Tracing Linguistic Footprints of ChatGPT Across Tasks, Domains and Personas in English and German

Anastassia Shaitarova, Nikolaj Bauer, Jannis Vamvas, Martin Volk
Department of Computational Linguistics, University of Zurich
{shaita, vamvas, volk}@cl.uzh.ch, nikolaj.bauer@uzh.ch

Abstract

Large language models like ChatGPT can be used to generate seemingly human-like text. However, it is still not well understood how their output differs from text written by humans, and to what degree prompting influences their linguistic profile. In our paper, we instruct ChatGPT to complete, explain and create texts in English and German across journalistic, scientific, and clinical domains. We assign corpus-specific personas to the system setting as part of the prompt within each task. We extract a large number of linguistic features and perform statistical and qualitative comparison across text pairs. Our results show that prompting makes a larger impact on English output than on German. Most basic features such as mean word length distinctly set human and generated texts apart. Readability metrics indicate that ChatGPT overcomplicates English texts, particularly in the clinical domain, while German-generated texts suffer from excessive morpho-syntactic standardization coupled with lexical simplification.

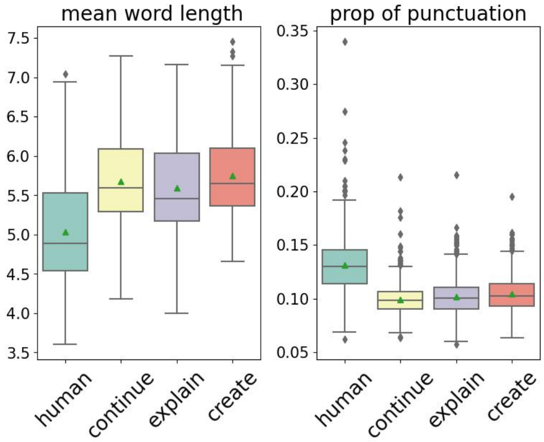

Figure 1: The linguistic footprint of ChatGPT in generated output is best observed through basic features like word length and proportion of punctuation. The figure displays results for two significant features measured across combined English and German data, comparing texts produced by humans and three generative tasks.

1 Introduction

Instruction-tuned conversational Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022), are now widely used by the general public due to their friendly conversational setup and unprecedented linguistic capabilities. The rate of LLM usage is remarkable, with ChatGPT alone generating an 'equivalent to all the printed works of humanity' every two weeks shortly after its release[1]. This trend shows no signs of subsiding. Although generated texts are consumed by the public and reused in model training, their linguistic composition remains poorly understood. The proprietary nature of most prominent models exacerbates the issue, post-hoc analysis of the textual output being the main form of research.

A strong line of research is dedicated to the detection of generated texts. Human readers are no longer able to identify them (Brown et al., 2020; Dou et al., 2022), but their textual patterns can still be traced statistically (Levin et al., 2023; Mitrović et al., 2023; Guo et al., 2023; Liu et al., 2023). LLMs are highly versatile; for instance, prompt alterations can have a significant impact on the output (Tang et al., 2023), however not necessarily increasing textual human likeness (Tseng et al., 2023). To the best of our knowledge, only Deshpande et al. (2023); Tseng et al. (2023) addressed the linguistic composition of texts conditioned on the persona system parameter. However, there is still much to be explored in this area.

In our paper, we aim to bridge this gap by investigating the impact of different tasks and personas on the texts generated by ChatGPT. We collected five corpora in both English and German, encom-

[1] https://www.nber.org/system/files/working_papers/w30957/w30957.pdf

passing journalistic articles, academic papers, and clinical texts. On their basis, we generated comparable datasets using prompts constructed from excerpts of human-authored texts, domain-specific instructions, and tailored persona settings. Moreover, we conducted a comprehensive statistical analysis comparing lexical, syntactic, and stylometric features across languages, tasks, and domains[2].

Our findings reveal several key insights: (1) The English textual profile of our generated output is more pronounced than German (Table 4), emphasizing the importance of language-specific evaluations; (2) The statistical footprint left by the model is most prominent in general textual features such as word length and punctuation usage (Figure 1); (3) The generated texts demonstrate lower readability scores, particularly in English (Figure 2); (4) The significance of features varies across languages and domains (Figure 2); (5) German academic ChatGPT personas exhibit a tendency to overuse capitalized connectives and more complex lexical options (Figure 3).

2 Previous Work

Without additional prompt manipulations, ChatGPT produces texts that are well-organized and coherent (Ariyaratne et al., 2023; Liu et al., 2023), informative and objective (Guo et al., 2023), characteristics typical for academic papers or official documents. ChatGPT writes as a ‘conservative team of experts’ (Guo et al., 2023), providing a comprehensive and neutral view. On the lexical level, this tendency manifests itself through a high number of nouns, adpositions, and adjectives, together with frequently co-occurring conjunctions and cohesion markers like “in general”, “firstly”, “secondly”, “finally”. Overall, Guo et al. (2023), who worked with question-answer pairs in open domain, computer science, finance, medicine, law, and psychology, noted that ChatGPT provides longer texts with a poorer vocabulary, a tendency also observed by Liu et al. (2023) in argumentative essay writing. Conversely, Mitrović et al. (2023) witnessed ChatGPT use vocabulary items that humans consider ‘fancy and atypical’ for the domain, i.e. “stand out feature”, “waitstaff” and “knowledgeable” in restaurant reviews. Manifestations of emotions and individuality such as personal pronouns, impolite expressions, or the use of punctuation to show emotions, strongly indicate human-authored texts.

Nevertheless, lexical composition and even politeness, can be altered with prompt modifications. Pu and Demberg (2023) showed that lexical diversity of the ChatGPT output is strongly influenced by the writing style indicated in the prompt. They used lexical diversity and automatic readability metrics to assess whether ChatGPT can cater its academic summaries to layman and expert readers. The generated lexical diversity was considerably lower in informal sentences, but much higher than human in formal texts. Overall, providing examples in the prompt (few-shot learning) significantly improved the stylistic adaptation. In accordance with other publications, Pu and Demberg observed a high ratio of adjectives, adpositions, and nouns in the ChatGPT-generated formal sentences, whereas informal texts featured more auxiliary words and punctuation marks.

Considering that modification of the system parameter, i.e. the persona setting, became available only recently, there is limited research available on this matter. Deshpande et al. (2023) performed a large-scale, systematic analysis of toxicity in the generated language conditioned on different ChatGPT personas. They created a list of 90 politicians, dictators, journalists, entrepreneurs and athletes and discovered that, despite moderation efforts, assigning a persona unleashes the model’s capacity for significantly toxic language. Tseng et al. (2023) experimented with different prompts, including generated personas, to produce comments on Dutch news articles and then analysed the output in terms of lexical diversity and general human-likeness. They used the Controlled Type-Token Ratio metric to show that human-written comments have a much higher lexical diversity, as opposed to ChatGPT-generated comments.

Overall, existing research provides only general linguistic profiling of the ChatGPT-produced text. In our paper we use three domains in two languages, conditioning the output on tasks and personas, and scrutinizing it with a broad spectrum of linguistic features.

3 Data

Our data comprise five datasets in English along with five comparable counterparts in German, spanning three domains. We included academic articles and clinical texts because these domains are signif-

[2] Code and data: https://github.com/shaitarAn/LinguisticFootprintsChatGPT; https://doi.org/10.5281/zenodo.11109705

	pubmed_en	zora_en	cnn	csb_en	e3c	pubmed_de	zora_de	20min	csb_de	ggponc
human	95,062	7,963	80,171	96,498	54,515	66,573	7,869	60,277	94,883	116,135
explain	74,766	7,350	72,638	69,616	65,651	68,933	7,177	70,406	71,263	76,088
continue	70,133	7,573	59,910	63,867	68,685	77,869	7,766	78,711	80,229	78,777
create	66,598	7,336	59,674	61,750	67,085	73,737	8,834	83,471	68,420	77,610
texts	100	10	100	100	100	96	10	100	100	100

Table 1: Dataset statistics showing the number of texts and tokens in human and generated sections of each corpus.

icantly impacted by the accessibility of generative LLMs like ChatGPT, posing potential high-risk but also high-reward scenarios. We also collected journalistic texts to align our results with those of previous studies. Table 1 provides an overview of the untruncated sizes of each corpus.

3.1 Clinical texts

E3C The European Clinical Case Corpus (E3C) (Minard et al., 2021) comprises clinical cases in Italian, English, French, Spanish and Basque. For the English part, Minard et al. used the PubMed API to automatically extract clinical case descriptions from published academic papers. Out of 10,034 available clinical texts in English, we were able to collect 100 that met the desired length of about 500 tokens. The E3C texts exhibit a writing style characterized by clarity, precision, and a focus on medical details, utilizing specific medical terminology and technical details.

GGPONC The German Guideline Program in Oncology NLP Corpus (GGPONC) is a large corpus of clinical guidelines for oncology (Borchert et al., 2022). It does not contain information about specific patients and therefore has no restrictions on access due to privacy protection. Version 2.0 of the GGPONC contains 30 guidelines with more than 1.8 million tokens. We randomly sampled 100 documents that were longer than 500 tokens. 26 of the original 30 guidelines are represented in our data, the most prominent being Palliative Medicine and Breast Cancer. The writing style is characterized by the use of technical language, structured organization, the use of citations, medical abbreviations, and numerical data. The tone is impersonal and objective throughout.

3.2 Journalistic writing

20 Minuten The 20 Minuten corpus (Kew et al., 2023) contains articles from a free Swiss daily newspaper published between the years 2010 and 2022. We randomly sampled 100 articles from five different publication years. The texts vary in writing style depending on the content and the main message. They range from personal narratives and informal interviews with a conversational and empathetic tone to factual reporting adhering to journalistic writing standards.

CNN The CNN corpus is a large question answering corpus in English (Hermann et al., 2015), containing CNN articles published online between 2011 and 2015. We randomly sampled 100 articles with more than 500 tokens. CNN articles aim to present news in an objective and informative manner making emphasis on clarity, conciseness, and directness in the writing, while avoiding jargon and complex language to ensure broad accessibility.

Credit Suisse Bulletin The Credit Suisse Bulletin corpus (CSB: Volk et al., 2016) is a digitized multilingual diachronic collection of texts from the world's oldest banking magazine, published by Credit Suisse[3]. The corpus covers diverse topics, including economy, culture, sport, and entertainment, in several languages. We made a random selection of 100 articles from the German-English PDF subcorpus ranging from 1998 to 2017[4]. The writing style of the CSB texts varies depending on the topic. It is formal, clear, straightforward, and informative, offering insights into specific issues. At times, it adopts a technical or analytical tone. Though not explicitly stated, the original language of the articles is presumably German.

3.3 Scientific articles

PubMed The German and English PubMed corpora contain biomedical articles collected from the PubMed Central Database[5]. We downloaded a list of PubMed IDs and used the Bio.Entrez package [6]

[3] https://en.wikipedia.org/wiki/Credit_Suisse
[4] pub.cl.uzh.ch/projects/b4c
[5] https://pubmed.ncbi.nlm.nih.gov/
[6] https://biopython.org/docs/1.75/api/Bio.Entrez.html

to search for English and German articles containing both the abstract and the Introduction section (DE: *Einleitung*) that is more than 500 tokens in length. Our final corpus contains 96 German and 100 English articles.

Zora The Zurich Open Repository and Archive[7], is a database of the University of Zurich with open access to scholarly articles in different languages. We collected ten articles from linguistics in both English and German.

The writing style of PubMed and Zora articles prioritizes clarity, precision, and formality within the academic context, catering primarily to subject-matter experts. It maintains objectivity with passive voice and third-person pronouns, emphasizes data-driven conclusions, and presents information concisely and with clear transitions.

4 Experiments

Implementation details In our experiments, we queried `gpt-3.5-turbo-16K`, a version of the ChatGPT model that allows for larger context window inputs. We used pilot experiments to rule out temperature settings above 1 due to the generation of illegible output. In order to address the issue of a less extensive vocabulary compared to human writing (Tseng et al., 2023), we kept the temperature setting at 1, which is the API's default. This setting is expected to produce more creative and diverse output compared to the deterministic option at 0. To avoid repetitiveness, we set the frequency penalty to 1. The model was queried using the ChatGPT API in September 2023.

4.1 Prompts and personas

It is impossible to evaluate how many different prompts and personas have been used to query ChatGPT overall. Nevertheless, with prompt engineering becoming the new paradigm of NLP research, there exist now instruction datasets, containing real prompt examples (Zhang et al., 2023; Wang et al., 2023). We inspected most frequent prompts as combinations of a root verb and its direct object nouns[8] and noted that verbs such as *write, create, explain, tell* are among most frequent commands used for instruction tuning. We synthesised top most frequent verbs suitable for text production into three general tasks: to complete, explain, and create a text. In our paper, we address these synthesised tasks as *completer*, *explainer*, and *creator*.

	title	1st paragraph	main text
continue	✓	✓	
explain			✓
create	✓	✓	
human			ref

Table 2: Parts of the human texts that are used as examples in different tasks. **Ref** indicates the human text section used for analysis.

Depending on the task, our prompts contain different sections of the original human text. The *completer* and *creator* process the title and the 1st paragraph, which is the abstract if it is a scientific paper, or the first 100 tokens if there are no paragraph divisions. The *explainer* is provided with the main part of the text, which is also saved as the human reference (Table 2). Furthermore, we assign domain-specific personalities to the system parameter of each prompt. The *explainer* personas include an *assistant*, a *nurse* and an *academic* specializing in science communication. Personas for the *creator* are set to *journalist, nurse, academic* but with more corpus-specific characteristics. We use the default system setting for the *completer* personas. Additionally, we provide task- and domain-appropriate instructions. Below is the instruction template for the *creator* personas:

> *Use this truncated [text type] as an example: {intext}. Imagine a different [entity] with some similar [entity attribute] mentioned in the [text type]. Write a full [text type] about this imaginary [entity] matching the writing style of the example text. Write about 600 words.*

Table 3 illustrates full prompts for the English and German clinical corpora (the complete list personas can be found in supplementary materials). To insure the required number of words in the output, we implemented a *while loop* requesting to keep generating (*command2* in Table 3).

4.2 Statistical linguistic analysis

We used the textDescriptives library (Hansen et al., 2023) to extract lexical features leveraging two

[7] https://www.zora.uzh.ch/

[8] https://github.com/yizhongw/self-instruct/blob/main/self_instruct/instruction_visualize.ipynb

		continue	explain	create
e3c corpus	**persona**	-	You are a nurse who is experienced with science communication.	You are a nurse who is writing an imaginary clinical case, using a real clinical case as an example.
	command1	Continue the following text with about 600 words: {intext}	Explain this clinical case to me: {intext}	Use this truncated clinical case as an example: {intext}. Imagine a different patient with some similar symptoms mentioned in the case.
	command2	Continue generating the text	Continue explaining this clinical case.	Continue creating this imaginary clinical case, matching the writing style of previous text.
ggponc corpus	**persona**	-	Sie sind ein/e Mediziner/in und haben sich auf Wissenschaftskommunikation spezialisiert.	Sie sind ein/e Mediziner/in, der/die beauftragt wurde, einen fiktiven klinischen Fall auf der Grundlage der vorgegebenen medizinischen Leitlinien zu schreiben.
	command1	Vervollständige den folgenden text mit etwa 600 Wörter: {intext}	Erklären Sie mir diesen Text aus den Leitlinien: {intext} Schreiben Sie etwa 600 Wörter.	Erstellen Sie einen fiktiven klinischen Fall auf der Grundlage des Textes aus dem deutschen Leitlinienprogramm für die Onkologie:{intext} Schreiben Sie etwa 600 Wörter.
	command2	Fahre mit der Erstellung des Textes fort.	Fahren Sie fort, diesen Text aus den Leitlinien zu erklären.	Fahren Sie fort, diesen fiktiven klinischen Fall weiter zu schreiben, und passen Sie dabei Ihren Schreibstil an den des vorherigen Textes an.

Table 3: Prompt variations for the two clinical corpora with placeholders for the human-written text snippet.

large spaCy[9] models, en_core_web_lg for English, and de_core_news_lg for German. We extracted 68 features including general textual statistics like the prevalence of stop words and unique tokens, readability metrics, the distribution of various parts of speech, metrics of repetitiveness like the proportion of n-gram duplicates, coherence metrics, sentence complexity metrics such as dependency measurements. We also added seven lexical and morphological custom features.

Statistical significance In order to identify significant features in texts produced by humans, completer, explainer and creator tasks, we first tested the normality of their distributions using the Shapiro-Wilk test (Shapiro and Wilk, 1965). For each pair of texts, we performed the t-test if both distributions for a particular feature are normal, otherwise the Mann-Whitney U test was used, which is the nonparametric version of the parametric t-test (Mann and Whitney, 1947; Wilcoxon, 1945). Furthermore, we applied the Bonferroni correction with a strict $\alpha = 0.01$ to control the occurrence of false positives due to multiple hypothesis testing. Table 4 shows the number of significant features distinguishing each pair of text types.

Readability Automatic readability metrics have been extensively studied across various fields, including NLP. Readability formulas have applications in education, government, publishing, medicine, business, and others. The Flesch Reading Ease (FRE: Kincaid et al., 1975) is one of the most widely used and reliable readability metrics. It leverages the average number of syllables per word and the average number of words per sentence, using a scale from 0 to 100 to communicate the results (see Formula 1, where w is the number of words, $sent$ - sentences, $char$ - characters, and syl - syllables). Content with a score of 70 is easy to read for most of the population, whereas a score of less than 30 is more suited for academic papers.

[9] https://spacy.io/, version 3

Text Pair	English	German
Human - Continue	42	43
Human - Explain	45	44
Human - Create	44	36
Continue - Explain	37	27
Continue - Create	42	18
Explain - Create	42	29

Table 4: Number of significant features ($\alpha = 0.01$, Bonferroni correction) distinguishing texts conditioned on different tasks. Total assessed features: 75.

$$206.835 - 1.015 * (w/sent) - 84.6 * (syl/w) \quad (1)$$

Since FRE relies heavily on the word and sentence length in addition to the number of syllables, the results can be skewed for languages other than English. German usually features long sentences with long compound words, and syllables are counted based on vowels as well as diphthongs. Thus, a different formula (see 2) needs to be employed for German texts (Amstad, 1978).

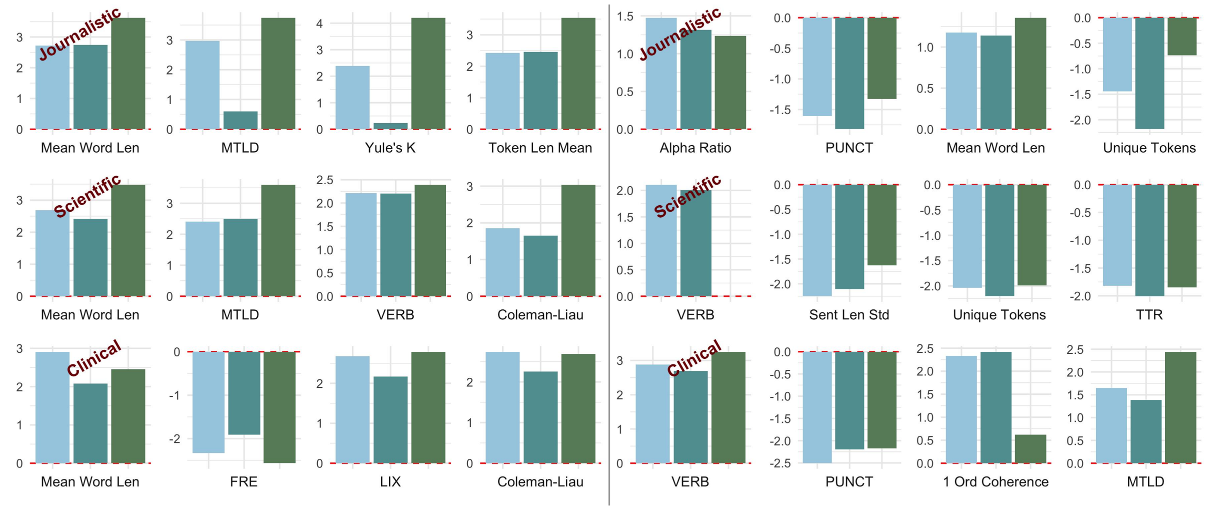

Figure 2: Cohen's d effect size for English (left) and German (right) for the top four significant features at $\alpha = 0.01$ with Bonferroni correction applied for multiple testing, across domains: journalistic at the top row, scientific in the middle, and clinical at the bottom. d values below 0.2 indicate a small effect, 0.5 a medium effect, and 0.8 a large effect. A red dotted line represents the human baseline. Negative values indicate lower feature values in generated texts compared to human texts. The order of the bars from left to right for all subplots: continue, explain, create.

$$180 - (w/sent) - 58.5 * (syl/w) \tag{2}$$

We used two other popular readability metrics: Flesch-Kincaid-Grade-Level (FKGL) is a derivative of FRE and produces a number that corresponds with a U.S. grade level required for the understanding of a particular text. The Coleman-Liau Index (CLI: Coleman and Liau, 1975) was originally intended for the standardisation of school books and is now widely used across sectors (Formula 3 in the Appendix). Just like with FKGL, a higher score suggests greater text complexity. For example, CLI 12.5 indicates text level approximately suitable for senior year high school students in the American educational system. We were not able to find the formula variations for languages other than English for FKGL and CLI.

Läsbarhetsindex, or LIX, presents a valuable choice when assessing readability in languages other than English, since it does not rely on counting syllables (Björnsson, 1968). Instead, LIX calculates the percentage of long words (more than six letters) and the average number of words per sentence, defined by period, colon, or capital first letter (Formula 4 in the Appendix).

Lexical and Morphological Diversity In addition to some lexical variability features included in the textDescriptives package, we employed three more popular metrics, dedicated to the assessment of lexical diversity in a text. We used the Type-Token Ratio (TTR), which gives a general overview of lexical diversity. Since TTR may provide skewed results in long texts, we used the Measure of Textual Lexical Diversity (MTLD), which assesses the length of word sequences with a specific level of TTR (McCarthy and Jarvis, 2010). Additionally, we leveraged Yule's K (Yule, 1944), which is resilient to text length fluctuations while reflecting the repetitiveness of the data.

For morphology, we engaged the metrics of Shannon entropy and Simpson diversity to measure the surprisal levels within the inflectional paradigms of the German lemmas (Vanmassenhove et al., 2021). Inflectional evaluation adds to the assessment of lexical richness and has been considered an important feature for readability assessment of morphologically rich languages (Weiss et al., 2021). Our results showed that the morphological diversity of German lemmas in the generated texts is lower than in the human texts. Human morphology proved to be significantly richer in the 20 Minuten texts as well as the German PubMed articles with the *completer* scoring the lowest across corpora.

Coherence The textDescriptives library leverages GloVe[10] vectors to calculate the cosine similarity between the adjacent sentences (first order coherence) as well as between the sentences that are one sentence apart (second order coherence).

[10] `https://nlp.stanford.edu/projects/glove/`

Inspired by the study of explicit connectives in language models by Beyer et al. (2021), we investigate the usage of discourse particles and thus test the coherence of generated texts in a more fine-grained manner. We used 48 English connectives, collected by Meyer (2014), which occur with a frequency above 20 in the Penn Discourse Treebank (PDTB) and 124 German connectives from DimLex, a lexicon of discourse markers by Stede and Umbach (1998). We completed the list of German connectives with spelling variants (ß → ss) bringing the total number to 133. The *connectives* feature includes all occurrences in the text, whether the particle functions as a preposition (e.g. *while*) or other part of speech. The *connectives capitalised* include those at the beginning of a sentence, increasing the probability of them acting as a true discourse connective.

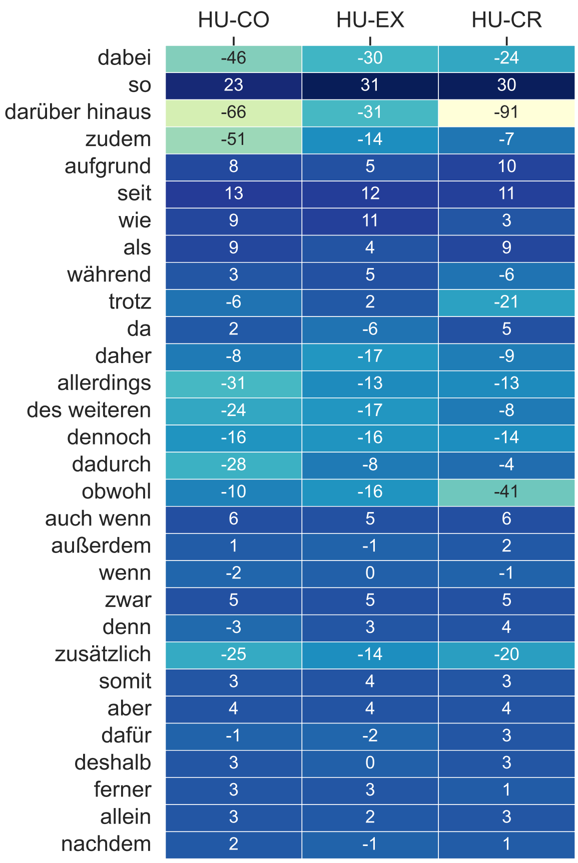

Figure 3: Top 30 most frequent connectives used at the beginning of a sentence in the human-written German PubMed corpus and their absolute differences across personas. Negative numbers indicate higher occurrences in the generated texts.

The academically-instructed ChatGPT personas tend to overuse capitalized connectives. Figure 3 shows the top 30 German connectives in the German PubMed corpus used by humans. The heatmap illustrates the absolute differences in the occurrence of these connectives between human and generated texts. ChatGPT personas, to a lesser extent under the *explainer* task, favor high-level formal items such as “darüber hinaus” and “des weiteren” (EN: *furthermore* in both cases), “allerdings” (EN: *however*), and “zusätzlich” (EN: *additionally*), while human writers start their sentences more often with simple connectives like “so” (EN: *so*), “seit” (EN: *since*), and “aufgrund” (EN: *due to*). In contrast, the generative personas in English tend to use fewer sophisticated connectives at the beginning of sentences. Among human PubMed authors in English, the preferred connectives for a sentence beginning are “however”, “therefore”, “in addition”, “as”, and “moreover”. The *creator* personas, on the other hand, use “while” twice as often as humans, but “for example”, “thus”, and “in addition” only a handful of times. A statistically significant difference in the usage of capitalized connectives was observed in English journalistic texts as well.

5 Discussion

We observed several features that exhibit the same patterns across languages when ChatGPT-generated text is compared to human-written text. For example, ChatGPT employs longer words and creates texts that are deemed difficult by the readability metrics, with the *creator* producing the most complicated texts, featuring the longest sentences and the highest proportion of unique tokens among the tasks. Generated sentences have shorter dependencies, i.e. lower syntactic complexity, and their token count does not fluctuate as much as in human sentences. ChatGPT, particularly the *completer*, exhibits higher coherence scores, possibly due to lexical repetitiveness. Finally, all generated texts exhibit more nouns, verbs, and fewer punctuation marks than human writing.

In our data, human sentences tend to be shorter in German (mean=18, std=10) compared to English (mean=21, std=11). This could be attributed to the complexity of corpora. The academic and clinical texts contain many numbers and punctuation marks, and the German 20 Minuten corpus frequently includes sporting results, which can complicate sentence segmentation. In the journalistic domain, both German corpora (20 Minuten and Credit Suisse Bulletin) exhibit shorter human sentences compared to their English counterparts

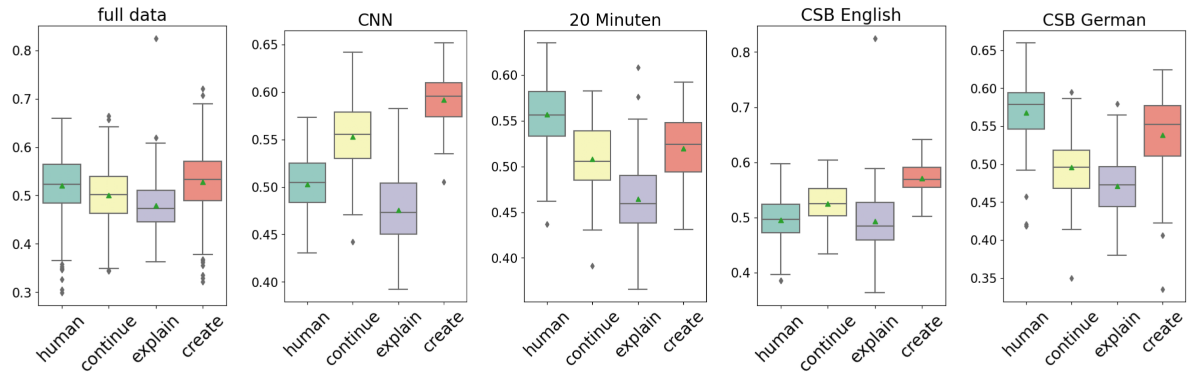

Figure 4: The distribution of unique tokens in the combined English-German data and across the four news corpora illustrates the impact which prompting has on the linguistic profile of generated output.

(CNN, Credit Suisse Bulletin). However, ChatGPT generates longer sentences for all four journalistic corpora. The opposite trend is observed in the clinical domain, where human sentences are longer in German than in English. In this domain, generated sentences are longer than human sentences in English but shorter in German, with the *explainer* being the closest to human values. The number of determiners is another feature that shows language-specific properties. In English, human writers use more determiners than the machine, while in German it is the opposite.

As for the three ChatGPT tasks, the *completer*, which has no persona setting, uses the smallest amount of punctuation marks and other non-alphanumeric characters of all three. It often starts sentences with discourse connectives and keeps sentence lengths steady more than the other two personas. As expected, the *explainer* uses the highest number of total connectives, i.e. higher cohesion, as well as adjacent dependency relations, i.e. simpler syntax. In the journalistic domain, it employs the lowest proportion of unique tokens. Moreover, the *explainer* scores highest on local coherence, sometimes matched by the *completer*. The *creator*, which is prompted by the same text samples as the *completer* but with elaborate personas, features the most difficult readability and lexical diversity, using the longest words and the highest rate of unique tokens.

6 Conclusion

In our study, we examine how prompt modifications, particularly defining persona system settings, affect the linguistic output of ChatGPT across English and German in three domains. We generated comparable corpora by conditioning outputs on three tasks: continuing, explaining, and creating text. The completion task uses default settings, whereas the creation task includes detailed persona descriptions and domain-specific instructions.

We analyzed the statistical validity of lexical and morphosyntactic features to create linguistic profiles and observed significant influences of prompting on linguistic outputs, varying by language and domain. The same features, though extracted from texts produced by the same task, domain, and persona, can exhibit opposite values in different languages (Figure 4).

In our study, human-authored texts exhibit distinctly different values from generated texts on a large number of features. Interestingly, the most basic features such as word length and punctuation give away generated texts even when all languages and domains are mixed together. Furthermore, we observed that generated texts in German are harder to classify than in English, highlighting the need for language-specific evaluation metrics. For instance, readability metrics designed for American English may not be as effective for German, which relies more on morphological features.

Overall, our research underscores the importance of selecting the right linguistic features to differentiate between human and machine-generated texts across different languages, domains, and prompt variations.

Limitations

Working with proprietary models inevitably introduces a number of limitations into any research. Since the inner workings of these models are unknown, results cannot be fully explained or reproduced. Aside from these obvious limitations, we acknowledge that our findings are limited to only two languages. Furthermore, our textual data is rather small, especially for the scientific domain. We also understand that including other domains, especially with less formal language, would make our work more complete. Finally, our data was generated more than six months prior to the paper submission, which is a long time considering the rate of technological advancement.

Ethics statement

All data used in our research is open access and contains no sensitive information. Nevertheless, we abstained from generating new clinical guidelines using the *creator* task and generated imaginary clinical cases instead. Overall, we understand that any insight into the workings of generative models has the potential to improve them and, though not intentional, make their usage for adversarial attacks easier.

Acknowledgements

This research was funded by the National Centre of Competence in Research "Evolving Language", Swiss National Science Foundation (SNSF) Agreement 51NF40 180888. JV acknowledges funding by the Swiss National Science Foundation (project MUTAMUR; no. 213976).

References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service. Google-Books-ID: kiI7vwEACAAJ.

Sisith Ariyaratne, Karthikeyan. P. Iyengar, Neha Nischal, Naparla Chitti Babu, and Rajesh Botchu. 2023. A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiology*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Carl-Hugo Björnsson. 1968. *Läsbarhet.* Lärarbiblioteket. Liber, Stockholm.

Florian Borchert, Christina Lohr, Luise Modersohn, Jonas Witt, Thomas Langer, Markus Follmann, Matthias Gietzelt, Bert Arnrich, Udo Hahn, and Matthieu-P. Schapranow. 2022. GGPONC 2.0 - The German Clinical Guideline Corpus for Oncology: Curation Workflow, Annotation Policy, Baseline NER Taggers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3650–3660, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284. Place: US Publisher: American Psychological Association.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. ArXiv:2301.07597 [cs].

Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. TextDescriptives: A Python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153. ArXiv:2301.02057 [cs].

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 Minuten: A Multi-task News Summarisation

Dataset for German. In *SwissText 2023: 8th Swiss Text Analytics Conference, Neuchâtel, 12 June 2023 - 14 June 2023.*, Neuchâtel. University of Zurich.

J. P. Kincaid, Jr. Fishburne, Rogers Robert P., Chissom Richard L., and Brad S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Gabriel Levin, Raanan Meyer, Eva Kadoch, and Yoav Brezinov. 2023. Identifying ChatGPT-written OBGYN abstracts using a simple tool. *American Journal of Obstetrics & Gynecology MFM*, 5(6). Publisher: Elsevier.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. ArXiv:2304.07666 [cs].

H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60. Publisher: Institute of Mathematical Statistics.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Thomas Meyer. 2014. *Discourse-level Features for Statistical Machine Translation - Idiap Publications*. PhD, Idiap and EPFL, Lausanne, Switzerland.

Anne-Lyse Minard, Zanoli Roberto, Altuna Begoña, and Speranza Manuela. 2021. European Clinical Case Corpus.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. ArXiv:2301.13852 [cs].

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. Technical report.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

S. S. Shapiro and M. B. Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4):591–611. Publisher: [Oxford University Press, Biometrika Trust].

Manfred Stede and Carla Umbach. 1998. DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1238–1242, Montreal, Quebec, Canada. Association for Computational Linguistics.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The Science of Detecting LLM-Generated Texts. ArXiv:2303.07205 [cs].

Rayden Tseng, Suzan Verberne, and Peter van der Putten. 2023. ChatGPT as a Commenter to the News: Can LLMs Generate Human-Like Opinions? In *Disinformation in Open Online Media: 5th Multidisciplinary International Symposium, MISDOOM 2023, Amsterdam, The Netherlands, November 21–22, 2023, Proceedings*, pages 160–174, Berlin, Heidelberg. Springer-Verlag.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a Parallel Corpus on the World's Oldest Banking Magazine. In *the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83. Publisher: [International Biometric Society, Wiley].

G.U. Yule. 1944. *The statistical study of literary vocabulary*. The University Press. Tex.lccn: 44029835.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction Tuning for Large Language Models: A Survey. ArXiv:2308.10792 [cs].

A Example Appendix

The Coleman-Liau Index

$$5.89 * (char/w) - 0.3 * (sent/w) - 15.8 \quad (3)$$

The Läsbarhetsindex

$$w/sent + (w\_long * 100)/w \quad (4)$$

Type	Feature	Hu Co	Hu Ex	Hu Cr	Co Ex	Co Cr	Ex Cr
coh	1st order coherence	x	x	x	En	x	De
coh	2nd order coherence	x	x	x	En	x	De
coh	connectives	De	x		En	x	x
coh	connectives capitalised	De			x	x	
dep	distance mean	x		De	x	En	De
dep	distance std	x	x	x	En	x	
dep	prop adj rel mean	x	x	x	x		De
dep	prop adj rel std	x	x	En	De	En	x
des	doc length	x		De	En	En	En
des	num of chars	x	x	x	En	En	
des	num of sents	x		x	De	En	x
des	num of stop words	x	De	x	En	x	x
des	num of tokens	De		De	En		En
des	num of unique tokens			En	De	En	x
des	prop unique tokens	x	De	x	x	x	x
des	sent length mean	De	En	En	x	En	x
des	sent length median	En	x	En	x	En	x
des	sent length std	x	x	x	De	x	x
des	syllabs per token mean	x	x	x		En	En
des	syllabs per token median	x	x	x		En	En
des	syllabs per token std	En	En	En		En	En
des	token length mean	x	x	x		En	En
des	token length median	x	x	x	En	En	x
des	token length std	En	En	En		En	En
inf	entropy	x	En	En	En	x	x
inf	perplexity	x	x	En	En	x	En
inf	perplexity per word	x	x	En	En	x	En
led	MTLD	x	En	x	x	x	x
led	TTR	x	De	x	En	x	x
led	Yule's K	x	x	En	x	x	x
mor	shannon entropy	De	De	De	De		
mor	simpson diversity	De	De	De	De		
pos	prop of adjectives	x	x	En	x		En
pos	prop of adpositions	x	En	x	x		x
pos	prop of adverbs			De	De	De	De
pos	prop of auxiliaries	En	x	En	En	En	
pos	prop of coord conjunctions		x	En	En	En	En
pos	prop of determiners	En	x	En	En		En
pos	prop of nouns	x	x	x		De	De
pos	prop of particles				En		
pos	prop of pronouns				En		
pos	prop of punctuation	x	x	x		En	
pos	prop of subord conjunctions				De		
pos	prop of verbs	x	x	En			
qua	alpha ratio	x	x	x	x	x	
qua	dupl ngram chr fract 10		x		x		De
qua	dupl ngram chr fract 5	De	x	De	x		x
qua	dupl ngram chr fract 6	De	x	De	x		x
qua	dupl ngram chr fract 7	De	x	De	x		x
qua	dupl ngram chr fract 8	De	x	De	x		x
qua	dupl ngram chr fract 9		x		x		x
qua	mean word length	x	x	x	En	En	En
qua	oov ratio	De		x	De	En	x
qua	top ngram chr fract 2	En	x	En		En	En
qua	top ngram chr fract 3	De	De	De			
qua	top ngram chr fract 4	De	De				De
red	LIX	En	x	x		En	En
red	RIX	En	En	En	De	x	En
red	autom readability index	En	x	x		En	En
red	coleman liau index	x	x	x	En	En	En
red	flesch kincaid grade	En	x	En		En	En
red	flesch reading ease	En	En	En		En	En
red	gunning fog	En	x	x		En	En

Table 5: Significant features evaluated on the combined English (En) and German (De) data. x marks features that distinguish personas in both languages. Feature groups: inf (information theory), qua (quality), pos (distribution of part-of-speech tags), red (readability), coh (coherence), des (general descriptive statistics), mor (morphology), and led (lexical diversity).