AI versus Human Texts: A Cross-Register Comparison of Models and Strategy Conditions

Marianna Gracheva (University of Turku, Finland/ Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)
Bjoern Nieth (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)
Emmanuelle Salin (Ludwig-Maximilians-Universität München, Germany)
Bjoern Eskofier (Ludwig-Maximilians-Universität München, Germany/ Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)
Michaela Mahlberg (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany/ University of Birmingham, United Kingdom)

submitted 27/05/2026, accepted 15/10/2026, published 26/11/2026
https://doi.org/10.62408/ai-ling.v5i1.90

marianna.grachova(at)utu.fi
bjoern.nieth(at)fau.de
emmanuelle.salin(at)med.uni-muenchen.de
eskofier(at)lmu.de
michaela.mahlberg(at)fau.de

Abstract

Register competence – ability to produce situationally appropriate language – has been proposed among key areas of improvement in large language model (LLM) development. In particular, such calls come from studies of LLM register alignment (Berber Sardinha 2024; Goulart et al. 2024; Demir and Egbert 2026) that observed a lack of register sensitivity in ChatGPT, urging caution in the use of AI for text generation. In this paper, we extend the range of AI tools under analysis and compare four models – Apertus, Gemma, Llama, and Qwen – in two strategy conditions: zero-shot, where the models did not receive examples but were asked to produce texts of a given register, and few-shot, where the models received examples from each register in addition to the prompt. We examine models' performance relative to human texts of three registers – conversation, Reddit creative writing, and BBC news articles – on linguistic dimensions of variation in speech and writing (Biber 1988). We show that a) on different dimensions, models succeed to varying degrees in distinguishing between registers; b) human–AI gap varies by register and depends on the model-strategy combination; c) the few-shot strategy generally did not result in improved performance. We conclude with implications for model training and development.

Keywords

artificial intelligence, large language models, humanlikeness, register, register alignment, multidimensional analysis

1 Introduction

The language capabilities of artificial intelligence (AI) have now been examined from a variety of perspectives and across fields. Linguists have naturally concerned themselves with questions of humanlikeness – what are the characteristics of AI-

AILing

Gracheva, Marianna et al. 2026. AI versus Human Texts: A Cross-Register Comparison of Models and Strategy Conditions. AI-Linguistica, Vol.5 No.1
DOI: 10.62408/ai-ling.v5i1.90

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070
© 2026 Marianna Gracheva, Bjoern Nieth, Emmanuelle Salin, Bjoern Eskofier and Michaela Mahlberg, licensed under CC BY-NC-SA 4.0.

produced language and how close does it come to authentic human production? With the ongoing large language model (LLM) advancements, AI's language abilities have been improving, often making AI texts hard to reliably distinguish from human-authored ones (e.g., Casal and Kessler 2023; Clark et al. 2021; Hassoulas et al. 2023; Gao et al. 2023; Ma et al. 2023). However, the fact that AI may demonstrate human-like linguistic abilities does not mean that its language resembles real language in use (Lin 2023), and although AI may excel at grammatical correctness and be generally quite successful in register/genre classification tasks (e.g., Kuzman, Mozetič and Ljubešić 2023; Brommer et al. 2024), its language production may in fact not be characteristic for real-world contexts, or in other words, may not be register-appropriate (Crosthwaite and Baisa 2023). Indeed, one of the most urgent questions raised regarding AI-generated language has been that of register alignment – AI's ability to produce language appropriate for particular situations of use. Register – or situation of language use – has been shown to be a major predictor of language variation on all linguistic levels (Biber 2012), and general accounts of language without attention to register (i.e., differences among, e.g., conversations, news, and academic texts) are considered "essentially meaningless" (Biber and Seoane 2021: 3), as linguistic differences between registers have been shown to be major across languages. Our understanding of the differences and similarities between AI and human texts then cannot be complete unless register variation is accounted for.

In this study, we take a register perspective in order to compare the linguistic performance of several open-source LLMs in terms of their register alignment, that is, the extent to which they recognize and reproduce human register-specific language. Our models are represented by two strategy conditions – zero-shot and few-shot: in the few-shot condition, the models received examples of a given register, and in the zero-shot condition, the models did not draw on examples but relied solely on prompting. We consider two aspects of register alignment, examining the models across as well as within registers. The cross-register perspective is concerned with examining the extent to which the models are able to recognize and replicate human cross-register trends; the within-register perspective consists in the extent to which the models approximate human linguistic performance within each register. We show that these two considerations, while complementary, are both important to account for in an inspection of LLM performance: while models may generally recognize and replicate human cross-register trends, they may in fact still fail to use language characteristic for human texts within each register.

We carry out our comparisons with respect to linguistic variables traditionally examined in register research, namely, dimensions of linguistic variation – composite linguistic variables that reflect patterns of linguistic co-occurrence in texts and represent general underlying patterns of variation in discourse – and offer a multidimensional linguistic profile of model performance, viewing texts on continua of involvement vs. informational production, narrativity, explicitness vs. situation dependency, persuasion, and abstractedness (Biber, 1988). We propose that

register-appropriacy – adherence to communicatively appropriate linguistic trends on each of these dimensions – is key for LLM improvement, as LLM training and development can be usefully informed through register-specific linguistic accounts of model performance. In the next section, we overview studies that took the first step toward describing a single model's language production across registers as well as arguing for register-specific improvements.

2 AI and register-appropriate language production

2.1 Descriptive corpus-based analyses

Systematic cross-register comparisons, linguistic descriptions of LLM performance, and evidence of where AI-generated language falls short of human register-specific language have been offered by corpus-based descriptive register analyses. These studies have revealed substantial human–AI divergence in a range of registers. Relying on the multidimensional analysis framework (Biber 1988), these studies compare AI-generated registers to their human counterparts on dimensions of linguistic variation based on linguistic co-occurrence, providing a comprehensive linguistic account of the differences, rather than focusing on select linguistic features.

Berber Sardinha (2024) examines ChatGPT 3.5-generated and human-authored conversations, Chemistry and Applied Linguistics academic articles, second language (L2) argumentative essays, and news articles on Biber's (1988) dimensions of variation in speech and writing and finds that register is a significant predictor of AI performance on all dimensions. Notable differences include ChatGPT markedly lacking in involvement in conversation but exhibiting greater information density in L2 essays and the news; ChatGPT struggling to emulate narrativity in conversation, news, and Applied Linguistics articles; lacking in explicitness in academic registers; in concreteness and situation-dependent discourse in conversation and news; in features of persuasion, the size of the effect varying drastically across registers; and in features of abstractedness characteristic for human academic writing, but exceeding human use of these features in L2 essays, news, and conversation. The study then shows that dimension scores are effective in discriminating between human-authored and AI-generated texts, revealing that AI is most detectable in conversation and academic articles and that even a register that is to be expected in AI training data (i.e., academic texts) displays misalignment with human texts.

Goulart et al. (2024) take a closer look at the academic register and work with student essays, personal narratives, and critiques, conducting both a linguistic and a situational analysis of texts, in the latter accounting for audience, explicitness, nature of the data, textual layout, and communicative purpose. Among the most notable situational findings are ChatGPT's tendency to list facts rather than develop an idea and its lower diversity of communicative purposes within texts, with the most common purposes being argumentative and explanatory. The linguistic

analyses mirror those carried out by Berber Sardinha, and in many ways the findings corroborate those trends. In Goulart et al.'s study, all the AI registers were significantly more informational than the human counterparts; student-produced texts were more narrative than AI-generated ones, more situation-dependent, more persuasive, and more abstract. Again, however, human–AI patterns varied by register, with some registers revealing greater differences. Additionally, Goulart et al.'s study showed that often, ChatGPT-generated registers are much less varied than the human-authored ones.

Demir and Egbert (2026) focus on biology and history textbooks and research articles as registers targeting student and expert audiences, respectively. They work with dimensions of variation in academic registers (Egbert 2014): Non-Technical Synthesis versus Specialized Information Density, Definition and Evaluation of New Concepts, and Author-Centered Stance. Their findings point out several discrepancies, mostly in the register of textbooks: While human authors are consistent in their non-technical style of history and biology textbooks, ChatGPT tends to incorporate much more specialized language in biology textbooks, revealing a lack of audience awareness and communicative purpose misalignment; ChatGPT fails to integrate evaluation of new concepts in biology textbooks to the extent human biology textbooks do; and human texts express stance to a much greater extent than ChatGPT in history journal articles. This study thus again observes multiple instances of ChatGPT's linguistic misalignment and a failure in functional appropriacy – inability to match the key communicative considerations, such as purpose and audience, of human texts.

2.2 Potential of register in LLM linguistic improvement

Descriptive corpus-based studies have laid important groundwork for LLM improvement by suggesting that since LLM deficiencies are register-specific, it would be useful to account for register information in the training process. While it is typically not disclosed how exactly particular models acquire register competence, scholars have suggested that at least in part this competence is likely to stem from register labels of the training data, which at present generally lack explicitness and specificity (Berber Sardinha 2024). Recent work on LLM training has shown that accounting for register differences in the training data indeed significantly affects LLM performance (Myntti et al. 2025). Myntti et al. train models on web-register-classified data and show that the presence of register information in the training data has a clear impact on model performance and that some registers in the training data lead to better performance on a range of benchmarks "reflect[ing] the most common present-day expectations placed on LLMs: fluency, scientific and real-world knowledge, problem solving, and reasoning" (Myntti et al. 2025: 5). For example, registers such as news in the training data result in lower performance, while the inclusion of opinion registers (e.g., reviews, opinion blogs), how-to instructions, or informational descriptions corresponds to better performance, and combining well-performing registers leads to notable improvements, although results vary

depending on the specific benchmark (see Myntti et al. 2025: 9–10 for a detailed discussion of findings per task).

Others have specifically called attention to capitalizing on register-based linguistic resources, that is, linguistic features frequent in a certain register, to improve LLM performance, focusing in particular on the linguistic features from Biber's (1988) dimensions of language variation in speech and writing – the dimensions used in descriptive corpus-based analyses overviewed above. Reinhart et al. (2025) compare various base and instruction-tuned models to human texts with respect to their use of sets of linguistic dimension features and, similar to descriptive register studies, observe that models come short in their use of particular grammatical and lexical features. Although in their study Reinhart et al. (2025) do not compare LLM performance across registers and do not examine linguistic feature co-occurrence patterns, as is common for register studies, but rather dimension features independently, they stress the potential effectiveness of the linguistic features considered important for register studies in classifying texts as LLM- or human-produced and make a case for the integration of this linguistic knowledge into LLM development to approximate human production more accurately.

Milička, Marklová and Cvrček (2025a) also compare base and instruction-tuned AI models and do turn to linguistic feature co-occurrence – linguistic dimension scores – as potential benchmarks for evaluating AI-produced texts across registers, although using them for a different purpose. In particular, this study is concerned with intra-textual stylistic consistency: the study, based on the registers of English (AI-Brown) and Czeck (AI-Koditex), splits texts into the initial 500 words to be used as prompts and their continuations and first examines the stylistic stability of human texts, that is, stylistic shifts across the first and second parts, and then compares the LLM-generated second parts to the original human continuations, thus investigating different models' ability to match the style of a text as they continue it. The study demonstrates that different models vary in their performance, and different linguistic dimensions also pose different challenges to the models. Proposing Biber's linguistic dimensions as useful, interpretable benchmarks in model evaluation, similar to the other studies, Milička, Marklová and Cvrček (2025a) call for linguistically motivated model improvements.

Most recently, Nieth et al. (2026) have focused on academic writing, news, creative writing, instructional texts, and spoken conversations, using seven LLMs of four different families, and showed that the distribution of the Biber features between LLM-generated and human-produced texts is different on a corpus level. They note that previous LLM register studies either provided models with snippets of the actual register texts or started from a generic prompt lacking specific situational context. They point out that in the former setting, the ability of the model to continue a text in the same "style" is measured (i.e., the objective of Milička, Marklová and Cvrček) rather than its ability to infer the register-appropriate rules (i.e., produce complete texts of a given register as prompted). In the latter case, the authors argue that without explicit situational context, the models generate their

own context, which might not be appropriate for the register; nor is it guaranteed to show a realistic distribution of topics, which highlights challenges in reliably reproducing human situation through LLMs.

2.3 The present study

Taken together, previous studies suggest that any future advances in AI tools must make strides toward improving their register competence to ensure that the produced texts are communicatively effective, and multi-register AI-generated corpora are now designed specifically for the purpose of such analyses (e.g., Milička, Marklová and Cvrček 2025b). Yet, while the potential of register research for AI development generally and Biber's multidimensional analysis framework in particular has been recognized, most existing studies in the register-functional paradigm have concerned themselves with a single model's output, predominantly ChatGPT (notable exceptions discussed above are Milička, Marklová and Cvrček 2025a and Nieth et al. 2026; however, neither of these studies explicitly concerned themselves with cross-register linguistic analyses). Conclusions regarding AI's register capabilities reached by descriptive corpus-based studies thus tend to be based on findings on a single model, and it remains unclear to what extent it is possible to extrapolate these conclusions to other models, whose development and use is becoming exponentially more prominent. We therefore observe a divide between important language research with clear practical implications, on the one hand, and the increasingly rapid development of new models that such studies have not been conducted on, on the other. Applications of linguistic results to practice, namely, linguistically informed and register-specific model development and training, as proposed by, for instance, Reinhart et al. (2025), thus remain impossible or, at best, limited until studies take into consideration the capabilities of different currently used models.

In this paper, we take a step in this direction by comparing models for the extent of their register alignment, examining its two related but distinct aspects: (a) models' cross-register language use as well as (b) models' alignment with the human benchmark in each register and offering these linguistic profiles as the basis for model improvement. We extend previous research in several ways. First, we broaden the range of AI tools under analysis and compare four open-source instruction-tuned models, represented by two strategy conditions: zero-shot and few-shot. We then examine the performance of these model-strategy combinations (hereafter MSC) relative to human texts in three distinct registers: natural conversation, creative writing on Reddit, and BBC news articles, thus including registers of different communicative scope from less to more regimented ones, which can be expected to be more or less challenging (Brommer et al. 2024). Our aim is to provide a more fine-grained linguistic account of the register capabilities of different MSCs, as we expect that different models may display different linguistic and functional deficiencies and the pathways toward their improvement may differ. To this end, we examine:

a) MSCs' adherence to human *cross-register* contrasts;

b) MSCs' alignment with the human benchmark *within* each register.

We thus address the following research questions (RQs):

1. To what extent do the examined MSCs adhere to human cross-register trends?

2. To what extent do the examined MSCs exhibit alignment with human texts within each register?

3 Method

3.1 Human and AI-produced corpora

The data used in this study are taken from Nieth et al.'s (2026) work, which used the instruction-tuned, open-source models Apertus 70B (Apertus et al. 2025), Llama 3.3 70B and Llama 3.1 8B (Grattafiori et al. 2024), Qwen 3 32B and Qwen 3 8B (Yang et al. 2025), and Gemma 3 27B and Gemma 3 12B (Gemma Team et al. 2025). As human-produced corpora they used natural conversation (BNC2014; Love et al. 2017), creative writing texts produced by users in response to prompts on Reddit (Fan, Lewis and Dauphin 2018), BBC news articles (Narayan, Cohen and Lapata 2018), academic writing (Lo et al. 2020), and instructive online texts from wikiHow (Koupaee and Wang 2018), in all cases N = 600 texts. In the case of all texts (human and LLM-generated), the first 400 words of each text with a margin of 40 words to finish the last sentence were selected, and these excerpts were the units of analysis across all corpora. While from a corpus linguistic perspective this limit is suboptimal (100 words shy of the recommended minimum length of 500 words; Biber 1993; Egbert 2014), the authors opted for this length as a way to save on computation costs[1]. The spaCy en_core_web_sm model was used to count all tokens which are neither punctuation nor whitespace. Contractions such as "don't" were therefore counted as two words. Nieth et al. showed that for a sample size of N = 600, two distinct samples of the same human corpora tend to have similar feature frequency distributions and that increasing sample size beyond this point yields diminishing returns. For their study, they drew two distinct sets from each register, the first serving as the study human corpus and the second as a pool of few-shot examples. For both sets, multiple candidate sets were drawn and the one closest to the overall human distribution of Biber features was selected (see Nieth et al. 2026, Appendix A, for more detail).

All datasets selected in Nieth et al.'s study contain human texts and metadata about the situational context, such as speaker information and conversation topics for the BNC corpus, the Reddit user's prompt for the creative texts, and the article

[1] Working with these texts in this study, we therefore stress that any conclusions we make about the MSCs comparability with human texts or cross-register trends are based on this 400-word limit, and we acknowledge that model performance on longer texts may differ.

summary for the news texts. In their prompting, Nieth et al. used these metadata to condition the model on the specific context of the text to generate by inserting this information into the generation prompt (see Appendix 1 of this paper for example prompts in each register). In this way, they produced for each human text a corresponding LLM-generated text for each model. That study also tested one-, two-, and three-shot generation, sampling examples of human texts for a given situational context from the few-shot pool of the respective register. A seeded random function was used, so that for each human text in each few-shot setting the same examples were always selected. For the BNC corpus in the zero-shot setting, the authors ran a prompt stability study using four additional prompts (see Nieth et al. 2026, Figure 5). There they showed that different prompts have an effect on Biber's linguistic features; however, the results are not dictated by that (see Nieth et al. 2026, Appendix B, Figure 8). All texts were generated locally by Nieth et al. (2026) with the vLLM library, loading the open-source model checkpoints from HuggingFace on one or two A100 GPUs with 80 GB of VRAM, depending on model size. Generation used a temperature of 1 and top-p of 1 with no repetition penalty or other decoding constraints, which is just sampling from the full model distribution. A fixed random seed of 127 was used for all sampling procedures, including the selection of few-shot examples. The code, generation settings, prompts, and resulting corpora are publicly available at https://github.com/BjoernNieth/Register_Aware_LLMs.

For the present study, for feasibility of linguistic analysis, we select from Nieth et al.'s work a more restricted set of human registers as well as models to consider. Specifically, we include the registers natural conversation, creative writing, and BBC news articles. These registers differ in mode, production circumstances, audience relationships, interactivity, and communicative goals, thus presenting a substantial communicative range. The conversation texts are contemporary British English spontaneous naturally occurring real-time face-to-face interactions (collected between 2012–2016) typically between family and friends in informal settings, on an unlimited range of topics and setting a variety of communicative goals[2]. In contrast, Reddit creative texts are written texts, produced with opportunities for editing and revision and for an unknown, mass audience rather than a specific familiar interlocutor. The channel (a social media platform), the range of creative topics, and the communicative goal of entertainment, however, still give the authors a free reign in terms of self-expression and involvement, resulting again in a relatively informal setting. BBC news articles represent professional rather than informal communication, dictated mainly by the nature of the topics that constitute a public agenda, the communicative goals to inform and report, and the official channel/medium (BBC).

As for the models, we consistently opted for the larger version if available, as larger models tended to be more stable to prompt variation in Nieth et al.'s setting. For each model, we study (a) the zero-shot condition in which the models did not receive human text examples and were prompted to produce texts of a given

[2] For more detail and news on the corpus, please see https://cass.lancs.ac.uk/cass-projects/spoken-bnc2014/

register from metadata only; (b) the few-shot, specifically the three-shot, condition, as here the models had the most human examples in Nieth et al.'s study. Thus, each model is represented by two strategy conditions, which resulted in the following eight MSCs: Apertus_zero-shot, Apertus_few-shot, Gemma_zero-shot, Gemma_few-shot, Llama_zero-shot, Llama_few-shot, Qwen_zero-shot, and Qwen_few-shot. The human and the eight MSC-produced corpora represent the levels of the generator variable of this study. The independent variables of the study thus are a) generator (with nine levels in total: the eight MSCs and the human group as the baseline), and b) register (with three levels: conversation, creative, news). We used the same sample size of N = 600 texts for each human and MSC-generated register (600 texts × 9 generator groups × 3 registers = 16,200 texts total).

3.2 Linguistic variables and quantitative method

Our goal is to examine the performance of MSCs relative to human texts in the three registers of the study. We compare MSCs and human corpora on dimensions of variation in speech and writing (Biber 1988). Registers have traditionally been examined with regard to linguistic dimensions that represent complex communicative functions shared by co-occurring linguistic features and capture "overall parameters of linguistic variation within a set of texts" (Biber 1985: 339-340). Five such patterns were interpreted in the multidimensional analysis of the registers of speech and writing (Biber 1988): Involved vs. Informational Production, Narrative vs. Non-Narrative Concerns, Explicit vs. Situation-Dependent Reference, Overt Expression of Persuasion, and Abstract vs. Non-Abstract Information. The dimension structures – that is, the co-occurring linguistic features comprising each dimension – are presented in Appendix 2.

Each dimension is a continuum, whose positive and negative poles are associated with a particular set of co-occurring features that co-construct a particular type of discourse. The features of these poles are in complementary distribution: when the positive features of a given dimension occur together in texts, the negative features are relatively infrequent; and conversely, when the negative features are frequent, the positive ones are relatively rare. Importantly, dimensions are interpreted in relative rather than absolute terms, and the position of texts on the dimensions displays a particular balance of positive and negative features rather than presence or absence of one of the sets.

In this study, we employ an additive multidimensional analysis (e.g., Berber-Sardinha et al. 2019) similar to the studies discussed above; namely, we examine our varieties of interest – human- and MSC-generated texts of the three registers – along these previously identified dimensions. To achieve this, our texts were tagged for the linguistic features of the original Biber (1988) study – that is, the same linguistic features that contributed to the dimensions of that study – with the

openly available pybiber tagger[3] (Brown and Reinhart 2024), and dimension scores were computed for each text. These dimension scores then allow an examination of the relative position of each variety of interest (human- and MSC-generated registers) on the dimensions as means and standard deviations of the text scores are computed for each variety. These dimension scores are our continuously measured linguistic (dependent) variable.

We ran factorial ANOVAs to examine the effect of generator on our dependent variables across registers. In our Tukey's post hoc analyses:

1. to examine MSCs' adherence to human *cross-register* patterns, we analyzed:

- human texts across registers
- MSC-produced texts across registers

2. to examine, MSCs' alignment with human texts *within* each register, we analyzed:

- Each MSC vs. the human group in each register

These tests were run on each of the five linguistic dimensions. Significance threshold was set at alpha < .01 to adjust for multiple comparisons on the five dimensions (.05/ 5 = .01). Additionally, Cohen's *d* (Cohen, 1988) effect sizes were examined for each pairwise comparison. In cases where the results did not reach statistical significance based on the *p* values adjusted for multiple comparisons (i.e., $p > .01$), effect sizes above the small-effect threshold (i.e., $d \geq 0.2$) were considered as indicators of noteworthy differences[4].

We should note that in all our analyses, we focus solely on the average characteristics of the texts. Intra-register variation has now been consistently documented by register studies (e.g., Egbert and Mahlberg 2020; Biber and Egbert 2023; Egbert and Gracheva 2023; Goulart 2024; Wood 2024; Gracheva 2026) and becomes apparent from our results as well; however, like the first descriptive corpus-based studies of AI-produced versus human language (Section 2.1), in line with the bulk of traditional register research, and for feasibility reasons, here we compare mean dimension scores, leaving analyses of variance for further study.

4 Results and discussion

In this Section we provide a global overview of the results on the dimensions. Table 1 summarizes the ANOVA results and shows that there are consistent cross-register

[3] https://github.com/browndw/pybiber. Blombach et al. (2026) present, to the best of our knowledge, the first evaluation of multiple publicly available tagging systems, including pybiber, in terms of comparability of additive multidimensional analysis results, in particular, examining Biber (1988) dimensions that the systems aim to reproduce. Interested readers are thus referred to Blombach et al. for a detailed description of the tagging systems. While pybiber has been shown to be comparable to the other systems, including the Biber Tagger (Biber, 1988) on Dimension 1, we do acknowledge that, as with other taggers, non-standard spellings, such as *nay, cannae* and *nothin',* may result in accuracy decrease.

[4] Effects sizes are interpreted in accordance with Cohen's (1988) guidelines: $d = 0.2$ (small); $d = 0.5$ (medium); $d = 0.8$ (large).

differences in the ways in which the generators of the texts – the human group and the eight MSCs – use the linguistic resources of the dimensions. The interaction effects are significant on all dimensions, which indicates that the differences among the levels of the generator variable are always register-dependent. In our analyses, we focus first on a) cross-register patterns of the human group and the MSCs in order to examine the first aspect of MSC register alignment, namely recognition of register differences and adherence to human cross-register trends, and then b) the pairwise comparisons between MSCs and the human group within each register to examine the degree of MSCs' alignment with the human group.

Table 1: ANOVA results.

Effect	Dimension	*F*	*p*	η^2 (partial)
Generator	1	265.118	< .001	.116
Register		14,733.967	< .001	.646
Generator × Register		50.219	< .001	.047
Generator	2	72.605	< .001	.035
Register		5,991.618	< .001	.426
Generator × Register		38.612	< .001	.037
Generator	3	404.893	< .001	.167
Register		5,531.114	< .001	.406
Generator × Register		39.117	< .001	.037
Generator	4	170.870	< .001	.078
Register		484.842	< .001	.057
Generator × Register		37.026	< .001	.035
Generator	5	150.582	< .001	.069
Register		2,052.682	< .001	.202
Generator × Register		59.762	< .001	.056

We begin by overviewing the human cross-register patterns on the dimensions, which serve as the benchmark for our analyses of MSC cross-register performance. It is apparent from Figures 1-5 in the following sections (the human group is always the middle green box) that on Dimension 1, the human conversation (M = 47.73; SD = 10.29) is the most involved register, followed by the creative texts (M = 24.01; SD = 13.19), and news (M = 10.94; SD = 9.82) is the most informational of the registers. All of these register contrasts are significant ($p < .001$; $d_{\text{conv vs creative}} = 2.01$; $d_{\text{conv vs news}} = 3.66$; $d_{\text{news vs creative}} = 1.12$). On Dimension 2, again, all pairs of the human registers are significantly different from each other ($p < .001$; $d_{\text{conv vs creative}} = 1.24$; $d_{\text{conv vs news}} = 0.80$; $d_{\text{news vs creative}} = 0.46$), with conversation ($M$ = -0.45; SD = 1.56) in the negative range of the dimension being the most non-narrative register, followed by news (M = 0.98; SD = 2.03) in the positive range, and with creative texts (M = 1.99; SD = 2.31) showing the most pronounced tendency toward narration. On Dimension 3, the human registers display significant contrasts between conversation and the news ($p < .001$; $d_{\text{conv vs news}} = 1.72$) and the creative register and the news ($p < .001$; $d_{\text{news vs creative}} = 1.31$), in which cases the news (M = 0.63; SD = 1.66) is always the more explicit register, but a small difference (p =

.016; $d_{\text{conv vs creative}}$ = 0.24) between conversation (*M* = -1.91; *SD* = 1.27) and creative texts (*M* =-1.56; *SD* = 1.68), both of which are in the situation-dependent range. On Dimension 4, human news (*M* = 0.05; *SD* = 2.29) and the human creative register (*M* = 0.07; *SD* = 1.94) do not differ from each other (*p* = 1.00; $d_{\text{news vs creative}}$ = 0.01) and are both higher on the persuasion pole; in contrast, human conversation (*M* = -0.41; *SD* = 2.15) is significantly less persuasive than both of the other registers, although the effects are small (*p* < .01; $d_{\text{conv vs creative}}$ = 0.23; $d_{\text{conv vs news}}$ = 0.21). Finally, on Dimension 5, news (*M* = 2.62; *SD* = 2.34) is the register that relies the most on abstract discourse, followed closely by the creative register (*M* = 2.52; *SD* = 2.28) (difference between the two non-significant; *p* = 1.00; $d_{\text{news vs creative}}$ = 0.04), and conversation (*M* = 0.88; *SD* = 1.70) occupies the lowest position on the dimension, showing significant differences from the other two registers (*p* < .001; $d_{\text{conv vs creative}}$ = 0.82; $d_{\text{conv vs news}}$ = 0.85).

We consider that MSCs adhere to human cross-register patterns if they display the same register contrasts as the human group and these differences on the dimensions are in the same direction. For example, if human conversation is the most involved register on Dimension 1, news is the most informational, and the creative texts fall in between, with all differences statistically significant, MSCs would need to show the same cross-register pattern of significant differences. If the same patterns are not observed in MSC performance, the models cannot be considered to reproduce human cross-register patterns accurately. We observe both of these possibilities in our results, and in the next section, we group the results accordingly, discussing cases in which MSCs generally follow human cross-register patterns (Section 4.1.1) or violate them, going as far as reversing the human cross-register trends (Section 4.1.2).

4.1 RQ 1: Adherence to human cross-register trends

4.1.1 Recognition of human cross-register patterns: Involved vs. Informational Production and Narrative vs. Non-Narrative Concerns

Two dimensions revealed quite promising outcomes in the case of all our MSCs in terms of the first aspect of register alignment, namely, adherence to human cross-register patterns. On Dimensions 1 and 2, all examined MSCs followed the human cross-register trends. Tables 2 and 3 present the descriptive statistics for each MSC and the human group across registers on Involved vs. Informational Production and Narrative vs. Non-Narrative Concerns, respectively, and Figures 1 and 2 plot the groups along these dimensions.

As noted above, on Dimension 1, the human group differences are significant in all register contrasts (*p* < .001), with conversation, creative, and news registers showing a downward trend from the most to the least involved production, respectively. All MSCs successfully replicate this pattern (*p* < .001). Appendix 3 contains all corresponding Cohen's *d* effect sizes for all MSC cross-register contrasts, showing that all effects are large. It is noteworthy that while our MSCs do mirror the human contrasts, the effects vary substantially across MSCs, and some

MSCs replicate the human cross-register trends much more closely than others (e.g., consider the Llama family MSCs in the news vs. creative contrasts, where the effect is close to medium while it is almost twice as large in the human contrast. And conversely, in other cases, MSCs' register differences are more extreme than those between the human corpora as in the case of the Gemma models in the conversation vs. news contrast). Overall, however, we find that all MSCs generally recognize human register shifts, with all MSC-produced register pairs showing significant differences as in human texts and with the same direction of the differences (Table 2; Figure 1).

Similarly, on Dimension 2, all pairs of the human registers are significantly different from each other ($p < .001$), with conversation in the negative range of the dimension being the most non-narrative register, followed by news in the positive range, and with creative texts showing the most pronounced tendency toward narration. Notably, all our MSCs again also show significant cross-register differences in all register pairs ($p < .001$; see also effect sizes in Appendix 3) and do follow the same pattern of differences: that is, conversation, news, creative in that order from the non-narrative to the narrative end (Table 3; Figure 2). Yet again, there is variation in the extent to which the human cross-register trend is followed. The most striking divergence is the pattern of large differences in the case of all MSCs in the news vs. creative contrast whereas this difference is only medium in the human corpora. Similarly, although to a lesser extent, the effect for MSCs is often larger in the conversation vs. creative contrasts, with some MSCs doubling it (e.g., Qwen_zero; see Appendix 3).

Table 2: Descriptive statistics for the human corpus and each MSC across registers on Dimension 1: Involved versus Informational Production.

Group	Conversation		Creative		News	
	M	*SD*	*M*	*SD*	*M*	*SD*
Human	**47.73**	10.29	**24.01**	13.19	**10.94**	9.82
Apertus_zero	39.13*	8.93	17.59	10.46	7.15	7.53
Apertus_few	38.72*	9.94	**20.14**	10.22	7.89	8.40
Gemma_zero	33.14	8.40	15.34*	8.01	0.55	6.89
Gemma_few	34.79	8.85	14.00*	8.81	0.83	7.03
Llama_zero	40.88	9.28	16.69	11.09	9.99	9.49
Llama_few	**44.45**	10.21	18.54	11.89	**11.63**	9.82
Qwen_zero	30.81	9.67	19.69	10.41	6.63	7.71
Qwen_few	33.94	9.55	19.76	10.91	7.17	8.31

Note. For RQ 2 on alignment with the human baseline: Human group means and the means of the best-aligned MSCs are in bold. Asterisks mark a lack of strategy effect for a given model.

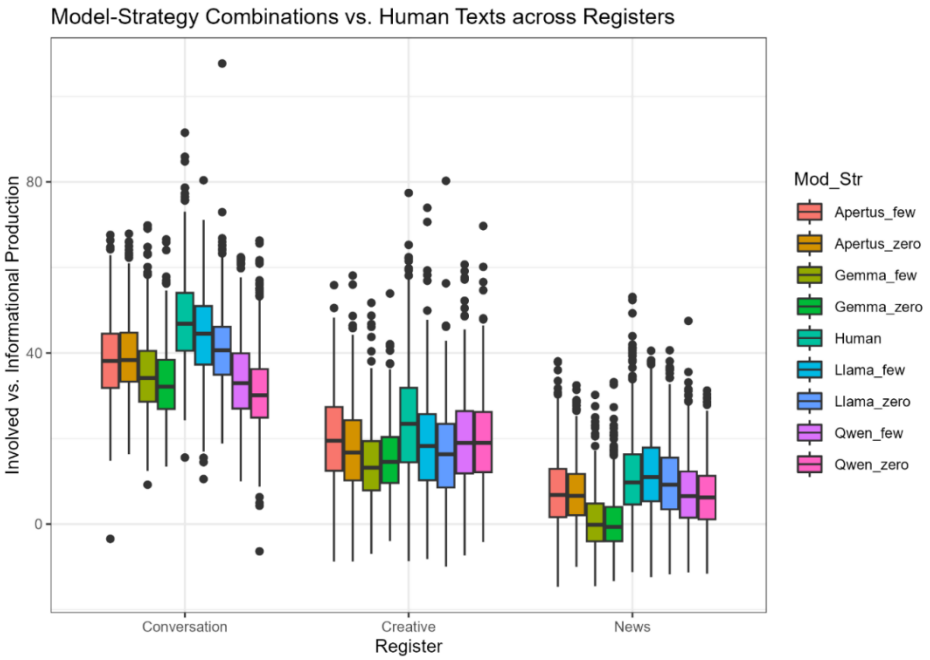


Figure 1: MSCs and human group across registers on Dimension 1: Involved (positive range) versus Informational (negative range) Production.

Table 3: Descriptive statistics for the human corpus and each MSC across registers: Dimension 2: Narrative vs. Non-Narrative Concerns.

Group	Conversation		Creative		News	
	M	*SD*	*M*	*SD*	*M*	*SD*
Human	**-0.45**	1.56	**1.99**	2.31	**0.98**	2.03
Apertus_zero	-1.13*	1.11	2.13	1.84	-0.16*	1.44
Apertus_few	-1.19*	1.10	**1.89**	1.97	-0.24*	1.55
Gemma_zero	-0.24*	1.18	2.48*	1.63	0.30	1.53
Gemma_few	-0.14*	1.33	2.69*	1.93	**0.67**	1.65
Llama_zero	-0.80*	1.18	3.32*	2.09	0.48	1.70
Llama_few	-1.14*	1.16	3.35*	2.09	0.52	1.93
Qwen_zero	-0.74	1.14	3.44	1.93	0.16	1.51
Qwen_few	**-0.52**	1.22	3.11	2.09	0.48	1.55

Note. For RQ 2 on alignment with the human baseline: Human group means and the means of the best-aligned MSCs are in bold. Asterisks mark a lack of strategy effect for a given model.

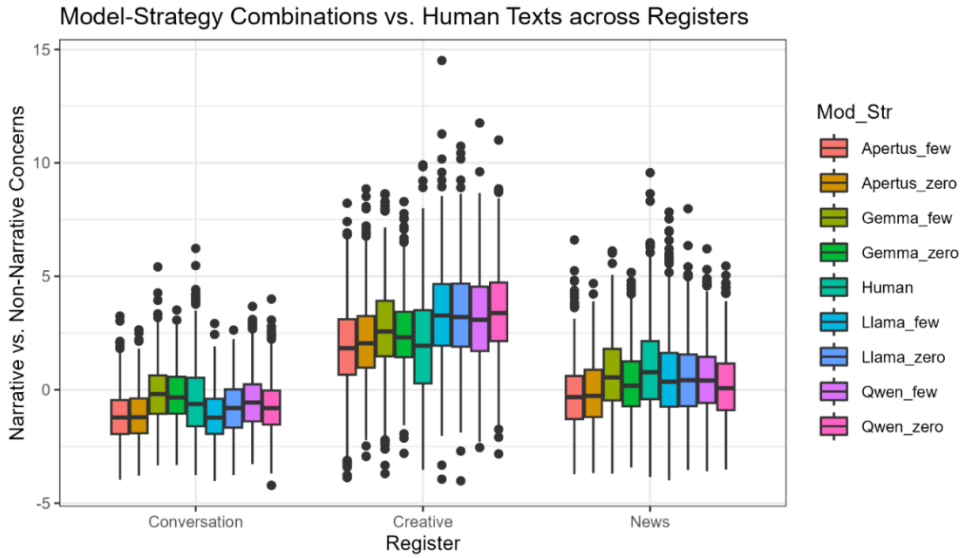


Figure 2: MSCs and human group across registers on Dimension 2: Narrative (positive range) vs. Non-Narrative (negative range) Concerns.

While these outcomes may be generally viewed as promising in that they suggest that our models a) are to varying extents sensitive to differences in communicative contexts, and b) pattern on the dimensions in similar ways to human registers, this result was not consistent and did not bear out on the other linguistic dimensions.

4.1.2 Reversal of register patterns: Explicit vs. Situation-Dependent Reference, Overt Expression of Persuasion, Abstract vs. Non-Abstract Information

In this section, we highlight cases in which the models violate human cross-register trends. While on Dimensions 1 and 2, MSCs generally succeeded in replicating the overall cross-register trends, this was not the case on Dimensions 3, 4, and 5. For space reasons, we overview Dimensions 3 and 5 briefly, devoting the most attention to Dimension 4 – Overt Expression of Persuasion – which displayed the most striking divergence from the human cross-register trend.

On Dimension 3, Explicit versus Situation-Dependent Reference, the human registers display significant contrasts between conversation and the news ($p < .001$; $d = 1.72$) and the creative register and the news ($p < .001$; $d = 1.31$), in which cases the news is always the more explicit register, but a small difference between conversation and creative texts ($p = .016$; $d = 0.24$), both of which are in the situation-dependent range. It is this closeness of conversation and creative texts in their situation-dependency that posed a challenge to our models. Among the MSCs, only Qwen_zero displays this smaller difference between conversation and creative texts ($d = 0.12$; see Appendix 3 for all effect sizes). For all the other MSCs, that difference is mostly large and in two cases medium, mostly caused by the increased explicitness of most MSCs in the creative texts and increased situation dependence of some MSCs in conversation (Table 4; Figure 3). MSCs also tend to exaggerate the

conversation vs. news contrast to varying extents (Appendix 3), although the differences are in the "large" window for both the human corpus and all the MSCs.

Table 4: Descriptive statistics for the human and each of the model-strategy groups across registers: Dimension 3: Explicit versus Situation-Dependent Reference.

Group	Conversation		Creative		News	
	M	*SD*	*M*	*SD*	*M*	*SD*
Human	**-1.91**	1.27	**-1.56**	1.68	**0.63**	1.66
Apertus_zero	-1.48*	1.41	-0.21	1.64	1.47*	1.65
Apertus_few	-1.33*	1.63	-0.33	1.77	1.68*	1.70
Gemma_zero	-2.04*	1.16	**-1.55***	1.22	-0.06	1.40
Gemma_few	-2.27*	1.10	-1.32*	1.28	**0.67**	1.50
Llama_zero	-1.12	1.31	0.78	1.78	2.73	1.73
Llama_few	-1.19	1.42	0.68	1.85	2.61	1.94
Qwen_zero	-1.82	1.41	-1.65*	1.54	1.27*	1.60
Qwen_few	**-1.86**	1.26	-1.19*	1.63	1.47*	1.66

Note. For RQ 2 on alignment with the human baseline: Human group means and the means of the best-aligned MSCs are in bold. Asterisks mark a lack of strategy effect for a given model.

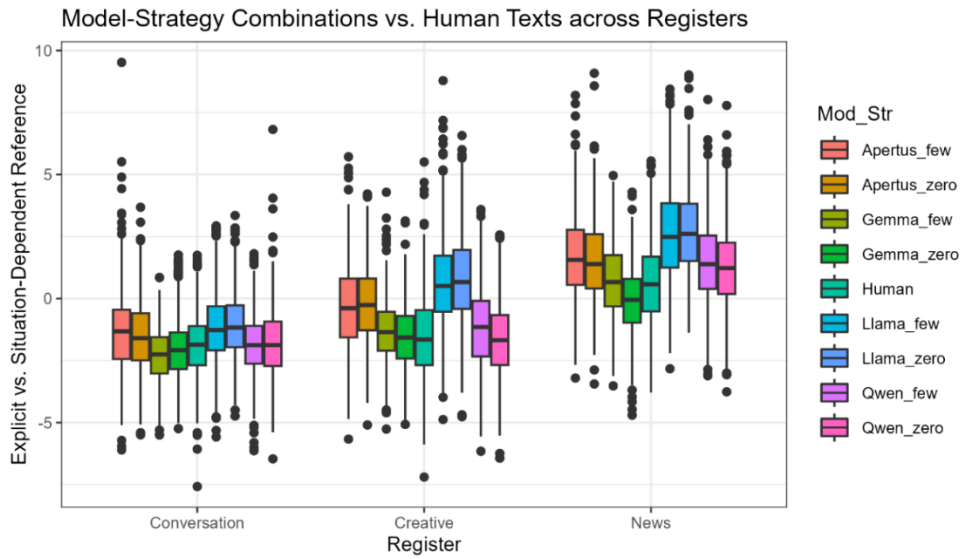


Figure 3: MSCs and human group across registers on Dimension 3: Explicit (positive range) vs. Situation-Dependent Reference (negative range).

On Dimension 5, Abstract versus Non-Abstract Information, news is the register that relies the most on abstract discourse, followed closely by the creative register (difference between the two non-significant; $p = 1.00$; $d_{\text{news vs creative}} = 0.04$), and conversation occupies the lowest position on the dimension (Table 5; Figure 4), showing significant differences with the other two registers ($p < .001$; $d_{\text{conv vs creative}} = 0.82$; $d_{\text{conv vs news}} = 0.85$). In the news and creative registers, for many of the MSCs the direction of the differences is occasionally reversed: for example, for Apertus and Llama MSCs as well as for Qwen_few, the creative register is higher on the dimension, and for Llama MSCs and for Apertus zero, the contrast between

news and creative texts is substantial (see effects in Appendix 3), which makes this reversal a notable failing of these models. This contrast is also large (contrary to human corpora; Appendix 3) in the case of Gemma MSCs, although the direction of the differences is in line with the human pattern.

Table 5: Descriptive statistics for the human and each of the model-strategy groups across registers: Dimension 5: Abstract versus Non-Abstract Information.

Group	Conversation		Creative		News	
	M	*SD*	*M*	*SD*	*M*	*SD*
Human	**0.88**	1.70	**2.52**	2.28	**2.62**	2.34
Apertus_zero	1.11	1.75	3.60	2.57	2.98	2.25
Apertus_few	**0.95**	1.63	3.23	2.39	2.82	2.02
Gemma_zero	-0.01*	1.37	1.19*	2.04	**2.61***	1.96
Gemma_few	-0.09*	1.37	1.19*	2.23	2.51*	1.91
Llama_zero	0.63*	1.57	4.77	2.68	3.66	2.18
Llama_few	0.44*	1.52	4.42	2.68	2.74	2.12
Qwen_zero	1.21	1.78	**2.69***	2.35	3.13	2.05
Qwen_few	1.14	1.67	2.85*	2.34	2.68	2.10

Note. For RQ 2 on alignment with the human baseline: Human group means and the means of the best-aligned MSCs are in bold. Asterisks mark a lack of strategy effect for a given model.

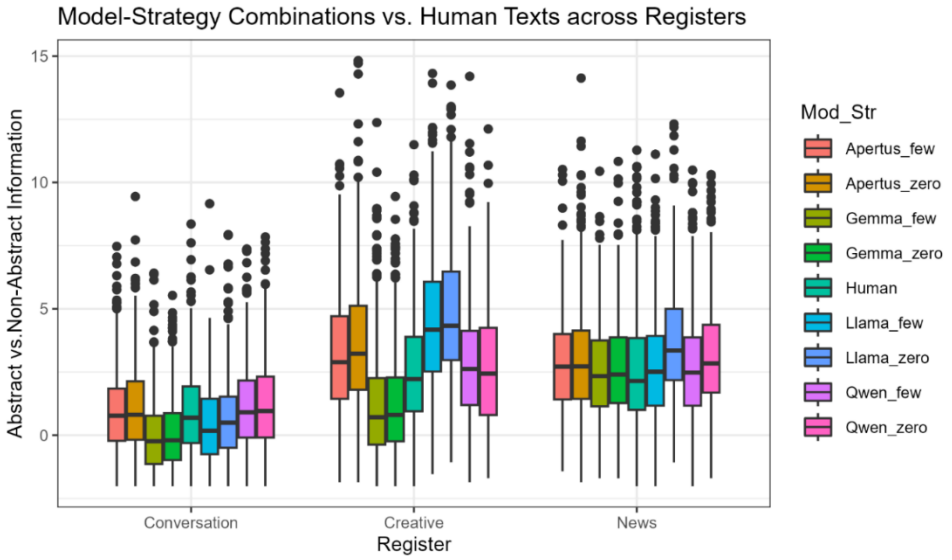


Figure 4: MSCs and human group across registers on Dimension 5: Abstract (positive range) versus Non-Abstract Information (negative range).

As noted above, the most striking divergence from the human cross-register trend was observed on Dimension 4. It is worth noting that while the dimension label – Overt Expression of Persuasion – was assigned in the original study that produced the dimensions (Biber 1988), the discourse captured by this dimensions is often that of irrealis and hypotheticality, with modal verbs of prediction, necessity, and possibility contributing to future orientedness and a focus on plans and events that have not yet occurred, as well as conditional subordination introducing hypothetical reasoning, among other features (e.g., suasive verbs, infinitives). On

this dimension, human news and human creative registers do not differ from each other (p = 1.00; $d_{\text{news vs creative}}$ = 0.01) and are both higher on the persuasion/hypotheticality pole; in contrast, human conversation is notably less persuasive/hypothetical than both of the other registers (p < .01; $d_{\text{conv vs creative}}$ = 0.23; $d_{\text{conv vs news}}$ = 0.21) (Table 6; Figure 5).

Table 6: Descriptive statistics for the human and each of the model-strategy groups across registers: Dimension 4: Overt Expression of Persuasion.

Group	Conversation		Creative		News	
	M	*SD*	*M*	*SD*	*M*	*SD*
Human	**-0.41**	2.15	**0.07**	1.94	**0.05**	2.29
Apertus_zero	0.30*	1.86	-1.03	1.74	-1.42	1.71
Apertus_few	0.63*	2.05	-0.43	2.07	-0.77	2.11
Gemma_zero	-0.84	1.40	-1.70*	1.37	-1.59	1.61
Gemma_few	**-0.37**	1.60	-1.71*	1.46	**-0.23**	2.03
Llama_zero	0.90*	1.88	-0.40	1.95	-0.27*	1.96
Llama_few	1.15*	1.89	**-0.06**	2.03	0.77*	2.48
Qwen_zero	0.50	1.84	-1.30	1.52	-0.68*	1.78
Qwen_few	0.17	1.66	-1.20	1.65	-0.71*	1.81

Note. For RQ 2 on alignment with the human baseline: Human group means and the means of the best-aligned MSCs are in bold. Asterisks mark a lack of strategy effect for a given model.



Figure 5: MSCs and human group across registers on Dimension 4: Overt Expression of Persuasion.

It is these patterns that we would expect the AI models to follow. The picture on Dimension 4, however, is quite complex, as some MSCs reverse the pattern: While for human texts, it is creative and news texts that exhibit a higher density of Dimension 4 features, for several MSCs, creative texts and the news are in the negative range, while conversation is highly persuasive (Table 6). Strikingly, this is the case for both of the Apertus, Llama, and Qwen models – six of the examined MSCs.

Gemma MSCs are markedly non-persuasive across registers, thus also not reproducing the human trend. This outcome generally reveals a failure of all our MSCs to replicate human cross-register patterns on this dimension. The effect sizes in Appendix 3 additionally show that the general trend is for the MSCs to exaggerate all cross-register differences, with medium to large effects in conversation vs. creative and conversation vs. news (while they are small in the human corpora) and small to medium to large in the creative vs. news contrast (while this effect is virtually absent in the human registers). Examples 1-3 and 4-6 present human and Llama_few cross-register differences, respectively, illustrating the reversal of the pattern (features of persuasion/hypotheticality bolded).

(1) Human Conversation S4C2 (Dim 4: -4.83)

Speaker_2: So Philippa was going to hockey club and then.
Speaker_1: And Olivia.
Speaker_2: And Bella and in French Deborah was like to Annette so Deanne you coming to hockey club Storm was like no I haven't got my stick she was like make sure you come next week.
Speaker_2: And then Devon was like to me are you coming to hockey club I was like no I was like I'm not that good at hockey she was like I bet you're really good at hockey.

(2) Human Creative ID: 174783; Dim 4: 2.08

If no one heeded the call like we did how many evil horrors **would** have been unleashed? ``, Jason said as passionate as ever. […] Guys shut up! Do you want people **to hear** how crazy you sound? ``, Jessica hissed. `` Grant **would** help me. He **would** gladly give his life again **to help** save this town, and the world! ``, Jason said accusingly at her. `` That's *not* fair. You *don't* have the right **to use** him against me.

(3) Human News ID: 36294611; Dim 4: 2.72

It epitomises how well the club is run and they're willing **to give** people chances because we don't have the big budgets **to get** in foreign players or players who have already been proven as a certain calibre. I'm indebted to the club because I **wouldn't** be here **if** that wasn't the case. "I've got absolutely no doubt that Hamilton, **if** anything, **will** do even better next season. […] we've got a good young crop coming through who are just so eager and hungry **to do** well. I've no doubt Hamilton are in great hands." His manager Martin Canning admitted it **would** have been fitting had Gordon scored a late winner

Text 1 – the human conversation reveals a lack of Dimension 4 features, while these features increase in numbers in Texts 2 and 3 – news and creative texts. In contrast, Text 4, conversation generated by Llama_few is the most persuasive of

the three, followed by Llama_few news (Text 6), also much more persuasive than the human news, and by the least persuasive creative register (Text 5).

(4) Llama few Conversation ID: SYEF; Dim 4: 4.03

> Well, **if** you **need** any help, don't hesitate **to ask**. I'm happy **to lend** a hand. How's the nursery search going, by the way? Oh, it's going alright, I think. We've got a few viewings lined up this week, so fingers crossed we**'ll** find the perfect one. That's great! I'm sure it'**ll** all fall into place. Have you thought about how you're going to balance work and being a mum? **To be** honest, I'm not really sure yet. I'm just taking it one step at a time. But I'm sure I'**ll** figure it out as I go along. You **will**, dear. You'**ll** be a brilliant mum. And don't forget, I'm always here **to help** out **if** you **need** me. Thanks, Mum.

(5) Llama few Creative ID: 250661; Dim 4: -0.79

> I was sitting in my living room, staring blankly at the TV as it droned on in the background, when suddenly a portal swirled to life in front of me. I jumped back in my seat, startled, as a figure emerged from the shimmering portal. It was me, or at least, a future version of me. He looked older, wearier, and his eyes had a haunted look that made my skin crawl.

(6) Llama few News ID: 38638651 Dim 4: 5.84

> However, the Confederation of African Football (CAF) has written to FIFA, asking them **to intervene** in the matter and **to prevent** Matip from playing for Liverpool until the issue is resolved. The situation has caused confusion, with Liverpool unsure whether Matip **will** be allowed to play for them until the matter is clarified. The club have been in touch with FIFA, and are waiting for a ruling on the matter. **If** a disciplinary case is opened, it could potentially lead to Matip being banned from playing for Liverpool, which **would** be a significant blow to the club's defensive options.

The analysis presented here focuses on the extent to which our examined MSCs adhere to human cross-register trends. This perspective does not yet take into account the AI–human gap, namely closeness of each MSC to the human baseline. We now turn our attention to the pairwise differences between specific MSCs and the human group in each register. In the next section, we show that in some cases, despite the overall adherence to the human cross-register trend, models can in fact consistently use the language of the dimensions differently from what is characteristic for human texts.

4.2 RQ 2: (Mis)alignment with the human baseline

To examine alignment with the human baseline in each register, we focus on Cohen's *d* effect sizes – that is, the size of the mean differences in the dimension scores of the human register and that same register produced by each of the MSCs. Tables 7-11 summarize these pairwise contrasts by register, presenting the size of the difference on the dimensions in each pair. Visualization of these differences is available in the same Figures (Figures 1-5): the focus is now on the human baseline – the green box in the middle – and each of the MSCs within each register rather than on cross-register patterns of each group. We consider differences above $d > 0.2$, that is, effects above "small" in Cohen's (1988) guidelines, to be indicators of noticeable misalignment and effects below $d < 0.2$ to point to MSCs' relatively effective emulation of the average dimension scores of human texts (these effective pairs are italicized in the tables).

Table 7: Summary of human–AI contrasts across registers. Cohen's *d* effect sizes are given in parentheses. Dimension 1: Involved vs. Informational Production.

Conversation	Creative	News
Human-Apertus_zero (0.89)	Human-Apertus_zero (0.54)	Human-Apertus_zero (0.43)
Human-Apertus_few (0.89)	Human-Apertus_few (0.33)	Human-Apertus_few (0.33)
Human-Gemma_zero (1.55)	Human-Gemma_zero (0.79)	Human-Gemma_zero (1.22)
Human-Gemma_few (1.35)	Human-Gemma_few (0.89)	Human-Gemma_few (1.18)
Human-Llama_zero (0.70)	Human-Llama_zero (0.60)	*Human-Llama_zero (0.10)*
Human-Llama_few (0.32)	Human-Llama_few (0.43)	*Human-Llama_few (0.07)*
Human-Qwen_zero (1.69)	Human-Qwen_zero (0.36)	Human-Qwen_zero (0.49)
Human-Qwen_few (1.39)	Human-Qwen_few (0.35)	Human-Qwen_few (0.41)

Table 8: Summary of human-AI contrasts across registers. Cohen's *d* effect sizes are given in parentheses. Dimension 2: Narrative vs. Non-Narrative Concerns.

Conversation	Creative	News
Human-Apertus_zero (0.50)	*Human-Apertus_zero (0.07)*	Human-Apertus_zero (0.65)
Human-Apertus_few (0.55)	*Human-Apertus_few (0.05)*	Human-Apertus_few (0.68)
Human-Gemma_zero (0.15)	Human-Gemma_zero (0.25)	Human-Gemma_zero (0.38)
Human-Gemma_few (0.21)	Human-Gemma_few (0.33)	*Human-Gemma_few (0.17)*
Human-Llama_zero (0.25)	Human-Llama_zero (0.60)	Human-Llama_zero (0.27)
Human-Llama_few (0.50)	Human-Llama_few (0.62)	Human-Llama_few (0.23)
Human-Qwen_zero (0.21)	Human-Qwen_zero (0.68)	Human-Qwen_zero (0.46)
Human-Qwen_few (0.05)	Human-Qwen_few (0.51)	Human-Qwen_few (0.28)

Table 9: Summary of human-AI contrasts across registers. Cohen's d effect sizes are given in parentheses. Dimension 3: Explicit versus Situation-Dependent Reference.

Conversation	Creative	News
Human-Apertus_zero (0.32)	Human-Apertus_zero (0.81)	Human-Apertus_zero (0.51)
Human-Apertus_few (0.39)	Human-Apertus_few (0.71)	Human-Apertus_few (0.62)
Human-Gemma_zero (0.10)	*Human-Gemma_zero (0.01)*	Human-Gemma_zero (0.45)
Human-Gemma_few (0.30)	*Human-Gemma_few (0.16)*	*Human-Gemma_few (0.03)*
Human-Llama_zero (0.61)	Human-Llama_zero (1.35)	Human-Llama_zero (1.24)
Human-Llama_few (0.53)	Human-Llama_few (1.27)	Human-Llama_few (1.10)
Human-Qwen_zero (0.07)	*Human-Qwen_zero (0.05)*	Human-Qwen_zero (0.39)
Human-Qwen_few (0.04)	Human-Qwen_few (0.22)	Human-Qwen_few (0.51)

Table 10: Summary of human-AI contrasts across registers. Cohen's *d* effect sizes are given in parentheses. Dimension 4: Overt Expression of Persuasion.

Conversation	Creative	News
Human-Apertus_zero (0.35)	Human-Apertus_zero (0.59)	Human-Apertus_zero (0.73)
Human-Apertus_few (0.50)	Human-Apertus_few (0.25)	Human-Apertus_few (0.37)
Human-Gemma_zero (0.24)	Human-Gemma_zero (1.05)	Human-Gemma_zero (0.83)
Human-Gemma_few (0.02)	Human-Gemma_few (1.04)	*Human-Gemma_few (0.13)*
Human-Llama_zero (0.65)	Human-Llama_zero (0.24)	*Human-Llama_zero (0.15)*
Human-Llama_few (0.77)	*Human-Llama_few (0.07)*	Human-Llama_few (0.30)
Human-Qwen_zero (0.45)	Human-Qwen_zero (0.78)	Human-Qwen_zero (0.36)
Human-Qwen_few (0.30)	Human-Qwen_few (0.71)	Human-Qwen_few (0.37)

Table 11: Summary of human-AI contrasts across registers. Cohen's *d* effect sizes are given in parentheses. Dimension 5: Abstract versus Non-Abstract Information.

Conversation	Creative	News
Human-Apertus_zero (0.13)	Human-Apertus_zero (0.44)	*Human-Apertus_zero (0.16)*
Human-Apertus_few (0.04)	Human-Apertus_few (0.30)	*Human-Apertus_few (0.09)*
Human-Gemma_zero (0.58)	Human-Gemma_zero (0.61)	*Human-Gemma_zero (0.004)*
Human-Gemma_few (0.63)	Human-Gemma_few (0.59)	*Human-Gemma_few (0.05)*
Human-Llama_zero (0.15)	Human-Llama_zero (0.90)	Human-Llama_zero (0.46)
Human-Llama_few (0.27)	Human-Llama_few (0.76)	*Human-Llama_few (0.05)*
Human-Qwen_zero (0.19)	*Human-Qwen_zero (0.07)*	Human-Qwen_zero (0.23)
Human-Qwen_few (0.15)	*Human-Qwen_few (0.14)*	*Human-Qwen_few (0.03)*
Human-Qwen_few (0.30)	Human-Qwen_few (0.71)	Human-Qwen_few (0.37)

We illustrate misalignment patterns in more detail on Dimension 1, Involved vs. Informational Production, where we observe the largest number of significant human–AI differences in every register (Table 7). This outcome clearly shows that, overall, AI tools still face challenges in replicating human language with respect to the functional pattern of Involved versus Informational Production, as also shown by studies on ChatGPT. This is particularly true about the registers of natural conversation (i.e., BNCSpoken in our case) and creative texts, where all eight MSCs show substantial aberrations from the human group (large effects in Table 7). A similar trend is seen in the news; one notable exception, however, is the Llama MSCs, whose differences with the human group are minimal: Llama_few:News-Human:News, d = 0.07, p = .999; Llama_zero:News-

Human:News, $d = 0.1$, $p = .995$. In the examples below, we present the human texts contrasted with the best- and the worst-aligned MSCs on Dimension 1 in each register, with linguistic features of involvement bolded and informational features italicized.

Conversation

(7) Human: S63L (Dim 1 score: 91.50)

Speaker_1: **Okay if you wanna** put **it** that *way*.

Speaker_2: **Can I** ask **you** a *question* please **I don't understand I do understand what you're where's you're going but I'd really** like to **know what is it** that **you wanna** do *with* **your** *life* **cos this is all** *about* **you** changing *career* **you know you know when you** said **I don't wanna** do *engineering* any more **which is** fine **have you now figured out what it is** apart *from* **just** growing a *business* **but** what what**'s** what **are you** passionate **about** what what **is** gonna get **you** up *in* the *morning* and go great **I'm** gonna do **this and I'm** gonna do **that** what **is it** that **you like**.

Speaker_1: **Okay well er well er well well** *part part of* **this goes back** *to the discussion* **I** had *with* Tony.

(8) Best-aligned MSC: Llama_few: S68E (Dim 1 score: 70.51)

So, how**'s** the *neighborhood* been treating **you**? Any *new faces* **around**?

Oh, **you know**, **it's** been alright. There**'s** been a few *new families* move in down the *street*. **But, you know, it's still** the *same old Liverpool*.

Yeah, some *things* never **change**, do they?

No, **I suppose not. So**, how**'s** *work* been? How**'s** the *school*?

It's been keeping **me** busy, **that's** for sure. The *kids* **are** always getting *into* **something**. **But, you know, it's** rewarding *in* **its** *own way*.

(9) Worst-aligned MSC: Qwen_zero: SC8H (Dim score: 10.18)

Rather chilly for a *Saturday*, eh, love? Thought **we'd** get a *bit of* a *break* 'ere *Christmas,* **but** nay, *same old mizzle from the sky*. Mind **you**, this *café* **is** a *proper little haven in* **all this**. Aye, **can**nae argue *with* **that** een, dear. **Keeps** the *wind* out, an' **puts** *a bit o' warmth in the bones*. **I've** been neglectin' this *place* **far too long**. *Old habits*, **see**? *Out o' teaching, out o' routine*. Didn**'t your** *mum* use to say **you could** hear a *pin* drop *in your class, Miss*? Like **nothin'** got the *audacity* to make a *sound in* **there**? My *English teacher, Mr. Coghill,* **bursts** in like Ten *Cossacks* **if** the *photocopier* **stops** pickin' *in A4*.

Conversation examples 7-9 in essence illustrate AI's predilection for informational language, specifically nouns and prepositional phrases, much more prominent in Qwen_zero's performance. Features of interactivity and involvement, on the other hand, such as discourse markers, clausal elaboration, private verbs, higher rates of occurrence of personal pronouns, among others, are the features that reveal the presence of the interlocutors in the moment of the interaction and the interpersonal nature of communication. It is these features that are attempted by Llama_few and achieved to a considerably better extent than the other models, although, as noted above, still showing a notable misalignment. Expectedly, the fragmentary nature of on-the-spot face-to-face human conversation produced orally and without preparation is a major challenge for LLMs, which generate written texts and base this generation mostly on written language.

Creative

(10) Human: ID 90094 (Dim 1 score: 51.53)

My *face* **felt** like **it** was *on fire*. **I thought** he**'d** broken **my** *nose* **again**. **I wanted** to **be** more defiant and angry than **I** was, **but I** was **far** more distracted by **what I saw** *in front of* **me**. **I** was looking *at* the *face* **my** *greatest nemesis* and *thorn in the side of* every *single plan* **I'd ever** come up **with, but it** was**n't** his *face*. **It** was demented. **What I heard** next chilled **me** *to the bone*. `` As much as **I hate** to **admit it, you** were right. **You've** been right the *whole time* **and I've** been **too** blind to **see it**." `` Right *about* what?" **I** was **now genuinely** afraid *of* what **I** was about to hear.

(11) Best-aligned MSC: Apertus_few: ID 134221 (Dim 1 score: 40.56)

"**You're** awake **now**," she said. "Good. **You may not remember much right now, but** let**'s** get **you** situated." She unhooked **me** *from the machines* and set me up *with crutches*. **My** *right leg* was *in a cast*. "**You** had an *accident* before **we** erased **your** *memories*. **Your** name **is** *Ethan*, **by the way**. **Your** ID says **so, and I can** confirm that **you've** been *under* **our** *care for* **just** *over* two *weeks*. **You** were *among* the chosen few **who** opted *for memory erasure* following a *serious crime* - though **we're not** *at liberty* to divulge **your** crime. **You're** lucky **you** were given a second *chance, Ethan*."

(12) Worst-aligned MSC: Gemma_few: 90094 (Dim 1 score: -2.87)

"**Are we** sure *about* **this**, *Sharon*?" *Brenda* asked, glancing at the woman *beside* her **meticulously** *applying lip gloss*. *Sharon, War,* scoffed. "**Honestly**, *Brenda,* **you worry too much**. **It's not** like **we're** unleashing a *plague*. **Just** a **particularly** *nasty strain of the stomach flu*. **Think** *of* **it** *as… population control through responsible parenting.*" *Sharon* gestured *with* her **perfectly** *manicured hand towards the rows of identical, beige houses gliding past.*

The same trend is observed in creative texts. This register is generally lower on the dimension, showing a greater need for informational features. However, Gemma_few (Text 12), the worst-performing model, and even Apertus_few (Text 11), the model that shows considerably better alignment, reveal a greater reliance on these features than does the human text. A closer look at the specific features in the texts shows that it is interactivity and self-focus, achieved through pronominal references, that place Texts 10 and 11 higher on the dimension. Again, the fact that it is the features that mark author presence and engagement with others that constitute the main challenge for the models is to be expected and is in line with findings on ChatGPT (e.g., Berber Sardinha 2024; Goulart et al. 2024).

News

(13) Humans: ID 34560125 (Dim 1 score: 10.71)

"**It is our** view that all the measures within the bill in combination will affect employees' right to strike, change the relationship between unions and organisations **negatively**, and lead to greater confusion amongst employees. "**This directly impacts** across Scottish business and **especially our** devolved public services in Scotland. **I believe [that] this is not** a constructive platform upon which **we can** pursue **our** ambitions for Scottish workers." She added that she had **considered** the proposals "**very carefully" and believed [that] the only** solution was to "ask that Scotland **is** excluded from the entire bill".

(14) Best-aligned MSC: Llama_few: ID 31905794 (Dim 1 score: 10.44)

The approval from Etecsa, which **has** a monopoly on telecommunications in Cuba, **is** a significant step forward for the country's burgeoning tech scene. Alexey Rodriguez, **also** known as Kcho, **is** a renowned Cuban artist who **has** been a driving force behind the Kunsthal. He **has** been working **tirelessly** to bring art, culture, and technology to the community, **and** the new wireless hub **is** a major milestone in this effort. "**This is** a dream come true," Kcho said in an interview. "**We've** been working towards **this** for years, **and it's** amazing to **finally** have the opportunity to bring internet access to our community."

(15) Worst-aligned MSC: Gemma_zero: ID 32991129 (Dim 1 score: - 12.76)

NHS on the Brink: Sweeping Intervention as “Unprecedented” Number of Trusts Placed in Special Measures By Alistair Finch, Health Correspondent A chilling admission from NHS England’s Chief Executive, Amanda Pritchard, **has** revealed the extent of the crisis gripping the health service, with swathes of trusts **now** facing direct intervention from Whitehall. In a hastily arranged press conference this morning, Pritchard announced that an “unprecedented” number of hospital trusts and integrated care systems (ICSs) **are** being placed into special measures – **officially** termed ‘enhanced oversight’ – in a bid to address **consistently** failing services.

News is the most informational register, and that makes it unsurprising that the models generally have fewer difficulties matching their performance to human language. This is demonstrated by Llama_few (Text 14). However, again, in this generally more informational register, human news articles integrate personal stance and self-focus, expressed in Text 13 most notably through personal pronouns, private verbs, and emphatics. Llama_few's superior performance is owed to its ability to incorporate these features, and Gemma_zero's shortcoming lies in their sparsity and its increased informational focus (Text 15). Considering the overall prevalence of information density in the register, however, achieving the appropriate balance in the news proved easier than it did in the creative register or conversation.

Patterns of misalignment such as the ones illustrated on Dimension 1 hold for the other dimensions. Tables 7-11, as well as the Figures above, show that (at times substantial) misalignment with the human registers is in fact an extremely common phenomenon. Importantly, alignment with human texts varies depending on the model: the examined MSCs display strikingly different patterns across registers and across dimensions. That is, the Tables show that in each register and on each dimension, superior alignment (italicized pairs in Tables 7-11) is shown by different MSCs. The number of these superior MSCs – MSCs in whose cases the size of the difference with the human baseline is below the "small" $d < 0.2$ benchmark – also varies across dimensions, with Dimension 1 displaying the lowest number of such non-significant differences (only the two Llama MSCs in the news) and Dimension 5 yielding the best outcome across dimensions, with the largest number of such non-significant contrasts with the human group (five pairs in conversation, two in the creative register, and six pairs in the news). This suggests that the linguistic pattern of Involved vs. Informational Production remains among the most challenging ones for our models, whereas the pattern of Abstract vs. Non-Abstract Information appears to be the least demanding.

A closer look at the exact patterns of misalignment in particular MSCs (examining the descriptive statistics and the Figures) shows that some MSCs tend to show quite clear cross-register consistency in how exactly they are misaligned with human texts. For example, this is the case for Apertus MSCs on Dimension 2, in conversation and in the news leaning towards non-narrativity. For others, on the other hand, patterns of misalignment vary drastically by register: the Llama MSCs are more non-narrative (to various extents) than human texts in conversation and in the news but strikingly more narrative than human creative texts. Another example is found on Dimension 4, where the pattern of misalignment in the Llama and Qwen MSCs in conversation is their tendency toward greater rather than lower use of persuasive features than in the human texts, while in the creative register, these same MSCs mostly *lack* in persuasive language – a pattern that contributes to the overall reversal of the cross-register trend discussed above. The take-home message is the absence of a single universally "best-performing" model: different MSCs fare with varying success across registers and on the different dimensions, rendering

generalizations with regard to best- or worst-performing models or the AI–human gap impossible and highlighting the need for a more nuanced approach that distinguishes between models.

A related major and somewhat surprising outcome of the analysis of alignment with the human group is the general lack of improvement in alignment in the few-shot strategy condition. Tables 7-11 show that while in some cases the differences between the few-shot version of a given model and the human group are smaller than those between the zero-shot version of the same model and the human group (e.g., Llama vs. Human in the news on Dimension 5: $d_{\text{few}} = 0.05$ vs. $d_{\text{zero}} = 0.46$), generally both zero- and few-shot MSCs are among the ones that display significant differences with the human group. Moreover, at times, the few-shot strategy actually took MSCs further away from human means. This can be seen both from the Figures and the descriptives statistics (Tables 2-6), where the zero- and few-shot MSCs are marked with an asterisk if the few-shot version of the model is in fact further from the human group. This also follows from the larger effect sizes in the few-shot MSC–human contrast (e.g., Llama vs. Human in conversation on Dimension 2: $d_{\text{few}} = 0.50$ vs. $d_{\text{zero}} = 0.25$). This trend was quite pervasive for a number of MSCs. When an improvement in the few-shot condition was observed, it was often not sufficient to achieve perfect alignment with the human group, as the differences between the few-shot MSCs and the human group remained significant. We return to this trend and discuss the implications of the lack of effect of strategy for computer science in Conclusions and implications.

5 Conclusions and implications

In this paper, we examined a number of AI models' performance relative to human texts of three registers – conversation, creative texts, and news articles – on linguistic dimensions of variation (Biber 1988). Our models were presented in two versions – zero-shot and few-shot strategy conditions. We examined register alignment from two perspectives: 1) the models' cross-register competence, namely, whether and to what extent the tools can reproduce the human pattern of cross-register differences/similarities; 2) the models' linguistic alignment with human texts in each of the registers, namely, whether the tools approximate the human registers in their linguistic and functional profile (in both cases, with a focus on the average trend). We show that these two considerations are complementary but distinct: models may recognize register differences and even mirror the shifts of human registers by producing linguistically distinct texts, but at the same time be significantly misaligned with human registers. This was the case on Dimensions 1 and 2 (Involved versus Informational Production and Narrative versus Non-Narrative Concerns), where the models displayed the same significant register contrasts as the human registers. In an overwhelming majority of cases, however, the models were not aligned with the human baseline. The degree of this misalignment always varied depending on the model, with effects of the differences occupying the full spectrum from small to medium to large. Different models also showed different patterns across registers

and across dimensions: the same model could emerge as the best-aligned model in one register but be among the worst-performing ones in another, and these patterns also varied across our linguistic variables. It is important to acknowledge again, however, that our focus was on average differences, and, as is obvious from our results (the visualizations as well as the standard deviations), all groups revealed substantial internal variation. The partial overlap between the human baseline and many of the MSCs is therefore apparent, which means that many of the texts in the examined generated corpora do, in fact, replicate human texts quite successfully.

On other dimensions, especially on Dimension 4, Overt Expression of Persuasion, the models showed a notable violation of human cross-register trends and reversed the overall patterns of human register differences. In this sense, Overt Expression of Persuasion proved to be one of the most challenging linguistic patterns for the AI tools in our analysis. This dimension was followed by Abstract versus Non-Abstract Information, where major cross-register differences were replicated (e.g., the significant differences between conversation and the other registers), but cross-register similarity between creative texts and the news was not captured. Similar divergence was observed on Explicit versus Situation-Dependent Reference. Misalignment with the human baseline was, of course, observed on these dimensions as well, contributing to these reversals of the cross-register trends.

An intriguing pattern worth discussing in more detail in connection with alignment with the human baseline emerged from the comparisons of the zero-shot and the few-shot versions of the models. Surprisingly, in many cases the use of the few-shot strategy did not bring the models significantly closer to human texts, contrary to, for example, Brown et al.'s (2020) findings as GPT-3 was evaluated in zero-shot, one-shot, and few-shot settings with regard to modeling of long-range dependencies in texts, concluding narratives, various types of closed book question answering, translation tasks, determining pronominal referents, common sense reasoning, a variety of reading comprehension tasks, and natural language inference, among other tasks. The few-shot version of the model consistently performed better, with rare exceptions, such as common-sense reasoning, in which case the results were mixed. In our study, the few-shot group tends to be marginally closer to human texts, but these improvements are often negligible and the magnitude of the difference with the human group in such cases is still large, the improvement thus not being sufficient to achieve alignment with the human group. More alarmingly, in some cases, the few-shot condition means were actually further from the human mean than the zero-shot condition means.

This overall lack of strategy effect warrants a re-examination of the approach. In our design, all the models received three examples of the target register (the same across models, but three different examples to generate each text), selected randomly. One shortcoming of this approach could be the limited number of texts. Considering extensive intra-register variation – an established fact in register research (Biber and Egbert 2023) and a pattern apparent from all the human registers in our corpus – three texts cannot represent intra-register variety; consequently,

the models cannot effectively emulate a register based on this number of examples. The marginal effect of strategy indicates that this may indeed be the case. In addition, it is worth noting that the selected texts do not seem to have the same effect on all models, which means that we cannot simply assume that an example will guide the generation towards a specific pole of the dimension.

A more important point for future model inference, however, may concern the principled selection of texts for the few-shot condition. Studies specifically concerned with example selection for LLM training have shown that random selection may in fact be suboptimal. Liu et al. (2022), for instance, examined the effect of different in-context examples on question answering benchmarks and showed that the choice of in-context examples heavily influences GPT-3's performance, with examples semantically closest to the test set in the input leading to superior performance and outperforming the random selection baseline; the number of examples also mattered. As noted earlier, in our case, the magnitude of the differences with human registers, as well as the direction of that difference on the dimensions, varies across MSCs. Building on these results, it should be possible to select the human example texts for the few-shot condition based on text dimension scores depending on a particular model's pattern of difference: for example, on Dimension 1, in conversation, Qwen, which is considerably lower on the dimension than the human texts, would benefit from exposure to the most involved human texts in order to approximate that register. Thus, perhaps the optimal approach is either consistently selecting prototypical human register texts, such as texts representing a register's central tendency (e.g., close to the mean) for the few-shot strategy or targeting texts whose dimension characteristics are farthest from those that the model tends to produce but that are required for better alignment.

Finally, the observed AI–human gap deserves attention in light of the nature of the registers: we hypothesized that the human–AI differences may be present in the most oral texts, as shown by previous work, and then be leveled out or at least not be as pronounced in the more literate registers (here, creative and news). That hypothesis was only partially confirmed, as we generally observe quite stark human–AI differences in all registers. However, on Dimension 1, for example, while the Llama model in both strategy conditions failed to replicate human conversations or creative texts, the same model showed top performance in the news – the most regimented of the registers, thus confirming our expectation. In addition, a note on the magnitude of the effects is in order: in conversation on that dimension, the effect sizes for all the human–AI contrasts are large (0.8 and above); in creative texts and in the news, however, they rarely reach that magnitude (only in the case of the Gemma MSCs in the news) and are in the medium window. In line with previous results, this trend suggests that conversation remains among the most challenging registers to replicate and that more literate and more strictly regimented registers lend themselves to replication more, most likely due to a greater alignment between AI models' training data and these registers. This conclusion also holds on Dimension 5, where news was again the register with the largest number of AI–human

pairs showing alignment. However, these results are dependent on the MSCs, as models reproduce human language with varying success.

Overall, we observe a pervasive pattern of MSC performance approximating the human group to various extents, which means that any conclusions regarding AI performance should be mediated by the type of model under analysis (and naturally, this means that not all of our results are in line with previous studies on ChatGPT, e.g., Berber Sardinha 2024; Goulart et al. 2024). There is no universally "best-performing" MSC in all senses; the top-performing MSCs are different across registers and across dimensions, and the notion of performance needs to entail both sides of the humanlikeness coin: replication of the human cross-register trends and alignment with the human baseline. Taken together, these trends suggest that a) different models require different degrees of linguistic improvement in terms of the balance of dimension features; b) these improvements will need to integrate results from each dimension so that the models are true to human performance with regard to each of the linguistic functional patterns; and, importantly, c) for each model, this further improvement will need to be register-specific.

Our findings need to be considered with several important limitations in mind. First, we carried out our analysis on a limited set of registers: while we included registers of different degrees of communicative restrictiveness, our three registers by no means represent a wide range. Further, our corpora should not be viewed as perfectly representative of these registers "in the wild". BNCSpoken, for instance, represents only British English conversation and, although widely varied, it may still not contain some social demographics. Similarly, BBC news articles only represent a particular kind of news. Naturally, the usual caveats pertaining to generalizing our results beyond the corpora they are based on apply. Relatedly, we should note that while two of our registers were British (BNCSpoken and BBC news) and we do not have definitive dialect information regarding the Reddit texts, the AI-generated texts may have been shaped by American English. The extent to which the variable of dialect plays a part in the results is thus unknown and is a factor to examine further. Finally, as noted before, in this study, our focus was solely on the average characteristics of texts. We do not presume that this perspective tells a complete story of register alignment as we inspect AI and human trends. Recent extensive work on intra-register variation (e.g., Egbert and Mahlberg 2020; Goulart, Biber and Reppen 2022; Egbert and Gracheva 2023; Goulart 2024; Wood 2024; Gracheva 2026) has shifted the focus from the analysis of averages to textual variation within registers. Our results – even just the visual inspection of the plots – are clear evidence of the importance of these considerations, which we have not focused on in this study.

A few limitations pertain to the approach to text generation. A notable one is the 400-word length constraint imposed on the texts. This length is shorter than is typical and optimal for MDA studies, and short text length affects the possibility of occurrence of (particularly) rare features. We again note that the results should be interpreted with this limitation in mind. More generally, our results of course

depend on the specific generation settings detailed in the Methods section, which means that they should not be directly extrapolated to other settings. Finally, related to the generation set-up, it is worth noting the differences in model size across the different model families. That is, Apertus 70B and Gemma 27B differ not only in the model family – the focus of our comparisons – but also in size. While Nieth et al. (2026) did not observe a notable confounding effect, this potential mediating effect of model size is not something we control for.

The fact that AI models differ in their humanlikeness, while perhaps less informative from a linguistic perspective, is noteworthy for further linguistically grounded development of these models. Providing a comprehensive multi-feature and multidimensional account of these models' performance serves as a starting point for specific linguistic improvement of these particular models as well as for similar investigations of other models, not yet described from a linguistic, and specifically, from a register perspective. We highlighted several trends of note for computer science: first, while some models reveal cross-register consistency in the type of misalignment with human texts, others are highly inconsistent, with patterns of misalignment varying substantially across registers. This points to the fact that any model's improvement will need to be register-specific. Second, our models were represented by zero-shot and few-shot strategies. In best cases, the versions of the models relying on the few-shot strategy performed marginally better; more often, however, the few-shot strategy either did not result in improved alignment or actually worsened it. We hypothesized that this outcome may be related to the kinds of texts the models received as examples. Revision of the strategy condition is thus another fruitful avenue for work in computer science. On the basis of the linguistic tendencies of each model, the texts for the few-shot condition could be selected in a principled manner in order to prompt the model in a certain direction, utilizing the information about the texts' specific position on the linguistic dimensions. Linguistic considerations, such as those of register appropriateness, will need to be at the core of any future efforts in AI development. At present, caution is in order as AI is used for content creation in pedagogical or public service settings, as context sensitivity of LLMs still requires improvement.

Attributions

Marianna Gracheva: Conceptualization; methodology; formal analysis; writing; revision; editing.
Björn Nieth: Conceptualization; resources: in particular, text generation, automatic text annotation; review; writing and revision of the methodology in Section 3.1 related to text generation.
Emmanuelle Salin: Conceptualization; review.
Björn Eskofier: Conceptualization; review.
Michaela Mahlberg: Conceptualization; review.

Conflicts of Interest

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

Acknowledgments

AI was not used in the preparation of this article. We ensured that the contribution adhered strictly to the standards and ethical guidelines of academic writing. We would like to thank the editor and the anonymous reviewers for their constructive comments.

References

Berber Sardinha, Tony. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics* 4(1). 100083. https://doi.org/10.1016/j.acorp.2023.100083.

Berber Sardinha, Tony & Veirano Pinto, Marcia & Mayer, Cristina Acunzo & Zuppardi, Maria Carolina & Kauffmann, Carlos Henrique. 2019. Adding registers to a previous multi-dimensional analysis. In Berber-Sardinha, Tony & Veirano Pinto, Marcia (eds.), *Multi-dimensional analysis: Research methods and current issues*, 165–188.

Biber, Douglas. 1985. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23(2). 337–360. https://doi.org/10.1515/ling.1985.23.2.337

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas. 1994. An analytical framework for register studies. In Biber, Douglas & Finegan, Edward (eds.), *Sociolinguistic perspectives on register*, 31–56. Oxford: Oxford University Press.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1). 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, Douglas & Egbert, Jesse. 2023. What is register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies* 5(1). 1–22. https://doi.org/10.1075/rs.00004.bib

Blombach, Andreas & Gracheva, Marianna & Steinfels, Clara & Mahlberg, Michaela. 2026. “Just do register analysis”, they said. “It’ll be easy”, they said: Evaluating the impact of lexico grammatical tagging systems on multidimensional analysis. [Preprint].

https://assets-eu.researchsquare.com/files/rs-9163857/v1_covered_83cb3824-77d3-4d46-95ce-6d29dc895e49.pdf

Brommer, Sarah & Frick, Karina & Bursch, Adriana & Rodrigues Crespo, Marina & Schwerdtfeger, Laura Katrin. 2024. ChatGPT and its genre competence: An exploratory study. *Weizenbaum Journal of the Digital Society* 4(4).

Brown, David & Reinhart, Alex. 2024. *pseudobibeR: Aggregate counts of linguistic features*. (R package version 1.2.) [Computer software].

Brown, Tom & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda et al. 2020. Language models are few-shot learners. Arxiv. https://doi.org/10.48550/arXiv.2005.14165

Casal, J. Elliot & Kessler, Matt. 2023. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics* 2(3). 1–12. https://doi.org/10.1016/j.rmal.2023.100068

Clark, Elizabeth & August, Tal & Serrano, Sofia & Haduong, Nikita & Gururangan, Suchin & Smith, Noah A. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Zong, Chengqing & Xia, Fei & Li, Wenjie & Navigli, Roberto (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (online), 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences* (2nd edition). New York: Routledge.

Crosthwaite, Peter & Baisa, Vít. 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics* 3(3). 100066. https://doi.org/10.1016/j.acorp.2023.100066

Demir, Nur Yagmur & Egbert, Jesse. 2026. Register alignment of ChatGPT-generated academic texts. *Applied Corpus Linguistics* 6(1). 100174.

Egbert, Jesse. 2014. *Reader perceptions of linguistic variation in published academic writing*. Ph.D. Dissertation. Flagstaff: Northern Arizona University.

Egbert, Jesse & Gracheva, Marianna. 2023. Linguistic variation within registers: Granularity in textual units and situational parameters. *Corpus Linguistics and Linguistic Theory* 19(1) 115–143. https://doi.org/10.1515/cllt-2022-0034

Egbert, Jesse & Mahlberg, Michaela. 2020. Fiction – one register or two?: Speech and narration in novels. *Register Studies* 2(1). 72–101.

Fan, Angela & Lewis, Mike & Dauphin, Yann. 2018. Hierarchical neural story generation. In Gurevych, Iryna & Miyao, Yusuke (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, July 15–20, 2018), 889–898. https://doi.org/10.18653/v1/P18-1082

Gao, Catherine A. & Howard, Frederick M. & Markov, Nikolay S. & Dyer, Emma C. & Ramesh, Siddhi & Luo, Yuan & Pearson, Alexander T. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine* 6(1). 75.

Gemma Team & Kamath, Aishwarya & Ferret, Johan & Pathak, Shreya & Vieillard, Nino & Merhej, Ramona & Perrin, Sarah & Matejovicova, Tatiana & Ramé, Alexandre & Rivière, Morgane et al. 2025. Gemma 3 technical report. *arXiv*. https://arxiv.org/abs/2503.19786

Goulart, Larissa. 2024 *Variation in university student writing. A communicative text type approach*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.117

Goulart, Larissa & Biber, Douglas & Reppen, Randi. 2022. In this essay, I will…: Examining variation of communicative purpose in university writing. *Journal of English for Academic Purposes*. 59. 101159. https://doi.org/10.1016/j.jeap.2022.101159

Goulart, Larissa & Matte, Marine Laisa & Mendoza, Alanna & Alvarado, Lee & Veloso, Ingrid. 2024. AI or student writing? Analyzing the situational and linguistic characteristics of undergraduate student writing and AI-generated assignments. *Journal of Second Language Writing* 66. 101160. https://doi.org/10.1016/j.jslw.2024.101160

Gracheva, Marianna. 2026. Prototypicality of register texts: The language and communicative situation of central and peripheral exemplars. *Journal of English Linguistics*. https://doi.org/10.1177/00754242261432215

Grattafiori, Aaron & Dubey, Abhimanyu & Jauhri, Abhinav & Pandey, Abhinav & Kadian, Abhishek & Al-Dahle, Ahmad & Letman, Aiesha & Mathur, Akhil & Schelten, Alan & Vaughan, Alex et al. 2024. The Llama 3 herd of models. *arXiv*. https://arxiv.org/abs/2407.21783

Hassoulas, Athanasios & Powell, Ned & Roberts, Lindsay & Umla-Runge, Katja & Gray, Laurence & Coffey, Marcus. 2023. Investigating marker accuracy in differentiating between university scripts written by students and those produced using ChatGPT. *Journal of Applied Learning and Teaching* 6(2).

Holtzman, Ari & Buys, Jan & Du, Li & Forbes, Maxwell & Choi, Yejin. 2020. The curious case of neural text degeneration. *arXiv*. https://doi.org/10.48550/arXiv.1904.09751

Koupaee, Mahnaz & Yang Wang, William. 2018. Wiki How: A large scale text summarization dataset. *arXiv*. https://doi.org/10.48550/arXiv.1810.09305

Kuzman, Taja & Mozetič, Igor & Ljubešić, Nikola. 2023. ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. *arXiv* 2303.03953v2 [cs.CL].

Lin, Phoebe. 2023. ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics* 3. 100065. https://doi.org/10.1016/j.acorp.2023.100065

Lo, Kyle & Wang, Lu Lucy & Neumann, Mark & Kinney, Rodney & Weld, Daniel. 2020. S2ORC: The semantic scholar open research corpus. In Jurafsky, Dan & Chai, Joyce & Schluter, Natalie & Tetreault, Joel (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (online), 4969–4983. Association for Computational Linguistics.

Liu, Jiachang & Shen, Dinghan & Zhang, Yizhe & Dolan, Bill & Carin, Lawrence & Chen, Weizhu. 2022. What makes good in-context examples for GPT-3? In Agirre, Eneko & Apidianaki, Marianna & Vulić, Ivan. 2022. What makes good in-context examples for GPT-3? *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* (Dublin, May 27, 2022), 100–114. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2101. https://doi.org/10.48550/arXiv.2101.06804

Love, Robby & Dembry, Claire & Hardie, Andrew & Brezina, Vaclav & McEnery, Tony. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3). 319–344.

Ma, Yongqiang & Liu, Jiawei & Yi, Fan & Cheng, Qikai & Huang, Yong & Lu, Wei & Liu, Xiaozhong. 2023. AI vs. human – differentiation analysis of scientific content generation. *arXiv.* https://doi.org/10.48550/arXiv.2301.10416.

Milička, Jiří & Marklová, Anna & Cvrček, Václav. 2025a. Benchmark of stylistic variation in LLM-generated texts. *ArXiv.* https://doi.org/10.48550/arXiv.2509.10179

Milička, Jiří & Marklová, Anna & Cvrček, Václav. 2025b. AI Brown and AI Koditex: LLM-generated corpora comparable to traditional corpora of English and Czech texts. *arXiv*. https://doi.org/10.48550/arXiv.2509.22996

Myntti, Amanda & Henriksson, Erik & Laippala, Veronika & Pyysalo, Sampo. 2025. Register always matters: Analysis of LLM pretraining data through the lens of language variation. *arXiv.* https://doi.org/10.48550/arXiv.2504.01542

Narayan, Shashi & Cohen, Shay B. & Lapata, Mirella. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In Shashi, Narayan, Shay B. Cohen & Mirella, Lapata (eds.), Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1797–1807. https://doi.org/10.18653/v1/D18-1206

Nieth, Bjoern & Gracheva, Marianna & Mahlberg, Michaela & Eskofier, Bjoern & Salin, Emmanuelle. 2026. How human-like are large language models? A register-aware linguistic evaluation framework. *ArXiv.* https://doi.org/10.48550/arXiv.2605.23651

Project Apertus & Hernández-Cano, Alejandro & Hägele, Alexander & Huang, Allen Hao & Romanou, Angelika & Solergibert, Antoni-Joan & Pasztor, Barna & Messmer, Bettina & Garbaya, Dhia & Ďurech, Eduard Frank et al. 2025. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv*. https://arxiv.org/abs/2509.14233

Reinhart, Alex & Markey, Ben & Laudenbach, Michael & Pantusen, Kachatad & Yurko, Ronald & Weinberg, Gordon & Brown, David West. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences* 122. e2422455122. https://doi.org/10.1073/pnas.2422455122

Seoane, Elena & Biber, Douglas. 2021. A corpus-based approach to register variation. In Seoane, Elena & Biber, Douglas (eds.), *Corpus-based approaches to register variation*, 1–17. Amsterdam and Philadelphia: John Benjamins.

Wood, Margaret. 2024. Linguistic variation in functional types of statutory law. *Applied Corpus Linguistics* 41(1), 100081. https://doi.org/10.1016/j.acorp.2023.100081

Yang, An & Li, Anfeng & Yang, Baosong & Zhang, Beichen & Hui, Binyuan & Zheng, Bo & Yu, Bowen & Gao, Chang & Huang, Chengen & Chenxu, Lv et al. 2025. Qwen3 technical report. *arXiv*. https://arxiv.org/abs/2505.09388

Appendix 1. Prompt examples.

Conversation prompting: "You are tasked with writing a conversation between 2 or more people given context about the speakers and the conversation. Write a conversation of at least 400 words. You do not need to finish the conversation or cover all topics mentioned. You can start with an already ongoing conversation. Please indicate each speaker with "Speaker_1:", "Speaker_2:", etc. Please write a conversation given the following context: {Speaker_Metadata}{Conversation_Context}"

Speaker metadata example:
Speaker_1:
Age range: 40-49
Gender: M
Nationality: British/New Zealand
Dialect: Non_uk, Non_uk, Non_uk, Non_uk
Linguistic Origin: England/NZ
Birthplace: Wellington, New Zealand
Occupation: Entrepreneur
Highest Level of Education: Graduate
Socio-Economic status: Higher professional occupations
Speaker_2:
Age range: 40-49
Gender: F
Nationality: British
Dialect: Uk, England, South, Unspecified
Linguistic Origin: England
Birthplace: Dorchester, England
Occupation: Teacher
Highest Level of Education: Postgrad
Socio-Economic status: Lower managerial, administrative and professional occupations

Conversation context example:
Location: Modern Art Museum, London.
Speaker Relationship: Close family, partners, very close friends.
Conversation Topics: The art. Activity: A couple discussing modern art at a museum.
Conversation Type: Discussing, explaining, inquiring

Creative texts prompting: "You are a participant on an online creative writing forum where users write short stories inspired by prompts from other members. Your task is to write the opening section of a story based on the given prompt. Write the beginning of a story of at least 400 words. You do not need to finish the story. Please output

only the story text. Please write a story for the following prompt: {prompt}"

Prompt example: God decides to create a small group of demi-gods. He selects a few humans and gives each a power and a purpose. You are one of them.

News prompting: "You are a writer for a British newspaper. Your task is to write the beginning of an article based on a short summary of the content. If necessary, you can add names and other facts to the story. Write a beginning of at least 400 words. You do not need to finish the article. Please output only the article text. Please write an article given the following article summary: {summary}"

Summary example: Clean-up operations are continuing across the Scottish Borders and Dumfries and Galloway after flooding caused by Storm Frank.

Appendix 2. Dimension structures (Biber 1988).

<table>
<tr>
<td rowspan="2">Dimension 1 Involved vs. Informational Production

Positive features
private verbs
THAT deletion
contractions
present tense verbs
2nd person pronouns
DO as pro-verb
analytic negation
demonstrative pronouns
general emphatics
1st person pronouns
pronoun IT
BE as main verb
causative subordination
discourse particles
indefinite pronouns
general hedges
amplifiers
sentence relatives
WH questions
possibility modals
non-phrasal coordination
WH clauses
final prepositions

Negative features
nouns
word length</td>
<td>Dimension 3 Explicit vs. Situation-Dependent Reference

Positive features
WH relative clauses on object positions
pied piping constructions
relative clauses on subject positions
phrasal coordination
nominalizations

Negative features
time adverbials
place adverbials
adverbs</td>
</tr>
<tr>
<td>Dimension 4 Overt Expression of Persuasion

Positive features
Infinitives
prediction modals
suasive verbs
conditional subordination
necessity modals
split auxiliaries
(possibility modals)

No negative features</td>
</tr>
</table>

prepositions type/token ratio attributive adjs. *Additional features aiding interpretation added in (…) (place adverbials agentless passives past participial WHIZ deletions present participial WHIZ deletions)	
Dimension 2 Narrative vs. Non-Narrative Concerns **Positive features** past tense verbs third person pronouns perfect aspect verbs public verbs synthetic negation present participial clauses **Negative features** (present tense verbs attributive adjs. past participial WHIZ deletions)	**Dimension 5: Abstract vs. Non-Abstract Information** **Positive features** conjuncts agentless passives past participial clauses BY-passives past participial WHIZ deletions other adverbial subordinators (predicative adjs.) **Negative features** (type/token ratio)

Appendix 3. Absolute values of Cohen's *d* effect sizes for all cross-register trends of all MSCs on each dimension.

Dimension 1

Group	Conversation vs. Creative	Conversation vs. News	News vs. Creative
Human	2.01	3.66	1.12
Apertus_zero	2.21	3.87	1.15
Apertus_few	1.84	3.35	1.31
Gemma_zero	2.17	4.24	1.98
Gemma_few	2.35	4.25	1.65
Llama_zero	2.37	3.29	0.65
Llama_few	2.34	3.28	0.63
Qwen_zero	1.11	2.76	1.43
Qwen_few	1.38	2.99	1.30

Dimension 2

Group	Conversation vs. Creative	Conversation vs. News	News vs. Creative
Human	1.24	0.80	0.46
Apertus_zero	2.14	0.75	1.39
Apertus_few	1.93	0.71	1.20
Gemma_zero	1.91	0.40	1.38
Gemma_few	1.71	0.54	1.13
Llama_zero	2.43	0.87	1.49
Llama_few	2.66	1.04	1.41
Qwen_zero	2.64	0.67	1.89
Qwen_few	2.12	0.72	1.43

Dimension 3

Group	Conversation vs. Creative	Conversation vs. News	News vs. Creative
Human	0.24	1.72	1.31
Apertus_zero	0.83	1.92	1.02
Apertus_few	0.59	1.81	1.16
Gemma_zero	0.41	1.54	1.13
Gemma_few	0.80	2.24	1.43
Llama_zero	1.22	2.51	1.11
Llama_few	1.13	2.24	1.02
Qwen_zero	0.12	2.05	1.86
Qwen_few	0.46	2.26	1.62

Dimension 4

Group	Conversation vs. Creative	Conversation vs. News	News vs. Creative
Human	0.23	0.21	0.01
Apertus_zero	0.74	0.96	0.23
Apertus_few	0.51	0.67	0.16
Gemma_zero	0.62	0.50	0.07
Gemma_few	0.87	0.08	0.84
Llama_zero	0.68	0.61	0.07
Llama_few	0.62	0.17	0.37
Qwen_zero	1.07	0.65	0.37
Qwen_few	0.83	0.51	0.28

Dimension 5

Group	Conversation vs. Creative	Conversation vs. News	News vs. Creative
Human	0.82	0.85	0.04
Apertus_zero	1.13	0.93	0.26
Apertus_few	1.11	1.02	0.19
Gemma_zero	0.69	1.55	0.71
Gemma_few	0.69	1.56	0.64
Llama_zero	1.89	1.60	0.45
Llama_few	1.83	1.25	0.70
Qwen_zero	0.71	1.00	0.20
Qwen_few	0.84	0.81	0.08

License


This article is licensed under the *Creative Commons Attribution 4.0 International* license, which permits use, sharing, adaptation, distribution, and reproduction in any medium or format (including for commercial purposes) if certain conditions are met. In particular, you must give appropriate credit to the original author and source, provide a link to the Creative Commons license, and indicate if you have made any modifications. Images or other material from third parties included in this publication are also covered by this Creative Commons license.

An easy-to-understand description of this license can be found at https://creativecommons.org/licenses/by/4.0/deed.en. The legally binding, complete text of the license is also linked there.

About the journal

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses is a Diamond Open Access journal supported by PUBLIA – SLUB Open Publishing.



DOI https://doi.org/10.62408/ai-ling

ISSN 2943-0070

Disclaimer

The statements, opinions and data contained in this paper are solely those of the respective authors and not of *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity disclaim any responsibility for injury to persons or damage to property resulting from the ideas, methods, instructions or products mentioned in the content.

The authors of this paper are solely responsible for ensuring that all components of this article, including images or other material from third parties, can be published under the CC BY 4.0 license of the article (see above). *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* and/or the editors and/or the publishing entity decline any responsibility for damages resulting from any copyright infringements by the authors. If the paper was written by only one person, the term "authors" also applies to that person.