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 élusive 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'autrice 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 élusive. 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 antécédentes (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 significatuves. 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’autrice 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’infirmer 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 none* 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 & Tonellotto, 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

Ramnial, 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