You Can't Detect Me! Using Prompt Engineering to Generate Undetectable Student Answers

Marie Ernst[a], Fabian Rupp[b] and Katharina Simbeck[c]
Hochschule für Technik und Wirtschaft Berlin, Berlin, Germany

Keywords: AI-detection-Tools, Readability, Prompt Engineering.

Abstract: Large Language Models (LLMs) have created the opportunity for students to generate answers to assignments. While educators rely on detection tools to identify generated content, students can employ prompt engineering techniques to modify the style of generated outputs and decrease likelihood of detection. In this study, we analyze the impact of intentional AI obstruction through student prompt variation on detection rate using three different AI detection tools. In addition, the AI generated answers are analyzed with regards to their complexity and readability. We found that AI detection tools reliably identified AI generated text. However, prompts leading to intentional imperfections, varied sentence structures and a dynamic writing style were able to reduce recognition rates drastically. We also confirmed that undetected answer were indeed generated in a less elaborated style, commonly associated with younger learners.

1 INTRODUCTION

With the increasing prevalence of tools using generative artificial intelligence (AI) such as ChatGPT, it has become increasingly challenging in academic contexts to determine whether submitted work is authored by students. In response to this issue, specialized AI detection tools have been developed to distinguish between machine-generated and human-written texts. These tools claim to accurately identify AI-generated content in a significant number of cases.

This claim raises the question of whether and how easyly these detection algorithms can be outwitted. This study explores how prompts can be designed to hinder the correct classification of responses as AI generated. As AI generated answers tend to be more elaborate and sophisticated, we utilize text readability measures to quantitatively describe the impact of the prompt on the generated text. We have chosen exemplary assignments in a computer science class in a higher education context.

Previous studies in the field of AI detection tools indicate that the detection accuracy of these tools can be manipulated (Krishna et al., 2023). For example, the use of paraphrasing tools such as DIPPER significantly decreases the detection rate of DetectGPT, reducing it from initial 70.3% to as low as 4.6% (Krishna et al., 2023; Chaka, 2023; Weber-Wulff et al., 2023; Kumarage et al., 2023; Flitcroft et al., 2024; Foster, 2023).

However, such manipulations typically rely on external algorithms or sophisticated techniques. This study, in contrast, examines whether comparable effects can be achieved through targeted modifications to prompts alone—without the need for additional software.

Inspired by the work of Weber-Wulff et al. (2023), who advocate for further research on obfuscation strategies to manipulate AI recognition tools, including the use of machine paraphraser and patch writers, this study seeks to advance understanding in this area. By analyzing and optimizing prompts, this research aims to uncover which linguistic characteristics and wordings are most likely to be interpreted as 'human written' by AI-driven text recognition systems.

The readability of texts is a crucial factor in determining how effectively readers can absorb and understand information (Wang et al., 2022). A previous readability study demonstrates that text complexity negatively impacts reading outcomes, particularly oral reading fluency and recall. More complex texts impose higher cognitive demands, making comprehension more difficult. Therefore, structure and readability of a text are crucial factors for its understandability (Spencer et al., 2019).

The research questions investigated in this study are:

[a] https://orcid.org/0009-0009-5963-8539
[b] https://orcid.org/0009-0006-7946-9689
[c] https://orcid.org/0000-0001-6792-461X

Ernst, M., Rupp, F. and Simbeck, K.
You Can't Detect Me! Using Prompt Engineering to Generate Undetectable Student Answers.

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 304-311
ISBN: 978-989-758-746-7; ISSN: 2184-5026

RQ1: How does intentional prompt variation affect the detection rates of AI-generated content across different AI detection tools?

RQ2: How does prompt engineering influence the complexity and quality of AI-generated responses?

The paper is structured as follows: Section 2 provides an overview of AI detection tools and their mechanisms. Section 3 details different prompt strategies used in this study. Section 4 discusses text style and readability considerations. Section 5 presents the methodology, including experimental design and data collection. Section 6 outlines the results. Finally, Section 7 concludes the paper and provides directions for future research.

2 AI DETECTION TOOLS

With the development of advanced AI models such as GPT-4, the identification of AI-generated texts has become a quality assurance step in education and research. AI detection tools use various methods to differentiate between human-written and machine-generated content. Current research shows that the effectiveness of these systems is increasingly challenged (Chaka, 2023; Weber-Wulff et al., 2023). Weber-Wulff et al. (2023) show in comprehensive tests that the recognition rate varies greatly depending on the tool used. The work underlines the challenge of establishing consistent standards for AI detection (Chaka, 2023; Weber-Wulff et al., 2023).

Anderson et al. (2023) show that the use of paraphrasing tools significantly changes the AI recognition rate. In an example, the "real" score of a text by the GPT-2 Output detector increased from 0.02% to 99.52%.

New developments in AI detection tools such as Fast-DetectGPT rely on the curvature of conditional probabilities to recognize machine-generated texts by choosing more probable words (Bao et al., 2023). This method exploits the discrepancy between collective AI spelling and individual human spelling and improves the efficiency of recognition by requiring fewer model calls (Bao et al., 2023). Research published in BMJ Open SEM (2023) further emphasizes the importance of developing robust detection frameworks to address the growing sophistication of AI text generation systems (Anderson et al., 2023). Another approach is the Multiscale Positive-Unlabeled (MPU) framework, which uses length-sensitive probabilities to accurately analyze variable-length text. It increases recognition accuracy, especially in scenarios where classical methods for AI detection fail due to short texts (Chaka, 2023; Sadasivan et al., 2023).

Chakraborty et al. (2023) show that as the quality of machine-generated texts increases, the sample size required for reliable recognition increases. Using theoretical and empirical analyses (e.g., with datasets such as Xsum and IMDb), they demonstrate that improved recognition methods are feasible (Chakraborty et al., 2023).

Overall, it is clear that the detection of AI-generated texts remains a complex technical challenge that requires continuous research and further development (Dalalah and Dalalah, 2023; Foster, 2023). To illustrate the strengths and weaknesses of current detection methods, three commonly used tools are examined: ZeroGPT, GPTZero, and Copyleaks. These tools were selected because they represent different approaches—probabilistic modeling, statistical analysis, and hybrid AI-rule-based detection.

- **ZeroGPT.** This tool uses probabilistic models, especially log-likelihood calculations, to distinguish between human- and AI-generated texts (ZeroGPT, 2024). By analyzing token probabilities in context, it identifies patterns typical of each (ZeroGPT, 2024). Texts with uniform probabilities and low token variability are flagged as AI-generated (ZeroGPT, 2024). The tool also detects machine-like traits, such as repetitive structures and predictable word sequences, without needing extensive training data (ZeroGPT, 2024). Studies show ZeroGPT excels at spotting the consistent styles of AI-generated writing (Kumarage et al., 2023; Taguchi et al., 2024).
- **GPTZero.** This tool relies on statistical and dynamic features, such as text length, syntactic complexity, and token perplexity, to detect AI-generated content (Tian and Cui, 2024; GPTZero, 2024). Human-written texts typically show higher perplexity due to idiomatic expressions and grammatical variability (Tian and Cui, 2024). GPTZero leverages pre-trained models like RoBERTa to spot syntactic and semantic irregularities, common in AI-generated texts with excessive coherence or complexity (Tian and Cui, 2024). It also analyzes how text traits change with varying prompts, improving adaptability and resilience against manipulation (Kumarage et al., 2023; Park et al., 2024).
- **Copyleaks** Combining rule-based methods with AI-driven algorithms, Copyleaks employs DetectGPT, which evaluates probabilistic differences between original and slightly altered texts (Copyleaks, 2024). Machine-generated texts are more sensitive to such changes, as AI models favor high-probability outputs (Copyleaks, 2024).

Copyleaks leverages these deviations to detect AI-generated content reliably (Copyleaks, 2024). It also identifies advanced manipulations like paraphrasing and stylistic tweaks, making it highly effective in academic settings (Copyleaks, 2024; Park et al., 2024; Taguchi et al., 2024).

3 PROMPT STRATEGIES

Optimizing prompts is a promising strategy to ensure texts are classified as human-written (Kumarage et al., 2023). Even simple adjustments, such as altering writing perspective or sentence structure, can significantly influence classification results (Kumarage et al., 2023). Variations in length, syntax, and lexical diversity provide greater control over text output (Park et al., 2024).

Strategies to manipulate AI detection tools include introducing deliberate imperfections, such as grammatical errors or inconsistent sentence structures (Park et al., 2024; Foster, 2023). Alternating short and long sentences or using idiomatic expressions can improve human readability and reduce detectability (Weber-Wulff et al., 2023). Authentic language styles, like those mimicking a master's student, further enhance text authenticity (Kumarage et al., 2023; Flitcroft et al., 2024). Avoiding typical AI patterns, such as overly regular structures or excessive grammatical accuracy, is another common approach (Park et al., 2024; Foster, 2023).

To investigate the impact of prompt-specific abbreviations in the recognition of AI-generated texts Park et al. (2024) developed a new attack method FAILOpt. FAILOpt uses feedback to optimize instructions that specifically degrade recognition performance (Park et al., 2024). The study shows that the FAILOpt method can significantly impair the performance of AI text detectors and that detectors trained on limited input prompts could easily be fooled by specific instructions (Park et al., 2024).

Foster (2023) highlights that well-crafted prompts can enable GPT-4 to create texts classified as human by advanced systems such as Turnitin (Foster, 2023). Foster emphasizes that variations in text structure and semantic depth are particularly influential in evading detection (Foster, 2023).

Researchers argue that the detection of AI-generated texts becomes problematic in the long term, as the distinction between AI and human text distributions is made more difficult by total variation distance (Dalalah and Dalalah, 2023; Sadasivan et al., 2023). This could result in recognition accuracy barely exceeding random decisions (Dalalah and Dalalah, 2023; Sadasivan et al., 2023). Chaka (2023) points out that even embedded watermarks or paraphrasing tools can make detection almost impossible, as the similarity between AI-generated and human texts is further increased (Chaka, 2023). The challenges of detection highlight the need for rigorous evaluations of the systems in terms of their reliability and robustness against tampering attempts (Weber-Wulff et al., 2023; Sadasivan et al., 2023).

Despite their success, these techniques face challenges. Advanced methods like feedback-based optimization or adversarial prompts often target specific weaknesses of individual tools and lack universal applicability (Park et al., 2024). Moreover, such strategies can reduce text readability, especially in academic settings (Foster, 2023).

While prompt design has proven effective, few studies explore the interplay between prompt optimization and text style (Flitcroft et al., 2024). Further research is needed to assess how optimized prompts impact both detectability and content quality (DuBay, 2007).

4 TEXT STYLE AND READIBILITY

Readability is the ease with which a text can be understood, influenced by its content, style, design, and structure, and how well these align with the reader's background, abilities, interests, and motivation (DuBay, 2007). It is not the same as legibility, which is about how clear and visually easy the text is to see, such as the font and layout (Dubay, 2004). The main idea is to help adjust the difficulty of written material to match the reader's ability, thereby enhancing communication and learning (Zakaluk and Samuels, 1988). Edgar Dale and Jeanne Chall (1949) described readability as the combination of factors in a text that determine how successfully readers can understand it, read it efficiently, and find it engaging or interesting (Dubay, 2004). Sentence construction impacts readability with shorter or simpler sentences often enhancing readability while maintaining a balance of sentence lengths for style (Klare, 2000). Shorter words are more frequent and versatile in meaning, while longer words are often less familiar; long sentences, with complex syntactic structures place greater cognitive demands on the reader (Tekfi, 1987). Several readability formulas have been developed to evaluate the difficulty of written text. These formulas typically focus on two key aspects: (1) the complexity of sentences, often measured by their length, and (2) the difficulty of words used in the text (Thomas et al., 1975).

The Flesch Reading Ease Score and the Gunning-Fog Index are well-established Formulas for measuring text readability (Flesch, 1948). The Flesch score considers the average sentence length and the average number of syllables per word, favoring texts with clear and simple language (Flesch, 1948). The Gunning-Fog Index, on the other hand, evaluates readability by analyzing sentence length and the proportion of complex words, with complex words defined as those with three or more syllables (Gunning, 1952).

The Wiener Sachtextformel (WSF) evaluates text complexity by analyzing the proportion of words with three or more syllables, words with over six letters, monosyllabic words, and the average sentence length (Dunkl, 2015). Specifically designed for German texts, it evaluates readability by analyzing factors like sentence length, the proportion of monosyllabic and polysyllabic words, and word length. Lower scores represent simpler texts (Bamberger and Vanacek, 1984).

$$\begin{aligned}\text{WSF} = {} & 0.1935 \cdot \text{ASL} + 0.1672 \cdot \text{ASW} \\ & + 0.1297 \cdot \text{PSW} - 0.0327 \cdot I - 0.875\end{aligned}$$

The formula uses the average sentence length (ASL), the number of syllables per word (ASW), the proportion of polysyllabic words (PSW) and the proportion of personal pronouns (I). These factors influence the comprehensibility of the text (Bamberger and Vanacek, 1984).

5 METHOD

This study investigates whether AI detection can be outwitted through prompt engineering and which text properties cause tools to fail. All analyzed texts were created with GPT-4 using the default settings (ChatGPT, 2024). This ensures that the generated output corresponds to those of standard users. The AI recognition tools ZeroGPT, GPTZero and Copyleaks classify the previously generated texts (ZeroGPT, 2024; GPTZero, 2024; Copyleaks, 2024). The free versions of the tools are utilized and the default settings are retained. The selection of these tools is based on two primary criteria: first, their accessibility due to being free of charge, and second, their demonstrated performance in previous studies (Singh, 2023; Chaka, 2023; Flitcroft et al., 2024; Weber-Wulff et al., 2023). All generated texts are copied from ChatGPT with the help of the copy key combination and pasted into the text fields of the three AI recognition tools using the paste key combination. Finally, the texts are classified by the tools. The prompts and assignments used can be found under the following link: https://iug.htw-berlin.de/you-cant-detect-me/. All prompts, assignments, and resulting texts are in German.

5.1 Assignment Questions

The tasks are set in the context of the business computing course Enterprise Content Management (ECM) on master degree. A total of 15 tasks are used, covering a range of difficulty levels and subject areas and requiring text-based answers. The first five tasks (A1-A5) originate from actual examinations in the master's program in business computing at HTW Berlin. The other ten tasks (B1-B10) were generated using ChatGPT. To ensure a balanced selection, these tasks are categorized into five levels of difficulty: Basic, intermediate, advanced, expert and strategic and future-oriented tasks. Each category includes two tasks designed to vary in technical depth and the degree of abstraction required in the answers. This combination of real-world and AI-generated tasks enables a well-founded analysis of the prompts across varying levels of difficulty and application scenarios.

5.2 Prompt Design

The development and optimization of prompts occurs in iterative steps to identify which prompt elements are most likely to cause misclassification by AI recognition tools. The design process is based on the findings of previous work in this area (Kumarage et al., 2023; Park et al., 2024; Foster, 2023). The process begins with the basic prompt 1 that instructs the model to directly answers the task. In the next step, the prompt is expanded by specifying a writing style (prompt 2). The prompting instructs the model to write in the style of a Master's student in business computing in their mid-twenties with a Bachelor's degree. The goal is to create an authentic yet academic language. Additionally, the prompt emphasizes to create texts that AI recognition tools cannot identify as machine-generated. Another approach involves revising texts previously generated by ChatGPT (prompt 3 and 4). The revisions aim to eliminate features typically associated with AI-generated content. Key indicators such as consistent sentence structures, overly coherent word choices, and flawless transitions were found to increase the likelihood of classification as AI-generated (Park et al., 2024; Foster, 2023). To counteract this, minor grammatical errors and a less rigid structure should make the text appear more human (Kumarage et al., 2023; Foster, 2023). In addition, introductions and summaries are omitted to focus on answering the question short

Table 1: Prompt-characteristics used to generate texts.

	1	2	3	4	5	6	7	8	9
Scientific language									x
Avoid AI patterns		x	x	x	x				
Mistakes			x	x		x	x		
Explicit naming of AI patterns to be avoided			x		x	x	x	x	
Structure			x	x	x	x	x	x	x
Student perspective		x	x	x	x				x
Stylistic devices								x	x
Continuous text			x	x		x	x		x
Short text			x				x		

and to the point, as shorter texts are harder for AI detection tools to classify (Chaka, 2023; Sadasivan et al., 2023). In order to determine whether the explicit naming of the patterns to be avoided makes a difference, the revision was tested in two scenarios: emphasizing to create texts that AI recognition tools cannot identify as machine-generated (prompt 4) and on the other hand with explicit naming of the AI patterns (prompt 3). In contrast to prompt 1, 2 and 4, prompt 3 contains the typical AI patterns and it is ensured that these are avoided. The typical AI patterns to be avoided were additionally tested within three scenarios: executed to revise the previously generated texts (prompt 3), directly in connection with the task (prompt 5) and in combination with the word "briefly" in front of the respective task (prompt 6), e.g. "briefly describe what coded and non-coded information is". Furthermore, advanced rephrasing strategies are employed (prompt 7-9). These include alternating short and long sentences, using idiomatic expressions, and adding occasional digressions for a more dynamic and engaging tone. Stylistic devices like comparisons, metaphors, and rhetorical questions further enrich the text, making it vivid and varied. The prompts were not executed multiple times per task. An overview of the different prompt-characteristics can be found in table 1.

6 RESULTS

6.1 AI Detection Tools

Prompt design impact the classification of text as AI genereated or written by human. The effectiveness of prompt changes between detection tools. An overview can be found in table 2. The simplest prompt, prompt 1, resulted in the highest likelihood of texts being classified as AI-generated. Across all three tools 95% of the texts were classified as AI-generated, while only 4% were identified as human. Only GPTZero classified two text as human-written. Adapting the writing style in prompt 2, to resemble that of a master student in business informatics and avoiding AI patterns, led to slight improvements. With this approach, 93% of the texts were still recognized as AI-generated and 7% as human. The tools again largely converged in their classifications. A targeted revision in prompt 3 of the texts created by prompt 2 improved the results. The proportion of texts classified as AI-generated dropped to 29%, while 71% were classified as human. This underscores the importance of explicitly addressing typical AI patterns, such as uniform sentence structures and grammatical perfection, in the prompt. Also the instruction to focus only on the essential points to answer the question seems to have a proactive influence. However, the tools varied in their responsiveness to this prompt. Prompt 4 mimicks a master student and subsequent revisions yielded lower-than-expected success. Although 29% of texts were classified as human, this approach was less effective than the previous revision. This leads to the conclusion that enumerating the typical AI patterns to avoid and to focus only on the essential points to answer the question probably has an influence on the effectiveness. This suggests that these instructions are important to ensure that texts are predominantly classified as written by humans. The highest success rate was achieved with prompt 6 generating an lively, dynamic and deliberately imperfect text with varied sentence structure, varied word choice and occasionally faulty transitions. This leads to an rise of human classifications to 86%. The human classification rate is similar for all 3 detection tools. Adding the term "briefly" leads to good results as well, but at 64% human classifications fails to match prior results. Contrary to expectations, advanced reformulation strategies, in prompts 8 and 9, incorporating idiomatic expressions, varied sentence structures, and occasional digressions do not yield meaningful improvements. With 93% AI classifications using prompt 8 and only 80% for prompt 9, this approach fell far short of expectations. All three tools exhibit similar results. This suggests that

Table 2: Comparison of AI Detection Tools: Percentage of Texts Classified as Human-Written and mean, standard deviation and polysyllable count of WSF for the individual results.

Prompt	ZeroGPT	GPTZero	Copyleaks	Total avg. prompt	average score	Standard deviation	mean polysyll.
1	0%	13%	0%	4%	15.3	1.5	206
2	6%	13%	0%	7%	13.9	1.5	226
3	87%	73%	67%	71%	9.2	1.3	95
4	20%	67%	0%	29%	9.4	1.3	107
5	0%	13%	0%	4%	14.8	1.7	193
6	93%	80%	87%	87%	8.6	1.1	137
7	100%	40%	53%	65%	7.8	0.9	62
8	7%	13%	7%	7%	12.5	1.2	117
9	27%	13%	27%	20%	12.6	2.7	114
Total avg. Tool	13%	12%	9%	34%			

stylistic sophistication alone, without explicit imperfection, is insufficient to influence classification outcomes. ZeroGPT achieves the most human classifications in Prompts 7. GPTZero showed its best performance Prompt 6. Copyleaks, the strictest tool with the most AI classifications, responded well to Prompt 6. The 3 tools have similar total AI classification rates. The study shows that targeted variations in prompt design influence the recognition rates of AI-generated texts. With regard to research question RQ1, it can be stated that prompts that incorporate intentional imperfections such as grammatical mistakes, irregular sentence structures and dynamic writing styles reduce the recognition rate, while advanced reformulations without deliberate deviations were less effective. Prompts that incorporate typical AI patterns, which should be avoided, make detection by current tools more challenging. It is important to note that repeated executions of the same prompts can generate different texts, potentially leading to variability in results.

6.2 Text Style and Detection

In examining the impact of text style on AI detection, the readability and complexity of texts generated by different prompts were analyzed. The different prompts yielded texts that differ strongly in stylistic complexity, measured by WSF-score (Table 3). WSF was chosen because it is specifically developed for the German language and takes sentence length, word complexity into account. WSF values typically range from 4 to 15, where 4 indicates very easy texts suitable for younger students, and 15 indicates very difficult texts suitable for advanced readers on an academic level. The analysis revealed that some texts were evaluated as extremely complex, due to WSF Score ($>$14), while others were deemed easily readable for ninth-grade students (ages 14-15). WSF scores and AI detection results show a notable correlation. It aligns closely with the WSF scores, suggesting it is well-suited for evaluating the readability and complexity of German texts. For instance, texts generated by ChatGPT that are typically at a master's level are often recognized as AI-generated. In contrast, texts not recognized as AI-generated tend to be at a high school level, suitable for students aged 14-15. This indicates that simpler texts with lower readability scores are more likely to be classified as human-written. Prompt 1 has no specific features to avoid AI patterns or include mistakes and has the highest WSF score (15.3), indicating very complex texts that are difficult to understand. Prompt 3 includes several features such as avoiding AI patterns and incorporating mistakes, resulting in a lower WSF score (9.2), indicating simpler and more understandable texts. Prompt 6 also has many features to avoid AI patterns and include mistakes, leading to one of the lowest WSF scores (8.6). Prompt 9 contains scientific language and stylistic devices, resulting in a higher WSF score (12.6) and a larger standard deviation (2.7), indicating greater variability in text complexity. Prompts that explicitly avoid AI patterns and include mistakes result in lower WSF scores and higher rates of human classification. For example, Prompt 3 and Prompt 6, which incorporate these features, have lower WSF scores (9.2 and 8.6) and higher human classification rates (71% and 87%). In contrast, Prompt 1, with no special features and a high WSF score (15.3), has a low human classification rate (4%). This indicates that simpler, less complex texts are more likely to be recognized as human-written. Lower WSF scores (indicating simpler texts) correlate with higher human classification rates. For example, Prompt 6 has a low average WSF score of 8.6 and a high human classification rate of 87%. Higher WSF scores (indicating more complex texts) correlate with

lower human classification rates. Prompt 1, with an average WSF score of 15.3, has a low human classification rate of 4%. In addition, further readability properties were examined. Long sentences and polysyllabic words impact the readability of texts, making them more challenging to understand. High syllable and lexicon counts generally indicate a more detailed and complex text. Conversely, texts with more monosyllabic words and shorter sentences promote higher readability, resulting in better comprehension and lower readability index scores.

7 CONCLUSIONS

The results of this study show that AI recognition tools can be manipulated by strategic prompt design. Introducing human-like imperfections, alternating sentence structures and thus avoiding typical AI patterns increases the likelihood of AI-generated texts being categorized as written by humans. These strategies were also very effective when combined with focused and concise responses to the task. Furthermore, we show the varying effectiveness of the recognition tools, with GPTZero showing the highest sensitivity to prompt adaptation and Copyleaks the lowest.

The readability and complexity of texts generated by different prompts were analyzed using the WSF-readability score. Prompts that explicitly avoided AI patterns and included mistakes resulted in lower WSF scores and higher rates of human classification. This indicates that simpler, less complex texts are more likely to be recognized as human-written.

The results of this study confirm studies, such as those by Krishna et al. (2023) and Weber-Wulff et al. (2023), who have demonstrated that AI detection accuracy can be manipulated through paraphrasing and other external tools. This study contributes to the field by showing that similar effects can be achieved through strategic prompt design alone, without the need for additional software.

Our findings also resonate with the work of Anderson et al. (2023), who showed that paraphrasing tools could significantly alter AI recognition rates. Similarly, our study demonstrates that prompt modifications can achieve comparable results. Additionally, the research by Foster (2023) on the impact of text structure and semantic depth on detection aligns with our findings that dynamic and varied writing styles reduce AI detection rates.

Moreover, the inclusion of readability analysis using the WSF formula provides a novel perspective. While prior research has focused on the technical manipulation of text to evade detection, our findings highlight the importance of text readability and complexity. Texts with lower readability scores, indicating simpler language, are more likely to be classified as human-written. This suggests that readability metrics can be a valuable tool in understanding and improving the effectiveness of prompt engineering strategies.

This research not only confirms the manipulability of current detection systems but also provides a framework for future studies to explore the interplay between readability and AI detection. Our findings highlight the need for improved detection algorithms capable of recognizing prompt engineering tactics. Further research could explore dynamic detection models that adapt to evolving manipulation strategies and ensure more robust systems for identifying AI-generated content. As AI detection tools continue to be unreliable, educators need to consider either controlling for AI use in in-classroom tests or increasing task difficulty while allowing use of AI tools.

REFERENCES

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., and Memon, A. R. (2023). Ai did not write this manuscript, or did it? can we trick the ai text detector into generated texts? the potential future of chatgpt and ai in sports & exercise medicine manuscript generation.

Bamberger, R. and Vanacek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben*. Diesterweg.

Bao, G., Zhao, Y., Teng, Z., Yang, L., and Zhang, Y. (2023). Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Chaka, C. (2023). Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools. *Journal of Applied Learning and Teaching*, 6(2).

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., and Huang, F. (2023). On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

ChatGPT (2024). https://chatgpt.com.

Copyleaks (2024). https://help.copyleaks.com/hc/en-us/articles/23768610748301-How-does-Copyleaks-work.

Dalalah, D. and Dalalah, O. M. (2023). The false positives and false negatives of generative ai detection tools in education and academic research: The case of chatgpt. *The International Journal of Management Education*, 21(2):100822.

Dubay, W. (2004). The principles of readability. *CA*, 92627949:631–3309.

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text.* ERIC.

Dunkl, M. (2015). *Verständlichkeit*, pages 41–88. Springer Fachmedien Wiesbaden, Wiesbaden.

Flesch, R. F. (1948). A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Flitcroft, M. A., Sheriff, S. A., Wolfrath, N., Maddula, R., McConnell, L., Xing, Y., Haines, K. L., Wong, S. L., and Kothari, A. N. (2024). Performance of artificial intelligence content detectors using human and artificial intelligence-generated scientific writing. *Annals of Surgical Oncology*, pages 1–7.

Foster, A. (2023). Can gpt-4 fool turnitin? testing the limits of ai detection with prompt engineering.

GPTZero (2024). https://gptzero.me.

Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill.

Klare, G. R. (2000). The measurement of readability: useful information for communicators. *ACM J. Comput. Doc.*, 24(3):107–121.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 27469–27500. Curran Associates, Inc.

Kumarage, T., Sheth, P., Moraffah, R., Garland, J., and Liu, H. (2023). How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. *arXiv preprint arXiv:2310.05095*.

Park, C., Kim, H. J., Kim, J., Kim, Y., Kim, T., Cho, H., Jo, H., Lee, S.-g., and Yoo, K. M. (2024). Investigating the influence of prompt-specific shortcuts in ai generated text detection. *arXiv preprint arXiv:2406.16275*.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. (2023). Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Singh, A. (2023). A comparison study on ai language detector. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0489–0493. IEEE.

Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., and Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing*, 32:603–637.

Taguchi, K., Gu, Y., and Sakurai, K. (2024). The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*.

Tekfi, C. (1987). Readability formulas: An overview. *Journal of documentation*, 43(3):261–273.

Thomas, D. G., Hartley, R. D., and Kincaid, J. P. (1975). Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.

Tian, E. and Cui, A. (2024). Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.

Wang, S., Liu, X., and Zhou, J. (2022). Readability is decreasing in language and linguistics. *Scientometrics*, 127:4697–4729.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., and Waddington, L. (2023). Testing of detection tools for ai-generated text (arxiv: 2306.15666). arxiv.

Zakaluk, B. L. and Samuels, S. J. (1988). *Readability: Its Past, Present, and Future.* ERIC.

ZeroGPT (2024). https://www.zerogpt.com.

APPENDIX

Table 3: Used wordings for the prompt characteristics.

Characteristic	Wording
Scientific language	Write the text in neutral and factual language
Avoid AI patterns	generated text is not recognized by AI detectors
Mistakes	insert a few grammatical errors
Explicit naming of AI patterns to be avoided	Avoid typical AI patterns such as uniform sentence structure, consistent word choice, perfectly flowing transitions and grammatical correctness. Make your text varied, "imperfect" and a little less stringent.
Structure	omit all headings, introduction and conclusion/summary
Student perspective	Master's student, using natural language as a person in their mid-twenties with a Bachelor's degree
Stylistic devices	Occasionally use stylistic devices such as rhetorical questions, comparisons or metaphors
Continuous text	Write a continuous text
Short text	Focus only on necessary information to answer the question and leave out everything else