Like a Human? A Linguistic Analysis of Human-written and Machine-generated Scientific Texts

Sergei Bagdasarov
Saarland University (Germany)
sergeiba@lst.uni-saarland.de

Diego Alves
Saarland University (Germany)
diego.alves@uni-saarland.de

Abstract

The purpose of this study is to analyze lexical and syntactic features in human-written texts and machine-generated texts produced by three state-of-the-art large language models: GPT-4o, Llama 3.1 and Qwen 2.5. We use Kullback-Leibler divergence to quantify the dissimilarity between humans and LLMs as well as to identify relevant features for comparison. We test the predictive power of our features using binary and multi-label random forest classifiers. The classifiers achieve robust performance of above 80 % for multi-label classification and above 90 % for binary classification. Our results point to substantial differences between human- and machine-generated texts. Human writers show higher variability in the use of syntactic resources, while LLMs score higher in lexical variability.

1 Introduction

The use of Large Language Models (LLMs) in research has become a common practice. Scenarios in which academia members resort to LLMs are varied, ranging from ideas generation and productivity enhancement to data analysis and writing (Panda and Kaur, 2024). 80.88% of researches surveyed by Liao et al. (2024) used LLMs in their academic activities, with 61% of researchers having at least once used LLMs for editing and 41%, for direct writing. A vast majority of scholars surveyed by Mishra et al. (2024) consider that LLMs will have an impact on various stages of the publication process.

State-of-the-art LLMs are capable of producing high-quality texts that are practically indistinguishable from human-created content for untrained individuals. This makes LLMs useful writing assistants, especially for researchers who are not native speakers of English. Yet numerous studies based on large enough amounts of data have shown that LLMs do write differently in comparison to humans according to certain measures.

In this study, we aim to analyze human-written texts (HWT) and machine-generated texts (MGT) – abstracts of academic papers – and identify linguistic features that can help tell them apart. While studies on this topic abound, they either do not focus specifically on academic texts, even though those might be present in the scrutinized corpora, or use the older GPT 3.5 model relying on the paper title and a short text snippet for abstract generation.

We will address these research gaps by using a large dataset of academic publications with full texts and human-written abstracts and resorting to a newer GPT-4o model (OpenAI, 2024). Moreover, we will complement our analysis with two open-source state-of-the-art models (Llama 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen 2.5 7B Instruct (Yang et al., 2025; Team, 2024)). The rationale behind this decision is that, although ChatGPT is the undisputed leader as a chat bot assistant, its use may be associated with data protection concerns (Ali et al., 2025; Novelli et al., 2024). Because of this, researchers might choose open-source LLMs, either running them locally or accessing them through in in-house university chat bot solutions. Therefore, the amount of scientific content potentially generated by open-source models might be increasing.

Furthermore, most studies comparing HWT and MGT use a predefined list of features. Instead, we will rely on Kullback-Leibler divergence, a measure rooted in information theory, to identify features that can reliably distinguish between HWT and MGT and then test these features in a classification task.

The remainder of the paper is structured as follows. Section 2 offers a brief overview of research on linguistic features in HWT and MGT. Section 3 describes the dataset and methodology, including

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 38–47, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-004

the procedure for feature selection. Then, Section 4 presents the results for text classification, showing the predictive power of the extracted features. Section 5 compares HWT and MGT across some of the selected features, while Section 6 contains a brief discussion of our findings. Finally, Section 7 offers some concluding remarks and an overview of future work plans.

2 Related Work

Due to easy accessibility and outstanding output quality, LLMs have become an integral part of many workflows, often being used to produce written content. This inevitably leads to the proliferation of machine-generated texts, making the study of synthetic language an important task.

A common approach for this consists in defining a set of features (e.g., sentence length, frequencies of words, part-of-speech categories or specific syntactic patterns, etc.) and comparing them in human-written texts (HWT) and machine-generated texts (MGT) (e.g., Zanotto and Aroyehun (2024); Culda et al. (2025); Muñoz-Ortiz et al. (2024); Georgiou (2024), etc.).

A general consensus in this field is that MGT differ considerably from HWT, especially when it comes to lexical variability, with human writers being characterized by higher lexical diversity (e.g. as measured by type-token ratio or amount of hapax legomena) (Zanotto and Aroyehun, 2024; Culda et al., 2025). At the same time, Opara (2024) pointed out that HWT manage to strike a balance between lexical richness and text length, while some LLMs seem either to overly restrict or expand their vocabulary.

Apart from that, LLMs have also been shown to overuse some stylistic vocabulary not related to the content of a text. Such overused lexical items, sometimes referred to as focal words, can be detected by comparing word frequencies before and after LLM era and typically include words like *delve, underscore, intricate, pivotal, showcase, meticulous*, etc. (Juzek and Ward, 2024; Kobak et al., 2025; Gray, 2024; Liang et al., 2024).

From the morphosyntactic perspective, MGT seem to use shorter sentences, showing however higher complexity in the constituency structure (Muñoz-Ortiz et al., 2024). MGT (at least those generated by GPT models) also tend to favor a more nominal style of writing, with higher proportion of nouns, nominalizations, phrasal coordination and determiners (Reinhart et al., 2025; Liao et al., 2023). Despite higher proportion of nouns, LLMs were found to use more general vocabulary in specialized registers, though, resulting in lower degree of specificity and higher readability scores in comparison to HWT (Liao et al., 2023).

Interestingly, HWT tend to convey more negative emotions on average, while LLMs produce more positive texts (Muñoz-Ortiz et al., 2024; Culda et al., 2025). This also holds for conversation-like texts, where LLMs exceed humans in some communicative processes, scoring higher in social behavior, politeness and attentional focus. However, they still do not reach the same level of authenticity as humans, at least in conversation (Sandler et al., 2024).

The studies reviewed above rely on a predefined set of features to explore differences between HWT and MGT. In contrast, in this study we introduce a more informed approach to feature selection based on Kullback-Leibler divergence. Moreover, we use three syntactic complexity measures, which, to the best of our knowledge, has not been done yet.

3 Data and Methods

3.1 Data

We use the ACL Anthology Corpus (Rohatgi, 2022) – a collection of ACL contributions ranging from 1950s to 2022. The reasons for choosing this dataset were twofold. First, it provides both abstracts and full texts, allowing us to generate abstracts based on full papers and compare them to HWT. Moreover, since the dataset is limited to publications prior to 2022, we ensure that no abstracts in it have been written by LLMs.

Due to the extensive size of the corpus, which would result in high material and computational costs, and the presence of noisy data, we selected a sample of papers that meet the following criteria: a) both full text and abstract are available; b) publication year: after 1999; c) language: English; d) length of the abstract: between 100 and 200 words; e) length of the full paper: only those within one standard deviation of the mean length among those in the interquartile range. After applying the filters, we obtained a subset of 10,393 papers with their abstracts.

3.2 Automatic Abstract Generation

We automatically generated abstracts based on the ACL papers using three LLMs: **gpt-4o-2024-**

08-06 (OpenAI, 2024), **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) and **Qwen2.5-7B-Instruct** (Yang et al., 2025; Team, 2024). The GPT model was prompted using the OpenAI API through the openai Python library[1]. To interact with the two open-source models, we used the huggingface[2] transformers text generation pipeline. All models were prompted using the temperature of 1. Except for output length, all other parameters were kept at default values. Number of output tokens was set to 400. Lower values approximating the desired abstract length of 200 words stipulated in the prompt resulted in a considerable amount of incomplete outputs by Llama. While the higher output length in the pipeline parameters helped mitigate the incomplete output problem, it caused Llama output to be longer as can be seen in Table 1. However, this was not critical for further analysis since all measures are normalized by the size of the subcorpora.

The prompt consisted both of a system message defining the model's role as academic writing assistant and describing the task as well as a user message providing the full text of a paper and giving the instruction to generate an abstract (see Appendix A). We included the full text to approximate the behavior users might exhibit when actually using the models for paper summarization since newer models feature much larger input context windows. The experiments were run at the end of June and at the beginning of July 2025.

Source	Tokens	Types	Sentences
Human	1,700,972	32,808	64,975
Llama	2,392,988	34,688	83,534
Qwen	1,524,521	29,388	60,018
GPT	2,034,978	32,880	76,710

Table 1: Comparison of abstract sources by tokens, types, and sentences.

3.3 Methods

We used **Kullback-Leibler Divergence (KLD)** (Kullback and Leibler, 1951) to compare HWT and MGT. KLD (Equation 1) is an information-theoretic measure that asymmetrically quantifies (in bits) the divergence between two probability distributions and allows us to identify the most distinctive features contributing to the divergence. A KLD value of 0 means that the distributions are identical, whereas a value larger than 0, in contrast, is indicative of a divergence.

We calculated KLD for (a) lemmas to capture how HWA and MGA diverge on the lexical level as well as (b) Universal Dependencies part-of-speech tags, and (c) dependency relations (de Marneffe et al., 2021) to analyze the syntactic differences. For the classification task, we used only POS tags and syntactic relations for which we calculated normalized document frequencies by dividing the count of a POS tag or dependency label by the total number of tokens in the document.

$$\mathrm{KLD}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \quad (1)$$

We complemented this initial set of features with variability measures since previous research has found substantial differences in variability of different elements in HWT and MGT (e.g.,Zanotto and Aroyehun (2024); Culda et al. (2025); Liao et al. (2023); Opara (2024), etc.). We operationalize variability with **Entropy**. Entropy is another information-theoretic measure and is calculated using the formula in Equation 2. It shows how much uncertainty there is in a system, with higher entropy values indicating higher uncertainty. In our context, higher uncertainty means more variability in language use. We calculated document-level entropy for lemmas, POS tags and dependency relations.

Additionally, we calculated the proportion of unique items in each selected POS category by dividing the number of unique items in a POS category by the total number of elements in the category.

$$H = \sum_{i=1}^{n} p_i \times \log_2 p_i \quad (2)$$

We also included three syntactic measures that are commonly analyzed when examining syntactic complexity: average dependency length (Gibson, 1998; Jiang et al., 2019), tree depth (Xu and Reitter, 2016), and average branching factor (Xu and Reitter, 2016).

Average dependency length (aDL) is calculated by measuring the distance between heads and their dependents in a syntactic dependency tree, ignoring punctuation. For each dependency in a sentence, the length is the absolute difference between

[1]https://pypi.org/project/openai/
[2]https://huggingface.co/

the positions (i.e. token indices) of the head and the dependent. These lengths are summed across all dependencies in the sentence and then divided by the total number of dependencies (i.e.number of tokens minus one). For example, in the parsed sentence with 8 tokens (excluding punctuation) shown in Figure 1, the sum of the dependency distances is 17. The average dependency length (ADL) is therefore calculated as $\frac{17}{8-1} = 2.43$.

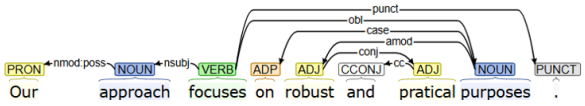

Figure 1: Example of parsed sentence.

Average branching factor (ABF) quantifies the mean number of immediate dependents (children) per internal node. It captures how syntactic constituents are organized: higher ABF values indicate greater syntactic parallelism, whereas lower values suggest more linear or sequential structuring. It is calculated by dividing the total number of children of internal nodes (i.e., the total number of nodes minus one) by the number of internal nodes.

Tree depth (TD) refers to the length of the longest path from the root node to any leaf node in a syntactic dependency tree, reflecting the degree of structural nesting.

As an example, consider the sentence "Anne lost control and laughed." Its constituency tree is: (ROOT (S (NP (NNP Anne)) (VP (VP (VBD lost) (NP (NN control))) (CC and) (VP (VBD laughed))) (. .))). The graphical representation of this tree is provided in Figure 10 in Appendix B.

The total number of nodes (internal nodes and leaves) is 19, the longest path from the root to any leaf (tree depth) is 6 (ROOT → S → VP → VP → NP → NN → leaf), and the average branching factor is the total number of children of internal nodes is 18 (i.e., total nodes minus 1) divided by the number of internal nodes (i.e., 13), giving an average branching factor of 1.39.

For each abstract (human- or machine-generated), these measures were calculated for each sentence and then divided by the number of sentences to obtain the average value per text.

Finally, we also included word and sentence length, average count of stop words per text and a readability measure[3]. Entropy, dependency relations counts and the three syntactic measures were calculated by custom Python scripts. For all other measures, we used the Linguistic Features ToolKit (LFTK) (Lee and Lee, 2023). In total, we obtained a set of 76 features.

We trained random forests classifiers both for binary and multi-label prediction to test the predictive power of the selected features. First, we only used the features extracted by KLD and then the complete set of features. To train the classifiers, we used the ranger[4] R package with the following settings: number of trees = 500, importance = impurity, classification = True, seed =123. For both classifiers, we used 80% of data for training and 20% for testing.

4 Classification Results

We first focused only on POS tags and dependency relations that have proven distinctive as per KLD (53 features in total). Then we used the full set of 76 features. As shown in Figure 2, these 76 features do allow to identify 4 clusters of texts. We can see an overlap between GPT and Qwen, on the one hand, and humans and Llama, on the other hand, indicating a greater linguistic similarity of these groups, in line with our KLD results. In turn, the clusters for GPT and Qwen are clearly distinguishable from human texts.

	Binary		Multilabel	
Model	**Acc.**	**F1**	**Acc.**	**F1**
NIR	.75	–	.25	–
KLD features only	.92	.89	.83	.83
All features	.94	.92	.88	.88

Table 2: Random forests results for binary (human vs machine) and multi-label (human vs GPT vs Llama vs Qwen) classification. NIR (no-information rate) shows the model's performance if the majority label is always assigned. Here it is used as baseline.

	Precision	**Recall**	**F1**
GPT	.87	.90	.88
Human	.89	.90	.89
Llama	.89	.87	.88
Qwen	.89	.86	.87

Table 3: Precision, recall, and F1-scores for each class in Random Forest classification based on the full set of features.

Using the two sets of features (the initial one obtained by KLD only and the complete one), we trained random forests classifiers both for binary

[3]Coleman-Liau Index (Coleman and Liau, 1975).

[4]https://cran.r-project.org/package=ranger

and multi-label prediction. As shown in Table 2, even the initial set of features allows for a robust predictive power, confirming that KLD is a suitable measure to identify relevant linguistic features. However, additional features do improve the model performance considerably. The prediction results are fairly similar across different classes, being slightly better for human texts and worse for Qwen-generated texts as measured by F1 score (see Table 3).

5 Feature Analysis

5.1 KLD

In line with previous findings, our KLD analysis indicated that HWT differ from MGT both lexically and syntactically (see Figures 3, 4 and 5). Llama was closest to humans on all three levels of comparison, while GPT and Qwen turned out to be more divergent from HWT, showing similar results.

The distinctive features extracted by KLD were very similar across all groups of comparison. On the lexical level, HWT are mostly characterized by function words (determiners, prepositions, particles, copula verbs and modal verbs), discourse markers (*however, therefore, thus*), adverbs (*usually, considerably*) and some abbreviations commonly used in academic writing (*i.e., e.g., etc.*). In contrast, MGT are characterized by the use of nouns, verbs and adjectives many of which have stylistic function and are considered "focal words" typically overused by LLMs (*highlight, underscore, demonstrate, introduce, incorporate, pivotal, significant*, etc.) (see Table 5). The verb *delve*, a prototypical example of such overused vocabulary, has also been identified as distinctive of all tested LLMs, especially the GPT model. However, its contribution to the overall divergence is relatively low, suggesting that it is not as overused anymore by GPT-4o as it was by GPT 3.5.

H x GPT	H x Llama	H x Qwen
be	be	we
the	this	be
of	in	of
a	paper	have
we	have	our

Table 4: Top 5 lemmas distinctive of humans compared to each of the LLMs.

KLD results for POS tags confirm the more extended use of function words in HWT observed at the lexical level (see Table 6). Besides, HWT are also characterized by the use of adverbs, proper nouns and numerals. Interestingly, adjectives are distinctive of HWT when compared to Llama, however, not when compared to GPT and Qwen. MGT

GPT x H	Llama x H	Qwen x H
enhance	and	and
future	our	author
and	approach	enhanced
highlight	include	demonstrate
potential	demonstrate	highlight

Table 5: Top 5 lemmas distinctive of each of the LLMs compared to humans.

H x GPT	H x Llama	H x Qwen
AUX	AUX	PRON
DET	ADP	AUX
ADP	ADV	ADP
PRON	DET	DET
ADV	ADJ	ADV

Table 6: Top 5 UD POS tags distinctive of humans compared to each of the LLMs.

GPT x H	Llama x H	Qwen x H
NOUN	NOUN	NOUN
VERB	CCONJ	PUNCT
ADJ	VERB	VERB
PUNCT	PUNCT	PROPN
PROPN	SCONJ	CCONJ

Table 7: Top 5 UD POS tags distinctive of each of the LLMs compared to humans.

H x GPT	H x Llama	H x Qwen
det	advmod	case
case	obl	obl
cop	case	cop
aux:pass	det	aux
advmod	aux:pass	advmod

Table 8: Top 5 UD relations distinctive of humans compared to each of the LLMs.

GPT x H	Llama x H	Qwen x H
obj	obj	obj
advcl	nmod:poss	punct
compound	conj	compound
amod	compound	advcl
punct	cc	amod

Table 9: Top 5 UD relations distinctive of each of the LLMs compared to humans.

Figure 2: t-SNE clustering of HWT and MGT based on a set of 76 features.

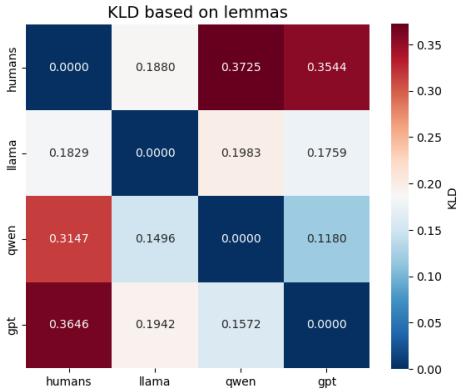

Figure 3: KLD values based on lemmas.

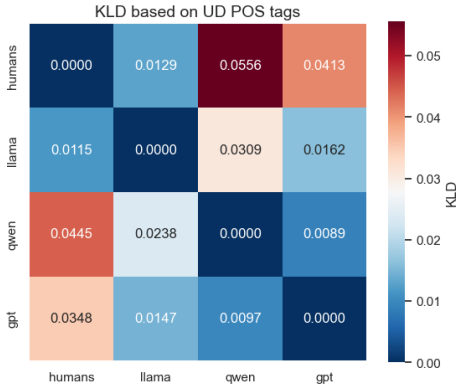

Figure 4: KLD values based on UD POS tags.

are characterized by POS tags labeling nouns and verbs as well as punctuation marks, coordinative conjunctions and subordinative conjunctions (see Table 7).

At the level of UD syntactic relations, we again see that one of the most distinctive features of HWT

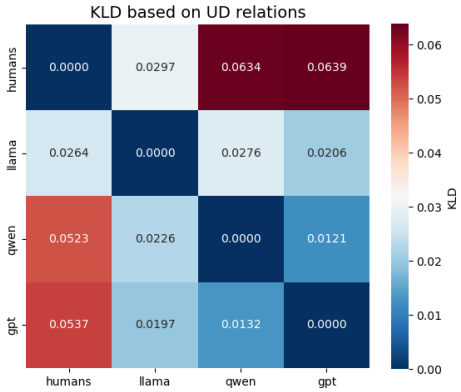

Figure 5: KLD values based on UD dependency relations.

are auxiliaries, determiners and adverbial modifiers. The *obl* and *case* labels point to a varied use of prepositional phrases as oblique arguments, adjuncts or nominal modifiers. LLMs, in turn, seem to favor more compact and dense constructions, at least in case of nominal modification, leading potentially to higher phrasal complexity. This is evident by adjective modifiers, compound nouns and possessive nominal modifiers being distinctive of MGT (see Table 9).

In terms of clausal features, HWT are especially characterized by passive constructions and finite relative clauses modifying either nouns or whole sentences. In contrast, LLMs tend to use more non-finite clausal modifiers (with the exception of Llama), adverbial clauses and clausal complements. In general, clausal subordination is more typically seen in MGT, which is reflected in the more pronounced use of the relation *mark*.

5.2 Word and Sentence Length

MGT contain longer words than HWT as measured by the number of syllables, which is in line with the overall prevalence of function words in HMW in comparison to MGT as indicated by KLD. In terms of sentence length (in words), Llama used the longest sentences as per median value. If measured in syllables, all LLMs used longer sentences because of a consistently longer word length. However, human abstracts show greater variability in sentence length, while all LLMs, especially Qwen, consistently produced sentences containing between approximately 20 and 40 words. (see Figure 7).

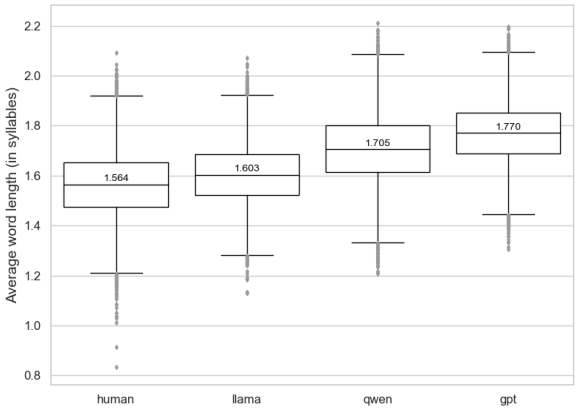

Figure 6: Average word length per document (in syllables).

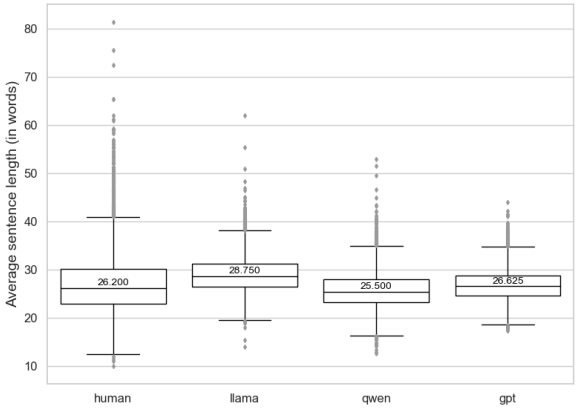

Figure 7: Average sentence length per document (in words).

5.3 Variability

At the level of lemmas, GPT shows the highest entropy, followed by Llama. Humans and Qwen are similar in terms of entropy. At the level of POS tags and dependency relations, humans have higher entropy than all LLMs, with Llama-generated texts being closest to HWT. In contrast, Qwen shows the lowest variability among the three LLMs across all three levels of comparison.

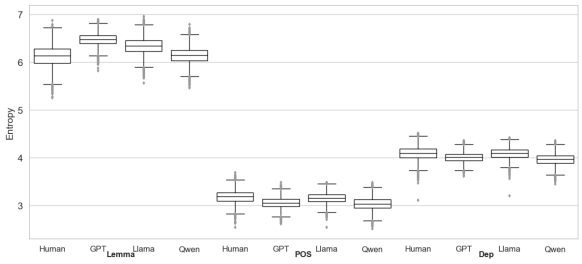

Figure 8: Document entropy for lemmas, UD POS tags and dependency relations.

5.4 Syntactic Complexity

Figure 9 shows the Kernel Density Estimation (KDE) analysis of the three syntactic measures considered in this study: tree depth, average branching factor, and average dependency length (each averaged per abstract). This method estimates the probability density function of a continuous variable, providing a smooth curve that represents the data distribution.

We observe different behaviours for each measure. GPT and Qwen tend to produce sentences with lower tree depth and average dependency length compared to humans, while Llama generates sentences with higher complexity according to these two measures. However, in terms of the average branching factor, all LLMs tend to produce sentences that exhibit greater syntactic branching.

One characteristic of scientific English is the frequent use of complex noun phrases, including pre-modifiers and compound constructions (Halliday and Martin, 2003; Degaetano-Ortlieb, 2021). These complex nominal phrase structures increase the ABF of sentences because the pre-modifiers are all at the same hierarchical level in the syntactic tree (i.e., children of the same NP node). Thus, it seems that language models tend to potentialize the usage of this type of NP.

However, the most striking characteristic of human abstracts compared to machine-generated ones is their greater variability in syntactic complexity. This is evident in the density plots, where human texts consistently cover a broader range with a less pronounced peak, while machine-generated texts exhibit much less variation.

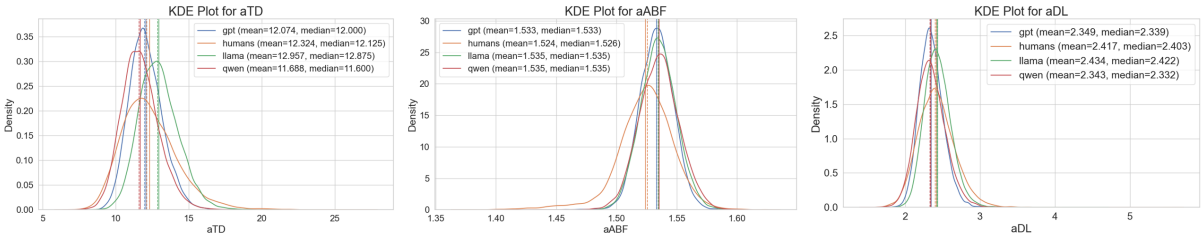

Figure 9: KDE plots of tree depth, average branching factor, and average dependency length (averaged per abstract).

6 Discussion

In general, our findings stand in line with previous research suggesting that HWT and MGT differ considerably both in terms of their lexical and syntactic features (cf. Culda et al. (2025); Zanotto and Aroyehun (2024); Georgiou (2024)). Given that GPT was considerably larger than the other two models and, therefore, was exposed to more training data, we expected the GPT model to produce more human-like content. Nevertheless, Llama was closest to humans both lexically and syntactically, at least as measured by KLD.

LLMs tend to use more complex phrasal structures, which is reflected in a higher average branching factor and a higher typicality of nominal premodifiers. Similar findings for other text registers and models (Muñoz-Ortiz et al., 2024; Reinhart et al., 2025) suggest that higher phrasal complexity is a general feature of LLM writing and is not attributable to the influence of our experimental setup.

Also in line with previous research, we have shown that LLMs exhibit lower morpho-syntactic variability (lower entropy of UD POS tags and syntactic relations as well as a narrower spread of values for syntactic complexity measures). This indicates a more repetitive use of patterns as opposed to a more varied use of syntactic resources by human writers.

Surprisingly, we found that MGT exhibit higher lexical variability than HWT, as measured by lemma entropy. This may suggest a general advancement in the lexical creativity of newer models. However, it could also be a consequence of our prompting settings — for instance, the models' exposure to full papers during abstract generation. Further analysis is needed to better understand this outcome.

7 Conclusion and Future Work

In this study, we analyzed the lexical and syntactic features of HWT and MGT. We obtained HWT from a large dataset of academic publications. MGT were generated by three state-of-the-art models (GPT-4o, Llama 3.1 and Qwen 2.5) based on the corresponding full papers.

Adopting a data-driven approach to feature selection, we employed KLD to identify features that effectively distinguish between HWT and MGT. The effectiveness of these features was evaluated using random forest classifiers, which demonstrated robust performance both when using only KLD-selected features and when incorporating an extended feature set.

Our results indicate that MGT still differ considerably from HWT, with Llama producing outputs that more closely resemble HWT than other LLMs. We observed that LLMs tend to generate more complex phrase structures than humans, yet exhibit less syntactic variability. In contrast, lexical variability as measured by entropy was higher in MGT.

Future research will explore whether alternative prompting configurations (e.g., varying temperature settings or employing few-shot prompting) lead to more human-like outputs. We also plan to extend our analysis to additional text types, models, and model sizes. Additionally, a more qualitative examination of LLM-generated abstracts and their perception by human readers would give us a more complete understanding of models' performance.

Limitations

LLM output is strongly influenced by the input data and prompt wording. Experiments based on other genres or using other prompting techniques might yield different results. Moreover, the dataset is limited to one scientific discipline (computational linguistics). So, we cannot account for linguistic divergence between different domains of scientific writing. Our findings might become outdated due

to constant model improvement and release of more powerful LLMs. Finally, although KLD has proven to be an effective feature selection technique, we may have missed some other relevant features that cannot be identified using KLD.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. 2025. Understanding users' security and privacy concerns and attitudes towards conversational ai platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*, page 298–316. IEEE.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

L. C. Culda, R. A. Nerişanu, M. P. Cristescu, D. A. Mara, A. Bâra, and S. V. Oprea. 2025. Comparative linguistic analysis framework of human-written vs. machine-generated text. *Connection Science*, 37(1).

Stefania Degaetano-Ortlieb. 2021. Chapter 11. measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific english. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.

Georgios P. Georgiou. 2024. Differentiating between human-written and ai-generated texts using linguistic features automatically extracted from an online computational tool.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan et al. 2024. The llama 3 herd of models.

Andrew Gray. 2024. Chatgpt "contamination": estimating the prevalence of llms in the scholarly literature.

Michael Alexander Kirkwood Halliday and James R Martin. 2003. *Writing science: Literacy and discursive power*. Routledge.

Jingyang Jiang, Peng Bi, and Haitao Liu. 2019. Syntactic complexity development in the writings of efl learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, 46:100666.

Tom S. Juzek and Zina B. Ward. 2024. Why does chatgpt "delve" so much? exploring the sources of lexical overrepresentation in large language models.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of llms in scientific papers.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. 2023. Differentiating chatgpt-generated and human-written medical texts: Quantitative study. *JMIR Med Educ*, 9:e48904.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Tapas Mishra, Egidia Sutanto, Rani Rossanti, et al. 2024. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Scientific Reports*, 14:31672.

Alba Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57:265.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law Security Review*, 55:106066.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. *ArXiv*, abs/2405.10129.

OpenAI. 2024. Gpt-4o system card.

Subhajit Panda and Navkiran Kaur. 2024. Exploring the role of generative ai in academia: Opportunities and challenges. *IP Indian Journal of Library Science and Information Technology*, 9(1):12–23.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8).

Shaurya Rohatgi. 2022. Acl anthology corpus with full text. Github.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Sergio E. Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models.

A Prompt

System message: You are an efficient writing assistant specialized in creating concise and accurate text summaries for scientific publications. I will provide you with the full text of a scientific paper from the field of computational linguistics. Your task is to read the paper and write a clear and concise abstract for it. Write the abstract from the author's perspective. The abstract should be between 100 and 200 words long. Do not include any additional text like "Abstract:" or "Here is the abstract:".

User message: Write an abstract for this scientific paper: [FULL TEXT OF PAPER]

B Constituency Tree

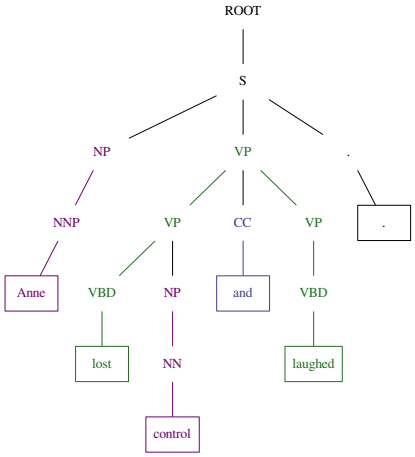

Figure 10: Example of constituency parsed tree