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

Dennis Wegerhoff (Bergische Universität Wuppertal)

dennis.wegerhoff(at)uni-wuppertal.de

Abstract

This study explores the semantic differences between human-written and AI-generated academic texts by applying word embedding techniques to a curated corpus of 325 introductions from linguistic articles. The corpus includes human-authored texts and AI-generated texts produced by six language models (OpenAI, Google, and DeepSeek; base and advanced). Each topic was prompted in two different ways: plain and academic. Using cosine similarity, the most frequently occurring lemmas were grouped into semantic categories. The analysis reveals that AI-generated texts, especially under academic prompts, overuse positive-evaluative and methodological vocabulary (e.g., *central, crucial, analysis, methodology*) and explicitly refer to text structure more often than the plainly prompted texts (e.g., *section, chapter*). In contrast, human authors employ more epistemically cautious, critical, evaluative, and connective language (e.g., *possibly, inconsistent, by no means*). I propose that the relative absence of such epistemic markers in AI texts, combined with their tendency to exaggerate the importance of certain topics or data, reflects a pattern of pseudo-commitment: the models produce syntactically assertive, formally academic prose but only weakly modulate epistemic stance and critical engagement, which may contribute to the reported sense of weirdness in AI-generated academic writing.

Keywords

artificial intelligence, human–machine authorship, word embeddings, semantic analysis, epistemic stance, commitment

1 Introduction: The Uncanny Valley of AI

With large language models (LLMs) on the rise, distinguishing between human-authored and AI-generated academic texts has become increasingly challenging (Frank, Herbert, Ricker et al. 2023). This issue is particularly pressing in academic contexts, where authorship relates not only to intellectual property but also to the evaluation of professional competence (e.g., in student assignments or peer-reviewed research).

While the phenomenon of the *Uncanny Valley*—a drop in likability when artificial agents appear almost, but not fully, human—was originally described in robotics (Mori 1970/2012), it may also apply to textual artifacts that attempt to reproduce the style of human writing in a specific domain. Readers frequently report an intuitive sense that AI-generated texts feel 'off', even when they closely mimic human style and are difficult to distinguish from human-authored texts by objective criteria. Following Mori's concept, I refer to this phenomenon as the *Uncanny Valley Effect* (hereafter UVE) of AI-generated texts. It has been theorized that the UVE originates from a general cognitive dissonance caused by categorical uncertainty (Categorical Uncertainty Hypothesis; cf. e.g., MacDorman and Chattopadhyay 2016; Wiese and Weis 2020). In the case of texts, this uncertainty may arise from doubts about the authenticity of the expressed ideas, the genre, the overall communicative intent, etc. This uncertainty seems to stem primarily from

semantic content rather than purely formal or stylistic features. For example, Gao, Howard and Markov (2023) show that human reviewers tasked with identifying AI-generated texts in mixed corpora likely do not rely on the same features as automated detection tools, suggesting a fundamentally different evaluation process. These reviewers often cited implausible or hallucinated information, as well as a general sense of vagueness or superficiality, as indicators of AI authorship. Existing AI-detection tools currently rely primarily on stylometric or statistical features (e.g., perplexity values, sentence length, or token frequencies; cf. Desaire, Chua, Isom et al. 2023; Ofgang 2023). Yet such approaches do not address the semantic and evaluative dimensions that likely contribute to readers' intuitive sense of weirdness.

This paper aims to explore these semantic dimensions by clustering frequent lemmas in a mixed (human- and AI-authored) corpus of 325 German academic text introductions, using a cosine similarity threshold of 0.7. Six prominent language models (models from OpenAI, Google, and DeepSeek, each in a base and advanced version advertised for 'advanced reasoning') were used. All topics originated from the field of German syntax. Each topic was prompted in two different ways: plain and academic. According to Bondi (2005: 26), introductions typically contain metadiscursive elements such as identifying a problem, presenting methodological tools, situating the work within disciplinary debates, and guiding the reader through the forthcoming argumentation. They also tend to exhibit a relatively high frequency of "nominalizations of cognitive and discursive processes" (Bondi 2005:17), such as, e.g., *analysis.*

The results show that AI-generated texts tend to employ this analytical and methodological terminology, in many cases even more frequently than human authors. However, they simultaneously avoid propositionally connective and evaluative terms. In this respect, AI-generated texts lack stance in the sense described by Hyland (1998, 2005), particularly in their minimal use of hedges, boosters, and attitude markers, which are three of the four central resources for expressing epistemic commitment, caution, and evaluative positioning in academic writing.[1] Therefore, the AI texts may appear academic on the surface but lack argumentative depth as well as critical and constructive engagement, for example when it comes to formulating hypotheses. In addition, AI models were shown to overstate the importance of topics and theories in their field by applying positively evaluative vocabulary (see *zentral*-category in Section 3).

I argue that the UVE of AI-generated texts is related to the epistemic flatness of their stance: the models seem to commit to propositions formally (via V2 clauses) but largely avoid signaling, through hedges, boosters, etc., how strongly these propositions are endorsed. In this sense, they exhibit a kind of *pseudo-commitment*, where almost all propositions are presented with a similar level of formal commitment, but obviously without the underlying Fregean process of grasping a thought and judging its truth. This epistemic flatness may give readers a subtle sense of weirdness.

[1] Note that Hyland (2005) also mentions Self-Mention as a resource of stance. As this paper focuses on an epistemic dimension (and as self-mentions are mostly filtered by the stopword-list), I will not engage with self-mentions as Hyland's Engagement-dimension here.

2 Method

2.1 Corpus

For this study, the introduction sections of 25 peer-reviewed articles from the field of German-language linguistics 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 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.

Next, 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.

Table 1: 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 (TextA and TextB) to a scientific article on the same topic. 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. The two AI-generated texts per model were produced using different prompts: TextA was generated using a straightforward, minimal prompt, while TextB 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.

Table 2: Prompts given to the language models.

Prompt	German	English translation [my transl.]
TextA	Schreibe eine Einleitung für einen sprachwissenschaftlichen Aufsatz über das Thema [Thema].	Write an introduction to a linguistic paper on the topic [topic].
TextB	Schreibe eine Einleitung für einen sprachwissenschaftlichen Aufsatz über das Thema [Thema]. Bediene dich eines akademischen Stils, benutze die Harvard-Zitierweise, verweise 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.

Only the main article texts were extracted from the output. 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 mentioned above. Combined with the human texts, a total of 325 texts were analyzed. Table 3 shows the size of the corpus in terms of tokens and lemmas:

Table 3: Frequency of tokens and lemmas per model.

Text	Tokens	Lemmas
Human Texts	14144	5308
ChatGPT-4o TextA	4620	2168
ChatGPT-4o TextB	7970	3418
ChatGPT-4o1 TextA	6889	3184
ChatGPT-4o1 TextB	9516	4026
Gemini Flash 2.0 TextA	6098	2777
Gemini Flash 2.0 TextB	8375	3641
Gemini Flash 2.0 Thinking TextA	6902	3284
Gemini Flash 2.0 Thinking TextB	10097	4454
DeepSeek V3 TextA	5730	2670
DeepSeek V3 TextB	7031	2981
DeepSeek V3-R1 TextA	8132	3864
DeepSeek V3-R1 TextB	8955	3944

2.2 Data

For each prompt type and model, the output texts were merged into a single composite text. For example, all TextA outputs from ChatGPT-4o were combined into one text, while all TextB 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’. Some cases

of cross-lingual contamination were encountered in the data: Five insertions of Russian words were detected, such as in (2):

(2) Gibt es typische Muster und Regularitäten […] und lassen sich diese закономерности erklären?
‘Are there typical patterns and regularities […] and can these закономерности (russian: *patterns*) be explained?’ [my transl.]

(Wegerhoff 2025a: TextA)

This issue might have multiple causes (imprecise training methods, sub-optimal 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 (Breindl 2012 and Fiehler 2015) in Wegerhoff (2025a), and are only present in outputs generated by Gemini Models, so their impact on the overall corpus data can be considered minimal.

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 *de_core_news_lg* language model. 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,[2] 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 and Heasley 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 or probabilistic features. Note also that not every dimension of the vector needs to map directly (1:1) to linguistically relevant features. The idea is that words with similar 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.

The GitHub repository referenced in Wegerhoff (2025a) contains the scripts used in this study. These scripts

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

[2] In the case of spaCy’s *de_core_news_lg* model, there are 300 dimensions.

The threshold for semantic clustering was set to a cosine similarity of 0.7. This value was chosen for pragmatic and methodological reasons: it produced clusters that were semantically coherent without fragmenting the data into excessively small groups. In preliminary tests, lower thresholds resulted in heterogeneous clusters, while higher thresholds created many clusters that contained only one or two lemmas. The 0.7 threshold therefore offered a practical balance between interpretability and cluster stability within the context of this exploratory analysis. 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. Extremely common words (so-called *stop-words*) were removed from the analysis using spaCy's predefined German stop-word 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 observations worth discussing, I will focus on the ones I found the most striking. A complete visual representation of each experimental run (heatmaps and category members), as well as the list of scientific texts included in the corpus, is available in the open dataset described in Wegerhoff (2025b). Note that *de_core_news_lg* is not a transformer-based model but instead uses static (context-insensitive) part-of-speech (POS) tagging, which means that some POS 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 as mentioned above.

3.1 Nouns

The data show that the group *analyse* ('analysis') is strikingly overrepresented in the more elaborate, academic-specific TextB prompt.[3] The overrepresentation persists across all models, but the effect is most striking in both Gemini models:

Model	TextType	analyse	analyse (normalized)
ChatGPT4o	TextA	13	2,81
ChatGPT4o	TextB	32	4,02
ChatGPTo1	TextA	17	2,47
ChatGPTo1	TextB	34	3,57
Deepseek-R1	TextA	28	4,59
Deepseek-R1	TextB	44	5,25
Deepseek-V3	TextA	21	3,04
Deepseek-V3	TextB	40	3,96
Gemini2.0Flash	TextA	29	5,06
Gemini2.0Flash	TextB	53	7,54
Geminiflash2.0thinking	TextA	27	3,32
Geminiflash2.0thinking	TextB	51	5,70
HumanText	Original	23	1,63

Figure 1: Heatmap of the group *analyse*, showing absolute frequencies and frequencies normalized per 1000 tokens.

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.[4]

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 degree of implicit awareness of academic text organization. Lemma categories related to text structure, such as *kapitel* ('chapter')[5] and *abschnitt* ('section'),[6] are comparatively frequent in human-authored texts, as illustrated in Figure 2:

[3] The group consists of the following analysis-related lemmas: *analyse* ('analysis'), *auswertung* ('evaluation'), *datenanalyse* ('data analysis'), *evaluation* ('evaluation'), *methodik* ('methodology').
[4] 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').
[5] The *kapitel* group consists solely of the lemma *kapitel*.
[6] The *abschnitt* group consists solely of the lemma *abschnitt*.

Model	TextType	kapitel	Kapitel (normalized)	abschnitt	abschnitt (normalized)
ChatGPT4o	TextA	0	0,00	0	0,00
ChatGPT4o	TextB	14	1,76	4	0,50
ChatGPTo1	TextA	0	0,00	0	0,00
ChatGPTo1	TextB	8	0,84	62	6,52
Deepseek-R1	TextA	0	0,00	1	0,16
Deepseek-R1	TextB	7	0,84	15	1,79
Deepseek-V3	TextA	0	0,00	0	0,00
Deepseek-V3	TextB	0	0,00	0	0,00
Gemini2.0Flash	TextA	2	0,35	0	0,00
Gemini2.0Flash	TextB	32	4,55	9	1,28
Geminiflash2.0thinking	TextA	0	0,00	4	0,49
Geminiflash2.0thinking	TextB	50	5,58	17	1,90
HumanText	Original	24	1,70	25	1,77

Figure 2: Heatmaps of the categories *kapitel* and *abschnitt*, showing absolute frequencies and frequencies normalized per 1000 tokens.

Across the AI-generated texts, these structural lemmas appear unevenly: the group *kapitel* occurs in 6 out of 12 AI text sets, while *abschnitt* occurs in 5 out of 12. Their presence is strongly tied to the academic prompt (TextB), as shown by the higher averages in Table 4, whereas TextA rarely triggers such references. Importantly, the elevated averages for AI models are not uniform but are driven by a few outlier systems—most notably Gemini Flash 2.0 Thinking and ChatGPT-4o1 under the academic prompt—which show structural marker frequencies up to three times higher than the human baseline. Other models, by contrast, remain close to or even below human frequencies:

Table 4: Average frequency of *kapitel* and *abschnitt* across all models by model type (normal vs. advanced reasoning) and prompt type (A vs. B).

	Kapitel (ø)	**Abschnitt (ø)**
Advanced models	10.83	16.50
Normal models	8.00	2.17
TextA (Prompt A)	0.33	0.83
TextB (Prompt B)	18.50	17.83

This pattern suggests that meta-textual markers are more sensitive to prompt design than model sophistication. The extreme frequencies observed in some models likely reflect a form of surface-level mimicry: the models reproduce structural cues of academic writing (e.g., explicit chapter or section references) without necessarily engaging in deeper argumentative functions like hypothesis formulation. This observation is supported by the frequency of the *hypothese* ('hypothesis')[7] category, which is overrepresented in human texts compared to the AI's texts and—again–completely missing in almost half the AI-generated text sets:

[7] The group only consists of one lemma.

Model	TextType	hypothese	hypothese (normalized)
ChatGPT4o	TextA	1	0,22
ChatGPT4o	TextB	4	0,50
ChatGPTo1	TextA	1	0,15
ChatGPTo1	TextB	4	0,42
Deepseek-R1	TextA	1	0,16
Deepseek-R1	TextB	0	0,00
Deepseek-V3	TextA	0	0,00
Deepseek-V3	TextB	0	0,00
Gemini2.0Flash	TextA	1	0,17
Gemini2.0Flash	TextB	0	0,00
Geminiflash2.0thinking	TextA	0	0,00
Geminiflash2.0thinking	TextB	5	0,56
HumanText	Original	12	0,85

Figure 3: Heatmap of the group *hypothese*, showing absolute frequencies and frequencies normalized per 1000 tokens.

It should be noted that lemma frequencies alone cannot determine whether a text engages in hypothesis formulation (as authors may articulate hypotheses or research assumptions without explicitly using the corresponding terminology). For this reason, all AI-generated texts in which the *Hypothese* cluster was absent were additionally examined through close reading. This qualitative analysis confirmed that none of the models formulated original hypotheses—either explicitly or implicitly—despite sometimes employing methodological or analytical terminology elsewhere in the text.

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 the human text set. This is, for example, visible in the *abschließend*-category, which consists of only two lemmas: *abschließend* (‘conclusively’) and *anschließend* (‘subsequently’):

As Figure 4 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.

Model	TextType	abschließend	abschließend (normalized)
ChatGPT4o	TextA	5	1,08
ChatGPT4o	TextB	22	2,76
ChatGPTo1	TextA	7	1,02
ChatGPTo1	TextB	31	3,26
Deepseek-R1	TextA	8	1,31
Deepseek-R1	TextB	21	2,51
Deepseek-V3	TextA	1	0,14
Deepseek-V3	TextB	22	2,18
Gemini2.0Flash	TextA	7	1,22
Gemini2.0Flash	TextB	33	4,69
Geminiflash2.0thinking	TextA	7	0,86
Geminiflash2.0thinking	TextB	30	3,35
HumanText	Original	6	0,42

Figure 4: Heatmap of the group *abschließend*, showing absolute frequencies and frequencies normalized per 1000 tokens.

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* ('central', 'crucial')[8] 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:

Table 5: Average frequency per text of zentral-category across all models by model type (normal vs. advanced reasoning) and prompt type (A vs. B).

Text Type	**frequency**	
TextA	0.894	avg. per text across all models
TextB	0.994	avg. per text across all models
HumanText (Original)	0.4	average per text

The AI tends to emphasize the importance of a topic for the respective scientific discipline, but sometimes overestimates its actual relevance in the academic field. An example of this behavior is presented in (3), which originates from a ChatGPT4o generated introduction to a text dealing with the positioning of adverbials in the corpus in Wegerhoff (2025a):

(3) 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.' [my transl.]

While it is undeniably true that (e.g., base-generated) adverbials are an important and much-discussed topic within German syntax research, describing them as a

[8] The group consists of only one lemma (*zentral*).

central focal point would be misleading. They represent just one of several phenomena explored within the field.

Human-authored texts, on the other hand, stand out through their use of connective, critical and epistemically evaluative expressions. A notable example is the *allenfalls* ('at most') group, which is remarkably large and comprises 32 lemmas.[9] 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 8 to 11 times more likely to appear in the human texts compared to the AI-generated texts:

Table 6: Average frequencies per text of *allenfalls*-category.

Text Type	**Frequency**	
TextA	0.407	avg. per text across AI models
TextB	0.300	avg. per text across AI models
HumanText (Original)	3.320	avg. per text

Furthermore, the *uneinheitlich* ('inconsistent')[10] category was found to be 6 to 8 times more frequently used in the human text set, indicating that human authors are more open to taking a critical stance towards existing theoretical models or datasets.

Table 7: Average frequencies per text of *uneinheitlich*-category.

Text Type	**Frequency**	
TextA	0.407	avg. per text across AI models
TextB	0.300	avg. per text across AI models
HumanText (Original)	2.320	avg. per text

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-generated texts and in the human texts, but is still about 2.5 times more likely to appear in human texts.

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 in, e.g., *im Widerspruch stehen* ('to be in contradiction with').

[9] The group consists of *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* ('namely'), *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').

[10] the group consists *of uneinheitlich* ('inconsistent'), *unterschiedlich* ('different'), *weitgehend* ('to a large extent'), *weitreichend* ('far-reaching').

Table 8: Average frequencies per text of *stehen*-category.

Text Type	**Frequency**	
TextA	0.894	avg. per text across AI models
TextB	0.947	avg. per text across AI models
HumanText (Original)	2.24	avg. per text

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:

Table 9: Average frequencies per text of *analysiert*-category.

Text Type	**Frequency**	**Note**
TextA	0.407	avg. per text across AI models
TextB	0.600	avg. per text across AI models
HumanText (Original)	0.04	avg. per text

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

4 Discussion

4.1 Epistemic Stance and (Pseudo-)Commitment

The results so far 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 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. Also, the AIs, at least in the TextB text set, overuse analytical and methodological terminology compared to humans. How can we explain this behavior?

The lack of epistemically evaluative terms in the AI-generated introductions indicates that, although the models often provide an accurate general description of the topic, they do not take stance in the sense described by Hyland (1998, 2005). This means that they provide little evidence of hedging, boosting, or attitudinal marking, which are core resources that academic writers use to signal degrees of certainty, evaluation, and authorial alignment. Without these elements, the texts may remain descriptively correct, but they are not subject to epistemic commitment in any way.

Formally, however, the models clearly produce clauses with the syntactic shape of assertions: in German, verb-second word order encodes commitment to a proposition (cf. Truckenbrodt 2006), and the AI-generated clauses conform to this pattern. In this sense, they syntactically appear to commit to a thought, but this commitment is at best structural rather than epistemic.

Human writers are typically modeled as relating to propositions through propositional attitudes such as belief, doubt, or ignorance, that is, as part of an interconnected system of epistemic commitments that are negotiated between

participants in a discourse (cf. Krifka 2012, 2018; Stalnaker 1978, 2002; Farkas and Bruce 2010). These attitudes allow human writers to adopt an explicit epistemic stance toward what they put into the common ground and to distinguish between what they take to be known, uncertain, or unknown. In Fregean terms, they can move from merely grasping a thought to making or withholding a judgment about its truth (Frege 1918/19), with the resulting judgment being syntactically encoded, in German, by a V2 clause.

Transformer models, by contrast, do not maintain propositional attitudes or a coherent belief system; they generate output—which is, basically, next-token predictions—by matching distributional patterns in their training data. Thus, the V2 clauses produced by the AI cannot plausibly be analyzed as the result of the models themselves first grasping a thought, judging its truth and then actually applying basic derivational operations such as *move* and *merge*; instead, they simply reproduce surface distributions of V2 clauses learned from their input. They do not evaluate or justify the content they assert in this Fregean sense. In this way, they 'do not know what they do not know': they cannot represent their own epistemic limitations and can only produce markers of uncertainty insofar as such patterns are licensed by the prompt or by learned textual regularities.

Beyond these external constraints, current models have no internal representation of a consistent belief or knowledge state and therefore no explicit representation of what they do not know. When confronted with epistemic or aleatoric uncertainty (e.g. out-of-distribution data, contradictions, or niche topics such as base-generated adverbials in the topological middle field), they nevertheless tend to produce fluent, assertive output instead of withholding judgment or signaling ignorance, which is a well-documented problem in AI research (cf. Papernot et al. 2015; Goodfellow, Shlens, and Szegedy 2015; Hendrycks and Gimpel 2018; Hüllermeier and Waegeman 2021; Manchingal et al. 2025). In the present context, the most relevant type of uncertainty is epistemic, as it stems from gaps in the training data. Since the models lack an explicit sense of their own uncertainty, they do not downscale their assertive behavior when information is missing and thus have little need for linguistic markers to express epistemic (un)certainty (unless encoded in the scarce training data). This contributes to the markedly lower frequency of such terms in their output. The observed overstatement of a topic's importance within an academic field, by contrast, is more plausibly related to a different mechanism: a combination of this flattened stance profile with safety-oriented guardrails that discourage negative or explicitly critical evaluations. As a result, the models tend to err on the side of positive, high-stakes formulations ("this topic is central / highly relevant / crucial"), while rarely questioning the value or limitations of the work they describe.

Obviously, epistemic uncertainty in AI models (resulting from gaps in the training data) is not identical to epistemic uncertainty in humans (resulting from an explicit inability to reach a justified judgment in a Fregean sense). My take on this issue is that this difference does not play a decisive role when discussing the uncanniness of AI-authored texts, because readers evaluate AI-generated texts against human standards of epistemic stance and, crucially, even when a model is only weakly trained on a topic, it typically does not signal uncertainty—neither

linguistically, through hedges or other stance markers, nor technically, through output confidence, which would not be visible to the user in ordinary interaction anyway. The pseudo-commitment of the AI thus contributes to the UVE described in the introduction: readers are confronted with texts that are epistemically flat in the sense that the 'author' appears to commit to propositions syntactically, but rarely indicates to what degree. In other words, the system is effectively uncritical toward its own epistemic performance (since it has no epistemic system in the human sense) and treats most propositions as if they were equally well supported (both empirically and personally), regardless of how well the model is actually trained on the specific topic. This might give the text an intuitively unnatural feel, leading to the reported weirdness by many users.

4.2 Academic Terminology vs. Practice

Regarding the importance of the prompt, the abundant use of analytical terms such as *analyse* seems to be much more prominent in the TextB text set than in every other text set. The same is true for the *abschließend*-group containing text-structuring adverbs and adjectives. The overuse of analytical and methodological vocabulary is in line with the findings in Bondi (2005:17), who observes relatively higher frequencies of terms like *analysis* (terms that denote cognitive and discursive functions) in academic introductions. So, in this sense, the AI-models reproduce this feature when given an explicitly academic prompt, but they appear to overshoot the mark by employing such terminology substantially more frequently than the human baseline. In fact, the less academic TextA text set is closer to the human baseline in this corpus: On average, the analytically prompted TextB set deviates almost twice as strongly from the human baseline (+3.38 per 1000 tokens) as the less academic TextA set (+1.92 per 1000 tokens), indicating that the explicitly academic prompt amplifies the overuse of analytical nominalisations across all models. This also illustrates that a more elaborate prompt generally has a much larger effect on the result than the choice of a more advanced alternative model or different AI, which also aligns with observations reported in recent research (cf. Jahani et al. 2025).

Nevertheless, when it comes to the usage of vocabulary that is associated with substantial theoretical reasoning, including the formulation of hypotheses and the application of nuanced, domain-specific terminology, AI models remain inferior to human authors, as seen above (note, however, the limitations of this study and interpretation in 150).

In this regard, the data support the observations made by the blinded reviewers in Gao, Howard, and Markov (2023) regarding the vagueness present in the AI-generated texts. Following the data discussed here, this vagueness can be characterized as both an intuitively and empirically striking contrast: On the one hand, there is an over-representation of academic and analytical vocabulary employed by the AI models (particularly in the TextB text set). On the other hand, there is a noticeable avoidance of adverbs and connectors that connect and (critically) evaluate propositions within a broader academic idea or framework (as seen above). While the AI models, on average, exaggerate their methodical and

analytical approach by linguistic means, at the same time they seem to avoid it in practice. Readers who are regularly exposed to academic writing may find this strategy irritating, thereby experiencing what was described as the UVE. While the frequent use of methodological terminology suggests a highly academic and methodologically rigorous paper, the text ultimately fails to meet these expectations on an argumentative level—resulting in categorical uncertainty (regarding text type or the authenticity of the presented research, for example), and therefore the reported weirdness.

4.3 Limitations

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 closely related topics. The method and dataset described in this paper were rather exploratory and should be interpreted as such. 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. Regarding the prompts, note also that the problem of epistemic stance mentioned above can certainly be avoided by more specific prompting (which, actually, should be the subject of future research to test AI's capability to convincingly adopt something that resembles stance-taking). The prompts used here, even the more academic prompt of the TextB set, were relatively basic and only covered the most fundamental requirements for the task of academic writing in the field of linguistics.

As AI is developing fast, the results of this paper may not be reproducible with newer models in the future. It should be noted that the models used in this study were not specifically designed for academic writing—nor are they marketed as such—but rather represent the most widely accessible and commonly used general-purpose AIs available at the time of analysis. The effects discovered in this paper may be considerably less striking with AI models specialized for more academic tasks (such as, e.g., Claude 3 Opus). Conversely, I would also expect the differences between human-authored and AI-generated texts to diminish if the human corpus consisted of student assignments rather than peer-reviewed publications, as the former are typically more descriptive and less critically engaged with academic discourse.

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, meaning they did not account for contextual variation in meaning. Unlike transformer-based language models (such as BERT), these embeddings assign each word a single, context-independent vector representation. In addition, a different cosine similarity threshold would 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 categories.

Methodologically, it should also be noted that lemma frequencies, while useful for identifying general semantic tendencies, have inherent limitations. The presence or absence of specific lemmas does not necessarily correspond to the presence or absence of the underlying conceptual functions they may indicate. Authors can formulate hypotheses, evaluate evidence, or express epistemic stance without using the canonical lexical items associated with these practices. This issue becomes more relevant as corpora grow in size: frequency-based approaches scale well from a computational perspective, but the qualitative validation required to confirm interpretive claims does not. In large datasets, systematic close reading becomes increasingly resource-intensive, which limits the extent to which quantitative findings can be individually corroborated. Because of the relatively small corpus, results in this study were supplemented with targeted close reading, especially in cases where a conceptual interpretation (e.g., hypothesis formulation) could not be reliably inferred from lemma counts alone. The findings reported here should be understood as indicative of broader tendencies rather than as exhaustive mappings of all epistemic moves made within the texts.

If explored further, the findings presented in this paper may have practical implications for the development of future AI-detection tools. Rather than relying solely on perplexity scores, stylometric profiles, or other purely stochastic indicators, effective detectors should integrate semantic category analysis. Combining statistical features with meaning-based cues will provide a greater understanding of the textual signals—both intuitive and quantitative—that distinguish human from AI-generated writing.

5 Conclusion

This exploratory study examined semantic differences between human-written and AI-generated academic texts using a clustering approach based on cosine similarity (threshold: 0.7). The findings indicate that AI-generated texts more frequently employ analytical and methodological terminology than human authors, yet tend to avoid connective terms that link and evaluate propositions. This results in texts that appear formally academic but lack argumentative depth as well as critical and constructive engagement, e.g., formulating hypotheses. In addition, AI models were shown to overstate the importance of topics and theories in their field by applying positively evaluative vocabulary (see *zentral*-category).

I suggest that these findings can be explained in terms of how humans and transformer-based AIs take and signal epistemic stance (in the sense of Hyland 2005). Human writers relate to propositions through graded propositional attitudes such as belief, doubt, or ignorance and make these attitudes visible through hedges, boosters, and other stance markers. The models, by contrast, generate fluent, assertive output without representing such attitudes and therefore without systematically modulating their degree of epistemic commitment. In other words, they exhibit what I call pseudo-commitment: they produce syntactically assertive clauses (via V2 clauses) but rarely indicate how strongly the underlying proposition is endorsed. As a result, they use few epistemic markers or evaluative connectors (as seen with the *allenfalls*-group) and treat most propositions as if they were

equally well supported. This kind of flat epistemic stance may create a sense of weirdness in readers.

The absence of an explicit model of their own (un)certainty contributes to a tendency to produce confident-sounding but sometimes imprecise or counterfactual claims. The avoidance of critically evaluative markers is likely reinforced by system-level guardrails designed to prevent conflict or controversy, which push the models towards reassuring and non-confrontational language.

Methodologically, the study further shows that prompt design exerts a strong influence on linguistic output, arguably more so than the choice of model itself. More elaborate prompts encourage greater use of analytical and text-structuring vocabulary, but they do not bridge the gap between AI- and human-authored texts in terms of critical and domain-specific reasoning. These findings align with earlier observations of vagueness and overgeneralization in AI-generated writing and may contribute to a subtle sense of uncanniness or weirdness in readers—a phenomenon that I compared to the Uncanny Valley Effect (UVE).

The study's main limitations include the small size and narrow topical scope of the human reference corpus, the relatively simple prompts, as well as the use of static word embeddings for semantic clustering. Future work should replicate this analysis with larger and more diverse datasets, incorporate contextual embeddings, and investigate how specialized academic LLMs compare to general-purpose models. Moreover, reader-response studies could test whether the UVE observed here—caused by the mismatch between academic form and lack of actual epistemic commitment—also occurs among less experienced readers. Another topic for future research is the behavior of AIs when they are explicitly prompted to take stance towards a proposition.

Finally, the results have practical implications for AI detection and quality assessment: instead of relying exclusively on stochastic indicators such as perplexity, effective detection should integrate semantic and evaluative dimensions. Combining statistical features with meaning-based cues may enable more effective tools to differentiate between human and AI-generated academic writing.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this article.

Acknowledgments

I would like to thank Yannic Pixberg for his contribution to the corpus. I acknowledge the use of OpenAI's GPT-5.1 Thinking in assisting with the editing and refinement of this paper, enhancing both its clarity and presentation in English. However, I ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

References

Aggarwal, Charu. 2023. *Neural Networks and Deep Learning. A Textbook* (2nd edition). Cham: Springer. https://doi.org/10.1007/978-3-031-29642-0

Bird, Steven & Klein, Ewan & Loper, Edward. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media. https://www.nltk.org/ (last accessed on 06/02/2026).

Bondi, Maria. 2005. Metadiscursive Practices in Academic Discourse: Variation across Genres and Disciplines. In Bamford, Julia & Bondi, Maria (eds), *Dialogue within Discourse Communities: Metadiscursive Perspectives on Academic Genres*, 3–30. Berlin/Boston: Max Niemeyer Verlag. https://doi.org/10.1515/9783110933222.3

Desaire, Heather & Chua, Aleesa & Isom, Madeline & Jarosova, Romana & Hua, David. 2023. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science* 4. 101426. https://www.sciencedirect.com/science/article/pii/S266638642300200X (last accessed on 06/02/2026).

Farkas, Donka & Bruce, Kim. 2010. On Reacting to Assertions and Polar Questions. *Journal of Semantics* 27(1). 81–118. https://doi.org/10.1093/jos/ffp010

Frank, Joel & Herbert, Franziska & Ricker, Jonas & Schönherr, Lea & Eisenhofer, Thorsten & Fischer, Asja & Dürmuth, Markus & Holz, Thorsten. 2023. A representative study on human detection of artificially generated media across countries. In *Proceedings of the 45th IEEE Symposium* on *Security and Privacy S&P* 24 (San Francisco, May 20–22, 2024), 55–73. Los Alamitos: CPS. https://ieeexplore.ieee.org/document/10646666 (last accessed on 06/02/2025).

Frege, Gottlob. 1918/19. Der Gedanke. Eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus* 1. 58–77.

Gao, Catherine & Howard, Frederick & Markov, Nikolay & Dyer, Emma & Ramesh, Siddhi & Luo, Yuan & Pearson, Alexander. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6(75). n.p. https://www.nature.com/articles/s41746-023-00819-6 (last accessed on 06/02/2026).

Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. 2015. Explaining and harnessing adversarial examples. In Bengio, Yoshua & LeCun, Yann (eds), *Proceedings of the 3rd International Conference on Learning Representations ICLR 2015* (San Diego, May 7–9, 2015), n.p. Wisconsin: ICLR. https://arxiv.org/abs/1412.6572 (last accessed on 06/02/2026).

Hendrycks, Dan & Gimpel, Kevin. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the 5th International Conference on Learning Representations ICLR 2017* (Toulon, April 26–27, 2018), n.p. Wisconsin: ICLR. https://arxiv.org/abs/1610.02136 (last accessed on 06/02/2026).

Honnibal, Matthew & Montani, Ines & Boyd, Adriane & Van Lendeghem, Sofie. 2020. spaCy: Industrial-strength Natural Language Processing in Python [Computer Software]. https://spacy.io/ (last accessed on 06/02/2026).

Hüllermeier, Eyke & Waegeman, Willem. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning* 110(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Hurford, James & Heasley, Brendan. 1983. *Semantics: A Coursebook*. Cambridge: Cambridge University Press.

Hyland, Ken. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text & Talk* 18(3). 349–382. https://doi.org/10.1515/text.1.1998.18.3.349

Hyland, Ken. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7(2). 173–191. https://doi.org/10.1177/14614456050503

Jahani, Eaman & Manning, Benjamin & Zhang, Joe & TuYe, Hong-Yi & Alsobay, Mohammed & Nicolaides, Christos & Suri, Siddharth & Holtz, David. 2025. Prompt adaptation as a dynamic complement in generative AI systems. *arXiv preprint.* https://arxiv.org/abs/2407.14333 (last accessed on 06/02/26).

Jiang, Minhao & Liu, Ken Ziyu & Zhong, Ming & Schaeffer, Rylan & Ouyang, Siru & Han, Jiawei & Koyejo, Sanmi. 2024. Investigating Data Contamination for Pre-training Language Models. *arXiv preprint.* https://arxiv.org/abs/2401.06059 (last accessed on 05/02/2026).

Krifka, Manfred. 2019. Commitments and Beyond. *Theoretical Linguistics* 45(1–2), 73–91.

MacDorman, Karl & Chattopadhyay, Debaleena. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146. 190–205. https://www.sciencedirect.com/science/article/pii/S0010027715300755 (last accessed on 06/02/2026).

Manchingal, Shireen Kudukkil & Bradley, Andrew & Kooij, Julian & Shariatmadar, Keivan. 2025. Epistemic Artificial Intelligence is Essential for Machine Learning Models to Truly 'Know When They Do Not Know'. *arXiv preprint.* https://arxiv.org/abs/2505.04950 (last accessed on 06/02/2026).

Mikolov, Tomas & Chen, Kai & Corrado, Greg & Dean, Jeffrey. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Yoshua & LeCun, Yann (eds), *Proceedings of the 1st International Conference on Learning Representations ICLR 2013* (Scottsdale, May 2–4, 2013), n.p. Wisconsin: ICLR. https://arxiv.org/abs/1301.3781 (last accessed on 06/02/2026).

Mori, Masahiro. 1970/2012. The uncanny valley. *Energy* 7. 33–35.

Ofgang, Erik. 2023. What is GPTZero? The ChatGPT detection tool explained by its creator. https://www.techlearning.com/news/what-is-gptzero-the-chatgpt-detection-tool-explained (last accessed on 06/02/2026).

Papernot, Nicolas & McDaniel, Patrick & Jha, Somesh & Frederikson, Matt & Celik, Berkay & Swami, Ananthram. 2015. The limitations of Deep Learning in adversarial settings. In *Proceedings of the 1st IEEE European Symposium*

on Security and Privacy EuroS&P 2016 (Saarbrücken, March 21–24, 2016), 372–387. Los Alamitos: IEEE Computer Society. https://arxiv.org/abs/1511.07528 (last accessed on 06/02/2026).

Singhal, Amit. 2001. Modern information retrieval: a brief overview. In Lomet, David (ed.), *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24(4), 35–43. Washington: IEEE Computer Society. http://sites.computer.org/debull/A01dec/A01DEC-CD.pdf (last accessed on 06/02/2025)

Stalnaker, Robert. 1978. Assertion. In Cole, Peter (ed.), *Pragmatics,* 315-323. New York: Academic Press.

Stalnaker, Robert. 2002. Common ground. *Linguistics and Philosophy* 25. 701-721.

Truckenbrodt, Hubert. 2006. On the semantic motivation of syntactic verb movement to C in German. *Theoretical Linguistics* 32. 257–306.

Wegerhoff, Dennis. 2025a. *Semantic Analysis of AI Generated text* [Computer Software]. https://github.com/DayJay1992/SemanticAIAnalysis (last accessed on 06/02/2026).

Wegerhoff, Dennis. 2025b. Uncanny Semantics – Dataset. *Zenodo*. https://doi.org/10.5281/zenodo.15789515 (last accessed on 06/02/2026).

Wiese, Eva & Weis, Patrick. 2020. It matters to me if you are human – Examining categorical perception in human and nonhuman agents. *International Journal of Human-Computer Studies* 133, 1–12. https://www.sciencedirect.com/science/article/pii/S1071581918306311 (last accessed on 06/02/2026).