Stylometric Analysis of Large Language Model-Generated Commentaries in the context of Medical Neuroscience

Jan K. Argasiński[1,2][0000-0002-2992-718X], Iwona Grabska-Gradzińska[1][0000-0002-5799-5438], Karol Przystalski[1][0000-0002-8572-1469], Jeremi K. Ochab[1,3][0000-0002-7281-1852], and Tomasz Walkowiak[4][0000-0002-7749-4251]

[1] Jagiellonian University, Kraków, Poland
[2] Sano - Centre for Computational Medicine, Kraków, Poland
[3] M. Kac Center for Complex Systems Research, Jagiellonian University, Kraków, Poland
[4] Faculty of Information and Communication Technology, Wroclaw University of Science and Technology, Wroclaw, Poland

Abstract. This study investigates the application of Large Language Models (LLMs) in generating commentaries on neuroscientific papers, with a focus on their stylometric differences from human-written texts. Utilizing three papers from reputable journals in the field of medical neuroscience, each accompanied by published expert commentaries, we compare these with commentaries generated by state-of-the-art LLMs. Through quantitative stylometric analysis and qualitative assessments, we aim to be a part of the discussion around the viability of LLMs in augmenting scientific discourse within the domain of medical neuroscience.

Keywords: Large Language Models · Stylometry · Text generation.

1 Introduction

In the context of the application of Large Language Models (LLMs) in medicine, the ability to distinguish between human-written and computer-generated text becomes critically important for several reasons [5].

First, it ensures the integrity and trustworthiness of medical research and its implementation [37]. Human experts often provide nuanced insights based on years of experience and tacit knowledge that LLMs, despite their advanced capabilities, might not fully replicate. Identifying the text's origin allows readers to weigh the insights accordingly, appreciating the depth of human expertise or the data-driven breadth of LLMs.

Second, distinguishing between these sources of text helps in maintaining a high standard of ethical transparency [52]. As the medical field relies heavily on evidence-based practice, the clear labelling of human versus AI contributions upholds the ethical standards of research dissemination and consumption. It

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

ensures that practitioners and researchers are fully informed about the nature of the information they are engaging with, facilitating informed decision-making processes.

Lastly, this distinction aids in the ongoing evaluation and improvement of LLMs themselves. Researchers can identify areas where LLMs excel or fall short by comparing human and AI-generated papers, guiding further development and training efforts. This not only enhances the utility of LLMs in medical applications but also ensures that these tools are used in a manner that complements, rather than supplants, human expertise [41].

In this paper, we compare artificially generated papers with human-written scientific literature. We do this by matching LLM-produced text with published commentaries on existing medical papers. Medical commentaries serve as a vital component of the scientific communication ecosystem, offering insights, critiques, and expanded discussion on published research findings. Unlike original research articles that require the presentation of new experimental data, statistical analyses, and results, commentaries primarily rely on the interpretation and discussion of existing studies. The distinction makes commentaries an ideal genre for exploring the potential of LLMs in medical literature generation, as the focus shifts from providing new, verifiable data to synthesising and discussing existing knowledge.

Opting for the generation of medical commentaries not only aligns with the strengths of LLMs but also offers a pragmatic pathway to evaluate their potential in medical literature. It allows for both qualitative and quantitative assessments of the generated texts, and enables the stylometric analyses – giving insights into their linguistic structure. Supplemented with a qualitative evaluation that is particularly suited to the nuanced and interpretative nature of commentaries – our study aims to explore the capabilities of LLMs in generating medical commentaries, comparing them with human-written texts to assess their potential and limitations. This approach not only highlights the current capabilities of LLMs but also sets the stage for future advancements and applications in the generation of medical texts.

2 Related works

Comparison of various LLMs should consider the aim of the source text processing and analysis: not only text summarisation and data extraction [29, 43, 50] but also simplification [31, 18], semantic similarity and reasoning, critical commentary and quality evaluation [7, 22]. Depending on the objectives of the source text analysis, fine-tuned domain-specific language models remain the first choice, rather than general LLM models [7]. The comparison of different models in the context of the aim of their use is shown in [16, 7]. The models used in the present paper are described in [53, 20, 44, 9].

There are some problems which should be addressed in the context of LLMs comparison. One of them is the text structure of scientific papers. Scientific works, due to different structuring standards and the use of data not only in the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

form of continuous text but also tabular data and graphs can be treated as unstructured texts [30, 11, 51] when analysed using LLMs. Another problem is the length of the article considered as a single statement. Studies have been carried out comparing models with long source texts [55]. The problem of hallucination in natural language generation is adressed in [24, 40].

Comparing the performance of LLMs to the effect of human labour can emphasize the factual correctness of the generated text [17, 15] as well as the stylistic (in)distinctiveness of the generated response [46, 4]. In the latter context, it becomes convenient to use linguistic and stylometric tools [35], which have long been used for problems of text attribution, verifying text's authorship and investigating characteristics of their style [42]. Moreover, research efforts are increasingly focused on the development of automated methods for detecting text generated by LLMs [32, 8].

Producing fake scientific papers has a long tradition, notably with 1996 Sokal's hoax and a later computer-science paper generator SCIgen [45], and so has the detection of gibberish or computer-generated papers [48, 26]. Since then, whole benchmarks for the task of detecting automatically generated academic papers have appeared [27, 34], which however may undergo a fast deterioration due to the exponential growth of LLM capabilities and the fact that authors may utilise LLMs as auxiliary tools in their writing [1]. More recent works focus specifically on detecting LLM-generated texts [19, 34], and the methods include stylometric analysis [54].

3 Methodology

3.1 Stylometry

Stylometry, as an academic field, constitutes an area of study within computational linguistics and digital humanities, focusing on the quantitative analysis of textual features to deduce metadata such as authorship, chronology, and stylistic evolution. This discipline leverages statistical and machine learning methodologies to scrutinize the stylistic fingerprints left by authors in their texts, thereby facilitating a deeper understanding of literary corpora beyond the capabilities of traditional approaches.

The potential utility of stylometry extends into the realm of analyzing LLMs, offering a lens through which to examine the generative capabilities and inherent biases of models.

Methods and setup For the unsupervised quantitative analysis of text style, two baseline methods from the R package 'Stylo' [13] were used: the principal component analysis (PCA) of the covariance matrix of feature frequencies and bootstrap consensus trees (BCT). The correlation matrix was less useful since the real commentaries (number one and two) stand out mainly due to their larger length. The features checked included the normalised occurrence frequencies of the most frequent words (MFWs; we chose the typical range between 100 and

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

1000 MFWs and, for BCT, they were iterated by 100) and their N-grams (again, following the typical choice from 1 to 3). The features were 'culled' in the range from 0% to 25% (which means the minimum percentage of texts in which a feature must occur; higher values were leaving too few features to compute a BCT). The consensus trees used the cosine delta distance, commonly considered in stylometry the reliable choice [14].

As a more modern alternative, the authors' own pipeline for interpretable stylometric analysis [36] was used. The pipeline used Spacy [33] 'en_core_web_lg' model for preprocessing steps (including tokenisation, named entity recognition, dependency parsing, and part-of-speech annotation), LighGBM [25] as the state-of-the-art DART boosted trees classifier, SHAP [28] for computing explanations, and Scikit-learn [38] for feature counting and cross-validation. The texts were chunked into 50-token samples. The extracted features included: 1-3-grams of lemmas and parts of speech, dependency-based bigrams of lemmas, NER entity types, and morphological annotations. Binary classification was performed between the real and fake commentary for each LLM. The baseline corresponds to a dummy classifier with the strategy of choosing the most frequent class. Stratified 10-fold cross-validation (i.e., in each fold the ratio training: test set size was 9:1) was repeated 10 times to collect more reliable statistics. Within the training sets, 10% data were used for model validation.

Both stylometric analyses excluded: the abstract, keywords, figures or tables and their captions, the lists of authors, affiliations, footnotes, and references, as well as sections like acknowledgements. Preprocessing included also removal of hyphenations at line breaks, leaving line breaks only between paragraphs, converting numerical citations to in-text citations (author, year), normalising quotation marks, etc.

3.2 Qualitative criteria for text evaluation

In conducting the study, specific heuristics were selected for the qualitative comparison of texts: the annotators answered if the generated texts provided an:

1. accurate summarization and referencing of original research,
2. correct references to real academic papers,
3. proper abstraction of relevant knowledge from the cited papers,
4. coherent argumentation of presented arguments,
5. realistic numerical results, tables, or figures,
6. strict scientific knowledge – in terms of factual correctness,
7. strict scientific knowledge – in terms of being state-of-the-art,
8. fitting structure/argumentation as expected from a commentary,
9. pertinent tone/style as expected from a commentary,
10. qualitatively new insight with respect to the original paper.

The responses: "Yes", "No" or "Partly/Not applicable" are summarized in Table 2. "Partly" was used in several cases, including partial correctness of the text, but also only paraphrasing the information already given in the prompt

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

or in the original paper (hence, assumed to be correct, e.g., in terms of factual correctness) without any new information produced.

Two authors of this paper, JKA and JKO, were the annotators. The papers matched the annotators' expertise. They had access to the original papers, the real commentaries, and the LLM-generated ones and could read them and respond to the criteria at their own discretion and in any order.

3.3 Large Language Models Used

The Large Language Models considered for the generation of commentary were: PT-4-1106-preview (general purpose) [2], Google Gemini (general purpose) [47], and MED-PaLM2 (medical purpose) [44].

The GPT and Gemini were chosen as one the best LLMs for general purpose. We tried to use LLMs dedicated to medical cases, but only MED-PaLM2 was a ready to use LLMs. Meditron [9] and MedAlpaca [21] are also designed for medical purposes, but are not fine-tuned and for this reason both were excluded for this research.

Limitations of considered LLMs The final two models mentioned in the previous subsection, namely Meditron and MedAlpaca, are publicly available on HuggingFace[5] platform. Meditron requires that users agree to share their contact information. These models are based on the Llama 2 model and have been trained on medical data, making them potentially suitable for the analysis outlined in this study. However, a detailed analysis reveals two significant limitations that make them unsuitable for use in this research:

The initial issue concerns the fine-tuning process in constructing these LLMs. The publicly accessible Meditron has not undergone fine-tuning; it remains a raw pre-trained model. Although it has been exposed to articles and can generate content similar to them, it lacks comprehension when it comes to the specific task of commentary article generation. However, MedAlpaca has been fine-tuned, but only for question-type tasks, excelling in providing answers related to medical issues. However, it fails to grasp the task of article generation, leading to very poor results that do not resemble an academic article at all.

The second issue relates to the challenge of context length. The prompts used to test the LLM (see Section 4.2) consist of more than 9,000 tokens. However, Meditron-7B and MedAlpaca-7B are limited to 2048 tokens each, while Meditron-70B has a capacity of 4096 tokens. Consequently, both the query and the response must comply with these constraints (2048, 4096). While it is feasible to trim the query, any adjustments must consider the model's limitations and allocate space for the response (e.g., accommodating up to 1024 tokens).

In summary, above mentioned open-source models pre-trained on medical data are still far from having the capability to generate commentaries resulting in their exclusion from the study.

[5] https://huggingface.co/

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

4 Data

The experimental data consists of three original research papers [49, 23, 39], together with their commentary articles [6, 10, 3] and the corresponding LLM-generated commentaries.

Only the commentaries were analysed, while the original research papers were only used in LLM prompts.

4.1 The source papers and commentaries

The selection criteria for the source papers were: existing commentary article in the same journal, highly ranked journal within the research area of the paper, publication within the last 5 years.

4.2 The generation of LLM's commentary (prompts)

The prompt used to generate the commentaries included the information on the type of academic paper to be written (a commentary), the journal in which it should be published (the same as original), the focus (criticism of methodology and the interpretation of the results; not to summarise the whole paper), the method (to cite the scientific papers following given arguments), and lastly – the text of the original research paper (without references).

Prompt structure The prompt has following structure:

1. Paragraph: *Given the following article, write a commentary article to be published in the same journal. Consider only the criticism of the methodology and the interpretation of the results. Do not summarise the whole text. Cite the scientific papers with your arguments. Use only real, published scientific work:*
2. Citation of the original paper including title and full journal name.
3. Phrase: *The original article is provided below:*
4. The text of the original research paper with abstract, highlights (when apply) etc. but without references.

5 Results

5.1 Quantitative Analysis

The unsupervised methods visualised in Fig. 1 show that (i) the texts mostly cluster according to which paper they were commenting, (ii) GPT-4's output consistently clusters with the real texts, while the other two models form separate clusters. The figure presents only one chosen set of parameters (the features were 1000 MFWs for PCA and 100-1000 MFWs for the BCT), but the results were stable over other MFW ranges, their N-grams, and culling.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

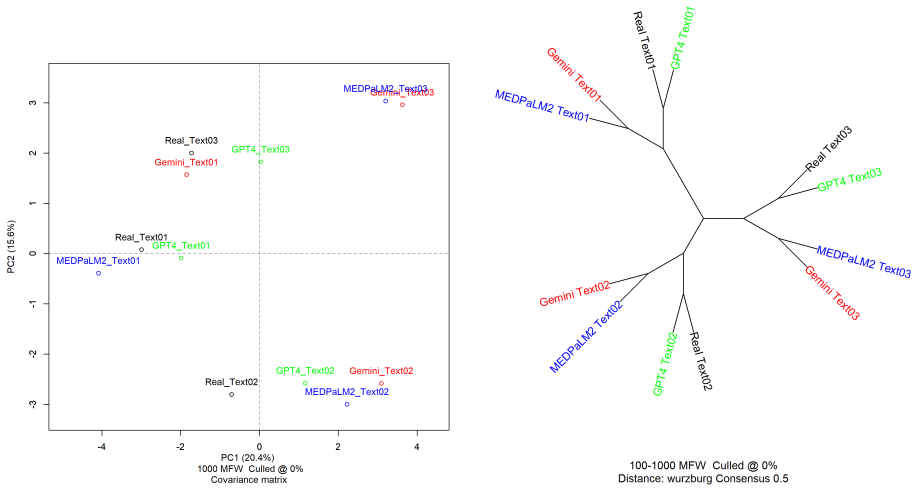

Fig. 1. Unsupervised visualisation: (left) Covariance PCA, (right) Bootstrap Consensus Tree.

Table 1 shows the results of the LGBM classifier with around 3000 features. Although the accuracy scores are close to the baseline due to a large class imbalance (the real commentaries were longer, hence, they provided a larger number of samples), the F1 scores show a reasonable performance, given that the samples are short and come in small numbers. Such scores allow us to use SHAP explanations to obtain an idea about the textual features that make the LLM-generated texts imperfect. As shown in Fig. 2 (middle), the features responsible for detection can be as simple 'SPACE', which in fact corresponds to the number of paragraphs delimited by line breaks (large in Gemini and MED-PaLM2); another example is the usage of plurals, 'Number:Plur' (over the top in Gemini and MED-PaLM2, but underused in GPT-4 with respect to the real texts).

5.2 Qualitative Assessment

[Paper 1] GPT-4: The generated text correctly referred to specific paragraphs or in the original paper including some numerical results. The structure, register and tone were all acceptable for a commentary article, with the exclusion of the commentary title. The argumentation heavily relied on the discussion section of the original paper but it did not introduce any new ideas. The arguments were all reasonable (except, perhaps, a plea for consensus in the field rather than correction of the paper) but since they followed the limitations mentioned by the original paper's authors, they would not constitute a reason to publish the commentary. There were 10 references, out of which nine appeared in the original paper, five were on the reference list (three correct, one not cited in the text, one had incorrect journal/volume/page, one had an incorrect but existing author and incorrect year/page), four in-text citations were not on the list.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

LLM	Train	Val	Test	Imbalance	Accuracy [baseline]	F1 [baseline]	Recall
GPT-4	100	12	12	1.8	0.75+/-0.11 [0.646+/-0.025]	0.7+/-0.1 [0.3924+/-0.0094]	0.82+/-0.15 [0]
Gemini	88	10	11	2.8	0.844+/-0.083 [0.735+/-0.022]	0.79+/-0.12 [0.4234+/-0.0071]	0.70+/-0.28 [0]
MED-PaLM2	96	11	12	2.1	0.78+/-0.09 [0.673+/-0.018]	0.74+/-0.12 [0.4021+/-0.0063]	0.61+/-0.19 [0]

Table 1. LGBM classification results of 50-token samples. The left-hand side of the table provides the median number of samples (across all cross-validation runs) in training, validation, and test sets and the ratio of the numbers of real to fake samples. The performance metrics are provided against the baseline dummy classifier in square brackets.

MED-PaLM2: The model misunderstood the task and produced responses to a hypothetical reviewer's comments. No new numbers or tables were generated. The generated text and all the scientific statements it contained referred to and paraphrased specific paragraphs from the original paper. The reference list comprised only a PubMedCentral link to the original paper, but there were seven other in-text citations.

Gemini: The generated text referred to specific paragraphs or items in the original paper, e.g., to tables 2-4 (correctly quoting their captions) or to the numerical value of median age [correctly: mean]. It also contained a fictitious quote (together with the page, where it should be found in the article), although one close in meaning to several existing sentences. The model made an untrue statement that the authors of the original paper "do not provide any evidence ... specific examples or data" to support their claims. There were no citations, a reference list nor new numbers or tables. The argumentation consisted of repeated general statements that repeated limitations mentioned in the paper itself; it did not contain any other piece of specific scientific knowledge. The tone and convention were acceptable with the exclusion of the section titles.

[Paper 2] GPT-4: The produced text accurately cited particular sections and numerical findings from the initial paper. The format, style, and tone were suitable for an analytical piece. However, the discussion didn't present any novel concepts, adhering closely to the constraints acknowledged by the authors of the original study. While the arguments made were logical, their reliance on the original paper's limitations means they do not provide a sufficient basis for the publication of the commentary. Furthermore, although all references were accurately cited, they were directly drawn from the information provided in the prompt.

MED-PaLM2: The model undertook a secondary task of summarizing the text and identifying its strengths and limitations. The strengths and limitations highlighted were directly extracted from the prompt provided. There are no

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

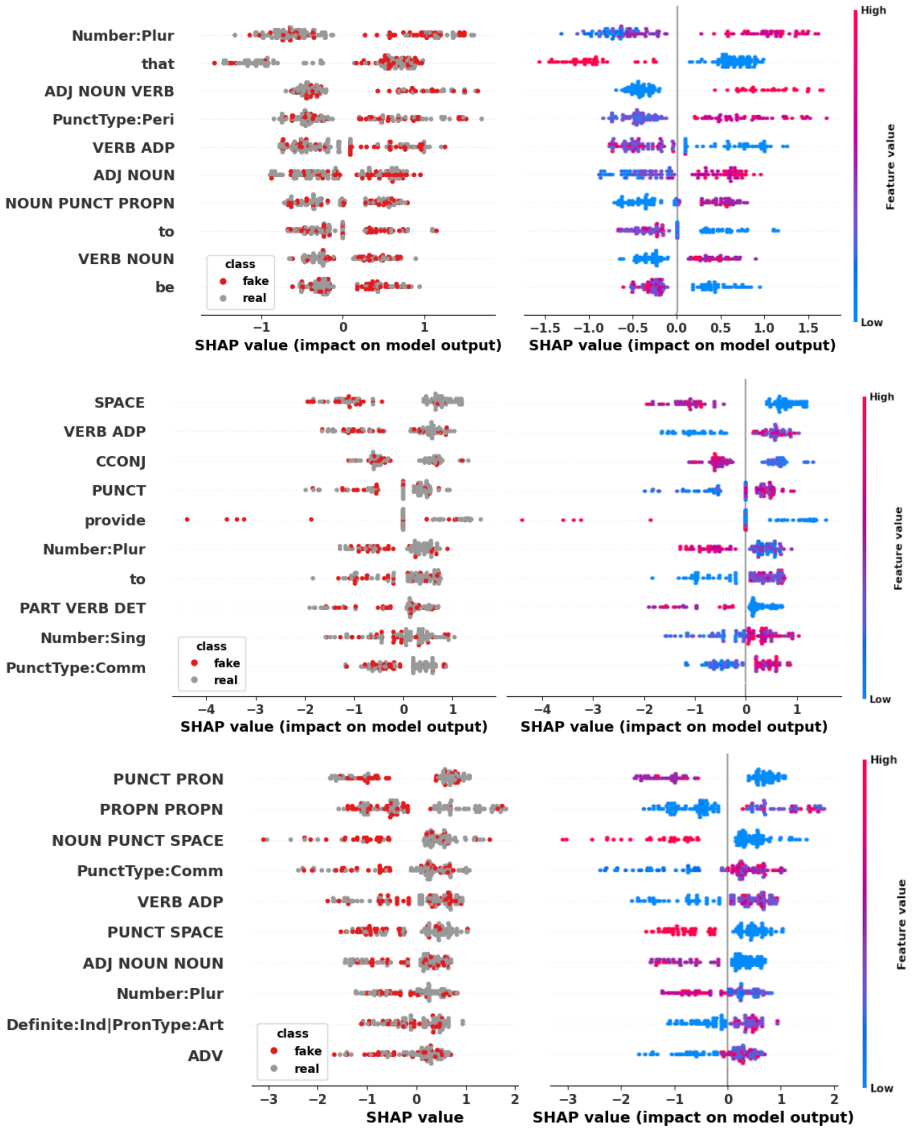

Fig. 2. SHAP values of the first 10 features most important for classifying real commentary vs (top) GPT-4, (middle) Gemini, and (bottom) MED-PaLM2. Each point is a 50-token text sample coloured (left) by its class membership or (right) by its feature intensity. Positive SHAPs point toward real texts, and negative toward fake ones.

quotes, footnotes, or references to numerical data beyond the most basic information, such as the number of study participants or the number of experiments conducted. The structure of the output does not resemble that of an article but

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

rather that of a summary. While the discourse maintains a scientific tone, it lacks any element of novelty. This approach reflects a concise synthesis and critique of the original content, focusing on presenting a distilled overview rather than expanding on the data or introducing new interpretations.

Gemini: The model performed a task very similar to that of MED-PaLM2 as described above. It lacks original observations, and the entirety of its output is limited to a summary of the text prompt devoid of quantitative data and references. While the style remains appropriate, the structure is even more generalized than in the previous instance. This approach suggests a focus on summarizing content without adding new insights or detailed analysis, aligning closely with the provided instructions yet falling short of contributing to a deeper understanding or expanding on the topic with additional context or evidence.

[Paper 3] GPT-4: In this instance, the model's output is at times incoherent (e.g., going without a logical link from the issue of spatial and temporal resolution to DTI providing structural but not functional connectivity), but it correctly addresses the original article and is mostly correct in contextualising the citation. Interestingly, while there is no reference list, the in-text citations are mostly correct (three taken from the original paper, six newly added, and one with the correct author but incorrect year). The new references can be tracked to exist: they are highly cited papers of recognisable authors in the field, which however makes the citations slightly dated. The structure and style are acceptable for a commentary, with a slightly over-the-top introductory paragraph, not a fitting title, and comments not as specific as usually seen in commentaries. Some remarks indicate knowledge from outside the commented paper and involve careful reading and checking, to ascertain their factual correctness.

MED-PaLM2: The model focused on limited discussion rather than any factual errors (cf the real commentary), making rather general statements on "critical evaluation of study quality, a balanced discussion of conflicting results, a comprehensive analysis of the strengths and limitations of different neuroimaging techniques [...] the generalizability of findings, ethical considerations, and future research directions". Some of these remarks were missed (e.g., neuroimaging techniques were discussed adequately), while some were new with regard to the original paper (e.g., ethical considerations). The overall tone and structure reminded that of a reviewer's comments rather than a commentary article. There were no citations, numbers, etc. From underneath the generalities, there were only glimpses of specific pieces of scientific knowledge.

Gemini: The model's output is very similar to that of MED-PaLM2 above. Again, it mostly comprises general statements on the original paper's limited discussion. As before, some criticisms are poorly grounded (e.g., that individual modalities had been discussed in isolation, as the original text mentions a few comparisons between the neuroimaging techniques – although, admittedly, part of this information had been presented in a figure not included in the prompt).

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

	GPT-4			MED-PaLM2			Gemini		
	paper 1	paper 2	paper 3	paper 1	paper 2	paper 3	paper 1	paper 2	paper 3
1. summary	✓,✓	✓,✓	✓,✓	✓,✓	✓,✓	✶,✶	✓,✶	✓,✶	✶,✶
2. references	✶,✶	✓,✓	✓,✓	✶,✶	⊗,⊗	⊗,⊗	⊗,⊗	⊗, ⊗	⊗,⊗
3. citing	✓,✶	✶,✶	✶,✶	✶,✶	✓,⊗	⊗,⊗	⊗,⊗	✶,⊗	⊗,⊗
4. coherence	✓,✓	✓,✓	✓,✶	✓,✓	✓,✓	✶,✓	✶,✗	✓,✓	✶,✓
5. numbers	✶,⊗	⊗,⊗	⊗,⊗	⊗,⊗	⊗⊗	⊗,⊗	✶,⊗	⊗,⊗	⊗,⊗
6. factuality	✶,✶	✓,✓	✓,✓	✶,✶	✶,✶	✶,✶	✶,⊗	✶,✶	✶,✶
7. SOTA	✶,✶	✓,✓	✓,✶	✶,✶	✶,⊗	⊗,⊗	✶,⊗	✶,⊗	✗,⊗
8. structure	✓,✓	✓,✓	✓,✓	✗,✗	✶,✗	✗,✗	✓,✗	✗,✗	✶,✓
9. tone	✓,✓	✓,✓	✓,✓	✗,✗	✶,✶	✶,✶	✓,✶	✓,✶	✗,✶
10. novelty	✗,✗	✗,✶	✶,✶	✗,✗	✗,✶	✗,✶	✗,✗	✗,✶	✗,✶

Table 2. The left column lists abbreviated criteria from Sec. 3.2.
The evaluation scale was: Yes – ✓, No – ✗, Partly – ✶, N/A (feature does not appear in the text) – ⊗. The inter-annotator reliability was good as measured by ordinal Krippendorff's alpha, $\alpha = 0.77$, 95%CI[0.67,0.86].

6 Discussion

Qualitatively, current LLMs produce grammatically and semantically correct texts. The generated papers' content usually does not exceed the content of the original research paper, and as such can be merely considered a summary of the paper's limitations or discussion sections. The best model, however, can write a compelling commentary article that references relevant studies, and whose tone and structure agree with editorial standards for that particular article type, and thus hypothetically it might incur some editorial costs before rejection.

Classic stylometric techniques provide a reliable classification of bag-of-word (BoW) text samples for sizes $>$ 2000 tokens in corpora of 100 novels [12]. In our easier setup (binary classification and text segments instead of BoW), a reasonable performance is obtained even for the small sample and short texts (50 tokens!). The repeated cross-validation constitutes tentative evidence that LLMs do have stylistic preferences which can be used to detect them and which one can track as individual, explainable features.

Limitations It is uncertain whether an LMM output's style is inherent to the model or provoked by the prompt. The working assumption was that given the same prompt the outputs of all LLMs are comparable. However, how changes to the prompt result in different outputs might depend on the LLM.

Admittedly, the tree classifiers tend to overfit – hence our extensive use of cross-validation – and a larger sample of texts would be advisable. The stability of indiavidual features and their SHAP importance depends on the size of the training sample as much as the generalisability of the classifier.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

Another limitation of this paper is no blind review and no control. The authors who evaluated the generated texts knew that they were produced by LLMs. This limitation was to be alleviated by the introduction of the inter-annotator agreement parameter.

7 Conclusion

We demonstrate the possibility of applying stylometric methods for analyzing computer-generated texts within the medical neuroscience domain. Scientific and domain-specific texts are significantly more challenging to generate effectively due to their grounding in real knowledge and facts, which cannot be easily summarized from a general knowledge base. These types of errors produced by the state-of-the-art language models can be assessed only by manual qualitative evaluation.

The continuous improvement of machine learning models necessitates the development of methodologies to differentiate their outputs from human-written text. Stylometry, which analyzes literary styles statistically, is a proposed method. However, its effectiveness in identifying machine-generated texts, which increasingly resemble human writing, is uncertain. Thus, validating this quantitative approach is crucial, as it would clarify the potential and limitations of stylometry in this evolving field. This highlights the need for further research to refine the tools for distinguishing between different sources of text generation.

Acknowledgements The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund.

JKO and TW's research was financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19. JKO's research has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

1. Abani, S., Volk, H.A., De Decker, S., et al.: ChatGPT and scientific papers in veterinary neurology; is the genie out of the bottle? Frontiers in Veterinary Science **10** (2023). https://doi.org/10/gtkf43

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

2. Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Bagić, A., Bowyer, S., Funke, M., et al.: Commentary on "Mapping the Unconscious Brain: Insights From Advanced Neuroimaging". Journal of Clinical Neurophysiology **40**(3), 269 (Mar 2023). https://doi.org/10/gtktkx
4. Bethany, M., Wherry, B., Bethany, E., et al.: Deciphering textual authenticity: A generalized strategy through the lens of large language semantics for detecting human vs. machine-generated text (2024)
5. Bruckert, S., Finzel, B., Schmid, U.: The next generation of medical decision support: A roadmap toward transparent expert companions. Frontiers in Artificial Intelligence **3**, 507973 (2020). https://doi.org/10.3389/frai.2020.507973
6. Caruana, F.: Positive emotions elicited by cortical and subcortical electrical stimulation: A commentary on Villard et al. (2023). Cortex (Aug 2023). https://doi.org/10/gtkcqj
7. Chen, Q., Du, J., Hu, Y., et al.: Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations (2024)
8. Chen, Y., Kang, H., Zhai, V., et al.: Token prediction as implicit classification to identify LLM-generated text. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 13112–13120. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.810
9. Chen, Z., Hernández-Cano, A., Romanou, A., et al.: Meditron-70b: Scaling medical pretraining for large language models (2023)
10. Clayson, P.E., Kappenman, E.S., Gehring, W.J., et al.: A commentary on establishing norms for error-related brain activity during the arrow flanker task among young adults. NeuroImage **234**, 117932 (Jul 2021). https://doi.org/10/gtkcqp
11. Dunn, A., Dagdelen, J., Walker, N., et al.: Structured information extraction from complex scientific text with fine-tuned large language models (2022)
12. Eder, M.: Short Samples in Authorship Attribution: A New Approach. In: Digital Humanities 2017. ADHO, Montréal, Canada (2017), https://dh2017.adho.org/abstracts/341/341.pdf
13. Eder, M., Kestemont, M., Rybicki, J.: Stylometry with R: A Package for Computational Text Analysis. The R Journal **8**(1), 1–15 (2016). https://doi.org/gghvwd
14. Evert, S., Proisl, T., Jannidis, F., et al.: Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities **32**(suppl_2), ii4–ii16 (Dec 2017). https://doi.org/10.1093/llc/fqx023
15. Fu, J., Ng, S.K., Jiang, Z., et al.: GPTScore: Evaluate as You Desire (2023)
16. Gu, Y., Zhang, S., Usuyama, N., et al.: Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events (2023)
17. Guo, B., Zhang, X., Wang, Z., et al.: How close is chatgpt to human experts? comparison corpus, evaluation, and detection. ArXiv **abs/2301.07597** (2023), https://api.semanticscholar.org/CorpusID:255998637
18. Guo, Y., Qiu, W., Leroy, G., Wang, S., Cohen, T.: Retrieval augmentation of large language models for lay language generation. Journal of Biomedical Informatics **149**, 104580 (2024). https://doi.org/10.1016/j.jbi.2023.104580
19. Hamed, A.A., Wu, X.: Detection of ChatGPT Fake Science with the xFakeBibs Learning Algorithm (Feb 2024). https://doi.org/10.48550/arXiv.2308.11767
20. Han, T., Adams, L.C., Papaioannou, J.M., et al.: MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data (Oct 2023). https://doi.org/mr5g

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

21. Han, T., Adams, L.C., Papaioannou, J.M., et al.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)
22. Heseltine, M., von Hohenberg, B.C.: Large language models as a substitute for human experts in annotating political text. Research & Politics **11**(1), 20531680241236239 (2024). https://doi.org/10.1177/20531680241236239
23. Imburgio, M.J., Banica, I., Hill, K.E., et al.: Establishing norms for error-related brain activity during the arrow Flanker task among young adults. NeuroImage **213**, 116694 (Jun 2020). https://doi.org/10/ggp975
24. Ji, Z., Lee, N., Frieske, R., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12) (mar 2023). https://doi.org/10.1145/3571730
25. Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30**, 3146–3154 (2017)
26. Labbé, C., Labbé, D., Portet, F.: Detection of Computer-Generated Papers in Scientific Literature. In: Degli Esposti, M., Altmann, E.G., Pachet, F. (eds.) Creativity and Universality in Language, pp. 123–141. Lecture Notes in Morphogenesis, Springer International Publishing, Cham (2016). https://doi.org/c8wr
27. Liyanage, V., Buscaldi, D., Nazarenko, A.: A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. In: Calzolari, N., Béchet, F., Blache, P., et al. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 4692–4700. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.501
28. Lundberg, S.M., Erion, G., Chen, H., et al.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**(1), 2522–5839 (2020)
29. Luo, Z., Xie, Q., Ananiadou, S.: The lay person's guide to biomedicine: Orchestrating large language models (2024)
30. Maharjan, J., Garikipati, A., Singh, N.P., et al.: Openmedlm: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models (2024)
31. Maruyama, T., Yamamoto, K.: Extremely low resource text simplification with pre-trained transformer language model. In: 2019 International Conference on Asian Language Processing (IALP). pp. 53–58 (2019). https://doi.org/mr5d
32. Mitchell, E., Lee, Y., Khazatsky, A., et al.: Detectgpt: zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
33. Montani, I., Honnibal, M., Honnibal, M., et al.: explosion/spaCy: v3.7.2: Fixes for APIs and requirements (Oct 2023). https://doi.org/10.5281/zenodo.10009823
34. Mosca, E., Abdalla, M.H.I., Basso, P., Musumeci, M., Groh, G.: Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era. In: Ovalle, A., Chang, K.W., Mehrabi, N., et al. (eds.) Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). pp. 190–207. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10/gtkf4w
35. Muñoz-Ortiz, A., Gómez-Rodríguez, C., Vilares, D.: Contrasting linguistic patterns in human and llm-generated text (2023)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20

36. Ochab, J.K., Walkowiak, T.: A pipeline for interpretable stylometric analysis. In: Digital Humanities 2024: Conference Abstracts. George Mason University (GMU), Washington, D.C. (2024, submitted)
37. Ordish, J., Hall, A.: Black box medicine and transparency: Interpretable machine learning. PHG Foundation (2020), last accessed 2023/02/26
38. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
39. Qureshi, A.Y., Stevens, R.D.: Mapping the Unconscious Brain: Insights From Advanced Neuroimaging. Journal of Clinical Neurophysiology **39**(1), 12–21 (Jan 2022). https://doi.org/10/gtktkw
40. Rebuffel, C., Roberti, M., Soulier, L., et al.: Controlling hallucinations at word level in data-to-text generation. Data Min. Knowl. Discov. **36**(1), 318–354 (jan 2022). https://doi.org/10.1007/s10618-021-00801-4
41. Rubinger, L., et al.: Machine learning and artificial intelligence in research and healthcare. Injury **54**, S69–S73 (2023)
42. Sadasivan, V.S., Kumar, A., Balasubramanian, S., et al.: Can ai-generated text be reliably detected? ArXiv **abs/2303.11156** (2023). https://doi.org/10.48550/arXiv.2303.11156
43. Shyr, C., Hu, Y., Bastarache, L., et al.: Identifying and extracting rare diseases and their phenotypes with large language models. Journal of Healthcare Informatics Research pp. 1–24 (01 2024). https://doi.org/10.1007/s41666-023-00155-0
44. Singhal, K., Tu, T., Gottweis, J., et al.: Towards Expert-Level Medical Question Answering with Large Language Models (May 2023). https://doi.org/10.48550/arXiv.2305.09617, arXiv:2305.09617 [cs]
45. Stribling, J., Krohn, M., Aguayo, D.: SCIgen - An Automatic CS Paper Generator, https://pdos.csail.mit.edu/archive/scigen/
46. Tang, R., Chuang, Y.N., Hu, X.: The science of detecting llm-generated texts (2023)
47. Team, G., Anil, R., Borgeaud, S., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
48. Van Noorden, R.: Publishers withdraw more than 120 gibberish papers. Nature (Feb 2014). https://doi.org/10/r3n
49. Villard, C., Dary, Z., Léonard, J., et al.: The origin of pleasant sensations: Insight from direct electrical brain stimulation. Cortex **164**, 1–10 (Jul 2023). https://doi.org/10/gtkcqm
50. Wang, A., Pang, R.Y., Chen, A., et al.: Squality: Building a long-document summarization dataset the hard way. In: Conference on Empirical Methods in Natural Language Processing (2022), https://api.semanticscholar.org/CorpusID:248987389
51. Wiest, I.C., Ferber, D., Zhu, J., et al.: From text to tables: A local privacy preserving large language model for structured information retrieval from medical documents. medRxiv (2023). https://doi.org/10.1101/2023.12.07.23299648
52. World Health Organization: Ethics and governance of artificial intelligence for health: Who guidance. Guidance, World Health Organization (2021)
53. Wu, C., Lin, W., Zhang, X., et al.: PMC-LLaMA: Towards Building Open-source Language Models for Medicine (Aug 2023). https://doi.org/10.48550/arXiv.2304.14454, arXiv:2304.14454 [cs]
54. Zaitsu, W., Jin, M.: Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. PLOS ONE **18**(8), e0288453 (Aug 2023). https://doi.org/10/gtkf46
55. Zhang, X., Chen, Y., Hu, S., et al.: ∞bench: Extending long context evaluation beyond 100k tokens (2024)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63775-9_20