Beyond Lemmas and Syntax: Comparing Human and LLM-Generated Scientific Abstracts

Sergei Bagdasarov, Diego Alves
Saarland University
Saarbrücken, Germany
sergeiba@lst.uni-saarland.de, diego.alves@uni-saarland.de

Abstract

In this study, we compare human-written (HWT) and machine-generated (MGT) abstracts of scientific papers, going beyond traditional lexical and syntactic analyses. We use an extensive corpus of publications on computational linguistics submitted to the Association of Computational Linguistics from mid 1950s to 2022. First, we generate abstracts with three state-of-the-art models (GPT-4o, Llama 3.1 and Qwen 2.5), providing the models with full texts of papers, and subsequently we compare these abstracts to those written by humans. We study the overall information content of abstracts, operationalised as surprisal, and the distribution of information in abstracts quantified as local Uniform Information Density (UID), both metrics related to the processing effort. Subsequently, we perform an extrinsic evaluation through topic modelling and clustering applying the BERTopic model. Our results show significant differences both in surprisal and UID, suggesting that abstracts generated by Llama are less cognitively demanding and show a more uniform distribution of information. Our topic modelling experiments show greater divergence between humans and LLMs than between LLM pairs. At the same time, Llama abstracts seem to be more semantically similar to those written by humans, standing in line with previous findings suggesting such similarity on lexical and syntactic level.

Keywords: generative LLMs, information theory, scientific English

1. Introduction

The integration of Large Language Models (LLMs) into academic research is now widespread. Scholars use these tools for a variety of tasks, including brainstorming ideas, enhancing productivity, analyzing data, and writing (Panda and Kaur, 2024). LLMs have changed the way people do research: a recent survey found that 80.88% of researchers use LLMs in their work, with 61% employing them for editing and 41% for direct writing (Liao et al., 2024). Moreover, a majority of academics further believe LLMs will fundamentally impact the scholarly publication process (Mishra et al., 2024).

State-of-the-art LLMs can produce high-quality texts that are often indistinguishable from human writing to an untrained eye, making them invaluable assistants, particularly for non-native English speakers. However, large-scale studies consistently reveal that LLMs possess distinct linguistic fingerprints, making their output quantitatively different from human prose (Zanotto and Aroyehun, 2024; Culda et al., 2025; Opara, 2024).

This study aims to compare human-written abstracts (HWT) with machine-generated ones (MGT) from academic papers, moving beyond traditional analyses that focus solely on lemmas, parts of speech, and syntactic patterns. Existing studies on this topic generally rely on older models like GPT-3.5 with limited contextual prompting and offer only a restricted linguistic analysis, concentrating on surface-level features such as hapaxes, overused words, and selected syntactic structures.

We address these gaps by analysing a large dataset of full academic publications paired with human-written abstracts. To create machine-written counterparts, we employ the advanced GPT-4o model (OpenAI, 2024) and complement our analysis with outputs from two leading open-source systems: Llama 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen 2.5 7B Instruct (Yang et al., 2025; Team, 2024). This multi-model approach is essential; while ChatGPT remains dominant, growing data protection concerns (Ali et al., 2025; Novelli et al., 2024) are steering researchers toward open-source alternatives, thereby increasing the likelihood that future scientific content will be generated by these models.

Our evaluation introduces two novel dimensions: (a) information-theoretic measures such as surprisal and Uniform Information Density (UID) to assess statistically significant differences in the cognitive processing of HWT versus MGT, and (b) extrinsic evaluation through topic modelling and clustering to examine the impact of MGT on NLP applications commonly used in the scientific domain.

The remainder of the paper is structured as follows. Section 2 provides an overview of research on linguistic features in HWT and MGT. Section 3 details the dataset and methodology. Section 4 presents the results of the information-theoretic measures and the topic modelling experiment. Section 5 discusses the findings. Finally, Section 6 concludes with remarks and directions for future work.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3823–3832
11–16 May 2026.

2. Related Work

Easy accessibility and user-friendly design of LLM-based chatbots have lead to LLMs being actively used in various domains from everyday activities to software development and academic research. With the amount of LLM-generated content rapidly increasing, the analysis of LLM-written texts and their differences from human-authored content has acquired a prominent role in the NLP community.

Such studies have been conducted for a variety of text registers. For instance, Muñoz-Ortiz et al. (2024) compared human-written news articles with those generated by three open-source LLMs. Similarly, Georgiou (2024) analysed ILETS essays written by humans and a GPT model. Sandler et al. (2024) used two GPT models to simulate oral conversations and compared them to real human interactions.

Focusing on scientific texts, Juzek and Ward (2024) and Kobak et al. (2025) analysed excess vocabulary associated with the style of writing characteristic of LLMs (e.g., words like *delve, pivotal, underscore*, etc.). A number of studies compared syntactic properties of academic texts written by humans and LLMs and performed quantitative linguistic analyses comparing distribution of part-of-speech tags, dependency relations and employing metrics like type-token ratio or lexical density (Liao et al., 2023; Culda et al., 2025; Berber Sardinha, 2024). Bagdasarov and Alves (2025) complemented such analyses with average dependency length, tree depth, and branching factor. Some studies, in turn, relied on mixed datasets that also contain academic texts, however without addressing this register specifically (Zanotto and Aroyehun, 2024; Opara, 2024; Reinhart et al., 2025).

While diverging in individual findings depending on exact text type and models selected, all previous studies conclude that HWT and MGT differ considerably both on lexical and syntactic level. However, we are not aware of a study that goes beyond linguistic analyses to consider how HWT and MGT handle the information and thematic content. Here, we aim to address this research gap.

3. Data and Methods

3.1. Data

In this study, we use the ACL Anthology Corpus (Rohatgi, 2022) to generate abstracts with LLMs. This corpus, containing ACL submissions ranging from mid 20th century to 2022, is a useful resource for our purposes as it contains both full texts of papers and their abstracts that can be used as baseline for comparison with LLM-generated texts. Apart from that, the dataset is limited to publications prior to 2022, ensuring that no abstracts were written with the use of 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.

LLM abstracts were generated using three state-of-the-art models: **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). Table 1 summarises the models' technical characteristics. The GPT model was accessed via OpenAI API using openai[1] Python library. The two open-source models were accessed through the huggingface transformers[2] Python library using the text generation pipeline. The models were prompted with the temperature set to 1, while default values were retained for all other settings. The prompt included both a system message and a user message and reads as follows:

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]`

Table 2 presents some statistics of both human abstracts and those generated by the three models. As can be seen, Llama was considerably more verbose than other LLM and humans both in terms of the number of words and sentences, often exceeding the output length stipulated in the prompt. In contrast, the abstracts generated by Qwen were more laconic.

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

Model	Parameters	Layers	Context
gpt-4o-2024-08-06	–	–	128K
Llama-3.1-8B-Instruct	8B	32	128K
Qwen2.5-7B-Instruct	7.61B	28	131K

Table 1: Specifications of the LLMs used for abstract generation.

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 2: Number of tokens, types and sentences of abstracts by source.

3.2. Information Content

We approach the informativity of HWT and MGT with surprisal – an information-theoretic measure that quantifies how predictable or unpredictable a word is in a given context (Shannon, 1948). Formally, surprisal is defined as the negative logarithm of a word's conditional probability given its context. (see Equation 1). A number of experimental studies have shown that surprisal is positively correlated with processing effort as measured, for example, by reading times (Demberg and Keller, 2008; Delogu et al., 2017; Wilcox et al., 2023).

$$\mathrm{S}(w_i) = -\log_2 P(w_i \mid w_1, w_2, \ldots, w_{i-1}) \quad (1)$$

We used the smallest GPT-2 model[3] with 124M parameters to calculate surprisal, while providing the model with full abstracts as context. The smallest GPT-2 model was chosen because multiple studies have systematically shown that larger models generate lower surprisal values, providing a poorer fit compared to both human reading times (Oh et al., 2022; Oh and Schuler, 2023; Oh et al., 2024) and neuroimaging data (Lin and Schuler, 2026). This pattern may arise because increases in training data and parameter count lead larger models to memorise even rarely occurring linguistic patterns and, therefore, underestimate the processing difficulty of lower-frequency words.

Surprisal values were extracted using `surprisal`[4] Python library. Word surprisal was subsequently calculated by summing the surprisal values of subword tokens. Then, we calculated mean surprisal for each abstract, excluding punctuation marks. Finally, we calculated mean surprisal values for each part of speech (PoS).

[3]https://huggingface.co/openai-community/gpt2

[4]https://pypi.org/project/surprisal/

3.3. Information Density

Apart from the mere information load of text, another factor influencing comprehension is the distribution of information. If we part from the premise that humans are rational writers striving for efficient communication, we would expect them to distribute information evenly in texts to avoid extreme peaks and troughs. This is known as Uniform Information Density (UID) hypothesis (Frank and Jaeger, 2008). Following Philipp et al. (2023), we operationalise information density as *local UID* by calculating negative average surprisal difference of neighbouring tokens for each abstract (see Equation 2). In this implementation, values closer to 0 indicate a more uniform distribution of information in a text.

$$UID_{LOCAL} = -\frac{1}{n}\sum_{i=1}^{n}(S(w_i) - S(w_{i-1}))^2 \quad (2)$$

Consider, for instance, Examples 1 and 2. As can be seen in Figure 1, the first sentence has a higher peak in the information content, also reflected in a more negative UID local value of -32.25 compared to -13.3 of the second sentence. This means that Example 1 has a more uneven, and therefore, less optimal distribution of information.

(1) *In this paper, we consider the problem of shifted label distribution.*

(2) *In this paper, we propose a novel news bias dataset.*

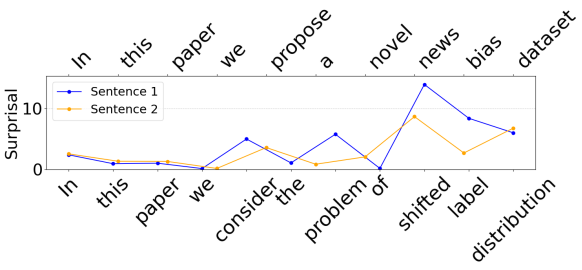

Figure 1: Example of more evenly (yellow) and less evenly (blue) distributed information in two sentences.

3.4. Topic Modelling

We conducted an extrinsic evaluation based on topic modelling to compare the thematic organisation of abstracts generated by LLMs with those written by humans. The goal was to assess how each system structures information into latent topics and to quantify their degree of similarity.

For each dataset (i.e., one per LLM and one for the human baseline), we applied BERTopic (Grootendorst, 2022), which combines transformer-based embeddings with clustering and class-based TF–IDF keyword extraction. Abstracts were first preprocessed by lowercasing, removing punctuation and digits, and filtering out stopwords using the NLTK English list extended with domain-specific terms (e.g., study, results, paper, method, analysis, data). This normalization reduced noise and improved topic interpretability.

We examined two experimental scenarios: (a) an unsupervised setting using BERTopic's default HDBSCAN clustering, which allows the algorithm to infer the optimal number of topics based on the intrinsic structure of each dataset; and (b) a controlled setting, where we replaced HDBSCAN with k-means, fixing the number of clusters to 20 to ensure consistent topic granularity across all models. In both scenarios, document embeddings were computed using the `all-MiniLM-L6-v2` model from SentenceTransformers (Reimers and Gurevych, 2019), which provides compact and semantically meaningful sentence-level representations.

For each configuration, BERTopic identified the ten most representative keywords per topic using a CountVectorizer configured with the same stopword list. Each abstract was assigned to one topic, and the resulting topic labels and keywords were saved for subsequent comparison.

After generating topic assignments and keywords for each model and for human-authored abstracts, we conducted a quantitative comparison of topic structures across all models by computing pairwise clustering similarity using Normalized Mutual Information (NMI). For each pair of models, the topic label arrays corresponding to the same abstracts were compared using scikit-learn's `normalized_mutual_info_score` function. NMI measures the agreement between two clusterings while being invariant to label permutations, and is normalized to range from 0 (no agreement) to 1 (perfect alignment). The resulting pairwise NMI values are presented in the format of a heatmap.

To complement this label-based comparison, we performed topic-level alignment based on keyword semantics. For each model, the top keywords of each topic were extracted and encoded into embeddings using the `all-MiniLM-L6-v2` model from SentenceTransformers (Reimers and Gurevych, 2019). Pairwise cosine similarity between topic embeddings of two models was computed, producing a similarity matrix where each entry represents the semantic closeness of a topic from one model to a topic in the other. For each topic in the first model, the most similar topic in the second model was identified, and the corresponding similarity score was recorded. The results were saved in a CSV file listing all aligned topic pairs and their cosine similarity scores.

4. Results

4.1. Surprisal and UID

We calculated the average surprisal values for each abstract. As shown in Figure 2, abstracts generated by GPT and Qwen exhibit slightly higher surprisal than human-written abstracts. In contrast, Llama produced texts with considerably lower information content. To further examine the effect of text source on surprisal, we fitted a mixed-effects model using the `lmerTest` package (Kuznetsova et al., 2017) in R 4.5.1 (R Core Team, 2025). The model used mean text surprisal as the response variable and text source as the predictor, with a random intercept for text ID. Possible effects of text length is minimised by the averaged response. We applied treatment coding to text source, setting *human* to be reference category. The mixed-effects model showed that the surprisal differences between all three LLMs and humans are statistically significant ($p < 0.001$) as illustrated in Table 3).

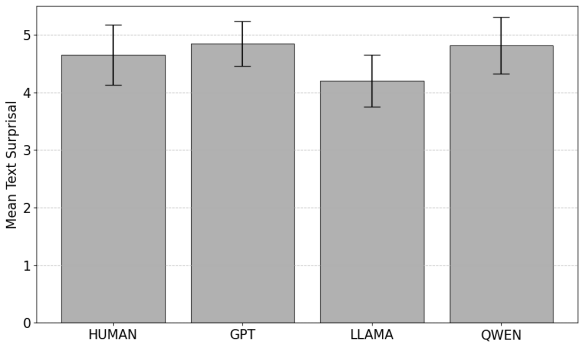

Figure 2: Grand mean surprisal for Humans, GPT, Llama, and Qwen. Error bars show standard deviation.

Effect	Estimate	Signif.
(Intercept)	4.646	***
SourceGPT	0.198	***
SourceLlama	-0.447	***
SourceQwen	0.166	***

Table 3: Effect of abstract source on mean surprisal. Significance: *** $p < 0.001$.

Looking at mean surprisal of PoS tags, Llama has the lowest values almost in all PoS categories (see Figure 4), which is in line with its overall lower surprisal on text level. Proper nouns (`PROPN`) show the highest values in all types of abstracts. This is plausible since words labelled with the `PROPN` tag include unique, often acronymic, names referring

to models (*XLMRoBERTa*, surprisal = 49.18), software (*EXMARaLDA*, surprisal = 47.33), datasets (*VSoLSCSum*, surprisal = 46.07), etc, that may be rare or non-existent in GPT-2's training data, resulting in surprisal values extremely larger than average. Among content words, proper nouns are followed by adjectives and verbs in all text sources in terms of their average surprisal. Interestingly, common nouns score relatively low.

Considering the distribution of information in abstracts, rather than the mere information content, we computed UID for each text. Being intrinsically negative, UID values closer to 0 indicate a more optimal distribution of information. As shown in Figure 3, the information content was more evenly distributed in abstracts generated by Llama, as compared to other LLMs and humans.

Similarly to how we proceeded with mean surprisal, we fitted a mixed-effects model using UID as response variable and text source as predictor, while also fitting random slopes for text ids. Again, confounding effects of text length is addressed by the averaged nature of UID. Here, however, we applied a sign inversion to UID values and subsequently log-transformed them to deal with heteroscedasticity of residuals. For the sake of readability, we report back-transformed estimates in Table 4. Again, all effects are statistically significant, but the difference between human-written abstracts and Llama abstracts seems to be much more practically relevant.

Effect	Estimate	Signif.
(Intercept)	-25.2	***
SourceGPT	0.45	***
SourceLlama	2.94	***
SourceQwen	-0.38	***

Table 4: Effect of abstract source on UID. Significance: *** $p < 0.001$.

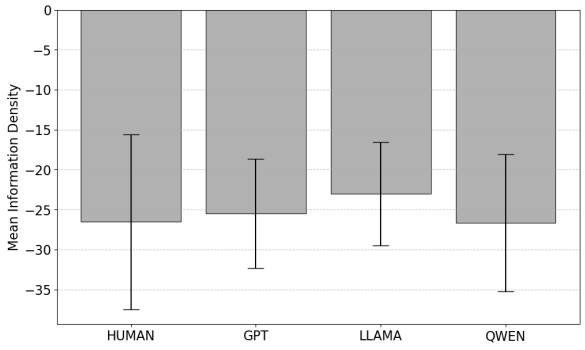

Figure 3: Mean UID. Values closer to 0 indicate a better distribution of information. Error bars show standard deviation.

4.2. Extrinsic Evaluation

As previously described, we also conducted a topic modelling experiment to assess the alignment between the MGT and HGT, evaluating the influence of MGT on NLP applications such as thematic clustering.

Two conditions were tested: (a) unsupervised, in which the clustering algorithm was free to determine the number of clusters for each dataset, and (b) controlled, in which the number of clusters was fixed at 20 for all datasets. Figure 5 shows the number of clusters identified for each LLM and the human abstract datasets.

In all cases, the number of clusters exceeds 140, with only small differences between humans, Llama, and Qwen. GPT abstracts generate the largest number of clusters, which could be due to its considerably larger vocabulary.

Figure 6 displays the NMI results for the pairwise comparison of clusters generated from each MGT and HGT dataset in both the unsupervised and controlled conditions.

NMI similarity scores are consistently higher in the controlled condition (0.59 to 0.66) compared to the unsupervised condition (0.53 to 0.61), as expected due to the smaller number of clusters considered. In both conditions, the lowest NMI scores correspond to the alignment between human and LLM clusters, indicating that human-authored abstracts and machine-generated texts tend to induce different topic distributions. Regarding the comparison among LLMs, in the unsupervised condition, GPT clusters are more aligned with Llama clusters, whereas this relationship is reversed in the controlled condition, though with a smaller difference. Additionally, in the controlled condition, the NMI scores between each LLM and human clusters show no notable differences.

These results suggest that MGT, even when semantically similar, produce patterns of thematic organisation that differ from HGT, affecting the resulting cluster structures. This divergence indicates that the use of LLMs in scientific writing may influence downstream NLP applications, such as topic modelling or clustering analyses, which rely on the statistical and semantic properties of the text.

Regarding the similarity of topics generated by MGT compared to HGT, Figure 7 shows the distribution of pairwise cosine similarity scores between human topics and each LLM in the controlled scenario, illustrating the semantic alignment of generated topics with human-authored ones.

Additionally, Table 5 complements this visualization by reporting the percentages of topics with similarity scores greater than or equal to 0.7, 0.8, and 0.9 for each LLM compared to humans.

The results presented in Figure 7 and Table 5 indicate that topics generated by Llama abstracts

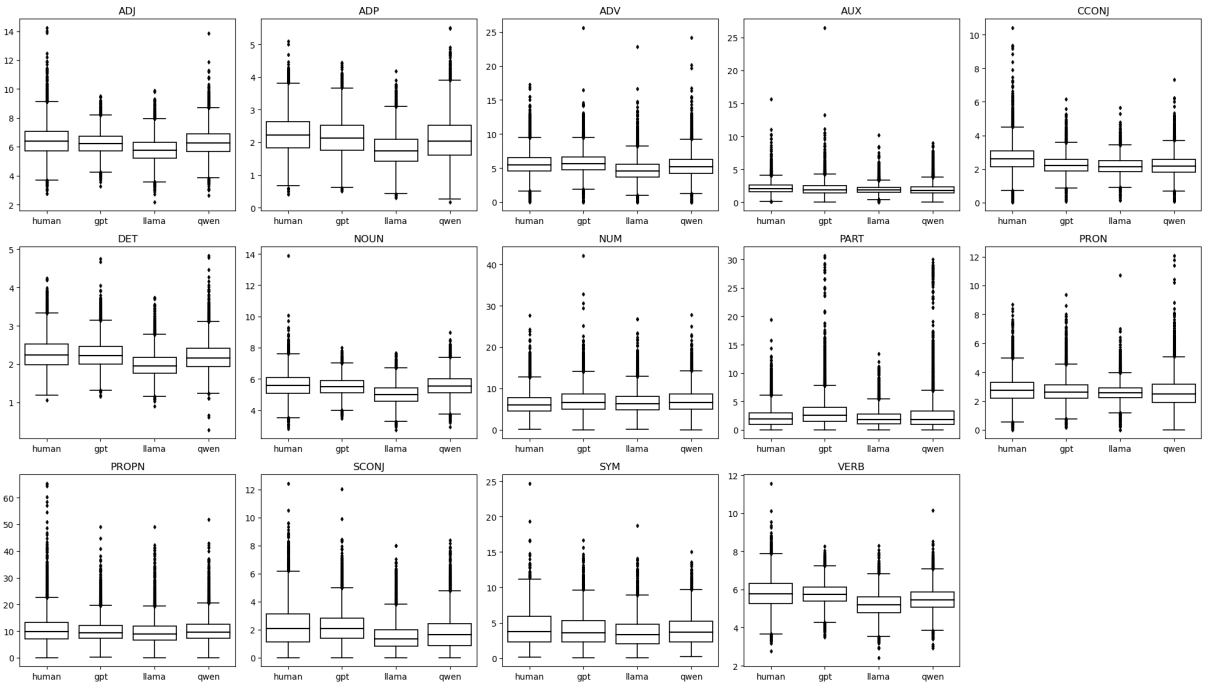

Figure 4: Mean PoS surprisal on text level.

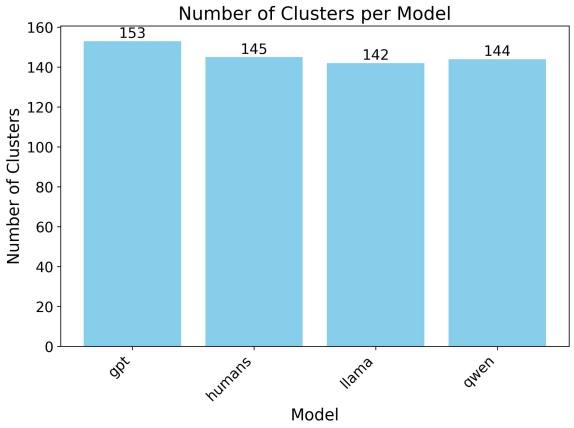

Figure 5: Number of clusters in the unsupervised condition for each LLM and human abstract dataset.

LLM	$\% \geq 0.7$	$\% \geq 0.8$	$\% \geq 0.9$
GPT	90.0	70.0	20.0
Llama	**95.0**	**85.0**	**40.0**
Qwen	90.0	80.0	30.0

Table 5: Percentages of topics with pairwise cosine similarity above or equal to 0.7, 0.8, and 0.9 for each LLM compared to HGT topics.

show the closest alignment with those extracted from HGT, followed by Qwen. This suggests that Llama tends to be more effective at producing text that preserves the thematic structure observed in human writing. Nevertheless, for all models, the vast majority of topics (i.e., at least 90% of the 20 topics) exhibit a cosine similarity of at least 0.7, indicating that even when differences exist, MGT generally capture the main semantic patterns present in human abstracts.

Table 6 lists the five human topics with the lowest similarity for each LLM, displaying the areas where the models diverge most from human topic representations.

Analysing the five topics with the lowest similarity to human topics for each LLM reveals some overlaps. GPT and Llama share three bottom topics, including areas related to speech, morphology, and semantic roles and metaphors. GPT and Qwen share two topics, touching on semantic roles and metaphors, as well as relation extraction and biomedical information, while Llama and Qwen also share two topics, including semantic roles and metaphors and discourse and story structure. The semantic roles topic has low similarity scores for all three models (lowest for Qwen, below 0.6), suggesting that in this specific area, the LLMs tend to produce abstracts with more lexical variation and less human-aligned expressions.

5. Discussion

Our results suggest that abstracts generated by Llama are less demanding in terms of cognitive processing in comparison to other models and humans as they show lower average surprisal and higher UID score. Lower surprisal might be due to Llama using a more general language vocab-

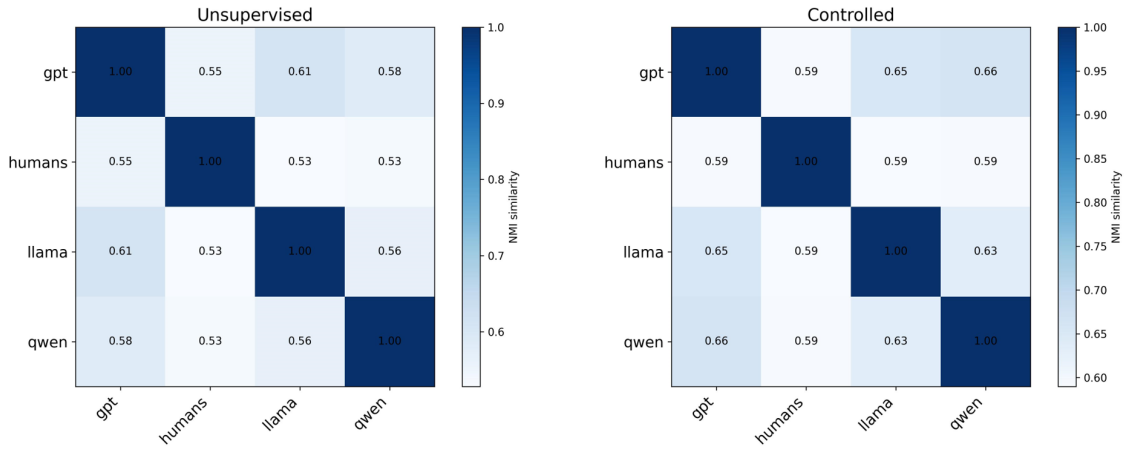

Figure 6: Normalized Mutual Information heatmaps showing cluster alignments between each LLM and human abstracts for the unsupervised (left) and controlled (right) conditions.

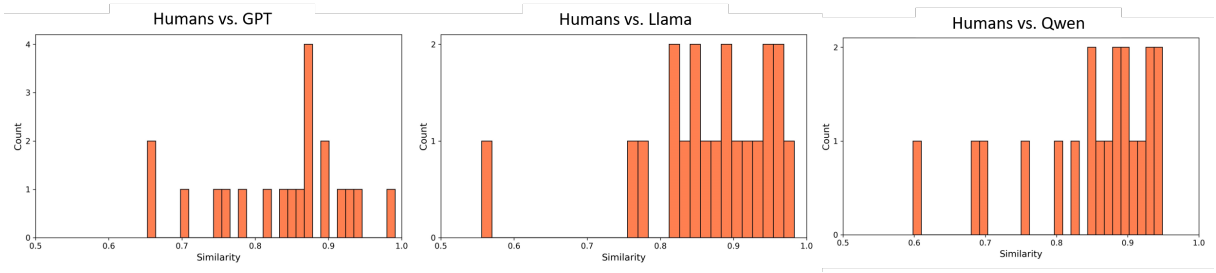

Figure 7: Histogram of cosine similarity scores between human topics and each LLM's topics. Higher values indicate greater semantic alignment with human-authored abstracts.

ulary. Better information distribution can also be explained by a less demanding vocabulary but also by the use of more conventional syntactic structures, reducing peaks and troughs in information content.

In the extrinsic evaluation via topic modelling and clustering, although Llama achieved a similar NMI compared to the other LLMs, it produced keywords that were semantically closer to HGT for the extracted topics. This suggests less semantic deviation from human-generated texts when compared to GPT and Qwen and stands in line with previous research suggesting stronger similarity between human writers and Llama on lexical and syntactic levels (Muñoz-Ortiz et al., 2024). However, although more than 90% of topics show semantic similarity between the LLMs and HGT, the NMI results indicate notable differences in the cluster classification of articles. This finding suggests that MGT may influence the outcomes of downstream NLP applications that rely on clustering or topic modelling.

6. Conclusion and Future Work

The aim of this study was to compare human-authored and machine-generated abstracts of scientific papers beyond traditional lexical and syntactic analyses. We investigated more than 10,000 abstracts generated by three state-of-the-art LLMs, comparing them with the original human-written abstracts from two perspectives: (a) information content and (b) topic distributions.

We operationalised information content as average abstract surprisal – an information-theoretic measure correlated with processing effort, which in this context reflects how much information per word is conveyed in each abstract. Additionally, we examined the evenness of information distribution, starting from the premise that a more uniform distribution enhances the reader's comprehension. Our results suggest that abstracts generated by Llama are less demanding in terms of cognitive processing.

For topic modelling, we applied the BERTopic model and evaluated both the consistency of generated clusters across categories (i.e. humans versus each LLM) and the similarity of key words used within each topic. Our findings indicate substantial divergence between humans and LLMs. However, in terms of semantic similarity of key words, Llama is closest to humans.

Future work will involve evaluations by human

LLM	Human Keywords	Similarity
GPT	social, media, news, detection, online, tweets, twitter, users, dataset, task	0.813
	speech, language, recognition, asr, system, error, corpus, model, text, automatic	0.795
	semantic, verbs, srl, role, metaphor, verb, syntactic, metaphors, expressions, constructions	0.783
	language, languages, morphological, word, chinese, segmentation, words, arabic, pos, tagging	0.757
	relation, extraction, information, relations, clinical, biomedical, knowledge, medical, entities, entity	0.728
Llama	translation, machine, alignment, statistical, smt, system, parallel, model, language, word	0.820
	speech, language, recognition, asr, system, error, corpus, model, text, automatic	0.813
	language, languages, morphological, word, chinese, segmentation, words, arabic, pos, tagging	0.779
	semantic, verbs, srl, role, metaphor, verb, syntactic, metaphors, expressions, constructions	0.757
	discourse, annotation, relations, story, structure, connectives, authorship, music, stories, text	0.556
Qwen	emotion, emotions, multimodal, sarcasm, sentiment, emotional, detection, tweets, model, irony	0.802
	relation, extraction, information, relations, clinical, biomedical, knowledge, medical, entities, entity	0.752
	image, visual, images, video, multimodal, captions, descriptions, language, dataset, caption	0.703
	discourse, annotation, relations, story, structure, connectives, authorship, music, stories, text	0.681
	semantic, verbs, srl, role, metaphor, verb, syntactic, metaphors, expressions, constructions	0.598

Table 6: Five human topics with the smallest similarity for each LLM. Similarity is measured as cosine similarity between model and human topics.

raters to assess the quality and perceived readability of the generated abstracts. We also plan to include models of varying sizes within the same family to examine the effect of model scale on writing quality, as well as to extend the analysis to other model families.

7. Limitations

This study presents several limitations that should be acknowledged. First, the analysis was performed on a limited number of language models. Including more recent and advanced models (e.g., GPT-5) could provide a broader and more up-to-date perspective on the current capabilities of LLMs.

Second, the outputs of large language models are highly influenced by the input prompts. Although we accounted for scientific English, our evaluation was restricted to a single disciplinary domain (that is, computational linguistics). Consequently, the findings may not fully generalise to other scientific fields.

Third, while the results indicate significant differences between MGT and HGT, with some models performing closer to humans according to certain measures, we have not yet conducted experiments involving human evaluation to determine whether these differences are perceptible to human readers.

Finally, we did not include automatic topic coherence metrics such as U-Mass or NPMI, which are commonly used as proxies for topic interpretability. Future work should incorporate human assessments of topic quality and examine their relationship to coherence metrics, text quality, and cognitive processing measures to provide a more comprehensive evaluation of model performance.

8. Acknowledgements

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

9. Bibliographical 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.

Sergei Bagdasarov and Diego Alves. 2025. Like a Human? A Linguistic Analysis of Human-written and Machine-generated Scientific Texts. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 38–47, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tony Berber Sardinha. 2024. AI-generated vs Human-authored texts: A Multidimensional Comparison. *Applied Corpus Linguistics*, 4(1):100083.

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).

Francesca Delogu, Matthew W. Crocker, and Heiner Drenhaus. 2017. Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, 161:46–59.

Vera Demberg and Frank Keller. 2008. Data from Eye-Tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2):193–210.

Austin F. Frank and T. Florian Jaeger. 2008. Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

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

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.

Maarten R. Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *ArXiv*, abs/2203.05794.

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.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

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.

Yi-Chien Lin and William Schuler. 2026. Surprisal from larger transformer-based language models predicts fmri data more poorly.

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.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

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.

J. Nathanael Philipp, Michael Richter, Erik Daas, and Max Kölbl. 2023. Are Idioms Surprising? In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 149–154, Ingolstadt, Germany. Association for Computational Lingustics.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Claude E Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

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.