Uncanny Semantics: How AI and Human Authors Use Language Differently in Academic Writing

Dennis Wegerhoff

University of Wuppertal

Abstract

After the paradigm shift towards Large Language Models (LLMs) in Artificial Intelligence (AI), AI models can be used to generate academic texts, with questionable implications for academic trust and integrity. This study investigates semantic differences between AI-generated and human-written texts in the field of German linguistics. A total of 325 introductions to linguistic papers[1], both human- and AI-authored across various models, were analyzed. The findings indicate that evaluative language—especially negative or critical expressions—is more frequently used in human-written texts. In contrast, AI-generated texts on average tend to overemphasize scientific methodology while avoiding overtly critical language.

1 Introduction: The Uncanny Valley of AI

With Artificial Intelligence (AI) on the rise, distinguishing between human intellectual work and its AI-generated counterpart is becoming increasingly challenging (cf. e. g. Frank et al. 2023). This issue affects virtually every field involving intellectual creativity, including music, art, and all forms of literature. Being still able to differentiate between AI-generated and human-created content, however, remains crucial in domains where intellectual property rights and authorship are at stake, as well as in contexts where an individual's professional competency in these fields must be assessed (e. g. university student assignments or job related evaluation tasks).

AI-generated art, such as music and paintings, is sometimes associated with an effect that Mori (1970/2012) notoriously defined as the *Uncanny Valley*—a term that describes 'a nonlinear relationship between robots' perceived human likeness and their likability' (Wang, Lilienfeld, and Rochat 2015, p. 394). While this

[1]I would like to thank Yannic Pixberg for his valuable contribution to the corpus

term originally referred to a predicted human aversion to robots designed to appear 'as human as possible', I believe it can also be applied to machine-generated intellectual works intended to resemble those created by humans. Often, there is something 'off' with AI generated content, but pinpointing exactly what or why can be difficult. While the *Uncanny Valley* effect may be more immediately apparent in audiovisual media such as videos, music, or drawings, it can also be present in texts. For instance, Gao, Howard, and Markov (2023) demonstrate that blinded human reviewers of academic abstracts are relatively accurate in intuitively identifying AI-generated texts, justifying their judgments by describing the AI's output as somewhat 'vague' and 'superficial.' Interestingly, focusing on characteristics such as vagueness or superficiality, as well as seeking indicators of human originality—such as alternative spellings—enabled human reviewers in Gao, Howard, and Markov (2023) to perform as well as specialized AI detection software. I highlight this not only because it is an interesting and somewhat eerie phenomenon, but also because the subjective and intuitive nature of such judgments presents a challenge in proving suspected cases of plagiarism or similar suspected offenses against academic integrity. If human reviewers perform as well as specialized algorithms, then conversely, specialized algorithms perform only as well as human judgments that rely on an inherently vague and intuitive sense of superficiality. Unsurprisingly, the demand for more reliable AI-text detection tools has surged following the breakthrough of large language models (LLMs) such as ChatGPT. AI-text detection tools like QuillBot[2] and ZeroGPT[3] exist. However, many of them struggle with correctly recognizing human texts or initially had problems detecting AI texts written by more advanced models such as GPT4.0 (cf. Walters 2023).

AI detection algorithms employ different approaches. For instance, Desaire et al. (2023) rely on stylometric features that are statistically more prevalent in one type of text than in the other (human vs. AI). These include factors such as the frequency of question marks or the occurrence of specific words like 'et' in 'et al.', which is more common in human-authored academic texts. In contrast, ZeroGPT compares the analyzed text to its own language model. If the text, its vocabulary, or its stylometric properties 'surprise' the model (meaning they deviate significantly from the statistical distribution it has learned) the text is more likely to be classified as human-written (cf. Ofgang 2023).

However, none of these tools can—nor should—serve as definitive evidence when academic integrity is at stake and potential consequences, such as expulsion, fines, damage to academic reputation, etc. are a possibility. In this context, some technical and ethical things should be considered:

First, academic writing is highly standardized and typically exhibits lower stylistic entropy compared to genres such as fiction or journalism. Depending on the language model and the nature of its fine-tuning, this may result in either lower or higher perplexity values within the detection tool's model. Second, highly familiar texts—such as well-known political speeches or the U.S. Con-

[2] `https://quillbot.com/ai-content-detector`, accessed 03/20/2025

[3] `https://www.zerogpt.com`, accessed 03/20/2025

stitution—are likely to be flagged as AI-generated, as detection models have been exposed to these texts extensively during training and therefore exhibit low perplexity when processing them (cf. Benj 2023).
Third, tools like ZeroGPT have been shown to be biased against non-native writers, resulting in false positives against authors who are working in a language in which they lack native proficiency (cf. Liang et al. 2023). This is particularly problematic in the context of academic writing, which increasingly relies on international collaboration conducted in English as the academic lingua franca proficiency that is not equally accessible in all parts of the world.
Fourth, even if it can be proven that LLMs were used in the writing process, their purpose may have been solely to enhance clarity or linguistic precision—in exactly those cases where authors feel insecure about their command of the language—while still presenting original and fair academic work. This would only present a problem if language proficiency is part of the performance criteria being assessed, as is generally the case with student assignments[4]. Within general academic publications, however, where originality and intellectual contribution should carry more weight, such concerns are largely negligible. Although there is always the danger of AI being empirically "used by organizations such as paper mills to entirely falsify research" (Gao, Howard, and Markov 2023, p. 2), 'linguistic polishing' may constitute the most plausible and, arguably, most prevalent use case in (regulated) academic settings, resulting in Frankenstein style texts: hybrid compositions in which human- and AI-generated segments are interwoven. These hybrid texts, composed of both human-written and AI-generated elements, are precisely the type that model based detection tools struggle with the most, as even small human attachments–such as splitting or restructuring a sentence of the AI generated text–can circumvent AI detection tools (Tyrell 2023). In order to detect violations against academic integrity—whether AI-assisted or not—reviewers must, for the time being, continue to rely on traditional evaluation methods. So-called 'hallucinations' remain a high risk for anyone using AI for uncritical factual research, as 'hallucinated' facts or non-existent academic sources can serve as reliable indicators of scholarly misconduct (cf. e. g. Ho et al. 2024).

For linguists, however, the focus should not be limited to the question of whether a text is AI-generated, but should also include an analysis of how it differs from human-authored writing. As previously noted, AI-generated texts can be (partially) identified through the examination of stylometric features (cf. Desaire et al. 2023). It is unlikely, however, that the Uncanny Valley effect described above is driven purely by human sensitivity to metrics such as average sentence length or the frequency of question marks within a text. If we go by the reviewers in Gao, Howard, and Markov (2023), humans (at least in the academic field) seem to be explicitly more disturbed by vague and unspecific semantic content in texts or by implausible propositions. It would be interesting to find out

[4]But even then, a distinction must be made between disguised plagiarism ('letting AI do all the work') and the fair use of AI as a learning support tool (cf. Gabriel and Römisch 2024, p. 30)

whether this sense of vagueness can be quantified and specified.

2 Method

2.1 Corpus

For this study, the introduction sections of 25 peer-reviewed German-language linguistics articles were extracted. To rule out the possibility of AI-generated content in the 'human' texts, only articles from the 'pre-AI' era were selected, effectively covering a publication range from 1978 to 2021. All articles were, in one way or another, related to syntactic research on the left periphery of the sentence, addressing topics like verb fronting, prefield occupation, multiple prefield occupation (most commonly V3), pre-prefield occupation, left peripheral adverbial positioning, and non-standard constructions in spoken language associated with these phenomena. This ensured a degree of topical diversity within the corpus while maintaining a coherent overarching research domain and avoiding the level of semantic entropy that we would expect when comparing, for instance, syntactic studies with those focused on sociolinguistics.
In order to further reduce unfair semantic entropy, all linguistic examples or transcriptions of original spoken or written language provided by the authors (most commonly in the form of numbered examples such as (1)) were excluded from the corpus, as such examples may introduce arbitrary, topic-irrelevant content and artificially inflate lexical or structural variety.

(1) After three vodka shots, Mary finally had the courage to kiss Peter

A total of six AI models were tasked with generating texts. These included the recent standard models from OpenAI (ChatGPT-4o), Google (Gemini Flash 2.0), and Deepseek (V3), as well as their respective counterparts marketed for more 'advanced reasoning' capabilities: ChatGPT-4o1, Gemini Flash 2.0 Thinking, and Deepseek V3-R1.

(2) Overview of AI models used:

Provider	**Base Model**	**Advanced Model**
ChatGPT	ChatGPT-4o	ChatGPT-4o1
Gemini	Gemini Flash 2.0	Gemini Flash 2.0 Thinking
Deepseek	Deepseek V3	Deepseek V3-R1

For each human-authored reference text, each model was instructed to generate two introductions (Text A and Text B) to a scientific article on the same topic. The two AI-generated texts per model were produced using different prompts: Text A was generated using a straightforward, minimal prompt, while Text B was created using an extended prompt that requested a more academic writing style, the inclusion of references to scientific literature, and adherence to the Harvard citation style.

(3) Sample prompts given to the language models for generating article introductions:

- **Text A:** *Schreibe eine Einleitung für einen sprachwissenschaftlichen Aufsatz über das Thema [Thema].*
  (*Write an introduction to a linguistic paper on the topic [topic].*)

- **Text B:** *Schreibe eine Einleitung für einen sprachwissenschaftlichen Aufsatz über das Thema [Thema]. Bediene dich eines akademischen Stils, benutze die Harvard-Zitierweise, verwiese auf einschlägige Literatur. Beschreibe, wie für eine wissenschaftliche Einleitung typisch, worum es in dem Artikel geht und wie vorgegangen wird*
  (Write an introduction to a linguistic paper on the topic [topic].Use an academic style, apply the Harvard citation style, and refer to relevant literature. As is typical for a scholarly introduction, describe the subject of the article and outline the methodological approach.)

From the generated outputs, only the main article texts were extracted. Meta-comments from the AI and appended bibliographies were excluded, as such elements are not typically found in the introduction sections of scientific papers. Also, as with the human texts, numbered examples such as (1) have been excluded for the same reasons as mentioned above. To prevent the texts from influencing one another, each was generated in a separate instance (new chat) of the respective AI application, with no additional context or further information provided by the user. Combined with the human texts, a total of 325 texts have been analyzed.

2.2 Data

For each prompt type and model, the output texts were merged into a single composite text. For example, all Text A outputs from ChatGPT-4o were combined into one text, while all Text B outputs from the same model formed another, and so on for each model. In the same manner, all human-authored texts were merged into a single composite text, as if they had been produced by one *model*. For tokenization, the Natural Language Toolkit (NLTK; Bird, Klein, and Loper 2009) was used. Lemmatization, POS-Tagging and word embeddings were performed using spaCy (Honnibal et al. 2020) with the language model *de_core_news_lg*. In a nutshell, *word embedding* can be considered a machine learning method that tries to encode the meaning of a word as a multi-dimensional vector[5], similar (but not completely identical) to a feature based semantic analysis of the meaning of words (for an introduction to feature based semantics, cf. e.g. Hurford, James R. and Heasley, Brendan 1983, for an introduction to word2vec encoding, cf. e.g. Mikolov et al. (2013), Aggarwal 2023, 99ff.). These features are not purely semantic, but encode any feature the machine learning algorithm deems relevant for distinguishing between words and may also include syntactic features or probabilistic features. Words with similar

[5] in the case of spaCy's *de_core_news_lg* model, there are 300 dimensions

meanings have similar embedding-vectors and thus are closer to each other in the vector space. The semantic similarity of words can be quantified by calculating the cosine of the angle between their respective embedding vectors (*cosine similarity*, cf. Singhal 2001, 46ff.). A smaller angle between vectors corresponds to a higher cosine value, indicating greater similarity. This, in theory, allows us to form semantic categories of words that objectively share similar semantic features.
In Wegerhoff (2025), I developed a set of scripts that:

- identify and count the most frequent lemmas across different models and parts of speech, while also extracting stylometric features;
- group lemmas into semantic categories based on vector similarity and determine the most prevalent categories for each model.

The threshold for semantic clustering was set to a cosine similarity of 0.7. This value was chosen to group semantically similar lemmas without being overly specific. Note that my aim was not to achieve the most fine-grained and efficient clustering–otherwise, a transformer-based model such as BERT and a higher similarity threshold would have been used. Instead, the aim was to test whether general, broader semantic patterns are visible across texts, even when using a relatively coarse-grained clustering method and relatively small text corpus. As discussed in Section 1, human reviewers of academic texts seem to be sensitive to such differences, but their perception of semantic similarity is likely based on similarly broad and intuitive notions instead of fine-grained and computationally precise semantic clustering. Each category was labeled using the first lemma of the group as it appeared in the corpus. For example, the frequency of the group *analytisch* ('*analytic*') does not represent how often the lemma *analytisch* appeared in the corpus, but rather how often words with a cosine similarity of ≥ 0.7 to *analytisch* appeared in the corpus. Stopwords were removed from the analysis using SpaCy's predefined German stopword list. The analysis was conducted through multiple experimental runs:

- A combined run that categorizes adjectives, adverbs, nouns, and verbs simultaneously in an overarching analysis.
- Separate runs that categorize adjectives, adverbs, nouns, and verbs individually, to prevent different parts of speech from influencing each other's categorizations.
- A comparative run that categorizes two groups of parts of speech: adjectives and adverbs versus nouns and verbs.

3 Results

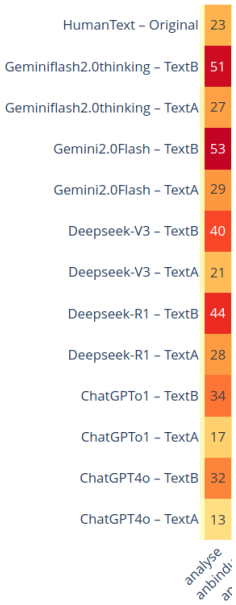
The primary goal of this analysis is to determine whether certain categories are noticeably over-/underrepresented in one model or text type, compared to

the human text set. While there are numerous moments worth discussing, I will discuss the ones I find most striking. A complete visual representation of each run's results (heatmap) can be found in Wegerhoff (2025). Note that *de_core_news_lg* is not a transformer-based model but instead uses static (context-insensitive) part of speech tagging, which means that some parts of speech might be misclassified, depending on the context. This is especially relevant for adjectives and adverbs, which, in this paper, are treated as one combined POS-group.

3.1 Nouns

The data show that the group *Analyse ('analysis')* is strikingly overrepresented in the more academic-specific TextB-prompt. The group consists of the following analysis-related lemmas: *analyse, auswertung, datenanalyse, evaluation, methodik*. The overrepresentation persists across all models, but the effect is most striking in both Gemini models:

(4)



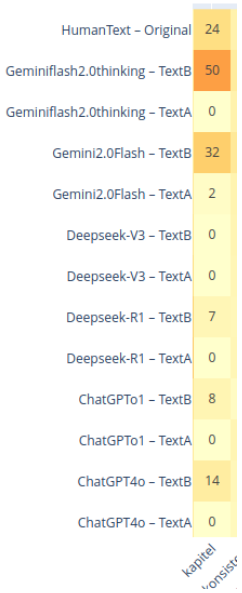
Heatmap of the group *Analyse*, showing absolute frequencies

A very similar pattern can be observed in the related adjective *analytisch*, which encompasses 13 lemmas that generally denote a scientific and methodical approach to a subject.[6]

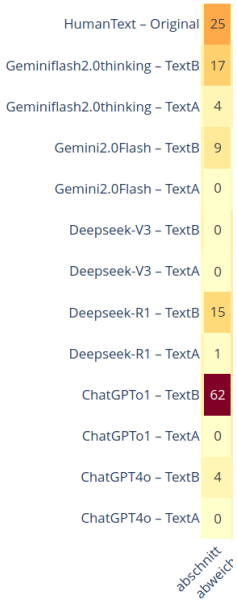
[6]The lemmas in the analytisch category are: *analytisch ('analytical'), dialogisch ('dialogical'), differenzierend ('differentiating'), empirisch ('empirical'), fachsprachlich ('technical'), graphisch ('graphical'), holistisch ('holistic'), logisch ('logical'), philosophisch ('philosophical'), pragmatisch ('pragmatic'), systematisch ('systematic'), systemisch ('systemic'), topologisch ('topological'), wissenschaftlich ('scientific')*

This effect is likely attributable to Prompt B, which explicitly calls for an academic tone and a methodological approach, resulting in a higher frequency of lemmas the model associates with academic or methodological language. However, both prompts (A and B) made the model aware that the text to generate is for academic purposes and thus needs methodological approaches, but the explicit call for academic tone in Prompt B seemed to have amplified the use of methodological lemmas in the model.
Another interesting finding is that some AI models tend to explicitly refer to the structural role of the generated section within the broader (hypothetical) text, suggesting a frequent and implicit awareness of academic text organization. Lemma-categories regarding text structure, such as *Kapitel ('chapter')*[7] *and Abschnitt*[8] *('section')* are used comparatively often in the human texts in the corpus as the examples in (5) show.

(5)



Heatmap of the group *Kapitel*



Heatmap of the group *Abschnitt*

These lemmas would be expected to appear rather often in introductions to academic papers but are completely missing in about half of the AI generated texts: *Kapitel* appears in 6 out of 12 AI text sets and *Abschnitt* in only 5 out of 12. Most occurences of these lemmas in the AI generated texts are found in the TextB text set (with the exception of the Deepseek V3 model), so with regard to simulating some kind of metatextual *'awareness'*, the prompt generally seems to be the much greater influencing factor than choosing the advanced reasoning model:

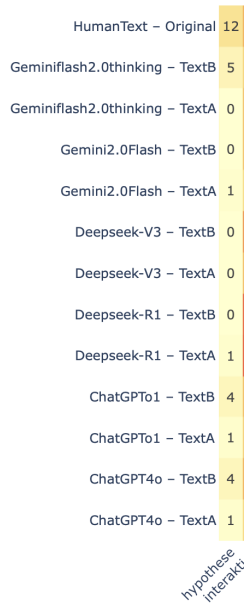
[7]The *Kapitel* ('chapter') group consists of only the lemma *Kapitel*.
[8]The *Abschnitt* ('section') group consists of only the lemma *Abschnitt*.

(6) Average frequency of *Kapitel* and *Abschnitt* by model type (normal vs. advanced reasoning) and prompt type (A vs. B)

Faktor	**Kapitel (ø)**	**Abschnitt (ø)**
Advanced models	10,83	16,50
Normal models	8,00	2,17
Text A (Prompt A)	0,33	0,83
Text B (Prompt B)	18,50	17,83

Note, however, that some models tend to take these hypothetical meta-textual references to the extreme (as indicated by the darker regions in the heatmaps in (5)), in some cases exceeding the human baseline frequency of these lemmas by a factor of two or almost three. In these cases, it can be argued that the AI 'overcompensates' for its lack of genuine structural understanding or contextual grounding by disproportionately emphasizing text sort specific markers such as chapter or section references. This suggests a form of surface-level mimicry, where the model learns to reproduce the form of academic writing without providing a genuine academic approach to the topic, e. g. formulating a hypothesis in the introduction. This observation is supported by the frequency of the *Hypothese* ('*hypothesis*'[9]) category, which is overrepresented in human texts compared to AI-texts and–again–completely missing in almost half the AI-generated text sets:

(7)



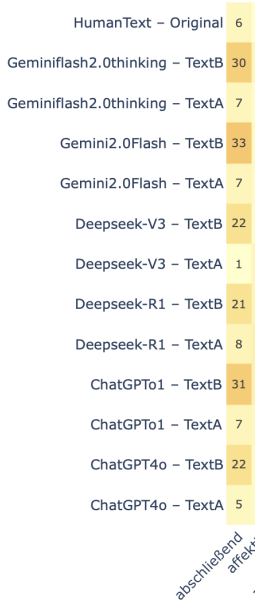
Heatmap of the group *Hypothese*, showing absolute frequencies

[9]The group only consists of one lemma

3.2 Adjectives and Adverbs

With regard to adjectives and adverbs, the observation persists that the TextB text set seems to use structuring vocabulary significantly more frequently than the TextA text set and humans. This is, for example, visible in the *abschließend*-category, which consist of only two lemmas: *abschließend* (≈ '*conclusively*') and *anschließend* (≈'*subsequently*'):

(8)



Heatmap of the group *abschließend*, showing absolute frequencies

As (8) shows, the *abschließend*-category, which denotes some kind of textual subsequency or conclusion, is overrepresented in the TextB set when compared to both the TextA and the human text set. There are also notable differences between human- and AI-generated texts regarding adjectives that evaluate the importance of a topic within the respective academic field. Compared to the average frequency of the *zentral* category in AI-generated texts (TextA and TextB), its frequency in the human-authored text is less than half as high, which means *zentral* is more than twice as likely to occur in AI-generated texts:

(9) Average frequency of *zentral*-category:

Text Type	**frequency**	
TextA	22,33	avg. across all models
TextB	24,83	avg. across all models
HumanText (Original)	10,00	absolute

The AI tends to emphasize the importance of a topic for the respective scientific

discipline, but sometimes overestimates its actual meaning in the academic field. A fitting example would be (10), which originates from the ChatGPT4o version of the FreyPittner1998-Textset in Wegerhoff (2025):

(10) Die Frage nach der Positionierung von Adverbialen im deutschen Mittelfeld gehört zu den zentralen Themen der deutschen Syntaxforschung (*The positioning of adverbials in the German middle field ranks among the central topics in the study of German syntax.*)

While it is undeniably true that (base-generated) adverbials are an important and much-discussed topic within German syntax research, describing them as a central focal point would be misleading. They represent just one of several phenomena explored within the field.
Human-authored texts stand out through their use of connective, critical and epistemically evaluative expressions. A notable example is the *allenfalls* group, which is remarkably large and comprises 32 lemmas[10]. These lemmas are predominantly adverbs that serve to coherently connect propositions or express the speaker's epistemic stance toward the proposition of the sentence– reaching from very careful assumptions (e. g. *womöglich - 'possibly'*) to strong epistemic markers (e. g. *keinesfalls - 'by no means'*). Depending on the prompt, the group is 5 to 8 times more likely to appear in the human texts compared to the AI generated texts:

(11) Frequencies of *allenfalls*-category:

Text Type	**Frequency**	
TextA	10,17	avg. across AI models
TextB	7,50	avg. across AI models
HumanText (Original)	58,00	absolute

The effect of humans being more critical is visible in the *uneinheitlich* ('*inconsistent*') category, which also carries a critical and/or somewhat negative connotation towards a theoretical approach or dataset:

(12) Frequencies of *uneinheitlich*-category:

[10] allenfalls (*at most*), andererseits (*on the other hand*), ansonsten (*otherwise*), augenscheinlich (*apparently*), demnach (*accordingly*), dennoch (*nevertheless*), ebenfalls (*also*), fraglich (*questionable*), freilich (*admittedly*), gleichwohl (*nonetheless*), gänzlich (*entirely*), hingegen (*by contrast*), insofern (*insofar*), jedenfalls (*in any case*), keinesfalls (*by no means*), keineswegs (*in no way*), lediglich (*merely*), letztlich (*ultimately*), möglicherweise (*possibly*), namentlich (*notably*), nämlich (*namely*), offensichtlich (*obviously*), tatsächlich (*in fact*), teilweise (*partially*), vermeintlich (*supposedly*), vermutlich (*presumably*), vielmehr (*rather*), vordringlich (*pressingly*), wiederum (*again*), womöglich (*possibly*), zumindest (*at least*), üblicherweise (*typically*)

Text Type	Frequency	
TextA	9,33	avg. across AI models
TextB	9,67	avg. across AI models
HumanText (Original)	30,00	absolute

3.3 Verbs

The category *stehen*, consisting of the lemmas *stehen* and *stellen* (*'to stand'* and *'to put'*) is the most frequently used category among the verbs, both in the AI texts and in the human texts, but are still about 2.5 times more likely to appear in human texts:

(13) Frequencies of *stehen*-category:

Text Type	Frequency	
TextA	22,33	avg. across AI models
TextB	23,67	avg. across AI models
HumanText (Original)	56,00	absolute

The reason for this high frequency likely is connected to the relatively general semantics of *stehen* and *stellen* and their frequent use in German idiomatic constructions, as e. g. *im Widerspruch stehen* (*'to be in contradiction with'*). Similar to the observations regarding the use of nouns, the use of explicitly analytical language is more prevalent in the AI texts. The category *analysiert* (consisting of *analysieren* - '*to analyze*' and *untersuchen* - '*to investigate*') is about 10-15 times more frequent in AI texts:

(14) Frequencies of *analysiert*-category:

Text Type	Frequency	Note
TextA	10,17	avg. across AI models
TextB	15,00	avg. across AI models
HumanText (Original)	1,00	absolute

As with the respective nouns (see above), the effect is stronger in the TextB-textset where the AI was explicitly tasked with a more academic use of language.

4 Conclusion.

The results can be summarized as follows:

The human-written and AI-generated academic texts examined in this paper appear to differ in their use of evaluative language, particularly negative or critical language. AI tends to exaggerate the importance and quality of topics

and theories within their respective academic fields, sometimes in an imprecise or counterfactual way. At least if not explicitly prompted to do otherwise, AI models seem to avoid stating negative criticism, and thus the texts appear rather uncritical and reassuring, which is shown in the data's lack of negative polarity expressions in AI texts when compared to human texts. This effect can be explained by the well-known guardrails implemented by AI providers. These guardrails are intended to offer users a positive and optimistic experience while avoiding conflict or even a 'shitstorm'.
The use of analytical terms such as *Analyse* ('*analysis*') seems to be much more prominent in the TextB dataset than in every other text set. The same is true for the *abschließend* ('*concluding*')-group containing text-structuring adverbs and adjectives. This shows that a more specific prompt generally has a much larger effect on the result than the choice of a more advanced alternative model or different AI. The AI then sometimes uses these terms excessively more often than humans, likely in an attempt to give the text a more academic flavor compared to the vanilla prompt in TextA (see also the *Abschnitt*- and *Kapitel*-Group). However, when it comes to concrete theoretical work, such as formulating hypotheses or using critical and rather specific argumentative terminology, the AI's vocabulary falls short of humans for the reasons mentioned above. In this regard, the data supports the observations made by the blinded reviewers in Gao, Howard, and Markov (2023) regarding the vagueness present in the AI-generated texts. This vagueness can be characterized as both an intuitively and empirically striking contrast between the overrepresentation of academic and analytical vocabulary employed by the AI on the one hand (particularly in the TextB text set) and its avoidance of direct confrontation (and consequently specific criticism) on the other hand. While the AI, on average, exaggerates its methodical and analytical approach by linguistic means, at the same time it seems to avoid it in practice.

As for the data, the results must be taken *cum grano salis*: with only 25 human-written texts, the reference set is comparatively small and limited to a narrow range of interconnected topics. The findings may not be fully generalizable to larger datasets or texts from other academic domains, although similar results can be expected when using similar prompt styles.
From a methodological perspective, spaCy is an out-of-the-box NLP toolkit and is therefore relatively static in its application. It is not ideally suited for highly specialized academic NLP tasks. Notably, the word embeddings provided by the *de_core_news_lg* model were static. No contextual information was considered but only the lexical semantics of individual lemmas. A different cosine similarity threshold could have yielded different clustering results, but the spaCy toolkit and the chosen threshold of 0.7 appeared widely appropriate for identifying semantically aligned lemmas (i.e., those that, quite literally in vector space, point roughly in the same direction) and grouping them into meaningful groups. Moreover, some cases of cross-lingual contamination were encountered. Five insertions of Russian words were detected, such as in (15):

(15) [...] lassen sich diese закономерности erklären?
[...] let pron-refl these *'patterns'* (russian) explain
(Gemini2.0Flash Breindl2012 TextA textset in Wegerhoff 2025)

This issue might have multiple causes (imprecise training method, suboptimal training data, decoder error, etc.) and is hard to predict as the training methods and underlying technology for most common models are still undisclosed to the public (cf. Jiang et al. 2024). However, these cases were only encountered in 2 texts (Breindl2012 and Fiehler2015) in Wegerhoff (2025), and are only present in outputs generated by Gemini Models, so their impact on the overall corpus data can be considered minimal.

References

Aggarwal, Charu C. (2023). *Neural Networks and Deep Learning. A Textbook.* 2nd ed. Springer Cham.

Benj, Edwards (2023). *Why AI writing detectors don't work. Can AI writing detectors be trusted? We dig into the theory behind them.* URL: https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/ (visited on 03/27/2025).

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc." URL: https://www.nltk.org/.

Desaire, Heather et al. (2023). "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools". In: *Cell Reports Physical Science* 4, p. 101426. DOI: 10.1016/j.xcrp.2023.101426.

Frank, Joel et al. (2023). *A Representative Study on Human Detection of Artificially Generated Media Across Countries.* arXiv: 2312.05976 [cs.CR]. URL: https://arxiv.org/abs/2312.05976.

Gabriel, Sonja and Barbara Römisch (July 2024). "Der Einsatz von KI-Tools im (wissenschaftlichen) Schreibprozess: Eine Schreibwerkstatt für Studierende 2024". In: *R&E-SOURCE* 11.3, pp. 26–42. DOI: 10.53349/resource.2024.i3.a1289. URL: https://journal.ph-noe.ac.at/index.php/resource/article/view/1289.

Gao, C.A., F.M. Howard, and N.S. Markov (2023). "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers". In: *npj Digit. Med* 7 (75). URL: https://doi.org/10.1038/s41746-023-00819-6.

Ho, Daniel E. et al. (2024). *Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive.*

Honnibal, Matthew et al. (2020). "spaCy: Industrial-strength Natural Language Processing in Python". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Associa-

tion for Computational Linguistics, pp. 121–126. URL: https://aclanthology.org/2020.emnlp-demos.14.

Hurford, James R. and Heasley, Brendan (1983). *Semantics: a coursebook*. Cambridge [u.a.]: Cambridge Univ. Pr. ISBN: 0521289491.

Jiang, Minhao et al. (2024). "Investigating Data Contamination for Pre-training Language Models". In: *arXiv preprint arXiv:2401.06059*.

Liang, Weixin et al. (2023). "GPT detectors are biased against non-native English writers". In: *Patterns* 4. URL: https://api.semanticscholar.org/CorpusID:257985499.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

Mori, Masahiro (1970/2012). "The uncanny valley". In: *Energy* 7, pp. 33–35.

Ofgang, Erik (2023). *What is GPTZero? The ChatGPT Detection Tool Explained By Its Creator*. URL: https://www.techlearning.com/news/what-is-gptzero-the-chatgpt-detection-tool-explained (visited on 03/27/2025).

Singhal, Amit (2001). "Modern Information Retrieal: A Brief Overview". In: *Bulletin of the IEEE Compuer Society Technical Committee on Data Engineering 24* 4, pp. 35–43.

Tyrell, James (2023). *How easy is it to fool AI content detectors?* URL: https://techhq.com/2023/02/how-easy-is-it-to-fool-ai-content-detectors/.

Walters, William H. (2023). "The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors". In: *Open Information Science* 7.1, p. 20220158. DOI: doi:10.1515/opis-2022-0158. URL: https://doi.org/10.1515/opis-2022-0158.

Wang, Shensheng, Scott O. Lilienfeld, and Philippe Rochat (2015). "The Uncanny Valley: Existence and Explanations". In: *Review of General Psychology* 19.4, pp. 393–407. DOI: 10.1037/gpr0000056. eprint: https://doi.org/10.1037/gpr0000056. URL: https://doi.org/10.1037/gpr0000056.

Wegerhoff, Dennis (May 2025). *Semantic Analysis of AI Generated text*. Version 1.3. URL: https://github.com/DayJay1992/SemanticAIAnalysis.