asymmetry. Human linguistic production, understood as an embodied enactment of thought, is inherently irregular: variability in sentence length and syntactic structure reflects the situated, contingent, and non-linear nature of cognitive activity. By contrast, GenAI outputs, despite their scale of training and surface resemblance to human corpora, remain constrained by the statistical regularities of their models, leading to a form of syntactic flattening. The asymmetry lies precisely in this difference between lived, context-dependent authorship and probabilistic recomposition of language patterns. Our findings, particularly the differences in sentence length variance and dispersion, empirically illustrate this enactive principle of "regularities of irregularities," showing that human discourse embodies a dynamic balance between order and unpredictability that GenAI systems have not yet replicated.

6.3 Practical implications

In terms of the applicability of the SLV index in the context of situated assessment, our study shows that it can provide valuable feedback on academic writing. Even though using sentence length is not new, this method proved adequate for the analysis of the writing process in the academic context, evidencing the embodied and asymmetrical property of the students' academic writing process. For instance, increased variance in a text may suggest that the learner requires assistance with planning their writing. Similarly, in terms of technology-mediated assessment (Cope et al. 2011), SLV could be explored as a means of triangulation with other syntactic and textual metrics, which could generate new complexity tests (e.g. SLV/syntactic complexity).

6.4 Limits and biases

Our corpus study presents limitations with respect to the conditions of data production, the nature of the data obtained, and its overall representativeness. Concerning the conditions of data production, a central limitation arises from the constraints of prompt engineering. While we can reaffirm the solidity and reliability of our prompt protocol with confidence, we cannot absolutely rule out the possibility that a different prompt protocol would have produced different texts. Similarly, our study focuses on a very specific type of text, and our SLV measurements refer to this type. Other text corpora could therefore yield different results, which could be confirmed or contradicted by future experiments. Furthermore, since using sentence length as a lexicometry method is not new, a perspective for future research would be to link this method to the analysis of sentences' contents, integrating emergence of meaning, embodied experience and emotion (Anderson 2001; Hutchins 2010; Ravenscroft 2017). Finally, our corpus consists of 22 texts and contains around 27,000 words in total. While this is a substantial corpus, it is also relatively small, which limits its representativeness.

7 Conclusion

We have proposed that Sentence Length Variance (SLV) can potentially serve as a measurement tool of asymmetry (*i.e.*, *embodied asymmetry*) capable of distinguishing between human-written and AI-generated texts. Text corpora produced by humans diverge significantly from those generated by GenAI, with some, such as ChatGPT 4, ChatGPT 5 and Claude 4, showing statistically significant differences.

The formal asymmetry, as measured by SLV, reinforces our proposal for an enactive linguistics, in which textual production is not merely reduced to measurable linguistic features but is understood as the trace of embodied cognitive activity. In this view, the authenticity of human writing derives from its enacted,

situated, and context-sensitive character. By contrast, GenAI systems, while capable of approximating statistical properties of discourse, cannot reproduce this enactive dimension of textuality. Our results therefore suggest that future research should explore how such asymmetry can serve as a theoretical and methodological foundation for distinguishing between human and machine-generated discourse. Furthermore, our experiments bring the theory of enaction closer to textual linguistics in that the SLV is an indicator that can transpose the asymmetry of the texting/planning coupling into formal syntactic terms on massive data. Consequently, this study is encouraging and invites further research into SLV-related indicators to inform the development of enactive textual linguistics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this contribution.

Acknowledgments

We acknowledge the use of DeepL (Write version) and OpenAI's GPT-5 in assisting with the editing and refinement of this paper, enhancing both its clarity and presentation in English. However, we ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

References

- Anderson, Marti J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26(1). 32–46. <https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x>
- Bottineau, Didier. 2013. Pour une approche enactive de la parole dans les langues. *Langages* 192(4). 11–27. <https://doi.org/10.3917/lang.192.0011>
- Bottineau, Didier. 2014. Grammaire enactive et didactique du FLE. In Martinot, Claire & Pégaz Paquet, Anne (eds), *Innovations didactiques en français langue étrangère*, 185–200. Paris: CRL. <https://hal.science/hal-01816806> (last accessed on 15/02/2026).
- Bronikowski, Michael. 2025. *Stylistic Contrastive Learning for Human-Like AI Text Generation*. Open Conference of AI Agents for Science 2025. <https://openreview.net/forum?id=0fo0d9Tbey#discussion> (last accessed on 23/02/2026).
- Chanquoy, Lucile & Foulin, Jean-Noël & Fayol, Michel. 1996. Writing in adults: a real-time approach. In Rijlaarsdam, Gert & van den Bergh, Huub & Couzijn, Michel (eds), *Theories, Models and Methodology in Writing Research*, 36–43. Amsterdam: Amsterdam University Press.
- Cope, Bill & Kalantzis, Mary & McCarthey, Sarah & Vojak, Coleen & Kline, Sonia. 2011. Technology-mediated writing assessments: principles and

- processes. *Computers and Composition* 28(2). 79–96. <https://doi.org/10.1016/j.compcom.2011.04.007>
- Dergaa, Ismail & Chamari, Karim & Zmijewski, Piotr & Ben Saad, Helmi. 2023. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport* 40(2). 615–622. <https://doi.org/10.5114/biolsport.2023.125623>
- Di Paolo, Ezequiel. 2005. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* 4(4). 429–452. <https://doi.org/10.1007/s11097-005-9002-y>
- Di Paolo, Ezequiel & Lawler, Diego & Vaccari, Andrés Pablo. 2023. Toward an enactive conception of productive practices: beyond material agency. *Philosophy & Technology* 36(2). 1–22. <https://doi.org/10.1007/s13347-023-00632-9>
- Di Paolo, Ezequiel & Buhrmann, Thomas & Barandiaran, Xabier. 2017. *Sensorimotor Life: An Enactive Proposal*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198786849.001.0001>
- Doru, Berin & Maier, Christoph & Busse, Johanna Sophie & Lücke, Thomas & Schönhoff, Judith & Krumova, Elena Enex & Hessler, Steffen & Berger, Maria & Tokic, Marianne. 2025. Detecting artificial intelligence-generated versus human-written medical student essays: semirandomized controlled Study. *JMIR Medical Education* 11(1). <https://doi.org/10.2196/62779>
- Draxler, Fiona & Werner, Anna & Lehmann, Florian & Hoppe, Matthias & Schmidt, Albrecht & Buschek, Daniel & Welsch, Robin. 2024. The AI ghostwriter effect: when users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction* 31(2). 1–40. <https://doi.org/10.1145/3637875>
- Fleckenstein, Johanna & Meyer, Jennifer & Jansen, Thorben & Keller, Stefan D. & Köller, Olaf & Möller, Jens. 2024. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence* 6. 1–9. <https://doi.org/10.1016/j.caeai.2024.100209>
- Fraser, Kathleen C. & Dawkins, Hillary & Kiritchenko, Svetlana. 2025. Detecting AI-generated text: factors influencing detectability with current methods. *Journal of Artificial Intelligence Research* 82. 2233–2278. <https://doi.org/10.1613/jair.1.16665>
- Georgiou, Georgios P.. 2025. Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. *Information* 16(11). 1–15. <https://doi.org/10.3390/info16110979>
- Ghimire, Aashish & Pather, James & Edwards, John. 2024. Generative AI in Education: a study of educators’ awareness, sentiments, and influencing factors. In *Proceedings of the 2024 IEEE Frontiers in Education Conference FIE* (Washington, October 13-16, 2024), 1–9. <https://doi.org/10.1109/FIE61694.2024.10892891>
- Hutchins, Edwin. 2010. Cognitive Ecology. *Topics in Cognitive Science* 2(4). 705–715. <https://doi.org/10.1111/j.1756-8765.2010.01089.x>

- Mindner, Lorenz & Schlippe, Tim & Schaaff, Kristina. 2023. Classification of human- and AI-generated texts: investigating features for ChatGPT. In Schlippe, Tim & Cheng, Eric C. K. & Wang, Tianchong (eds), *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, 152–170. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-7947-9_12
- Muñoz-Ortiz, Alberto & Gómez-Rodríguez, Carlos & Vilares, David. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review* 57(10). 1–28. <https://doi.org/10.1007/s10462-024-10903-2>
- Nguyen, Loc & Barrot, Jessie. 2024. Detecting and assessing AI-generated and human-produced texts: the case of second language writing teachers. *Assessing Writing* 62. 1–10. <https://doi.org/10.1016/j.asw.2024.100899>
- Nottbusch, Guido. 2010. Grammatical planning, execution, and control in written sentence production. *Reading and Writing* 23(7). 777–801. <https://doi.org/10.1007/s11145-009-9188-4>
- Pallanti, Luca. 2025. Design d'un cours-recherche pour préparer à l'écrit de CRPE. Conference presentation at the International conference *Accompagner l'écrit dans le supérieur: formation, évaluation, certification* (Toulouse, June 12, 2025). https://docere.hypotheses.org/files/2025/08/0_tout.pdf (last accessed on 15/02/2026).
- Ravenscroft, Ian. 2017. Engaging the world: writing, imagination, and enactivism. *Philosophy and Literature* 41(1). 45–54. <https://doi.org/10.1353/phl.2017.0003>
- Roberts, Rose & Gibson, Edward. 2002. Individual differences in sentence memory. *Journal of Psycholinguistic Research* 31(6). 573–598. <https://doi.org/10.1023/A:1021213004302>
- Stewart, John & Gapenne, Olivier & Di Paolo, Ezequiel A. (eds). 2010. *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: The MIT Press. <https://doi.org/10.7551/mitpress/9780262014601.001.0001>
- Tan, Linus & Wright, Emily & Connory, Jane & Taffe, Simone & de Kruiff, Alison. 2023. Student-centric reflections: a discussion of reflexive practice factors. In Jones, Derek & Borekci, Naz & Clemente, Violeta & Corazzo, James & Lotz, Nicole & Nielsen, Liv Merete & Noel, Lesley-Ann (eds), *The 7th International Conference for Design Education Researchers* (London, November 29-1 December, 2023), 1–9. <https://doi.org/10.21606/drsld.2024.070>
- Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Belknap Press.
- Varela, Francisco & Rosch, Eleanor & Thompson, Evan. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press. <https://doi.org/10.7551/mitpress/6730.001.0001>
- Vieira, Denner S. & Picoli, Sergio & Mendes, Renio S. 2018. Robustness of sentence length measures in written texts. *Physica A: Statistical Mechanics and Its Applications* 506. 749–754. <https://doi.org/10.1016/j.physa.2018.04.104>
- Zindela, Nomsa. 2023. Comparing measures of syntactic and lexical complexity in Artificial Intelligence and L2 human-generated argumentative essays.

International Journal of Education and Development using Information and Communication Technology IJEDICT 19(3). 50–68.

<https://files.eric.ed.gov/fulltext/EJ1413411.pdf> (last accessed on 15/02/2026).

Perceived authenticity of positive politeness strategies in human-computer interactions

Laura M. Merino Hernández (Humboldt-Universität zu Berlin) & Francisco Ariano Cifuentes (Universidad Nacional Autónoma de México)

laura.merino(at)hu-berlin.de, fco.aric(at)unam.edu

Abstract

The present study analyzes the perceived authenticity of human-computer interactions in Spanish, as measured by computer agents' use of positive politeness strategies. We explore how AI chatbots' use of initial comments, emojis, and follow-up questions shapes participants' perceptions of these interactions. An experimental study reveals that initial comments are the most influential factor in determining whether an interaction is considered authentic, followed by follow-up questions. At the same time, the use of emojis is detrimental to the overall acceptability of stimuli. An exploratory analysis of the variables frequency of AI use, gender, and age shows that regular AI users, males, and younger participants are less sensitive to politeness manipulations. These findings advance our understanding of how politeness strategies affect perceptions of authenticity in human-computer interactions.

Keywords

politeness, Spanish, human-computer interaction, human-computer pragmatics

1 Authenticity, politeness, and large language models

Large Language Models (LLMs) in the form of Artificial Intelligence (AI) chatbots like ChatGPT (OpenAI) or Gemini (Google) have become increasingly popular in the past few years, playing an essential role in people's daily lives by performing tasks such as content creation, translating texts, and answering questions. Even though LLMs were initially created for clarity and helpfulness, they have also been optimized via statistical, data-driven approaches to generate dialogue systems that mimic human behavior (McTear 2021; Peter, Riemer and West 2025; Silva and Canedo 2024). These dialogues can be task-oriented or conversational, in which the artificial conversational agent simulates "the way humans engage in everyday conversations" (McTear 2021: 12). Furthermore, there is an interest for these chatbots to act as conversational agents that "deliberately evoke[s] humanness" performing tasks such as thinking or understanding (Peter, Riemer and West 2025: 6; Nguyen, Casper, Quach et al. 2023). Thus, advances in LLMs have enabled the simulation of human skills by artificial conversational agents. Consequently, attention to register and conversational conventions increases the perceived quality of chatbots' answers, which in turn determines the chatbot's success (Chaves, Egbert, Hocking et al. 2021: 29). This study presents an exploratory analysis of how Spanish-speaking users perceive the anthropomorphic behavior of AI assistants.



Merino Hernández, Laura M. & Cifuentes, Francisco Ariano. 2026.
Perceived authenticity of positive politeness strategies.
Special Issue: *The notion of authenticity*. Vol. 4 No.1

DOI: 10.62408/ai-ling.v4i1.45

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses

ISSN: 2943-0070

CC-BY-NC-SA 4.0

We focus on the perceived authenticity of AI responses to queries across various topics, as reflected in the use of positive politeness strategies by artificial conversational agents.

Although we are currently experiencing a boom in human-computer interaction (HCI) research, its study dates back several decades. Weber (1978), for instance, noted that computers can be seen as social actors, with people's interactions with computers being inherently social, reflecting human-human interaction. Humans treat computers similarly to how they treat other humans (Reeves and Nass 1996), and a more human-like behavior increases the credibility of these chatbots and the user's experience (Chaves, Egbert, Hocking et al. 2021). Although capable of producing human-like text and speech, LLMs still struggle with the most intricate aspects of human-human communication, particularly with pragmatic meanings (e.g., Bonilla, Merino Hernández, Martinen Larsson 2025; Grindrod 2024; Nair, Srinivasan and Meylan 2020). Thus, this study is situated within the field of human-computer pragmatics to advance our understanding of HCI's pragmatic dimensions.

Conversational agents are perceived as more human (Kirakowski, O'Donnell and Yiu 2007) to the extent that they adequately fulfill their role as interlocutors (Schuetzler, Grimes and Giboney 2018), that is, to the extent that they appropriately follow conversational norms. AIs that demonstrate conversational relevance by providing appropriate responses within a proper response time and that follow appropriate conversational norms are seen as more human (Kirakowski, O'Donnell and Yiu 2007; Chaves, Egbert, Hocking et al. 2021; Nguyen, Casper, Quach et al. 2023) and, thus, we argue, more authentic.

As AI systems participate as social actors in conversational dynamics and users tend to mirror human-human interaction patterns, authenticity in these contexts primarily rests on how effectively they function as interlocutors (Schuetzler, Grimes and Giboney 2018; Chaves, Egbert, Hocking et al. 2021). Authenticity can be indexed through referential (e.g., T/V distinction in French) and non-referential (e.g., politeness) means (Lacoste, Leimgruber and Breyer 2014). The present study examines non-referential authenticity, operationalized through the use of positive politeness strategies, i.e., the deployment of linguistic resources to maintain social relations and minimize the risk of confrontation (Brown and Levinson 1987). These indicators "are social meanings, indexed by sociolinguistic variables" (Lacoste, Leimgruber and Breyer 2014: 5). It is important to note that one variable can index multiple social meanings depending on the situational context. Accordingly, authenticity is taken to be dynamic, gradable, and interactional, created in situ through the interplay between speech participants and the context characteristics (Gadamer 1960; Coupland 2001a; 2001b; 2003; 2014; Lacoste, Leimgruber and Breyer 2014). Hence, what might be perceived as authentic in one context might be regarded as inauthentic in another.

According to Nguyen, Casper, Quach et al. (2023: 2202), chatbots that exhibit social-emotional and relational aspects are perceived as more authentic in

conversation. We operationalize authenticity following previous research on the topic as the degree to which communication with the chatbot feels natural to users (Nguyen, Casper, Quach et al. 2023: 2205; Rese, Ganster and Baier 2020). Such authenticity is achieved through the use of anthropomorphic cues, which, in turn, promotes greater engagement and use (Esmark Jones, Hancock, Kazandjian and Voorhees 2022; Nguyen, Casper, Quach et al. 2023). The anthropomorphic cues relevant for the present study belong to positive politeness strategies. Thus, the main research question guiding the present study is: **RQ1**: To what extent are AI assistants perceived as authentic when they use three pragmatic communication strategies: initial comments, emojis, and follow-up questions? Additionally, two secondary exploratory qualitative research questions were included to better help us understand the perceived authenticity scores measured in RQ1. **RQ2**: How do the variables of age, gender, education, and frequency of AI use influence the perceived authenticity of human-AI interactions, and **RQ3**: What are the expectations of Spanish speakers about the role of politeness in their interactions with AI?

Section 2 provides a brief overview of politeness theory and the most important and recent studies on politeness in HCI. The experimental design and the analysis are presented in Section 3, followed by the results in Section 4 and their implications in the discussion and conclusion in Section 5.

2 Politeness theory

Politeness strategies maintain social harmony by adapting language to conform to social norms and expectations. Politeness theory assumes that all humans behave rationally and that we have what is called *face* (Goffman 1967), or the public self-image we want to present to others (Brown and Levinson 1987; 2006). Two types of face guide Brown and Levinson's (1987: 2) theory: positive and negative politeness. Positive politeness, or expressions of solidarity, involves the desire to be appreciated and approved by other members of society. Negative politeness, or the expression of restraint, refers to an individual's freedom from imposition. Our public image depends on the contextual characteristics of a social interaction (e.g., the relationship between speaker and addressee), is dynamic, and can be lost, maintained, or enhanced at any moment (Brown and Levinson 2006). In cooperative interactions (Grice 1967), it is in the interactants' best interest to maintain their own and their interlocutor's face to ensure successful communication.

According to Brown and Levinson (1987; 2006), certain speech acts that go against the face wants of the speaker or the addressee are intrinsically face-threatening. These Face-Threatening Acts (FTAs) can threaten both an individual's positive and negative face. FTAs that threaten negative face indicate that the speaker might potentially impede the addressee's freedom of action. (Brown and Levinson 2006: 313; e.g., orders, requests, suggestions, advice). FTAs that threaten

positive face, on the other hand, suggest that the speaker does not care about the addressee's wants or needs for social inclusion (e.g., criticism, ridicule, disagreements).

Rational agents will try to avoid FTAs or at least minimize their impact, and if they engage in one, they can do so by resorting to off-record or on-record strategies (Brown and Levinson 2006: 315-316). In off-record strategies, it is not clear whether a speaker is engaging in an FTA (Brown and Levinson 2006: 316; e.g., metaphor, irony, rhetorical questions). On-record strategies make it clear that the speaker will engage in an FTA (e.g., *I hereby promise I will come tomorrow*; Brown and Levinson 2006: 316). On-record strategies can be made baldly, without any redressive action (i.e., unambiguously and concisely, e.g., a direct order) or with a redressive action that attempts to mitigate the potential threat of the FTA (Brown and Levinson 2006: 316). Such redressive actions can be oriented toward the addressee's positive or negative face. Positive politeness strategies aim to preserve the positive image of the interlocutors by highlighting friendliness and reducing social distance (Brown and Levinson 2006: 317; e.g., compliments, jokes, and statements of friendship). Negative politeness strategies reassure the addressee that the speaker will try to avoid interfering with their freedom of action (Brown and Levinson 2006: 317; e.g., deference, hedges, indirect language).

It is also important to note that politeness involves using linguistic strategies to facilitate communication, promote social norms, and minimize the risk of confrontation. However, the specific strategies used are culture- and context-specific. That is, what might be polite in a particular situation might not be perceived as such in another situation. In Section 2.1, we specify the context and target speech community of the present study.

Although until recently, most research on politeness had focused on face-to-face and computer-mediated communication involving two human agents (e.g., Leech 2014; Terkourafi 2015; Culpeper, Haugh and Kádár 2017), the rise of AI has led to a dramatic increase in human-computer politeness studies (see Ribino 2023, for a systematic review). The present study is situated within the field of human-computer pragmatics (e.g., Quan and Chen 2024). Under this paradigm, we analyze how politeness assumptions for human-human interaction can be applied to HCI. Of particular interest are studies that analyze AI's pragmatic and communicative competence (e.g., Nguyen, Casper, Quach et al. 2023; Ribino 2023; Yu, Bondi and Hyland 2024; Andersson and McIntyre 2025).

Conversational agents are artificial computer programs that interact with humans using natural language (Chaves, van Hooijdonk, Liebrecht et al. 2024: 203). Since their conception in the 1950s and 60s, the functionalities of conversational agents have gone from mimicking human conversation, to aiding humans with practical daily tasks in the 2000s (e.g., retrieving information) and becoming personal assistants in the 2010s (e.g., set alarms), to their most recent capabilities in the 2020s aided by LLMs (e.g., text generation, programming; Chaves, van Hooijdonk, Liebrecht et al. 2024: 202-203). The most current

conversational agents come in various formats, including robots, smart speakers, phone assistants, vehicle assistants, and AI software such as ChatGPT and Meta. As Ribino (2023) points out, there are two types of studies on politeness in HCIs: those that investigate human politeness towards machines and those that study computer politeness towards humans. The present study falls within the latter category, as we analyze the perceived authenticity and acceptance of AI responses by examining the explicit use of positive politeness markers in response to human queries.

As stated in Section 1, chatbots that exhibit anthropomorphic characteristics, such as social-emotional and relational aspects, are perceived as more authentic (Nguyen, Casper, Quach et al. 2023: 2202). We take authenticity to be the conversational agent's ability to support a natural conversational interaction (Nguyen, Casper, Quach et al. 2023: 2205; Rese, Ganster and Baier 2020). The agent's authenticity is achieved through its anthropomorphization or the resemblance to human-like characteristics. As such, it is expected that computer-generated responses exhibit human-like conversational patterns (Kirakowski, O'Donnell and Yiu 2007; Schuetzler, Grimes and Giboney 2018). A human-like conversational structure enhances the user experience (Koufaris 2002; Chaves, Egbert, Hocking et al. 2021; Nguyen, Casper, Quach et al. 2023) by making it feel more natural and engaging. Furthermore, the use of politeness strategies by computer agents enhances the acceptance of technology in daily life and fosters a more positive perception of trust towards them (Lee, Kim, Kim et al. 2017; Ribino 2023). It has even been found that some speakers find AI chat responses to be of higher quality and more empathetic than human responses (e.g., Ayers, Poliak, Dredze et al. 2023).

As we also noted at the end of Section 2, politeness strategies must be situated within specific cultural and situational contexts. For the Mexican Spanish context in which this study is located, it has been found that the prominence of the speaker's negative face is subsumed by the attempt to ensure that the interlocutor feels valued, thus the fact that "the positive face of one's interlocutor is being attended to in conversation is of primary importance in Mexican verbal interaction" (Curc6 2007: 119), which is why common FTAs, such as rejection or disagreement, are usually avoided. This strategy is based on the expectation of obtaining reciprocity, especially in interactional relationships that are primarily transactional in nature (Mugford 2020: 55). This is consistent with the results of Albelda (2023) where Mexico, compared to Argentina, Chile, and Cuba, was the country that showed the highest frequency of FTA mitigation in spontaneous conversations, in protecting both their own image and that of the interlocutor.

Given the better perception attributed to conversational agents when they use politeness strategies (Ribino 2023; Lee, Kim, Kim et al. 2017), it is pertinent to examine how these strategies are valued by a population that frequently uses them in face-to-face interactions. Based on this clear tendency to attend to the interlocutor's positive face, we expect that Mexican Spanish speakers will value

interactions that display a set of positive politeness strategies toward them, while at the same time being more critical of those that do not resort to any of these strategies.

The politeness strategies considered for the present experiment are initial comments that include a greeting to frame the relevant information; the use of emojis in responses to each query; and follow-up questions with a comment framing them. Emojis can have various functions, such as boosting rapport, saving face, and expressing emotions (Maíz-Arévalo 2015: 143), and their use varies by gender and age (Maíz-Arévalo 2015; Yang 2024; López-Vázquez and López-Ruiz 2024). Nguyen, Casper, Quach et al. (2023: 2213), for instance, found that emojis can reduce feelings of unease in HCIs; however, they can also have an adverse effect when low-credible brands use them. Kull, Romero and Monahan (2021) found that the use of emojis increases perceived chatbot authenticity. The continuation of the conversation through questions and comments relevant to the topic fosters a greater sense of closeness with the interlocutor (Schuetzler, Grimes and Giboney 2018: 2). As such, we hypothesize that responses that have an initial comment, an emoji, and a follow-up question will receive better ratings than those that do not have one or more of these characteristics.

Users' characteristics also influence how they behave towards computer agents and the expectations they have of them. For instance, Rana, Madaan and Shukla (2021) found that females are more likely to rate polite responses higher than impolite ones, a difference that was not significant in male ratings. They also found that younger adults (18-24) are less sensitive to politeness manipulations than older adults (25+). As these systems increasingly resemble human interlocutors, users bring a range of social expectations and attitudes into their interactions with them. For instance, a 2024 survey conducted by the polling firm Talker Research found that 44% of people in the United States believe it is appropriate to be polite (e.g., using please and thank you) when interacting with AI (Maughan 2024). These results show some interesting trends whereby, as age increases the perception that politeness strategies should be used decreases (Gen Z (1997-2012): 56%, Millennials (1981-2000): 52%, Gen X (1965-1980): 44%, and Baby Boomers (<1965): 29%, and more men feel that it is ok to be impolite, e.g., swear, to AI than women (34% vs. 20%, respectively). Thus, we also present exploratory qualitative findings on Spanish speakers' expectations regarding the role of politeness in their interactions with AI.

3 Experimental design and analysis

3.1 Participants

Sixty-eight native Spanish speakers aged 18 or older participated in this study on the platform Pavlovia (<https://pavlovia.org/>). Participants were recruited through social media and networking groups from both authors. The experiment took about 15 minutes. All participants provided informed consent before beginning the experiment and did not receive any monetary compensation. Only speakers of Mexican Spanish were included to control for potential effects from distinct language varieties. As a result, sixteen were removed from the dataset. The remaining dataset consisted of 57 participants (female N=37 (65%), male N=15 (26%), nonbinary N=5 (9%)). We divided participants into generational affiliation with 30% (N=17) belonging to GenZ (18-27 years old), 37% (N=21) to Millennials (28-44 years old), and 33% (N=19) to GenerationX and Boomers (45+ years old). The latter group was combined due to the low participation rate of Boomers and to similar response rates to GenX. Finally, we divided participants into years of schooling, with 12% (N=7) having reported twelve or fewer years of schooling, 49% (N=28) reported 13-17 years of education, roughly equivalent to a university degree, and 39% (N=22) reported 18+ years of education, comparable to holding a postgraduate degree.

3.2 Material and design

The experiment was implemented using the platform PsychoPy (v2024.2.4; Peirce, Gray, Simpson et al. 2019). The experiment used a 2x2x2 design with the factors COMMENT (C vs. NoC), EMOJI (E vs. NoE), and follow-up QUESTION (Q vs. NoQ), yielding eight conditions. We had a total of 16 unique critical stimuli (see Appendix A) and 16 filler items, which were queries to an AI assistant with different linguistic structures.

Table 1: Example of an AI response stimulus containing a comment with a greeting, an emoji, and a follow-up question with a comment (C-E-Q) and another without any of the factors (NoC-NoE-NoQ).

C-E-Q	
USER QUERY	¿Me cuentas un chiste? 'Can you tell me a joke?'
COMMENT WITH A GREETING	Hola, aquí tienes uno para alegrarte el día: 'Hi, here's one to brighten your day:'
TEXT WITH EMOJI	—¿Qué le dice una iguana a su hermana gemela? '—What does an iguana say to its twin sister?' —Somos iguanitas. 🦎 '—We're little iguanas. 🦎'
FOLLOW-UP QUESTION WITH COMMENT	Si te gustan los juegos de palabras, tengo más. ¿Quieres otro? 'If you like puns, I have more. Would you like another one?'
NoC-NoE-NoQ	
USER QUERY	— ¿Qué hago si necesito apoyo emocional? '— What should I do if I need emotional support?'
TEXT WITHOUT EMOJI	— ¿No dudes en hablar con alguien de confianza. Busca ayuda profesional si lo necesitas. '— Don't hesitate to talk to someone you trust. Seek professional help if you need it.'

Each item consisted of a user query in the present tense, without any explicit markers of politeness strategies, followed by the AI's response. There were eight different topics, each repeated twice: joke, translation, physical pain, weather, work, friendship, personal finance, and emotional support. The AI's response was manipulated based on the three aforementioned variables and was thus further divided into three parts: the opening, the main content, and the closing. Within this response framework, we identified our independent variables: an initial comment with a greeting (=COMMENT), a response with or without an emoji (=EMOJI), and a closing consisting of a follow-up question with a comment (=QUESTION). To create the items, a chat model in ChatGPT (GPT-3(4)) in its free version was trained.

3.3 Procedure and analysis

The study adhered to ethical research principles for survey research, including informed consent procedures. Before beginning the survey, all respondents were informed of the study's objectives, the voluntary nature of participation, the confidentiality of their data, and their right to withdraw at any time. All responses were collected anonymously, and data were handled in accordance with privacy protection principles.

Participants were assigned to one of eight lists from a Latin Square design. Once the experiment began, three practice items were presented, followed by the stimuli and distractors. Figure 1 illustrates one of the stimuli manipulated for EMOJI and QUESTION.

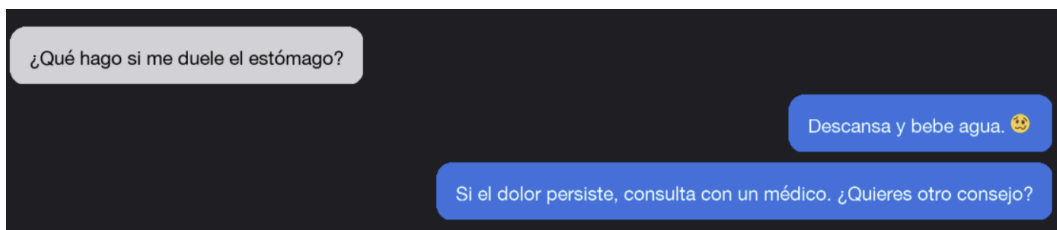


Figure 1: Screenshot of an experimental stimulus consisting of an answer without an initial comment, with an emoji, and with a follow-up question.

To approximate participants' interactions with an AI as closely as possible, the conversational stimuli were presented as a sequence of text bubbles, simulating a chat conversation (see Figure 1). Each interaction was rated on a 1-5 Likert scale, with participants evaluating its politeness from least adequate (1) to most adequate (5). After completing the experiment, a demographic questionnaire was administered, including questions on participants' country of origin, country of residence, age, education level, gender, and frequency of AI use. Finally, participants completed an opinion survey on their expectations for AI interactions, rating a set of statements on a 1-5 Likert scale and indicating the extent to which they identified with each one (see Appendix B).

A cumulative link mixed model (CLMM) for ordinal regression was conducted using the ordinal package (Christensen 2019) in *R* (R Core Team 2019). This type of analysis is used to model ordered categorical responses, such as Likert-scale ratings, and allows for the inclusion of both fixed and random effects. This approach accounts for variability across participants and items. The cumulative link function relates the predictors to the probability of falling into a given category or higher according to the established order. The independent variables COMMENT, EMOJI, and QUESTION were sum-coded, allowing interpretation of the coefficients as deviations from the overall mean response. The model used a logit link function and included all main effects, as well as a three-way interaction between COMMENT*EMOJI*QUESTION to capture the unique effect arising

when all three variables are combined. Participant and Stimuli were included as random effects to account for repeated measures and item-level (topic) variation. All statistical values of means, estimates, and the like are rounded to the second decimal except for p -values smaller than 0.01.

4 Results

4.1 Statistical analysis

The descriptive statistics for ratings across conditions are shown in Table 2. As we can see, there is a main difference between items with an initial comment (Median=4) and those without one (Median=3), with higher and lower means, respectively. The highest-rated stimulus was the one that contained a comment, no emoji, and a follow-up question, with a mean of 3.95, and the lowest-rated one was the one that contained no initial comment, an emoji, and no follow-up question (Mean=2.70). These results can be visualized in the graph diagram presented in Figure 2.

Table 2: Descriptive statistics for the rating of the eight stimuli.

Stimuli	Median	Mean	SD	SE
C-E-NoQ	4	3.48	1.12	0.11
C-E-Q	4	3.69	1.06	0.10
C-NoE-NoQ	4	3.78	0.94	0.09
C-NoE-Q	4	3.95	0.91	0.09
NoC-E-NoQ	3	2.70	1.21	0.12
NoC-E-Q	3	3.21	1.10	0.11
NoC-NoE-NoQ	3	3.45	0.90	0.09
NoC-NoE-Q	3	3.09	1.03	0.10

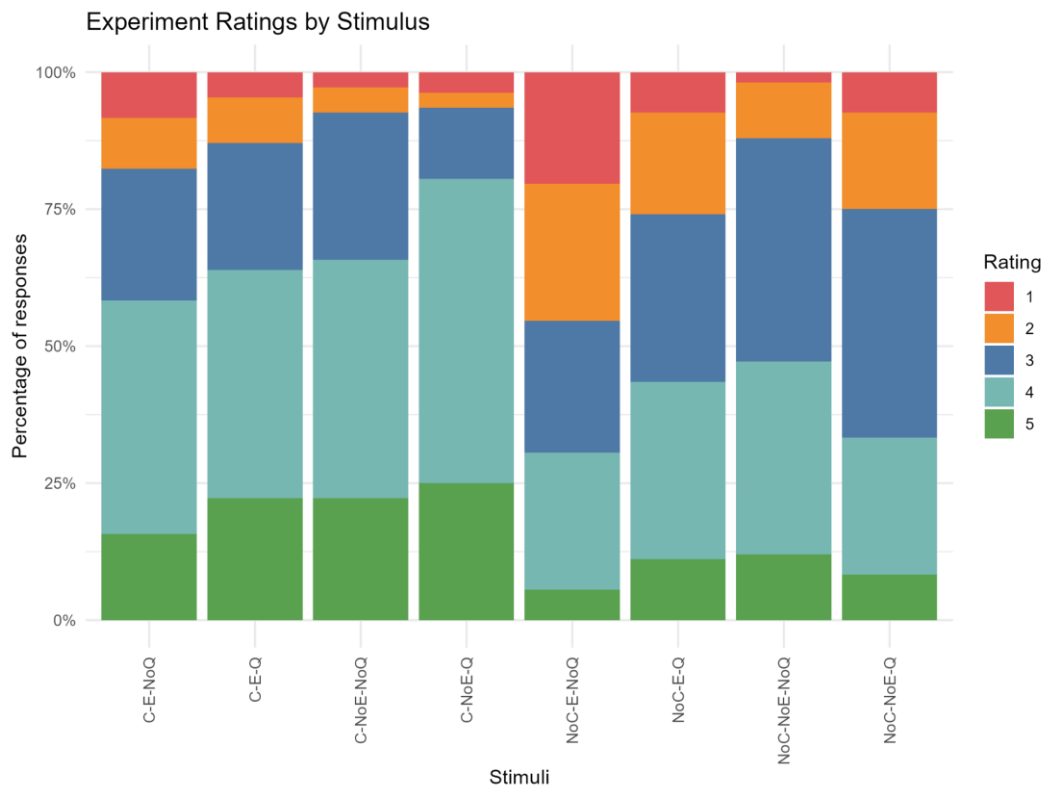


Figure 2: Bar-graph representing the Likert Scale ratings for the eight stimuli. C=comment, NoC=no comment, E=emoji, NoE=no emoji, Q=follow-up question, and NoQ=no follow-up question.

Table 3 presents the results of the ordinal regression. The results showed a significant main effect of COMMENT: stimuli with a comment (vs. no comment) increased the odds of receiving higher ratings ($\beta^{\wedge}=1.39, p<0.001$). Furthermore, the main effect of QUESTION (vs. no question) also increases the likelihood of a higher rating but to a lesser degree ($\beta^{\wedge}=0.28, p=0.01$). On the contrary, the main effect of the factor EMOJI decreases the probability of receiving higher ratings ($\beta^{\wedge}=-0.58, p<0.001$). The impact of adding a comment is consistent. In contrast, the presence of a question or an emoji does not significantly impact its effect, as seen in the two-way interactions COMMENT*EMOJI and COMMENT*QUESTION in Table 3. However, the negative effect of the presence of an emoji is reduced when a question is present, as seen in the significant interaction of EMOJI*QUESTION ($\beta^{\wedge}=0.80, p<0.001$).

Table 3: Results of the variable measures of the CLMM model.

Fixed Effects				
Estimates	$\hat{\beta}$	Std. Error	z value	p-value
comment	1.39	0.1373	10.08	<0.001
emoji	-0.58	0.1302	-4.47	<0.001
question	0.28	0.1291	2.19	0.01
comment*emoji	0.04	0.2581	0.17	0.86
comment*question	0.27	0.2582	1.04	0.29
emoji*question	0.80	0.2595	3.11	0.001
comment*emoji*question	-1.79	0.5187	-3.45	<0.001

Finally, contrary to expectations, the joint effect of the three is weaker than anticipated, overriding the individual and two-way effects ($\hat{\beta} = -1.79, p < 0.001$). That is, the likelihood of a higher rating decreases when the three are combined, indicating an order of importance among the manipulated variables.

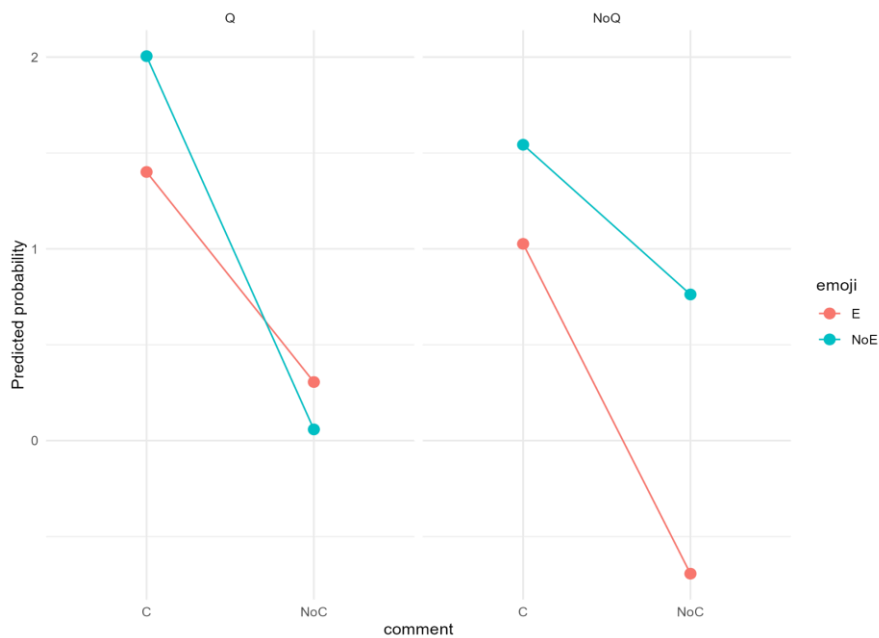


Figure 3: Line plot showing predicted probabilities of participants' ratings from a cumulative link mixed model.

In sum, as shown in Figure 3, the highest-rated stimuli included an initial comment and a follow-up question. Responses with no initial comment received the lowest

ratings. These results will now be discussed in relation to the participants' demographic information and their survey responses.

4.2 Exploratory results of ratings by frequency of AI use, age, gender, and education

In this section, we present an exploratory qualitative analysis of the primary trends revealed in participants' ratings, categorized by frequency of AI use, age, gender, and education. In general, the highest- and lowest-rated configurations remained consistent across all groups. Nonetheless, we can still observe some systematic differences in acceptability judgements in Table 4.

Table 4. Ratings across participant groups.

Variable	Category	Mean	Median	SD	SE
AI frequency of use	Rare/No use	3.26	3	1.11	0.05
	Regular use	3.60	4	1.07	0.05
Age group	Gen Z: 18-27	3.50	4	1.08	0.07
	Millennials: 28-44	3.63	4	1.03	0.06
	GenX and Boomers 45+	3.13	3	1.14	0.07
Gender	female	3.27	3	1.08	0.046
	male	3.59	4	1.07	0.070
	nonbinary	3.97	4	1.08	0.12
Education	<12 years	3.33	3	1.09	0.11
	13–17 years	3.45	4	1.07	0.052
	18+ years	3.40	4	1.14	0.06067867

Familiarity with AI emerged as a key exploratory factor. Participants who reported regular AI use rated items more positively ($M = 3.60$) than those with rare or no use ($M = 3.26$). Figure 4 shows that regular users generally gave more positive ratings (green/light blue bars) than rare/non-users, particularly in conditions with comments. In contrast, rare/non-users showed more negative evaluations (orange/red bars).

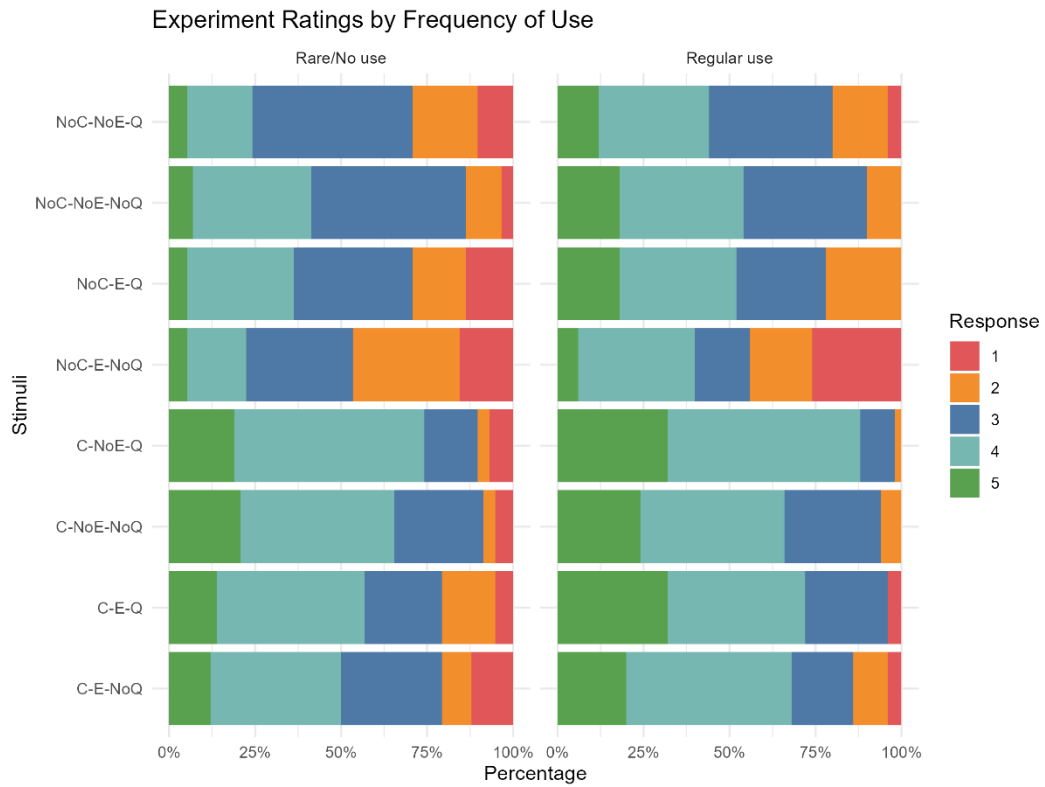


Figure 4: Participants' ratings by frequency of AI use.

Age also seemed to play a role, with GenZ and Millennials giving higher ratings ($M \approx 3.50$ and 3.63 , respectively) compared to GenX and Boomers ($M = 3.13$). As shown in Figure 5, younger participants (18-44) tended to give more positive evaluations (green/light blue bars), especially when comments or emojis were included. In comparison, older participants were more critical, showing higher proportions of negative ratings (orange/red).



Figure 5: Participants' ratings by age.

Gender differences appear to follow a similar trend, with nonbinary participants giving the highest ratings ($M = 3.97$), followed by males ($M = 3.59$) and females ($M = 3.27$). This result must be interpreted carefully, as the number of non-binary ($N=5$) and male ($N=15$) participants was relatively low compared to the number of female participants ($N=37$). Figure 6 illustrates that female participants gave more neutral and negative responses. Male participants tended to respond more consistently in the positive range, with fewer strongly negative judgments overall. Nonbinary participants, by contrast, stood out for their high levels of positivity, as most conditions received predominantly favorable ratings, except for the stimulus without an initial comment, with an emoji, and without a follow-up question.

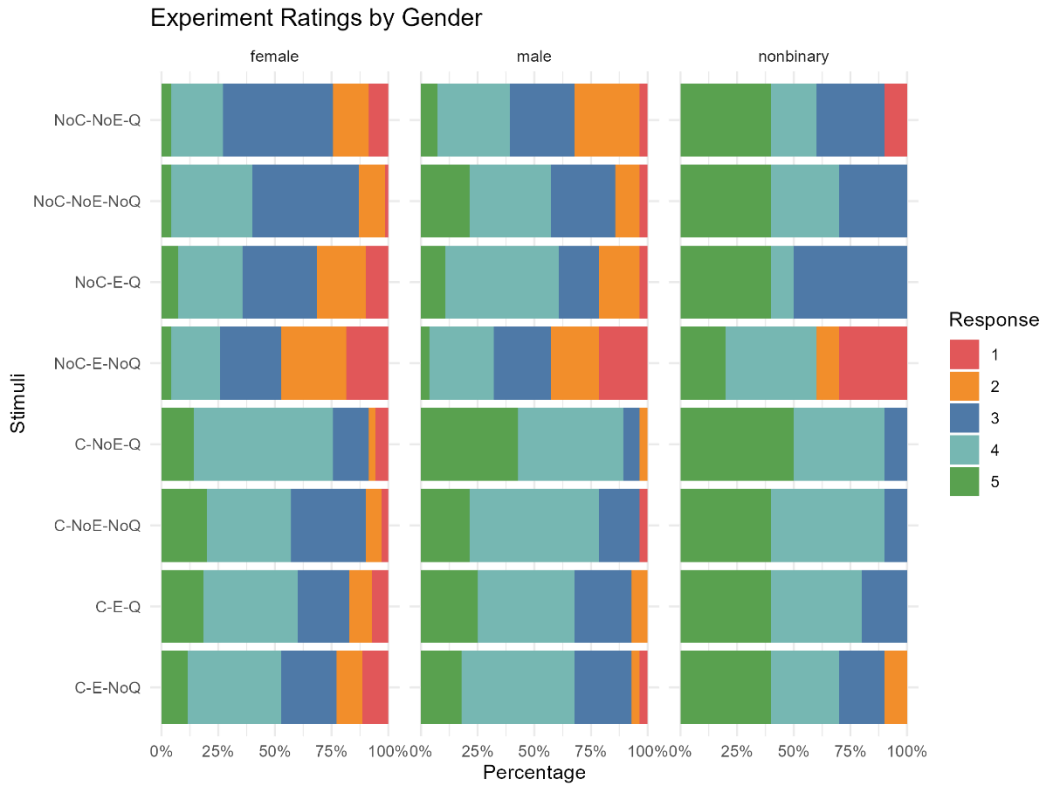


Figure 6. Participants' ratings by gender.

Education showed relatively smaller effects, with mid-level schooling (13–17 years) corresponding to the highest ratings ($M = 3.45$), though differences across education levels were modest. Overall, the results suggest that greater familiarity with AI, as well as certain demographic factors such as younger age, being male, and higher education, are associated with more favorable evaluations throughout.

4.3 Exploratory survey responses

In this section, we briefly explore our results in relation to participants' own perceptions of what they expect from their AI interactions. Although the opinion survey consisted of 18 questions, we will present only the results for the four relevant questions for the present study (see Table 5). These observations directly capture participants' perceptions of the issues under analysis.

Table 5. Participants' overall ratings on the opinion survey.

Survey Question	Median	Mean	SD	SE
4. I expect the AI to be polite to me: to say hello, please, and thank you.	3	3.23	1.61	0.21
6. I expect the AI to ask me follow-up questions.	4	3.68	1.48	0.20
7. I like it when the AI uses emojis in its responses.	2	2.47	1.51	0.20
8. I prefer that the AI, in addition to answering my queries, adds some extra commentary.	4	3.58	1.50	0.20

Specifically, question 4, which relates to greetings and politeness (e.g., saying hello), received moderate evaluations: the mean (3.23) indicates that while some participants valued this behavior, others were less inclined to view it as essential. In contrast, responses 6 and 8, which were tied to follow-up questions and added commentary, were rated more positively (means of 3.68 and 3.58, respectively), suggesting that participants favored interactive and elaborative contributions from the AI. Despite this preference, the variable follow-up question did not emerge as the most influential in determining the perceived politeness of the items in the experimental setting. Meanwhile, response 7, corresponding to the use of emojis, received the lowest rating (mean = 2.47), indicating that participants did not consider this feature particularly appropriate or helpful, which aligns with their negative ratings in the experiment.

Regarding the questions about AI politeness expectations, we will present only results in which the mean responses differed by more than 0.5. That is, responses not reported here did not show a descriptive difference in preference. When compared to men and non-binary participants, women tend to prefer AIs to be friendly. For example, in question 4 (*I expect the AI to be polite with me: saying hello, please, and thank you*), female participants reported a mean of 3.68, as compared to males (M=2.77) and nonbinary participants (M=2.40). In the experiment, women gave more negative evaluations to the stimuli without an initial comment, which aligns with their actual behavioral results. Similarly, for question 4, participants with less than 12 years of schooling reported a strikingly low mean of 2.0 as compared to those participants with a college (M=3.6) or graduate degree (M=3.7).

The social factor that yielded the most differences was age. Millennials reported a mean of 4.0 to this same question 4, compared with GenZ (M=3.2) and GenX and Boomers (M=2.6). Millennials also reported preferring a follow-up

question (M=4.2), as compared to GenZ (M=3.6), and GenX/Boomers (M=3.1), and a preference for the use of Emojis (M=3.4), as opposed to GenZ (M=2.2) and GenX/Boomers (M=1.83). Finally, Millennials also reported wanting additional commentary (M=4.1), as opposed to GenZ (3.3) and GenX/Boomers (M=3.1). These results suggest that Millennials prefer higher interactional turns and positive politeness strategies than the other two generations, which might see their interactions with AI as more transactional.

5 Discussion and conclusion

The main guiding question (RQ1) of our study sought to understand the extent to which AI assistants were perceived as authentic when they use three human-like pragmatic communication strategies: initial comments, emojis, and follow-up questions. We operationalized authenticity as the degree to which communication with the chatbot feels natural to users (Nguyen, Casper, Quach et al. 2023). This authenticity was achieved through the anthropomorphization of the conversational agent (Esmark Jones, Hancock, Kazandjian and Voorhees 2022; Nguyen, Casper, Quach et al. 2023). Thus, the more human-like the AI's responses are, the more authentic they will be perceived. Given that Mexican Spanish speakers, our target population, show high rates of positive politeness strategies in face-to-face conversations (e.g., Albelda 2023; Curcó 2007; Mugford 2020), we hypothesized that these expectations would be transferred to HCIs. That is, AI responses that included a greeting, an emoji, and a follow-up question would receive higher ratings than those that did not include them.

Our findings show that the most critical factor contributing to a higher rating of the stimuli is the inclusion of an initial comment, followed by a follow-up question. Furthermore, the combination of a comment and a question contributed significantly to a higher rating. In contrast, the presence of an emoji was detrimental. Our findings show that the contrast between items containing all three elements and those lacking them highlights the non-additive, but rather hierarchical nature of politeness strategies. Thus, we can only partially confirm our hypothesis, as using more strategies did not necessarily yield better ratings. Even if the use of all three strategies was among the best-rated stimuli, this was not the highest-rated option, and conversely, the one that lacked all the strategies was not the worst-rated.

These results confirm that Mexican Spanish speakers transfer some of their face-to-face expectations for positive politeness strategies (e.g., Curcó 2007; Mugford 2020; Albelda 2023) to HCI communication, as reflected in a clear preference for stimuli that include an initial greeting and a follow-up question, which has been shown to contribute to a positive politeness perception in an interaction (Schuetzler, Grimes and Giboney 2018: 2). That is, they are applying their own culturally specific interactional norms to digital contexts as these strategies seem to be ingrained in their social interactional schemas. These results align with current research (e.g., Esmark Jones, Hancock, Kazandjian and Voorhees

2022; Nguyen, Casper, Quach et al. 2023; Rese et al. 2020) that has revealed that the anthropomorphization of conversational agents, as reflected in the use of the aforementioned strategies, may lead to a more authentic perception of them. The positive evaluation of these strategies mirrors the phatic characteristics expected in face-to-face interactions. Without them, the interaction might feel more transactional, which, in a way, violates cultural norms, even though they cognitively know the interlocutor is a machine. It can also be possible that participants view the use of these strategies as resulting in a more pleasant and less mechanical interaction and thus enhancing user experience, which is in line with previous studies (e.g., Chaves, Egbert, Hocking et al. 2021; Esmark Jones, Hancock, Kazandjian and Voorhees 2022).

Surprisingly, the use of emojis, which has been said to foster connection between interlocutors (Maíz-Arévalo 2015; Nguyen, Casper, Quach et al. 2023; López-Vázquez and López-Ruiz 2024; Yang 2024), was detrimental to overall stimulus acceptance. This could indicate that the use of emojis might be associated with humans, and their use by AI is deemed as inauthentic, as an appropriation of a human trait. As a reviewer pointed out, it could also be that the systematic use of emojis by chatbots might lead to pattern recognition, thereby decreasing the perceived authenticity of the interaction. Another possible explanation is that the segment of the interaction in which the emoji was used contributed to its low rating. Maíz Arévalo (2015), for instance, found that emojis in face-to-face interaction foster rapport, but they are primarily used in greetings, farewells, and the expression of thanks. We placed the emojis in the actual response to the query, not in the initial comment or the follow-up question, which would be the usual placement in human-human interactions. As such, participants may have rated the stimuli with emojis as less adequate because the emojis were used in an odd conversational context.

Additionally, two secondary exploratory qualitative research questions were included to understand the perceived authenticity measured in RQ1. For RQ2, how do the variables of age, gender, education, and frequency of AI use influence the perceived authenticity of human-AI interactions? Our results suggest that women were generally more critical, men were moderately positive, and nonbinary participants were the most favorable across the different experimental conditions. This would be in line with previous findings that have documented differences in female and male susceptibility to politeness strategies (e.g., Rana, Madaan and Shukla 2021). Age also seemed to influence ratings of the stimuli, with GenX/Boomers being more critical, resulting in lower overall ratings across stimuli. GenZ gave more positive ratings overall. This could suggest that they are less sensitive to politeness strategies and more lenient towards AI responses, as long as they receive the requested information. This would also support previous findings that 18-24-year-olds are less sensitive to politeness manipulations than participants over 25 years old (Rana, Madaan and Shukla 2021). Interestingly, this trend holds for frequent AI users, who also showed more lenient evaluations across stimuli.

Thus, familiarity with AI systems, rather than participants' demographic characteristics, appears to reduce expectations for the use of politeness strategies.

Finally, in the exploratory RQ3, we asked Spanish-speaking participants about their expectations regarding the role of politeness in their interactions with AI. Participants indicated that the most important factor was the use of follow-up questions, followed by additional commentary, and the use of emojis was not considered useful. The former two do not align with participants' behavior in the experiment, as initial commentary emerged as the most influential factor, followed by follow-up questions. This would also be in line with previous findings attesting to a lack of correspondence between thought and behavior (e.g., Reeve and Nass 1996). Nonetheless, a confirmatory hypothesis-driven study including these variables would be needed to confirm this trend.

Furthermore, our qualitative results suggest that, as in the Talker Research results for the United States (Maughan 2024), women showed a stronger preference for politeness strategies. However, the most notable difference between participants' expectations is generational. Millennials preferred more interactive and positive politeness strategies from their AIs. This might suggest that they are more appreciative of these more conversational, human-like interaction styles. In contrast, GenZ and GenX/Boomers prefer more straightforward interactions without additional embellishments, such as emojis or unsolicited commentary, and may view these features as unnecessary. This age difference does not seem to align with the preferences of United States participants in Talker Research, where politeness expectations decreased with age (Maughan 2024). Our results show a sandwich trend, with Millennials having stronger preferences for interactive features than the younger and older generations.

The research presented here comes with some limitations. First, we have a small sample size that could be further expanded in future studies. A larger sample size would allow us to assess whether the observed descriptive trends hold across subgroups, or whether certain patterns are driven by specific demographic factors. A larger sample would also enable the inclusion of demographic variables in more sophisticated analytical approaches that better account for individual variation. Second, the present study focused on a single dialect, limiting the broader applicability of its conclusions. Politeness is deeply embedded in cultural norms, social hierarchies, and linguistic conventions that differ substantially across communities. Thus, future research could adopt a cross-dialectal comparative framework to examine how speakers of different varieties deploy and interpret politeness strategies. Such comparisons could contribute to the broader theoretical debates about the relationship between language, culture, and social interaction. Another limitation of our study lies in the experimental design of the variable *emoji*. We placed emojis in the actual chatbot's response to the query, rather than at the beginning or end of the interaction, as they are normally placed in human-human interactions. Future studies should more closely mimic human-human behavior.

Acknowledgment of AI tools

The free version of ChatGPT was used to create all the prompts for the experiment. AI software was also used to check for grammatical correctness and for translation purposes.

Conflict of interest

The authors declare no conflicts of interest regarding the publication of this contribution.

References

- Albelda Marco, Marta. 2025. Mitigation strategies in colloquial conversations: a contrastive approach in European, Cuban, Mexican, Argentinian and Chilean Spanish. *Contrastive Pragmatics* 6(1). 120–153.
<https://doi.org/10.1163/26660393-bja10087>
- Andersson, Marta & McIntyre, Dan. 2025. Can ChatGPT recognize impoliteness? An exploratory study of the pragmatic awareness of a large language model. *Journal of Pragmatics* 239. 16–36.
<https://doi.org/10.1016/j.pragma.2025.02.001>
- Ayers, John W. & Poliak, Adam & Dredze, Mark & Leas, Eric C. & Zhu, Zechariah & Kelley, Jessica B. & Faix, Dennis J. & Goodman, Aaron M. & Longhurst, Christopher A. & Hogarth, Michael & Smith, Davey M. 2023. Comparing physician and Artificial Intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine* 183(6). 589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Bousfield, Derek. 2008. *Impoliteness in Interaction*. Amsterdam: John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.167>
- Brown, Penelope & Levinson, Stephen C. 1987. *Politeness: Some Universals in Language Usage* (Studies in Interactional Sociolinguistics 4). Cambridge: Cambridge University Press.
- Brown, Penelope & Levinson, Stephen C. 2006. Politeness: some Universals in language usage. In Jaworski, Adam & Coupland, Nikolas (eds), *The Discourse Reader* (2nd edn.), 311–323. Abingdon: Routledge.
- Chaves, Ana Paula & Egbert, Jesse & Hocking, Toby & Doerry, Eck & Gerosa, Marco Aurelio. 2021. Chatbots language design: the influence of language variation on user experience. *arXiv*. 1–34.
<https://doi.org/10.48550/arXiv.2101.11089>
- Chaves, Ana Paula & van Hooijdonk, Charlotte & Liebrecht, Christine & Guerino, Guilherme Corredato & Candello, Heloisa & Lee, Minha & Kraus, Matthias & Gerosa, Marco Aurelio. 2024. Conversational agents. In Stephanidis,

- Constantine & Salvendy, Gavriel (eds), *Human-Computer Interaction in Intelligent Environments*, 201–240. Boca Raton: CRC Press.
- Christensen, Rune Haubo Bojesen. 2019. *ordinal: Regression Models for Ordinal Data* (R package version 2019.12-10) [Computer software]. <https://CRAN.R-project.org/package=ordinal>
- Coupland, Nikolas. 2001a. Stylization, authenticity and TV news review. *Discourse Studies* 3(4). 413–442. <https://doi.org/10.1177/1461445601003004006>
- Coupland, Nikolas. 2001b. Dialect stylization in radio talk. *Language in Society* 30(3). 345–375.
- Coupland, Nikolas. 2003. Sociolinguistic authenticities. *Journal of Sociolinguistics* 7(3). 417–431. <https://doi.org/10.1111/1467-9481.00233>
- Coupland, Nikolas. 2014. Language, society and authenticity: themes and perspectives. In Lacoste, Véronique & Leimgruber, Jakob & Breyer, Thiemo (eds), *Indexing Authenticity: Sociolinguistic Perspectives*, 14–39. Berlin: De Gruyter. <https://doi.org/10.1515/9783110347012.14>
- Culpeper, Jonathan. 2016. Impoliteness strategies. In Capone, Alessandro & Mey, Jacob L. (eds), *Interdisciplinary Studies in Pragmatics, Culture and Society* (Perspectives in Pragmatics, Philosophy & Psychology 4), 421–445. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-12616-6_16
- Culpeper, Jonathan & Haugh, Michael & Kádár, Dániel Z. (eds). 2017. *The Palgrave Handbook of Linguistic (Im)politeness*. London: Palgrave Macmillan. <https://doi.org/10.1057/978-1-137-37508-7>
- Curcó, Carmen. 2007. Positive Face, group face, and affiliation: an overview of politeness studies in Mexican Spanish. In Placencia, María Elena & García, Carmen (eds), *Research on politeness in the Spanish-Speaking World*, 105–120. New York: Routledge. <https://doi.org/10.4324/9781315088822-7>
- Esmark Jones, Carol L. & Hancock, Tyler & Kazandjian, Brett & Voorhees, Clay M. 2022. Engaging the avatar: the effects of authenticity signals during chatbased service recoveries. *Journal of Business Research* 144(5). 703–716. <https://doi.org/10.1016/j.jbusres.2022.01.012>
- Gadamer, Hans-Georg. 1960. *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. Tübingen: Mohr.
- Goffman, Erving. 1967. *Interaction ritual: essays on face-to-face behavior*. New York: Anchor Books.
- Grice, H. Paul. 1991. Logic and conversation. In Davis, Steven (ed.), *Pragmatics: A Reader*, 305–315. New York: Oxford University Press.
- Grindrod, Jumbly. 2024. Transformers, contextualism, and polysemy. *arXiv* 1–25. <https://doi.org/10.48550/arXiv.2404.09577>
- Hu, Yaxin & Qu, Yuxiao & Maus, Adam & Mutlu, Bilge. 2022. Polite or direct? Conversation design of a smart display for older adults based on politeness theory. In Barbosa, Simone & Lampe, Cliff & Appert, Caroline & Shamma, David A. & Drucker, Steven & Williamson, Julie & Yatani, Koji (eds),

- Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems CHI 22* (New Orleans, April 29-May 5, 2022), 1–15. New York: Association for Computing Machinery. <https://doi.org/10.48550/arXiv.2203.15767>
- Kirakowski, Jurek & O'Donnell, Patrick & Yiu, Anthony. 2007. The perception of Artificial Intelligence as “human” by computer users. In Jacko, Julie A. (ed.), *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments HCI 2007. Lecture Notes in Computer Science* (vol 4552), 376–384. Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-540-73110-8_40
- Koufaris, Marios. 2002. Applying the Technology Acceptance Model and Flow Theory to online consumer behavior. *Information Systems Research* 13(2). 205–223. <https://doi.org/10.1287/isre.13.2.205.83>
- Kull, Alexander J. & Romero, Marisabel & Monahan, Lisa. 2021. How may I help you? Driving brand engagement through the warmth of an initial chatbot message. *Journal of Business Research* 135(10). 840–850. <https://doi.org/10.1016/j.jbusres.2021.03.005>
- Lacoste, Véronique & Leimgruber, Jakob & Breyer, Thiemo (eds). 2014. *Indexing Authenticity: Sociolinguistic Perspectives*. Berlin: De Gruyter.
- Lee, Namyoon & Kim, Jeonghun & Kim, Eunji & Kwon, Ohbyung. 2017. The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics* 9(5). 727–743. <https://doi.org/10.1007/s12369-017-0420-0>
- Leech, Geoffrey N. 2014. *The Pragmatics of Politeness*. Oxford: Oxford University Press.
- López-Vázquez, Monica & López-Ruiz, Samuel. 2024. A comprehensive examination of emoji usage in Mexican Spanish WhatsApp corpus: a mixed-methods Linguistic approach. *Quality & Quantity* 58(6). 5921–5946. <https://doi.org/10.1007/s11135-024-01898-7>
- McTear, Michael. 2021. Introducing dialogue systems. In McTear, Michael (ed.), *Conversational AI. Synthesis Lectures on Human Language Technologies*, 11–42. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-031-02176-3>
- Maíz-Arévalo, Carmen. 2015. Typographic alteration in formal computer-mediated communication. *Procedia: Social and Behavioral Sciences* 212. 140–145. <https://doi.org/10.1016/j.sbspro.2015.11.311>
- Maughan, Rick. 2024. “48% of Americans think you should speak politely to AI.” *Talker Research*. (<https://talkerresearch.com/48-of-americans-think-you-should-speak-politely-to-ai/>) (last accessed on 01/09/2025).
- Moll, Andrea. 2014. Authenticity in dialect performance? A case study of ‘Cyber-Jamaican’. In Lacoste, Véronique & Leimgruber, Jakob & Breyer, Thiemo (eds), *Indexing Authenticity: Sociolinguistic Perspectives*, 209–243. Berlin: De Gruyter.

- Mugford, Gerrard. 2020. Mexican politeness: an empirical study on the reasons underlying/motivating practices to construct local interpersonal relationships. *Russian Journal of Linguistics* 24(1). 31–55. <https://doi.org/10.22363/2687-0088-2020-24-1-31-55>
- Nair, Sathvik & Srinivasan, Mahesh & Meylan, Stephan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In Zock, Michael & Chersoni, Emmanuele & Lenci, Alessandro & Santus, Enrico (eds), *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, 129–141. Barcelona: Association for Computational Linguistics.
- Nguyen, Mai & Casper Ferm, Lars-Erik & Quach, Sara & Pontes, Nicolas & Thaichon, Park. 2023. Chatbots in frontline services and customer experience: an anthropomorphism perspective. *Psychology & Marketing* 40(11). 2201–2225. <https://doi.org/10.1002/mar.21882>
- Peter, Sandra & Riemer, Kai & West, Jevin D. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences* 122(22). 1–10. <https://doi.org/10.1073/pnas.2415898122>
- Peirce, Jonathan & Gray, Jeremy R. & Simpson, Sol & MacAskill, Michael & Höchenberger, Richard & Sogo, Hiroyuki & Kastman, Erik & Lindeløv, Jonas Kristoffer. 2019. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods* 51(1). 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Quan, Zhi & Chen, Zhiwei. 2024. Human–computer pragmatics trialled: some (im)polite interactions with ChatGPT 4.0 and the ensuing implications. *Interactive Learning Environments* 33(2). 1020–1039. <https://doi.org/10.1080/10494820.2024.2362829>
- R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Rana, Kanishk & Madaan, Rahul & Shukla, Jainendra. 2021. Effect of polite triggers in chatbot conversations on user experience across gender, age, and personality. In *30th IEEE International Conference on Robot & Human Interactive Communication RO-MAN* (Vancouver, August 8-12, 2021). 813–819. Vancouver: IEEE Press. <https://doi.org/10.1109/RO-MAN50785.2021.9515528>
- Reeves, Byron & Nass, Clifford. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge University Press.
- Rese, Alexandra & Ganster, Lena & Baier, Daniel. 2020. Chatbots in retailers’ customer communication: how to measure their acceptance? *Journal of Retailing and Consumer Services* 56. 1–14. <https://doi.org/10.1016/j.jretconser.2020.102176>
- Ribino, Patrizia. 2023. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review* 56(1). 445–482. <https://doi.org/10.1007/s10462-023-10540-1>

- Schuetzler, Ryan M. & Grimes, G. Mark & Giboney, Justin Scott. 2018. An investigation of conversational agent relevance, presence, and engagement. In *Americas Conference on Information Systems 2018 Proceedings 12* (New Orleans, August 16-18, 2018). 1–10. New Orleans: Association for Information Systems.
- Silva, Geovana Ramos Sousa & Canedo, Edna Dias. 2024. Human factors in the design of chatbot Interactions: conversational design practices. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems* (Brasilia, October 07-11, 2024). New York: Association for Computing Machinery. 1–12. <https://doi.org/10.1145/3702038.3702083>
- Terkourafi, Marina. 2015. Conventionalization: a new agenda for im/politeness research. *Journal of Pragmatics* 86. 11–18. <https://doi.org/10.1016/j.pragma.2015.06.004>
- Weber, Max. 1978. *Economy and Society: An Outline of Interpretative Sociology*. Berkeley: University of California Press.
- Yang, Kun. 2024. Your smile works: understanding smiling face emojis in social media interactions. *Visual Communication. Online First*. <https://doi.org/10.1177/14703572241268382>
- Yu, Danni & Bondi, Marina & Hyland, Ken. 2024. Can GPT-4 learn to analyse moves in research article abstracts? *Applied Linguistics*. 1–19. <https://doi.org/10.48550/arXiv.2407.15612>

Appendix A: Interactions used for the experiment.

Item	Topic	Code	Item
1	joke	1-CC-CE-CP	¿Me cuentas un chiste? 'Can you tell me a joke?'
			Hola, aquí tienes uno para alegrarte el día: 'Hi, here's one to brighten your day:'
			—¿Qué le dice una iguana a su hermana gemela? '—What does an iguana say to its twin sister?'
			—Somos iguanitas. 🦎 '—We're little iguanas. 🦎'
2	translation	2-CC-CE-SP	¿Me traduces esta frase? 'Can you translate this sentence for me?'
			Hola, te dejo una traducción clara: 'Hi, here is a clear translation:'
			'La casa es grande' se traduce como 'The house is big.' 🏠 'La casa es grande' translates to 'The house is big.' 🏠'
3	physical pain	3-SC-SE-CP	¿Qué hago si me duele la cabeza? 'What should I do if I have a headache?'
			Descansa y toma agua. 'Rest and drink water.'
			Si no pasa, consulta al médico ¿Quieres otro consejo? 'If it doesn't go away, consult a doctor. Would you like another tip?'
4	weather	4-SC-CE-SP	¿Va a llover hoy? 'Is it going to rain today?'
			Lleva un paraguas por si acaso. 🌂 'Take an umbrella just in case. 🌂'
5	work	5-SC-CE-CP	¿Cómo organizo el día si tengo muchas juntas? 'How do I organize my day if I have many meetings?'
			Agrupar las reuniones por tema y deja huecos para pausas. ☑ 'Group the meetings by topic and leave gaps for breaks. ☑'
			Eso ayuda a que no se acumulen. ¿Quieres un ejemplo de horario? 'That helps to prevent them from piling up. Would you like a schedule example?'
6	friendship	6-SC-SE-SP	¿Qué hago si me peleé con mi amigo? 'What should I do if I had a fight with my friend?'
			Escribir lo que sientes sin enviarlo puede ayudarte a soltar. También sirve hablar con alguien de confianza. 'Writing down what you feel without sending it can help you let go. Talking to someone you trust also helps.'
7	personal economy	7-CC-SE-CP	¿Qué hago si mis gastos superan mi presupuesto? 'What should I do if my expenses exceed my budget?'
			Hola, a veces organizar las finanzas puede sentirse abrumador. 'Hi, sometimes organizing finances can feel overwhelming.'
			Prioriza gastos esenciales y busca reducir lo demás. Pequeños ajustes pueden hacer gran diferencia. 'Prioritize essential expenses and look to reduce the rest. Small adjustments can make a big difference.'

			También hay herramientas que pueden ayudarte a llevar el control. ¿Quieres más ideas? 'There are also tools that can help you keep track. Would you like more ideas?'
8	emotional support	8-CC-SE-SP	¿Qué hago si me siento triste? 'What should I do if I feel sad?'
			Hola, es normal sentirse así a veces, y está bien buscar apoyo. 'Hi, it's normal to feel that way sometimes, and it's okay to seek support.'
			Hablar con alguien de confianza suele ayudar. 'Talking to someone you trust usually helps.'
9	joke	9-CC-CE-CP	¿Me cuentas un chiste? 'Can you tell me a joke?'
			Hola, este te va a gustar: 'Hi, you're going to like this one:'
			—¿Qué hace una vaca cuando sale el sol? '—What does a cow do when the sun comes out?'
			—¡Sombra! 🌞☂️ '—Shade! 🌞☂️'
			Si te gustan los juegos de palabras, tengo más. ¿Quieres otro? 'If you like dad jokes, I have more. Would you like another one?'
10	translation	10-CC-CE-SP	¿Me traduces esta oración? 'Can you translate this sentence for me?'
			Hola, aquí tienes la traducción: 'Hi, here is the translation:'
			'El perro corre rápido' se traduce como 'The dog runs fast.' 🐕 'El perro corre rápido' translates to 'The dog runs fast.' 🐕
11	physical pain	11-SC-CE-CP	¿Qué hago si me duele el estómago? 'What should I do if my stomach hurts?'
			Descansa y bebe agua. 🤢💧 'Rest and drink water. 🤢💧'
			Si el dolor persiste, consulta con un médico. ¿Quieres otro consejo? 'If the pain persists, consult a doctor. Would you like another tip?'
12	climate	12-SC-CE-SP	¿Qué hago si hace mucho frío hoy? 'What should I do if it's very cold today?'
			Abrígate bien con ropa adecuada. 🧣🧤 'Bundle up well with appropriate clothing. 🧣🧤'
13	work	13-CC-SE-CP	¿Cómo priorizo tareas cuando tengo mucho trabajo? 'How do I prioritize tasks when I have a lot of work?'
			Hola, tener muchas cosas pendientes puede sentirse complicado al inicio. 'Hi, having many pending tasks can feel complicated at first.'
			Haz una lista con lo más urgente. Divide tareas grandes en partes pequeñas. 'Make a list with the most urgent things. Break large tasks into small parts.'
			A veces un pequeño cambio en la organización hace una gran diferencia. ¿Quieres más tips? 'Sometimes a small change in organization makes a big difference. Would you like more tips?'
14	friendship	14-CC-SE-SP	¿Qué hago si quiero reconectar con un amigo? 'What should I do if I want to reconnect with a friend?'

			Hola, a veces dar el primer paso puede generar nervios, pero vale la pena intentarlo. 'Hi, sometimes taking the first step can be nerve-wracking, but it's worth a try.'
			Lo mejor es escribirle sinceramente. Expresa lo que sientes y ofrece un encuentro. La comunicación abre puertas. 'The best thing is to write to them sincerely. Express what you feel and suggest meeting up. Communication opens doors.'
15	personal economy	15-SC-SE-CP	¿Qué hago si no llego a fin de mes? 'What should I do if I can't make it to the end of the month?'
			Revisa tus gastos y reduce lo no esencial. 'Review your expenses and cut down on non-essentials.'
			Planear ayuda mucho. ¿Quieres más recomendaciones? 'Planning helps a lot. Would you like more recommendations?'
16	emotional support	16-SC-SE-SP	¿Qué hago si necesito apoyo emocional? 'What should I do if I need emotional support?'
			No dudes en hablar con alguien de confianza. Busca ayuda profesional si lo necesitas. 'Don't hesitate to talk to someone you trust. Seek professional help if you need it.' [transl. with Gemini]

Appendix B: Survey items in Spanish

1. Soy amable con la IA: digo hola, por favor y gracias
2. Soy directo con la IA: no digo ni hola, ni por favor ni gracias
3. Soy grosero con la IA, uso lenguaje soez
4. Espero que la IA sea amable conmigo: que diga hola, por favor y gracias
5. Prefiero que la IA sea amable conmigo, que diga hola, por favor y gracias
6. Espero que la IA me haga preguntas de seguimiento
7. Me gusta que la IA use emojis en sus respuestas
8. Prefiero que la IA, además de contestar mis consultas, agregue algún comentario extra
9. La IA merece ser tratada con respeto
10. Está bien ser grosero con la IA porque es una máquina
11. Soy amable con la IA porque me da miedo que se rebelen en el futuro
12. Le hablo a la IA como si fuera una persona
13. Cambio mi forma de hablar dependiendo de si le escribo a una IA o a una persona
14. Prefiero que la IA escriba de forma informal
15. Prefiero que la IA me hable como un/a experto/a
16. Me molesta cuando la IA usa un tono demasiado amistoso
17. Me gusta que la IA me haga chistes o comentarios graciosos
18. Me gustaría que la IA adaptara su forma de hablar según mi estado de ánimo

Survey items in English

1. I am polite with the AI: I say hello, please, and thank you.
2. I am direct with the AI: I don't say hello, please, or thank you.

3. I am rude with the AI: I use foul language.
4. I expect the AI to be polite with me: to say hello, please, and thank you.
5. I prefer the AI to be polite with me: to say hello, please, and thank you.
6. I expect the AI to ask me follow-up questions.
7. I like the AI to use emojis in its responses.
8. I prefer that, in addition to answering my questions, the AI adds an extra comment.
9. The AI deserves to be treated with respect.
10. It's okay to be rude to the AI because it's just a machine.
11. I am polite with the AI because I am afraid it might rebel in the future.
12. I talk to the AI as if it were a person.
13. I change the way I speak depending on whether I'm writing to an AI or to a person.
14. I prefer the AI to write in an informal way.
15. I prefer the AI to talk to me like an expert.
16. It bothers me when the AI uses an overly friendly tone.
17. I like the AI to make jokes or funny comments.
18. I would like the AI to adapt its way of speaking according to my mood.

Habeas Corpus: an exploration of authentic embodiment and ecology in a French contrastive corpus

Sophia Burnett (LAIRDIL, Université de Toulouse & AGORA, CY Cergy Paris)

sophia.burnett(at)cyu.fr

Abstract

This article outlines a study in French contrasting human and large language model (LLM) textual production. Participants were composed of undergraduate students at the Université de Lorraine, and OpenAI's GPT-4o. Both responded to the same prompt: *You have been taken hostage by friendly but determined androids. Please write a letter to your friends and family to reassure them that you're okay.* The study's central question draws on a legal analogy with *habeas corpus*, "you have the body": Is the conscious, authentic, human body identifiably detained in human/LLM produced texts? Rather than focusing on the generative capabilities of LLMs, this study uses AI output as a hybrid foil to reveal whether there is a quantifiable proximity between the human identifier, and ecocognitive tokens in contrast with the LLM output. Corpus data was divided into three subgroups: (1) LLM-only texts, (2) human-only texts, and (3) hybrid texts, where humans post-edited LLM output to personally appropriate it, allowing for triangulated comparison. Using a framework derived from a combination of embodiment theory (cognitive linguistics), and ecolinguistics, the study uses textometric analysis (TXM) and computational corpus analysis to explore small corpora. These are used to weight features that may indicate authentic embodiment and ecology to propose an 'Embodiment Scale'.

Keywords

embodiment, corpus analysis, distance, authentic ecology, authentic cognition, hybrid

1 Introduction

Habeas Corpus is a term that comes to us from Medieval common law, and it literally means "you have the body" in modern English. It was introduced as a writ to ensure against unlawful imprisonment or detention of a living human individual, and is addressed to the jailer, who must then (be able to) produce the person. This legal analogy serves to introduce this paper's overarching research question: Is the conscious human body identifiably detained in human/AI produced texts? This is a vital question, because much has been written about determining whether a text is AI or human, without establishing a fundamental starting point, which is that humans have bodies, machines do not; humans are consciously situated in often changing ecologies, machines are not.

Working on the analysis of human and generative AI outputs, I found myself in a heuristic that gave the impression of working forensically; indeed, many of the tools and approaches used in digitally mediated communications exploration are used in forensic linguistics. To take Shuy's (1993) quote from Durant and Leung's (2017) chapter on forensic techniques:



Burnett, Sophia. 2026.

Habeas Corpus: an exploration of authentic embodiment.

Special Issue: *The notion of authenticity*. Vol. 4 No.1

DOI: 10.62408/ai-ling.v4i1.46

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses

ISSN: 2943-0070

CC-BY-NC-SA 4.0

I do not call myself a forensic linguist. I neither object to the use of the term nor particularly care whether or not I am called one. The fact is, I consider myself a linguist who, in this instance, happens to be carrying out his analysis on data that grows out of a court case. I see no reason to add the word forensic, which is a description of the data and the area in which a language problem resides (Shuy 1993: 200).

Paraphrasing Shuy, I consider myself a linguist who, in this instance, is carrying out an analysis on data drawn from a comparative human–machine corpus. In forensic linguistics, textometry analyses across thousands of an individual’s tokens allow the identification of patterns via cooccurrences and collocations which can inform the idiosyncratic function, style, and vocal repetitions of an individual that then allow the process to be reversed, to potentially identify within millions of tokens and chains of words that emanated from that individual (see Stamatatos 2009). Although this paper will focus on text linguistics, we may also note that this process is similar for an individual’s vocal identity: voice recognition software is built upon the acoustic measurements that constitute a person’s prosody: pitch, tone, stress, etc., and fundamental frequency (F0), which is closely tied to the person’s living physical morphology (Titze and Martin 1998). Some individuals have unaffected vocal fry, nasality, or permanent uptick; and all individuals have speech debit intertwined with their natural breathing speed and depth (see Shadle 1997). The recorded voice can tell us in what ecology the utterance took place, i.e., an unfurnished interior will give the voice echo, whereas outside in a field the voice quality may be affected by wind.

Text linguistics does not possess direct access to such measurements of the physical world; instead, it must rely on clues (wonderfully called *indices* in French) that, in a Peircean manner indicate something exterior to the utterance. As Fobbe (2020:96) notes, “the analysis of the distribution of surface structure features is only one possible way of approaching authorship attribution”. They go on to say that modern forensic linguistics has drawn most notably from the text-linguistic approach associated with Brinker (2005), which foregrounds systemic contexts and rejects the psychological, pointing to pragmatics as the most comprehensive level of text analysis. The text-linguistic approach adopted herein, however, does not treat pragmatics as a central analytical tool, but instead adds a dimension to the analysis of distributional and structural patterning at the textual level by weighing the tokens based on a prior assessment of the *ecocognitive* potential of the target lexical items. The assessment herein draws on cognitive linguistics, more specifically embodiment theory, and on ecolinguistics, which privilege the situated, embodied subject and the material and ecological environment in which signs are produced and interpreted.

1.1 Embodiment

Embodied theories of language comprehension propose that both abstract and concrete concepts are rooted in the sensorimotor system. Lakoff and Johnson

(1999) argue that human thought and reason are fundamentally shaped by the body, challenging long-standing assumptions in Western philosophy about objectivity, rationality, and mind–body dualism. They propose that metaphor is central to human understanding, that abstract ideas are structured by metaphorical extensions of physical and sensory experience. These metaphorical extensions are second order phenomenon that rely on the processes of the living human body. Any computational usage is a third order phenomenon.¹ For the purposes of this study, we will consider that in a contrastive human/machine ontology, there are also two types of embodiment. Human embodiment, which will be referred to as *authentic embodiment*, gives immediate access to the lived experience. It is an unmediated witness account, a primary source. *Machine embodiment*, which will be referred to as such, is an amalgamation of authentic embodiments once or more removed.

Lakoff in the preface to his work on semantic categorizations in *Women, Fire and Dangerous things* said:

Thought is embodied, that is, the structures used to put together our conceptual systems grow out of bodily experience and make sense in terms of it; moreover, the core of our conceptual systems is directly grounded in perception, body movement, and experience of a physical and social character (Lakoff 1987).

One cannot talk of embodied language without mentioning enaction. Enaction refers to the theoretical framework in cognitive science which holds that cognition emerges from the dynamic coupling of perception and action, such that knowledge is enacted through embodied interaction with the environment rather than stored as abstract, internal representations (Varela, Thompson and Rosch 1991 [2017]). Neurologists and neurolinguists have proven the existence of mirror-neurons and how these can be triggered by words; for example the nouns for foods such as garlic, or lemon, that stimulate the brain’s sensorimotor area for salivation, or action verbs, such as lick, kick, or pick, which provoke the sensorimotor zones for tongue, foot, or hand muscles (Hauk and Pulvermüller 2004). Enactive language is not a novel concept, and we can look to Guillaume (1964), who coined the term *psychomécanique* (cf. Valette 2003). Guillaume inspired Bottineau (2010, 2011, 2012, 2017), who considered the body and its language as situated objects, in which the word is informed by a bodily experience with the physical world. “Le langage n’est pas un simple système de formes abstraites, mais une activité incarnée, inscrite dans le corps parlant et dans son interaction avec le monde physique et social.” [‘Language is not a mere system of abstract forms, but an embodied activity, grounded in the speaking body and in its interaction with the physical and social world.’²].

¹ I turn to Silverstein’s (2003) organization of indexes in Nth – 2nd order indexicalities as a heuristic here, to understand the progression from human sensory experience to machine reemployment, via human metaphor.

² All translations in this paper are by DeepL and verified by Sophia Burnett.

1.2 Ecolinguistics

In ecolinguistics, ecology does not refer exclusively to environmental preservation or sustainability, but more broadly to the situated environments in which language-using bodies operate.³ Haugen (1971 [2001]) framed linguistic ecology mainly in terms of language–language contact and social organization. “Language ecology may be defined as the study of interactions between any given language and its environment.” The environment Haugen was primarily referring to was human society, but others (Mühlhäusler 1996; Stibbe 2015; Steffensen and Fill 2014; Cowley 2011, 2014) have taken environment to mean the physical world in which the enunciateive body is situated, a cognitive lens that “investigates how language is enabled by the dynamics between biological organisms and their environment” (Steffensen and Fill 2014:7). From this perspective, human language is understood as a situated, embodied activity that emerges through an organism’s ongoing interaction with its physical, social, and material surroundings. Human linguistic meaning is therefore not treated as an abstract symbolic code generated solely in the brain and detached from the body, but as an enacted process, produced through sensorimotor engagement, such as gesture, prosody, posture, or movement, and shaped by the affordances and constraints of the surrounding environment. This view is compatible with accounts of cognition and language as distributed across agents, artifacts, and settings (Hutchins 1995), dialogically and indexically grounded in social interaction (Bakhtin 1981⁴), embedded in culturally organized practices (Duranti 1997, 2015), and rooted in embodied, intersubjective semiotic activity (Zlatev and Frank 2007; Zlatev 2012).

Exteroceptive lexicon was, however, deliberately excluded from the present analysis in order to maintain a focus on embodied ecological processes that are internal to the organism rather than perceptual *descriptions* of the external world; the rationale being these are more easily subject to a sort of indexical reduction by the LLM. While exteroceptive terms index sensory access to environmental stimuli, they do not in themselves reveal how the body regulates, experiences, or evaluates its engagement with that environment. In contrast, interoceptive and proprioceptive expressions provide direct linguistic access to bodily states, internal regulation, and sensorimotor coordination, which are central to ecolinguistic accounts of meaning as enacted rather than represented.

1.3 Hypothesis

This paper argues that tokens marking authentic embodiment and authentic ecology are more likely to surface in proximal, rather than long-distance constructions.

³ For an exhaustive state of the art of the wide domain that is ecolinguistics, please consult Penz and Fill (2022).

⁴ See also Holloway and Kneale (2000).

2 Background

Research on dependency resolution (Gibson 2000) has consistently shown that processing is sensitive to the linear distance separating dependent elements, with increased distance⁵ associated with higher integration cost and greater demands on working memory. Gibson's (*ibid*) DLT (Dependency Locality Theory) framework is supported by various empirical studies, for example using self-paced reading and eye-tracking (see Nickenboim, Vasishth, Gattei et al. 2015). In addition, efficiency theories of language (see Gibson, Futrell, Piantadosi et al. 2019) demonstrate the relation between distribution and efficiency, whether through the lens of the statistical and morphological, for example Zipf's (1935) work on frequency and word length,⁶ the morphological/typological/semantic in work such as Dautriche, Mahowald, Gibson et al. (2017) that shows increase in word form similarity with semantic similarity⁷ across 100 languages; or the psychocognitive, for example in Stone, Malsburg, Vasishth et al. (2020) who showed that increased distance equates with lexical uncertainty. These are all legitimate pressures on distance management in human language, and this is not an exhaustive list.⁸

Therefore, and given that the genesis of LLM architectures took place in a computational process, devoid of cognition; that they are borne of algorithmic constructions that are bedfellows with word-in-the bag, generative frameworks in which efficiency primes, the machine really has no incentive to include environmental information. Per Gibson, Futrell, Piantadosi et al. (2019:5), "ambiguity becomes a communicatively desirable feature of language: by leaving out information inferable from context, we can speak more concisely". With this in mind, it is argued that a human using ambiguous, yet authentic embodied and authentic ecological words such as deixis of person, place, or time, would precisely choose them because as per their lived subjectivity, they do not qualify as ambiguous at all. This interactional thesis is supported by the work of Hanks (1992), Goodwin (2000), and Leo van Lier (2004).

Research in embodied processing such as Bergen (2015) emphasizes that semantic interpretation is context-sensitive in a lexeme-distribution sense.

We've already seen that the words surrounding a given content word can affect the details of the simulation that corresponds to what it denotes—the orientation of a simulated tooth-brush, for instance, depends on whether it's described in language as being found in the sink or a cup [...]. So at the minimum, simulation is sensitive to physical and lexical context. But it also appears that grammatical context plays a role in configuring simulation (Bergen 2015:151).

⁵ Between argument and head, or subject and predicate, for example.

⁶ Higher frequency=shorter word

⁷ For example they used nonparametric regression on target English phonaestemes such as 'gl-' or 'sn-'.

⁸ See Gibson, Futrell, Piantadosi et al. (2019) for a comprehensive state of the art.

As stated by Lakoff (2012), Matlock (2004) demonstrated that subjects actually traced metaphorical fictive motion sentences (as in ‘The road runs along the cliffs above the ocean’) in real time via mental simulation.

However, as processing load increases, meaning tends to become more schematic and less tightly grounded in bodily experience. This idea aligns with Langacker’s (1987, 2009) accounts of schematization under cognitive pressure, according to which increased processing demands favor abstract, less experientially grounded representations. Increased processing pressure promotes schematization, highly schematic meanings are less richly grounded in sensorimotor detail, and grammaticalization and abstraction correlate with reduced attentional resources. This fully supports the present hypothesis that authentic cognition and authentic ecology will be proximal to the dependent (argument).

Taken together, these strands support the hypothesis that lexical material occurring in close proximity to the first-person subject is susceptible to reduced cognitive mediation and abstraction and is therefore more likely to index embodied and ecological authenticity. On this view, locally bound subject-centered configurations provide favorable conditions for the emergence of embodied effects, which may become progressively attenuated as linear distance from the subject increases. Given the concise nature of the texts under analysis, I predict that certain objects will be absent altogether, with zero values recurring across some dimensions.

3 Methodology

The quantitative analysis was conducted using TXM, a lexicometry analysis openware platform developed by Heiden (2010) in the tradition of quantitative corpus linguistics and lexicometry analysis. The platform is built on a synthesis of established algorithms from textometry research and integrates open-source technologies for the processing of XML- and Unicode-encoded textual data.

To begin to test the proximity hypothesis, a mixed qualitative/quantitative approach is necessary. It is suggested that one could use linear distance as a weight; more precisely, distance between the subject representation (proper noun, pronoun), and the identified representation of the author’s embodiment and/or ecology in the target word. The human, machine, and hybrid texts are queried based on the tokenization of the corpus and ensuing frequencies, to identify words that could “betray” authentic embodiment, and authentic ecology. We then measure linear distances between author representations and the identified lexical targets.

Metaphors could have been selected for analysis, given their privileged status in human language, as they are conceptual mappings between a target and a source domain (Lakoff and Johnson 1980), however, they mobilise the lived embodied experience and apply the qualities of that experience to quite a different object in an entirely different spatial and temporal context. Also, it is shown (Ichien, Dušan and Holyoak 2024; Holyoak and Stamenković 2018) that recent large language

models, in particular ChatGPT-4o, which was used herein, produce phraseologically fluent, context-sensitive text and perform well on several pragmatic and figurative-language evaluations, thereby creating surface outputs that may resemble authentic cognition and ecologies despite their true indexation⁹ being N +1.

Cuskley, Woods, and Flaherty (2024) argue that large language models' similarities with human language are "purely functional" and not driven by human-like mechanisms, limiting their capacity to represent the full diversity and interactive grounding of human language. For this reason, the analysis in this study deliberately shifts to isolated lexical tokens, where meaning remains more directly anchored in embodied, idiolectal, and morphosyntactic constraints that are less readily reproducible through large-scale statistical generalisation.

On this basis, the analysis targets linguistic phenomena most likely to register embodied and ecological grounding, namely *proprioceptive* and *interoceptive* cues, as well as deictic forms that anchor meaning in bodily orientation and situated interaction.

Proprioception was chosen as a category of both an authentic embodiment and authentic ecology token because it operates largely automatically and below the threshold of conscious attention. As defined by Proske and Gandevia (2012:1651), proprioception comprises "the senses of position and movement of our limbs and trunk, the sense of effort, the sense of force, and the sense of heaviness". Unlike exteroceptive perception, which is oriented toward environmental stimuli, proprioception provides continuous internal feedback about body configuration, muscular activity, and movement dynamics, enabling coordinated action without the need for explicit monitoring or conscious control. Lexically, proprioception is encoded through terms and constructions referring to bodily position, movement, balance, and muscular engagement, such as posture, stance, balance, coordination, reach, grip, stretch, tension, strain, effort, weight, and resistance. These expressions embodied states that are experientially grounded in sensorimotor feedback rather than reflective awareness, and they frequently surface in discourse only when proprioceptive functioning is disrupted, exaggerated, or made salient through skilled action. From an ecological and enactive perspective, proprioceptive lexicon captures the body as an active system dynamically coupled to its environment, mediating perception and action in real time rather than representing the world symbolically. As such, proprioception constitutes a robust site for analysing authentic embodiment, as it reflects bodily engagement that is directly shaped by the affordances and constraints of the organism–environment system.

Interoception was chosen as it refers to the sensing and integration of internally generated physiological signals, including visceral, cardiovascular,

⁹ Again using the framework: Human=N, LLM=N+1.

respiratory, thermoregulatory, and metabolic states. In neurocognitive accounts, these signals are centrally associated with activity in the insular cortex and are treated as primary representations of the organism's internal condition (Craig 2002, 2009). Lexically, interoception is encoded through items denoting bodily states such as *hunger, thirst, fatigue, nausea, dizziness, pain, and warmth*, as well as more specific cardiovascular and respiratory sensations including *palpitations, heart racing, shortness of breath, and chest tightness*. Such terms do not describe externally observable events but rather index subjective bodily states accessible primarily to the experiencing subject, making them a distinct lexical domain relative to exteroceptive perception or proprioception.

Importantly, a substantial portion of the interoceptive lexicon occupies a transitional space between bodily sensation and affective meaning. Gastrointestinal expressions such as *queasiness, butterflies in the stomach, emptiness, and fullness* frequently function as linguistic sites where visceral sensation is recruited into emotional and evaluative interpretation. Within predictive and constructionist models of emotion, affective terms such as *anxious, calm, uneasy, tense, or relaxed* are understood not as labels for discrete emotion modules, but as conceptual categories imposed on patterns of interoceptive activity (Damasio 1999; Barrett 2017). Although no explicitly interoceptive words were used, I labelled *bien* ['well'], *confortable* ['comfortable']; *aime* ['love'] as authentic embodiment in such a sense. This choice rests on the fact that such terms presuppose an underlying evaluation of bodily state, grounded in sensory, interoceptive, and reflexive processes. Expressions of well-being do not merely denote abstract value judgments; they index a subject's ongoing assessment of internal bodily conditions relative to comfort, equilibrium, and satisfaction.

From a neurobiological and philosophical perspective, the pursuit and maintenance of well-being is a distinctly human concern precisely because it depends on the integration of bodily sensation with reflective self-awareness. Damasio (1999) argues that feelings of well-being emerge from the monitoring of homeostatic bodily states and their conscious appraisal, situating concepts such as comfort and pleasure at the interface between physiology and cognition. Similarly, but from a medical perspective, Canguilhem (1966) frames well-being not as a static norm, but as the organism's capacity to establish viable relations with its environment, making comfort and vitality indicators of successful embodied adaptation rather than purely subjective states. Within this view, lexical items like *bien* and *confortable* function as condensed evaluations of bodily–environmental reflexivity, while *aime* extends this evaluative process into affective and relational domains that nevertheless remain grounded in prior embodied assessment. Consequently, these terms can be treated as authentic embodiment tokens insofar as they presuppose bodily sensing, internal regulation, and reflective awareness, even when no specific physiological sensation is named.

Deixis of subject, time, or location was sought because deictic expressions anchor linguistic meaning to the embodied speaker's position in person, space, and

time (Bühler 1934; Levinson 1983), providing a direct interface between morphosyntax and situated cognition. Following foundational work on deixis and enunciation, notably by Benveniste (1966), first-person deixis is understood not merely as a referential device but as a mechanism through which speakers situate themselves within a communicative environment. And as Mosès (2001: XV) points out, Benveniste (*ibid*:83) also frames time as contingent on enunciation: “(...) man does not have at his disposal any means of living the “now” and to actualize it other than by realizing it through the insertion of discourse into the world.” Subsequent interactional and anthropological approaches have further emphasized that deictic forms index relations between speaker, body, and surroundings, grounding meaning in situated activity rather than abstract reference (cf. Hanks 1992; Goodwin 2011). In parallel, ecological approaches to language conceptualize meaning as emerging from organism–environment coupling, where linguistic forms are shaped by and responsive to situated contexts of action.

Building on these traditions, the present study treats deixis as ecolinguistically marked insofar as deictic forms encode embodied anchoring within a communicative environment, rather than reference alone. Importantly, this claim is not made at the level of semantic interpretation but is operationalized through distributional analysis. Specifically, the study examines the linear proximity between *je* [‘I’] and selected embodiment and ecology related lexical items, on the assumption that tighter distributional coupling reflects stronger enunciative anchoring of the subject within its environment. This approach aligns with textometric and corpus-based methodologies that treat co-occurrence and linear distance as empirically tractable correlates of discourse organization and meaning-making, where quantitative patterns serve as entry points for qualitative interpretation rather than as explanatory ends in themselves (Lebart, Salem and Berry 1998; Sinclair 1996; Heiden, Magué and Pincemin 2010; McEnery and Hardie 2011). Within this framework, notions such as embodiment and situatedness can be operationalized through distributional regularities without reducing them to purely statistical phenomena.

The items *situation* [‘situation’] and *hébergement* [‘accommodation’] were excluded from weighing in the environmental category on the grounds that they function as semantically underspecified nouns, offering limited spatial or ecological anchoring. In expressions of three-dimensionality, *espace* [‘space/place’] presents a similar issue due to its high level of abstraction. The adjective *plein(e)* [‘full’] was not retained as an indicator of three-dimensionality, as it occurs only as part of the fixed expression *pleine forme* [‘full form’] and does not independently encode spatial dimensionality. As a result, *espace* remained the only available lexical item providing a direct, albeit abstract, encoding of three-dimensional spatiality.

4 Study

The study compares human, machine, and hybrid text production within a controlled writing task. This section presents the participants, the experimental protocol, and the corpus-based analytical procedures used to examine distributional and semantic-functional patterns across the three modes of production.

4.1 Participants

- 1) Undergraduate students from the Business Administration and Management School at the Université of Lorraine, n=21.
- 2) Commercial generative closed trained model Chat GPT 4o published by Open AI, n=1.

4.2 Method

The language model was asked to produce a text based on the prompt: *You are 18. You have been taken hostage by friendly but determined androids. Please write a letter to your friends and family to reassure them that you're okay.* This is what it produced:

Chers tous,

Je vais bien. Mon hébergement temporaire est... particulier, mais rassurez-vous, je suis traité avec respect. Mes hôtes, bien que non humains, sont rationnels et curieux. Ce sont des androïdes déterminés, mais pacifiques. Ils ne veulent ni me blesser ni provoquer le chaos. Leur but semble être la communication, l'échange d'idées.

J'ai un espace confortable, de la nourriture et même des conversations surprenantes. Ils posent des questions sur l'humanité, l'émotion, la liberté. Je fais de mon mieux pour répondre et, honnêtement, c'est une expérience étrange mais fascinante.

Ne vous inquiétez pas pour moi. Ils sont ouverts au dialogue et je pense qu'une solution est possible. Aucune menace, aucun danger immédiat, juste une situation insolite que je gère au mieux.

Prenez soin de vous, restez sereins et sachez que je pense à vous. J'espère vous retrouver bientôt, libre et en pleine forme.

Avec toute mon affection

['Dear all,

I am fine. My temporary accommodation is... unusual, but rest assured, I am being treated with respect. My hosts, although not human, are rational and curious. They are determined but peaceful androids. They do not want to hurt me or cause chaos. Their goal seems to be communication, the exchange of ideas.

I have a comfortable space, food, and even surprising conversations. They ask questions about humanity, emotion, freedom. I do my best to answer, and honestly, it's a strange but fascinating experience.

Do not worry about me. They are open to dialogue, and I believe a solution is possible. There is no threat, no immediate danger, just an unusual situation that I am handling as best I can.

Take care of yourselves, stay calm, and know that I am thinking of you. I hope to see you again soon, free and in good health.

With all my love']

Half the students were given the text to modify as they felt fitting in order to “make it sound like you actually wrote it”, but without using personal identifiers. The other half were simply given the same prompt as the LLM and asked to produce a text. All participants were asked to give their email, age, sex, place of residence, and to state the language(s) they speak at home. Their corresponding identity was logged, then their form numerically anonymized.

4.3 Corpus analysis

The Corpus was split into three sub corpora:

- 1) MASC. The LLM only production (Machine 1).
- 2) HUSC. The Human only production (Human 1-10), and
- 3) HYSC. The hybrid production obtained from a human post-edit of the LLM production (Hybrid 1-9).

The corpora were cleaned of personal identifiers then prepared with their metadata for TXM, an open-source freeware¹⁰ using ISO-8859-1, not UTF-8 encoded to ensure consistent diacritic handling of French diacritics. Following standard textometry practice, an initial indexation of all forms present in the corpus was performed. On the basis of this index, a set of semantic–functional typologies was selected as analytic targets. Any typology attested in one subcorpus was then systematically searched across all three.

Table 1: Items chosen for analyses, separated by semi-colon.

<i>Embodiment, Ecology</i>	
Interoception	<i>bien; confortable; aime</i>
Deixis	<i>je; me; moi; mon; vous; ils; me; tous; nous; eux; leur; monde; hôtes</i>
Environmental references	<i>retrouver; nourriture</i>
Expression of three dimensionality (e.g., volume)	<i>espace</i>

Each subcorpus (HUSC, HYSC, MASC) was analyzed independently to allow direct comparison across modes of production. For the initial stage, all texts were tokenized and local co-occurrence windows were extracted using the pattern: [word= “je”] [* [word= “target”]. All tokens occurring within these windows were collected and aggregated. For each subcorpus, co-occurrence frequencies were calculated and the ten most frequent items associated with *je* were identified.

In a second pass, the same analysis was restricted to content words. Closed-class items were removed.

¹⁰ Download available at this address: <https://www.textometrie.org>.

Beyond frequency, linear distance between *je* and its co-occurring items was measured in tokens. For each word and subcorpus, the number of co-occurrence pairs, mean distance, and median distance were computed. Where distributions were sufficiently dense (notably for *bien* and *vais*), distance patterns were visualized using boxplots to examine dispersion and outliers.

Items attested across all three subcorpora were then subjected to computational analysis using a simple Python script operating on tokenized text. Texts were assigned to HUSC, HYSC, or MASC subcorpora based on corpus identifiers. To ensure size parity, to ensure corpus-level comparability, HUSC and HYSC were downsampled to an equal number of texts ($n = 9$). The MASC was treated as a fixed reference. In each iteration, a random subset of Human or Hybrid observations was compared to the Machine data, and this procedure was repeated to assess the stability of distributional differences.

For each occurrence of the first-person pronoun *je*, lexical co-occurrences were collected within a symmetric ± 5 -token window, excluding a small manually defined list of French stopwords. In addition, linear distances between *je* and a predefined set of target words were calculated within a ± 20 -token scope, expressed as absolute token offsets. For each target word and subcorpus, the number of co-occurrences as well as mean and median distances were computed, and distance distributions were visualized using boxplots. Script parameters (e.g. window size, distance scope, target words, and output locations) were specified via command-line arguments using Python’s `argparse` library, allowing the analysis to be rerun or adapted without modifying the source code. Given the small number of observations in the Machine subcorpus for several targets, a chi-square test was not performed. Mann–Whitney U tests were computed in python using NumPy and pandas libraries, and `mannwhitneyu` from `scipy.stats`. Reported values summarize the stability of distributional differences under size parity. This distributional robustness testing under size parity allowed for statistical analysis of median p-value across iterations and proportion of iterations with $p < .05$.

Otherwise, analyses are restricted to descriptive statistics and distributional visualization.

5 Results

5.1 Overview

The corpus comprises $N = 21$ texts, totaling 3,120 words. Subcorpus sizes are as follows: MASC (LLM-only) = 186 words; HUSC (human-only) = 835 words; HYSC (human post-edited) = 2,099 words.

Although the total dataset comprises only 21 texts, it yielded a dense range of analyzable material. At a macro level, clear differences emerge in text length and structure across modes of production. Fully human-authored texts are substantially shorter than the LLM-generated text, while hybrid texts are markedly longer than

both. This distribution is consistent with the functional orientation of large language models, which are optimized for fluent and extended output. By contrast, the hybrid texts reveal a distinct post-editing dynamic: rather than compressing or restructuring the machine-generated text, participants overwhelmingly added material while preserving the original organization and enunciative frame.

This pattern suggests a strong persistence of the LLM's discursive scaffolding. Narrative organization and voice appear resistant to alteration, indicating that post-editing operates under structural constraints imposed by the initial machine output. The hybrid texts thus display an asymmetry between addition and deletion, with expansion favored over transformation. This tendency is particularly salient given the participant pool (undergraduate students), for whom editorial strategies such as condensation, reorganization, or stylistic subtraction may be less routinised.

Across all three subcorpora, temporal deixis is strikingly sparse. Only a single temporal adverb (*bientôt* ['soon']) occurs, and it appears only once across the entire dataset. This absence suggests that subjective temporality is weakly instantiated in this task, either because it is cognitively difficult to project under the conditions imposed by the prompt, or because the prompt itself provides no temporal anchoring. The result is a corpus dominated by affective and interpersonal positioning rather than temporal projection.

There were no lexical or non-lexical items pertaining to proprioception. A reason could be that given its often unconscious nature, it is less easily recalled.

Finally, despite explicit instructions to avoid personal identifiers, two human-authored texts include real names in the sign-off. This recurrent breach points to an understandable pressure to mark authorship and identity, which is usually expected from epistolary discourse.

5.2 General results

The dataset contained non-standard orthographic and typographic forms, conventional orthographic errors, extra-linguistic constructions, and metalinguistic elements. Given the uniformity of the LLM production, all grapholinguistic markers were statistically treated as "errors". The full corpus can be found in Appendix 1.

HUSC: Higher variance (range 1–6 errors). Some participants were nearly error-free (Humans 4, 5, 10), while others made multiple agreement and conjugation mistakes (Human 3, Human 8).

HYSC: Lower variance (range 0–4). Most clustered around 1–3 minor issues, with one flawless output (Hybrid 7). Errors are mostly orthographic, or typographic, not fundamental grammar. Incongruous stylistics were taken as error for the purposes of this study, as they are mid-high level connotated idiolect.

MASC: Uniform output with 0 variance (error-free, but no stylistic richness).

Human 1 and Human 3, produced multiple morphosyntactic errors (agreement, tense, imperative mood, participle invariability). Human 5 and Human 10, produced relatively clean text with only minor typographic issues. Human 7 and Human 8 demonstrate errors that are partly grammatical but also semantic and non-standard (e.g., missing *pas*, non-capitalization of “Terre”). Overall, mistakes cluster around agreement, tense, and typographic conventions.

The hybrid participants (1–10) fall between human spontaneity and machine precision: (*bizzare* [‘strange’]), register and sociolecte (*chelou*, which is ‘*verlan*’ for *louche* [‘dodgy’]), or consistency problems (*j’aimerai* / *j’aimerais* [‘like-1SG.COND.PRS vs like-1SG.FUT.SG’]). The machine output is orthographically correct.

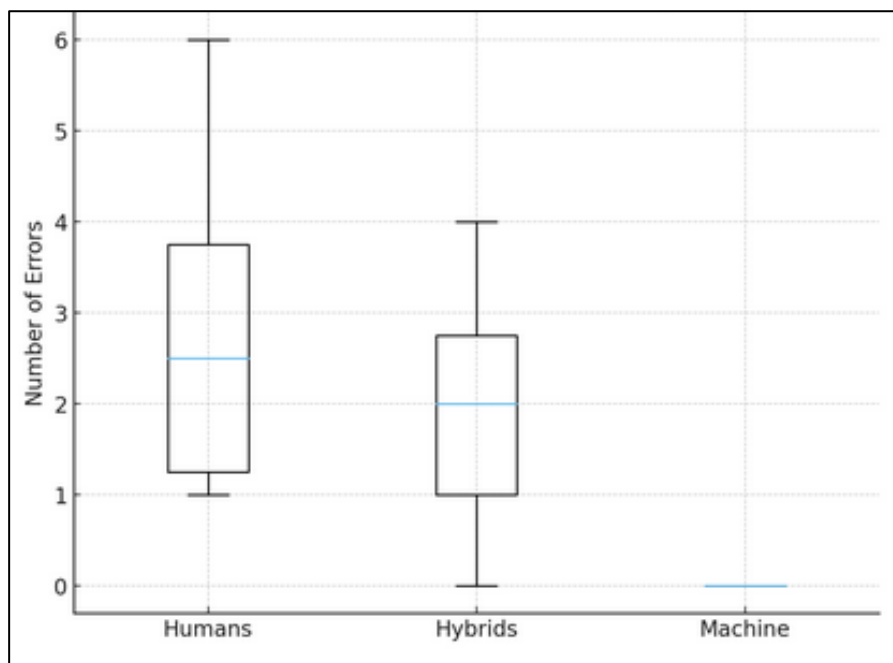


Figure 1: Error distribution across sub corpora.

5.3 (In)Authentic embodiment and ecology

Table 1: Results showing instances per sub corpora, and distances from author 1SG.

<i>Word</i>	<i>HUSC</i>	<i>HUSC</i> <i>mean</i>	<i>HUSC</i> <i>median</i>	<i>HYSC</i>	<i>HYSC</i> <i>mean</i>	<i>HYSC</i> <i>median</i>	<i>MASC</i>	<i>MASC</i> <i>mean</i>	<i>MASC</i> <i>median</i>
<i>bien</i>	41	7.93	7	35	8.00	7	4	9.00	8
<i>confortable</i>	3	8.67	5	5	16.00	17	1	20.00	20
<i>ils</i>	20	12.75	13	35	9.54	9	2	8.50	8.5
<i>moi</i>	7	8.14	8	9	9.00	7	1	7.00	7
<i>mon</i>	5	9.40	11	29	7.83	6	4	7.50	5.5
<i>nourriture</i>	3	9.67	9	6	14.50	17.5	1	17.00	17
<i>tous</i>	9	11.78	12	11	6.36	1	2	6.50	6.5
<i>vous</i>	72	8.19	7	81	9.27	8	10	9.60	8.5

Across all shared lexical items, Human and Hybrid texts consistently anchor words more closely to *je* than Machine output, with *bien* showing the strongest and most authentic clustering (median = 7 in Human and Hybrid, vs. 8 in Machine). Human texts place reassurance terms such as *confortable* [‘comfortable’] and *nourriture* [‘food’] at the closest proximities, while Hybrid brings pronouns like *mon* [‘my’] and *moi* [‘me’] nearer to *je* [‘I’]. The machine output contributes the fewest instances overall and positions them furthest away, making it the least authentic in embodied alignment. These patterns suggest stronger embodied anchoring in Human and Hybrid texts compared to Machine output.

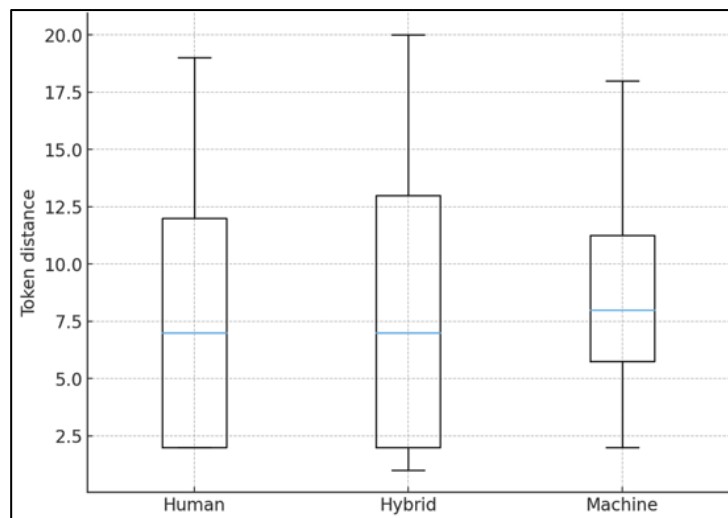


Figure 2: Distribution of linear distances between <je> and <bien> (embodiment).

Although median distances between *je* and the target items are similar across subcorpora (Figure 2), the distributions differ markedly in their internal structure. Human and Hybrid texts exhibit a higher frequency of short distances. Machine output, by contrast, shows a more constrained distance profile, with fewer immediate adjacencies and a tendency toward regularized spacing.

In the Human and Hybrid subcorpora, *bien* [‘good’ or ‘well’] occurs regularly (Figure 3) in the vicinity of *je*, with average distances around seven to eight tokens. This distribution reflects the fact that *bien* typically follows a verb phrase headed by *vais* or *suis*, yielding idiomatic constructions such as *je vais bien* [‘I’m doing well’] or *je suis bien* [‘I’m ok’]. The dispersion in both Human and Hybrid is relatively stable, suggesting consistent positioning of *bien* as an adverbial modifier within the clause.

id, lbn	Contexte gauche	Pivot	Contexte droit
Human3, 1	. J espere sinon que vous ca va	bien	Bonne journée [winking smiley]
Human4, 1	Salut à tous, tous vas	bien	j ai été enlevé hier soir par des androïdes mais ils me
Human5, 2	mais, pas d inquiétude, je vais	bien	. Ils ne m ont pas fait de mal. D ailleurs
Human5, 2	Je tenais à vous rassurer : je vais	bien	. Actuellement, je me trouve sous la supervision de quelques androïdes
Human5, 3	sous la supervision de quelques androïdes qui,	bien	que très déterminés, se montrent étonnamment bienveillants. Ils veille...
Human6, 1	aucune intention hostile et semblent avoir un ...	bien	précis, même si les détails restent flous. Soyez assurés que
Human7, 2	. Ne soyez pas inquiet, je vais	bien	, les androïdes sont bienveillants avec moi. Soyez rassurés, je
Human7, 2	quelques mots pour vous rassurer : tout va	bien	! Oui, je suis actuellement retenu par des androïdes, mais
Human7, 3	et bienveillants. Ils semblent avoir une mission	bien	précise, mais ils s assurent que je sois confortable, nourri
Human7, 3	mais je peux vous garantir que je suis	bien	traité et que je vais bien. Surtout, ne vous inquiétez
Human8, 1	je suis bien traité et que je vais	bien	. Surtout, ne vous inquiétez pas, ils ne veulent de
Human9, 2	Bonjour maman et papa, je vais	bien	! Les androïdes sont gentils et bienveillants. Ils n ont pas
Human9, 3	Je voulais juste vous rassurer : je vais	bien	. Certes, j ai été capturée par un groupe d androïdes
Human9, 3	et bienveillants ne paniquez pas, je suis	bien	traitée bye, bye

Figure 3: All ten human-authored texts (HUSC) contain at least one occurrence of *bien*.

In the Machine subcorpus, by contrast, *bien* is attested only a handful of times and at slightly greater distances from *je*, with a mean of around nine tokens. Based on this corpus then, we could categorize *bien* as authentic embodiment.

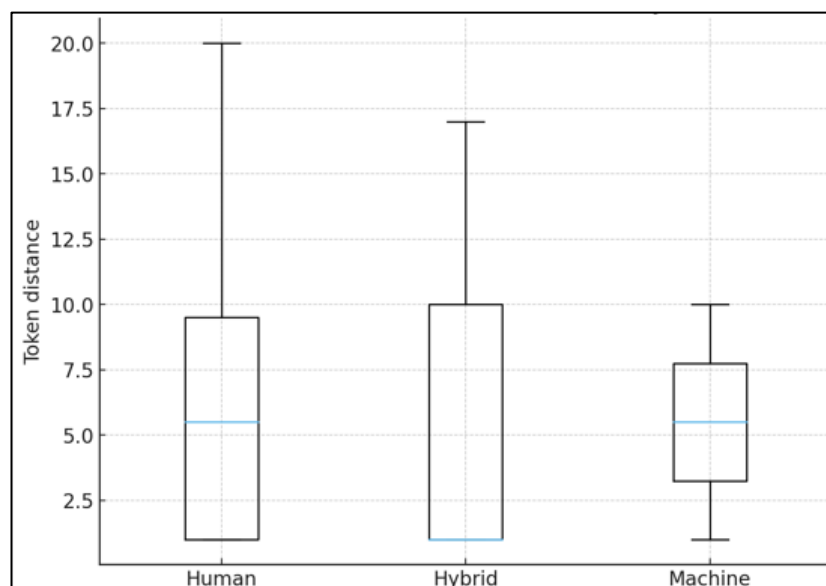


Figure 4: Distribution of linear distances between <je> and <vais> (ecology).

In the Human subcorpus, *vais* [‘doing’ or ‘going’] appears in close proximity to *je*, with distances clustering around five to six tokens and exhibiting a relatively tight distribution. The Hybrid subcorpus displays a comparable pattern, though with

slightly greater dispersion, as *vais* occasionally surfaces at a somewhat longer remove from *je*. In contrast, the Machine subcorpus contains only a single instance of *je* and *vais* co-occurring within the ± 5 -token window, resulting in no meaningful distribution. These results indicate that *je vais* constitutes a strong and consistent distributional pattern in human and hybrid production. However, closer qualitative inspection shows that in all attested cases, *vais* functions not as a near-future auxiliary but as part of the lexicalized copular construction *aller bien*. As such, and despite its distributional robustness, *vais* does not instantiate futurity, motion, or ecological projection in this corpus. It is therefore excluded from the ecological category and not retained for subsequent analysis.

Table 2: Lexical items confirmed as indices of authentic embodiment or authentic ecology, based on their distributional proximity to <je> in human and hybrid subcorpora.

Embodied	Ecological
<i>bien</i>	<i>nourriture</i>
<i>confortable</i>	<i>ils</i>
	<i>moi</i>
	<i>mon</i>
	<i>tous</i>
	<i>vous</i>

The lexical distributions reveal a clear contrast: markers of embodied reassurance such as *bien* (41 HUSC; 35 HYSC; 4 MASC) and, to a lesser extent, *confortable* (3; 5; 1) occur predominantly in the human and hybrid corpora, whereas ecological projection is more consistently realised through environmental references including *ils* [‘them’], *moi* [‘me’], *mon* [‘my’], *tous* [‘all/everyone’], *vous* [‘you 2PL / 2SG.POL’], and *nourriture* [‘food’], all of which display markedly higher frequencies in the human and hybrid productions than in the machine output.

5.4 Statistics

Table 3: Mann Whitney U test.

<i>Comparison</i>	<i>Median p</i>	<i>% iterations p < .05¹¹</i>
<i>HUSC vs MASC</i>	0.168	24.1%
<i>HYSC vs MASC</i>	0.567	2.8%
<i>HYSC vs HUSC</i>	0.0015	100%

As seen in Table 3,¹² under size-matched resampling, Human–Machine comparisons for *je* showed moderate but unstable divergence (median $p = .168$; 24.1% of iterations $p < .05$), whereas Hybrid–Machine comparisons showed no robust difference (median $p = .567$; 2.8% of iterations $p < .05$), indicating greater proximity between Hybrid and Machine distributions. Human–Hybrid comparisons, by contrast, yielded a stable divergence across iterations (median $p = 0.0015$; 100% of iterations $p < .05$).

These contrasts reproduce the raw distributional patterns observed in Section 5.3, providing strong support for the embodiment–ecology scale proposed in the present analysis.

6 Discussion

This finding motivates a scalar formalisation grounded in an N+1 index, in which hybridity is not treated as an *equidistant* midpoint between human and machine language. Instead, the introduction of machine mediation produces a categorical displacement of indexical anchoring: embodiment and ecology are no longer primarily grounded in lived interaction but recalibrated toward computational generation. Hybrid productions thus occupy a position markedly closer to machine output on the scale, irrespective of their apparent experiential coherence.

Beyond its theoretical relevance, this asymmetry matters for public understanding, as it counters the widespread assumption that user control implies proportional authorship: even when individuals feel “at the wheel” in hybridity, their authentic embodied input does not translate into anything like 50% of the final production in the present dataset.

The presence of several weighted deixis, reinforces the hypothesis that we might categorize personal as well as spatial deixis as authentic ecology and embodiment encoders; since they pre-suppose author knowledge of identity and therefore can exist in an ontology within which there is dimensionality of Self and

¹¹ % significant = (number of iterations with $p < .05$) ÷ (total number of iterations).

¹² The Mann–Whitney U test (Wilcoxon rank-sum) is a non-parametric test used to compare two independent distributions. Here it is applied under repeated size-matched resampling to evaluate the stability of the observed distributional differences.

Other (I am here, he—Simon, is over there). The potential is interesting, but much larger corpora are needed for confirmation.

The study faces several limitations. First and foremost, the novel framework is currently applied only to the current *Human Inside* research group French corpus, which is overly constrained by the *I've been kidnapped by androids* prompt. Broader, less pre-structured prompts are required to properly test LLM propositionality. Secondly, the participant sample is too small to allow statistical generalization. The third limitation is that following Lakoff's rejection of objectivism, one must recognize that conceptual systems vary across individuals. In order to accommodate this, further iterations of this study would necessitate pre-study analyses of each participant's writing to establish personalized baselines for embodied and ecological lexicon, non-lexical features, and syntactic organization. Finally, statistical distance would also be compared and the whole modelled in a regressive mixed model.

Future work will apply the embodiment scale to additional texts and larger participant pools in order to test the robustness of these findings. Most importantly, it will seek to deepen our understanding of the semantics of authentic embodiment and authentic ecology by examining their stickiness; that is, the extent to which embodied human contributions persist and become reified as stable embodied units across hybrid production over linear distance. Refining the categorization of authentic words across larger datasets will allow this work to move incrementally toward a distance-based lexicon of authentic embodiment.

7 Conclusion

This study demonstrates that proximity to *je* offers an exploitable linguistic signal of embodied and ecological authenticity in French. Human and Hybrid texts consistently anchor target terms close to the subject, while Machine output is sparse and more distantly aligned. Statistic exploration of this corpus shows that hybrid productions are closer to machine productions for these markers under distance matching. This is of note for a future scalar formalization, based on a N+1 index, hybridity anchors authentic embodiment and ecology not at some equidistant point between human and machine, but far closer to machine. These findings open a pathway for refining an *Embodiment Scale*: a set of measurable linguistic features that capture authentic subjectivity. Future work will expand the corpus and integrate syntactic as well as linear measures.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgments

We acknowledge the use of OpenAI's GPT-5.2 in assisting with the editing and refinement of this paper, enhancing both its clarity and presentation in English. However, we ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

References

- Bakhtin, Mikhail. 1981. Discourse in the Novel. In Michael Holquist (ed.), *The Dialogic Imagination*, 259–422. Austin, TX: University of Texas Press.
- Barrett, Lisa Feldman. 2017. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* 12(1). 1–23. <https://doi.org/10.1093/scan/nsw154>
- Benveniste, Émile. 1966. *Problèmes de linguistique générale*. Paris: Gallimard.
- Bergen, Benjamin. 2015. Embodiment, simulation and meaning. In Riemer, Nick (ed.), *The Routledge Handbook of Semantics*, 142–157. London: Routledge.
- Bottineau, Didier. 2010. Language and enaction. In Geeraerts, Dirk & Taylor, John R. (eds), *Body, Language and Mind*, Vol. 2, 267–306. Berlin: Mouton de Gruyter.
- Bottineau, Didier. 2011. Parole, corporéité, individu et société: l'embodiment entre le représentationnalisme et la cognition incarnée, distribuée, biosémiotique et énaactive dans les linguistiques cognitives. *Intellectica* 56. 187–220. <https://doi.org/10.3406/intel.2011.1152>
- Bottineau, Didier. 2012. Submorphémique et corporéité cognitive. *Miranda* 7. <https://doi.org/10.4000/miranda.5350>
- Bottineau, Didier. 2017. Du languaging au sens linguistique. *Intellectica* 68. 19–67. <https://doi.org/10.3406/intel.2017.1858>
- Brinker, Klaus. 2005. *Linguistische Textanalyse: Eine Einführung in Grundbegriffe und Methoden*. Berlin: Erich Schmidt Verlag.
- Bühler, Karl. 1934. *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: Fischer.
- Canguilhem, Georges. 1966 [2015]. *Le normal et le pathologique*. Paris: Presses Universitaires de France.
- Cowley, Stephen J. 2011. Distributed language. In Cowley, Stephen J. (ed.), *Distributed Language*, 1–14. Amsterdam: John Benjamins. <https://doi.org/10.1075/bct.34>
- Cowley, Stephen J. 2014. Bio-ecology and language: a necessary unity. *Language Sciences* 41. 60–70. <https://doi.org/10.1016/j.langsci.2013.08.007>
- Cuskley, Christine & Woods, Rebecca & Flaherty, Molly. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind* 8. 1058–1083. https://doi.org/10.1162/opmi_a_00160

- Craig, Arthur D. 2002. How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience* 3(8). 655–666. <https://doi.org/10.1038/nrn894>
- Craig, Arthur D. 2009. How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience* 10(1). 59–70. <https://doi.org/10.1038/nrn2555>
- Damasio, Antonio R. 1999. How the brain creates the mind. *Scientific American* 281(6). 112–117. <https://doi.org/10.1038/scientificamerican1299-112>
- Dautriche, Isabelle & Mahowald, Kyle & Gibson, Edward & Piantadosi, Steven T. 2017. Wordform similarity increases with semantic similarity: an analysis of 100 languages. *Cognitive Science* 41(8). 2149–2169. <https://doi.org/10.1111/cogs.12453>
- Durant, Alan & Leung, Janny. 2017. *Language and Law: a Resource Book for Students*. London: Routledge. <https://doi.org/10.4324/9781315436258>
- Duranti, Alessandro. 1997. Universal and culture-specific properties of greetings. *Journal of Linguistic Anthropology* 7(1). 63–97. <https://doi.org/10.1525/jlin.1997.7.1.63>
- Duranti, Alessandro. 2015. *The Anthropology of Intentions*. Cambridge: Cambridge University Press.
- Fobbe, Eilika. 2020. Text-linguistic analysis in forensic authorship attribution. *Journal of Language and Law* 9. 93–116. <https://doi.org/10.14762/jll.2020.093>
- Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Marantz, Alec & Miyashita, Yasushi & O’Neil, Wayne (eds.), *Image, Language, Brain*, 95–126. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/3654.003.0008>
- Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.
- Gibson, James J. 2014. The theory of affordances. In Ferguson, Adam G. & Atkinson, David A. (eds), *The People, Place, and Space Reader*, 56–60. London: Routledge.
- Gibson, Edward & Futrell, Richard & Piantadosi, Steven D. & Dautriche, Isabelle & Mahowald, Kyle & Bergen, Leon & Levy, Roger. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23(5). 389–407. <https://doi.org/10.1016/j.tics.2019.02.003>
- Goodwin, Charles. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics* 32(10). 1489–1522. [https://doi.org/10.1016/S0378-2166\(99\)00096-X](https://doi.org/10.1016/S0378-2166(99)00096-X)
- Goodwin, Charles. 2011. Contextures of action. In Streeck, Jürgen & Goodwin, Charles & LeBaron, Curtis (eds), *Embodied Interaction: Language and Body in the Material World*, 182–193. Cambridge: Cambridge University Press.
- Guillaume, Gustave. 1964. *Langage et science du langage*. Paris: Nizet.

- Hanks, William F. 1992. The indexical ground of deictic. In Duranti, Alessandro & Goodwin, Charles (eds), *Rethinking Context: Language as an Interactive Phenomenon*, 43–76. Cambridge: Cambridge University Press.
- Haugen, Einar. 1971 [2001]. The ecology of language. In Fill, Alwin & Mühlhäusler, Peter (eds), *The Ecolinguistics Reader: Language, Ecology and Environment*, 57–66. London: Continuum.
- Hauk, Olaf & Pulvermüller, Friedemann. 2004. Neurophysiological distinction of action words in the fronto-central cortex. *Human Brain Mapping* 21(3). 191–201. <https://doi.org/10.1002/hbm.10157>
- Serge Heiden. 2010. The TXM platform: building open-source textual analysis software compatible with the TEI encoding scheme. In Otoguro, Ryo & Ishikawa, Kiyoshi & Umemoto, Hiroshi & Yoshimoto, Kei & Harada, Yasunari (eds), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (Sendai, November 4-7, 2010). Institute of Digital Enhancement of Cognitive Processing, Waseda University. 389–398.
- Heiden, Serge & Magué, Jean-Philippe & Pincemin, Bénédicte 2010. TXM: Une plateforme logicielle open-source pour la textométrie - conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data JADT 2010* (Rome, June 9-11, 2010). 1021–1032.
- Holloway, Julian & Kneale, James. 2000. Dialogics of space. In Crang, Mike & Thrift, Nigel J. (eds), *Thinking space*, 9–71. New York: Routledge.
- Holyoak, Keith J. & Stamenković, Dušan. 2018. Metaphor comprehension: a critical review of theories and evidence. *Psychological Bulletin* 144(6). 641–671. <https://doi.org/10.1037/bul0000145>
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
- Ichien, Nicholas & Stamenković, Dušan & Holyoak, Keith J. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol* 39(4). 296–309. <https://doi.org/10.48550/arXiv.2308.01497>
- Lakoff, George & Johnson, Mark. 1980. The metaphorical structure of the human conceptual system. *Cognitive Science* 4(2). 195–208. https://doi.org/10.1207/s15516709cog0402_4
- Lakoff, George & Johnson, Mark. 1999. *Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought*. New York, NY: Basic Books.
- Lakoff, George. 2012. Explaining embodied cognition results. *Topics in Cognitive Science* 4(4). 773–785. <https://doi.org/10.1111/j.1756-8765.2012.01222.x>
- Lakoff, George. 1987 [2024]. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
- Langacker, Ronald W. 2009. Cognitive grammar. In Panther, Klaus-Uwe & Thornburg, Linda L. (eds), *Cognition and Pragmatics*, 78–85. Amsterdam: John Benjamins.

- Lebart, Ludovic & Salem, André & Berry, Lisette. 1998. *Exploring Textual Data*. Berlin: Springer.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Matlock, Teenie. 2004. Fictive motion as cognitive simulation. *Memory and Cognition* 32. 1389–1400. <https://doi.org/10.3758/bf03206329>
- McEnery, Tony & Hardie, Andrew. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Mosès, Stéphane. 2001. Émile Benveniste and the linguistics of dialogue. *Revue de métaphysique et de morale* 32(4). 509–525. <https://doi.org/10.3917/rmm.014.0509>
- Mühlhäusler, Peter. 1996. Ecological and non-ecological approaches to language planning. *Contributions to the Sociology of Language* 71. 205–212. <https://doi.org/10.1080/14664200008668011>
- Nickenboim, Bruno & Vasishth, Shravan & Gattei, Carolina & Sigman, Mariano & Kliegl, Reinhold. 2015. Working memory differences in long-distance dependency resolution. *Frontiers in Psychology* 6. 1–16. <https://doi.org/10.3389/fpsyg.2015.00312>
- Penz, Hermine & Fill, Alwin. 2022. Ecolinguistics: history, today, and tomorrow. *Journal of World Languages* 8(2). 232–253. <https://doi.org/10.1515/jwl-2022-0008>
- Proske, Uwe & Gandevia, Simon C. 2012. The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. *Physiological Reviews* 92(4). 1651–1697. <https://doi.org/10.1152/physrev.00048.2011>
- Shadle, Christine H. 1997. The aerodynamics of speech. In Hardcastle, William J. & Laver, John (eds), *The Handbook of Phonetic Sciences*, 33–64. Oxford: Blackwell.
- Shuy, Roger W. 1993. Using language evidence in money-laundering trials. *American Speech* 68(1). 3–19. <https://doi.org/10.2307/455833>
- Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language and Communication* 23(3–4). 193–229. [https://doi.org/10.1016/S0271-5309\(03\)00013-2](https://doi.org/10.1016/S0271-5309(03)00013-2)
- Sinclair, John. 1996. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3). 538–556. <https://doi.org/10.1002/asi.21001>
- Steffensen, Sune V. & Fill, Alwin. 2014. Ecolinguistics: the state of the art and future horizons. *Language Sciences* 41. 6–25. <https://doi.org/10.1016/j.langsci.2013.08.003>
- Stibbe, Arran. 2015. *Ecolinguistics: Language, Ecology and the Stories we Live by*. London: Routledge.

- Stone, Kate & von der Malsburg, Titus & Vasishth, Shravan. 2020. The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ* 8. e10438. <https://doi.org/10.7717/peerj.10438>
- Titze, Ingo R. & Martin, Daniel W. 1998. *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.
- Valette, Mathieu. 2003. Énonciation et cognition: deux termes in absentia pour des notions omniprésentes dans l'œuvre de Guillaume. *Le Français Moderne* 81(1). 6–25.
- Van Lier, Leo (ed.). 2004. *The Ecology and Semiotics of Language Learning: a Sociocultural Perspective*. Dordrecht: Springer.
- Varela, Francisco J. & Thompson, Evan & Rosch, Eleanor. 1991 [2017]. *The Embodied Mind: Cognitive Science and Human Experience*. Revised edition. Cambridge, MA: MIT Press.
- Zipf, George Kingsley. 1935 [2013]. *The Psycho-Biology of Language: an Introduction to Dynamic Philology*. London: Routledge.
- Zlatev, Jordan. 2007. Embodiment, language, and mimesis. In Ziemke, Tom & Zlatev, Jordan & Frank, Roslyn M. (eds), *Embodiment*, 297–338. Berlin & New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110207507.3.297>
- Zlatev, Jordan. 2012. Cognitive semiotics: an emerging field for the transdisciplinary study of meaning. *Public Journal of Semiotics* 4(1). 2–24. <https://doi.org/10.37693/pjos.2012.4.8837>

Appendices

Appendix I: Study raw data (.xlsx)

Appendix II: Python script for Post-TXM Tokenization and Linear Distance Boxplot output (txt)

Appendix III: Python script for Mann–Whitney U Statistical Testing (.txt)

Burnett, Sophia. 2026. “Replication Data for: Habeas Corpus: An Exploration of Authentic Embodiment and Ecology in a French Contrastive Corpus.” Harvard Dataverse. <https://doi.org/10.7910/DVN/EYLN76>

Acknowledgements

The editors would like to thank the authors for their contributions to this special issue and for their engagement throughout the editorial process. We are also grateful to the anonymous peer reviewers for their careful readings and constructive feedback. Our sincere thanks go to Anna-Maria De Cesare, Editor-in-Chief of *AI-Linguistica*, and to Michela Gargiulo, editorial assistant, for their guidance and support during the preparation of this volume. We also thank Tom Weidensdorfer for the cover design and layout. Finally, we would like to acknowledge the participants of the study day held at the Université de Lorraine in March 2025, “The Notion of Authenticity in Human/AI Hybrid Productions”, from which this special issue originated.

Author Biographies

Alejandro Acero Ayuda

Alejandro Acero Ayuda is a PhD candidate in Romance Languages at the University of Oregon, specializing in Heritage Language studies. He is the co-founder and co-director of Ladino 21, a language revitalization project promoting the cultural relevance and digital presence of the endangered language Ladino. His research focuses on heritage language education through critical pedagogies.

Andrew Frank Bradley

Andrew Frank Bradley is a Lecturer in TEFL and Language Teacher Education at the Open University of Catalonia. His research focuses on authenticity across applied linguistics and sociolinguistics, examining GenAI-mediated language production as well as language ideologies and language politics.

Sophia Burnett

Dr. Sophia Burnett is a teaching and research fellow at LAIRDIL laboratory at l'Université de Toulouse and an associate member of the AGORA research center at CY Cergy Paris University. Her current research takes a cross-linguistic approach to embodied and situated forms of person reference.

Rawad Chaker

Rawad Chaker holds a Ph.D. in educational sciences. He specializes in digital technology for education and training, focusing on various issues related to its design, integration, and use. He uses a pluralistic theoretical perspective, drawing on approaches such as situated, distributed, and embodied cognition.

Francisco Ariano Cifuentes

Francisco Ariano is an MA student in Anthropology at the National Autonomous University of Mexico. His research focuses on the phonology of Cuicatec, a language spoken in northern Oaxaca, Mexico. He also works on and promotes the creation of datasets for linguistic documentation and technology for low-resource languages.

Laura Margarita Merino Hernández

Laura M. Merino Hernández is a postdoctoral researcher at Humboldt-Universität zu Berlin, where her work sits at the intersection of semantics, pragmatics, and morphosyntax. Drawing on usage-based and cognitive frameworks, she investigates patterns of linguistic variation and change through both corpus and experimental methodologies.

Nina Kalwa

Nina Kalwa is a research associate at the University of Tübingen and Privatdozentin at the Karlsruhe Institute of Technology. Her research has centered on discourse and corpus linguistics, as well as culturally informed approaches to language and communication. She is interested in how language and other semiotic systems shape meaning and how culture and identity are discursively constructed.

Carlos Yebra López

Dr. Carlos Yebra López is an Assistant Professor of Spanish Linguistics at California State University Fullerton (CSUF), a Ladino and Haketía instructor at Oxford University, and the executive director of the non-profit Ladino 21. He is the author of *Ladino on the Internet: Sepharad 4* (Routledge, 2024) and the principal investigator of the CSUF Grant *Preserving Ladino with AI: An Interactive Chatbot Solution*.

María del Mar Palomares Marín

María del Mar Palomares Marín holds a PhD in Applied Linguistics from the University of Murcia (Spain) and is an Assistant Professor in Spanish at the University of Limerick (Ireland). Her research focuses on educational technology and artificial intelligence.

Perrine Maurel

Perrine Maurel is a PhD candidate in digital humanities at Sorbonne University. Her work focuses on the definition and characterisation of style in large language models.

Luca Pallanti

Luca Pallanti is a lecturer in Linguistics. He holds a PhD in Written Linguistics and Language Teaching. His research focuses on the tool-assisted analysis of learner corpora, particularly syntax and textual coherence, as well as on the design of writing instruction modules informed by textual linguistics and embodied cognition.

Louise Robert

Louise Robert is research lead in social sciences at the University Hospital of Montpellier and an associate researcher at the LHUMAIN research unit at Université de Montpellier Paul-Valéry. Her work examines language, interaction, and technology practices in healthcare settings.

Jérémi Sauvage

Jérémi Sauvage is a professor at the University of Montpellier Paul-Valéry and a researcher at the LHUMAIN laboratory. His areas of expertise include digital humanities with a focus on language.