

Comparison of the effectiveness of selected tools for detecting texts generated by artificial intelligence

Marcin Brodacki*, Małgorzata Plechawska-Wójcik

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

This paper compares the effectiveness of selected publicly available tools for detecting AI-generated text: GPTZero AI Detector, ZeroGPT AI Detector, Copyleaks AI Detector, and Scribbr AI Detector. The experiment included 150 text samples in Polish and English, divided into human-written texts, AI-generated texts, and texts paraphrased using AI. ZeroGPT achieved the highest overall accuracy, equal to 64.0%, and the highest F1-score, equal to 77.7%. GPTZero achieved an accuracy of 58.7%, while Scribbr and Copyleaks achieved lower results, equal to 44.0% and 42.7%, respectively. Fully AI-generated texts were detected more effectively than AI-paraphrased texts. The overall accuracy was similar for Polish and English samples, equal to 52.7% and 52.0%, respectively.

Keywords: artificial intelligence; text detection; LLM; AI detector; machine-generated text

*Corresponding author

Email address: s95370@pollub.edu.pl (M. Brodacki)

Published under Creative Common License (CC BY 4.0 Int.)

1. Introduction

In recent years, there has been rapid development in artificial intelligence technology, particularly in the area of large language models (LLMs). These models enable the generation of text with a high level of coherence, grammatical correctness, and semantic consistency [1]. As a result, it is becoming increasingly difficult to distinguish between content created by humans and text automatically generated by artificial intelligence systems [2, 3].

The aim of this article is to compare the effectiveness of selected publicly available tools for detecting AI-generated text: GPTZero, ZeroGPT, Copyleaks AI Content Detector, and Scribbr AI Detector. The study covered applying selected AI text detectors to verify a prepared set of text samples.

The scope of the study includes an analysis of the performance of selected detectors on human-written texts, AI-generated texts, and paraphrased texts. The study included samples in Polish and English, and the results were evaluated based on correct classifications, false positives, false negatives, and ambiguous cases.

The article adopts three research hypotheses. H1 assumes that detectors achieve higher accuracy in identifying texts fully generated by AI models than texts written by humans that were then paraphrased using AI. H2 assumes that the effectiveness of individual tools differs significantly. H3 assumes that detection results depend on the language, and the detector's effectiveness in Polish texts may be lower than in English, consistent with the observation that models are often trained primarily on English texts.

2. Related works

Weber-Wulff et al. [4] analysed 12 publicly available tools and two commercial systems used in the academic community. The authors investigated whether the tools could recognise human texts, texts generated by

ChatGPT, and texts that had been translated, edited, or paraphrased. The results indicated that the detectors were not sufficiently accurate and reliable, and that modification techniques significantly degraded their performance.

Perkins et al. [5] investigated simple techniques for bypassing GenAI detectors. Their results showed that even simple text modifications can degrade the detection tools' performance. The authors reported a 17.4% decrease in accuracy after implementing these modifications.

Kartelj, Mladenović, and Vujičić Stanković [6] investigated the detection of paraphrased texts by ChatGPT in English and Serbian. The authors tested a number of classifiers, and the results showed better performance in English than in Serbian. The study highlights that paraphrasing is not only a technical issue, but also a linguistic one, as detection efficiency may depend on the resources and characteristics of the given language.

Schaaff, Schlippe, and Mindner [7] studied the classification of human, AI-generated, and AI-paraphrased texts in various languages and domains. The authors developed a multilingual corpus containing educational texts in English, French, German, and Spanish, as well as news texts. The study showed that the detection of AI-generated and paraphrased texts can be performed in multiple languages, but its effectiveness depends on the text type and language.

Liang et al. [8] described the problem of detector bias against people writing in English as a foreign language. In their study, detectors often marked TOEFL texts, i.e., texts written as part of the Test of English as a Foreign Language, as AI-generated texts.

In the context of this article, these findings justify the inclusion of paraphrased texts in the experiment, since testing only unmodified AI-generated texts does not fully reflect real conditions of AI tool use. The reviewed studies also indicate that detection effectiveness may

depend on the detector, the language of the sample, and the type of analysed text.

3. Materials and methods

3.1. Research material

The research material was divided into 15 thematic sets. Each set covered a single topic and was prepared in two language versions: Polish and English. This structure limited the impact of thematic differences on the study results and allowed for a comparison of detector performance across languages.

Each set contained five types of text: human-written text, ChatGPT-generated text, Claude-generated text, human text paraphrased by ChatGPT, and human text paraphrased by Claude. Each sample was approximately 250 words long. A total of 150 text samples were prepared, including 30 human-written texts, 60 AI-generated texts, and 60 AI-paraphrased texts.

Human-written texts served as control material. They were obtained from sources published before 2017, before the widespread use of modern large-scale language models. A coherent fragment of approximately 250 words was selected from each text. The style, vocabulary, and sentence structure were not altered. Only technical elements such as headlines, captions, advertisements, links, or fragments unrelated to the actual content were removed.

3.2. Tools

All measurements were conducted between 6 May 2026 and 9 May 2026 using the free browser-based versions of the tools. Since these tools may be updated by their providers, the results refer to the state of the detectors during that measurement period.

3.2.1. GPTZero AI Detector

GPTZero AI Detector [9] is a web-based tool for detecting texts generated by large language models. The study used the free version of the tool, accessible through a web browser. The classification result indicates whether the text was written by a human, generated by AI, or is a mixture of both. The manufacturer states that the tool can indicate the use of AI at the level of the entire document, sentences, and individual words.

3.2.2. ZeroGPT AI Detector

ZeroGPT AI Detector [10] is a web-based AI-based content detection tool. The tool allows to paste text into a form and then returns an overall score indicating the likelihood of AI-generated content. According to the manufacturer, the detector analyses linguistic and statistical signals, such as token patterns, entropy, and classification features learned from human and AI-generated text collections.

3.2.3. Copyleaks AI Content Detector

Copyleaks AI Content Detector [11] is a web-based tool for detecting AI-generated content. The tool analyses text using linguistic and statistical patterns characteristic of both human-written and AI-generated content. It returns

a percentage score indicating how much of the text is likely AI-generated and can indicate mixed content.

3.2.4. Scribbr AI Detector

Scribbr AI Detector [12] is a web-based tool for detecting AI-generated text. The provider does not disclose the exact detection model used, but explains that the tool compares the analysed text with patterns associated with AI-generated content, such as predictable word choices and limited variation in sentence structure and length. The result is presented as a percentage from 0% to 100%, and the tool can also indicate fragments that likely contain AI-generated or AI-processed content.

3.2.5. ChatGPT (GPT-5.5 Instant)

ChatGPT [13, 14] was used to prepare one group of generated texts and one group of paraphrased texts. The producer describes ChatGPT as an AI assistant that can be used for various tasks, such as writing, learning, planning, mathematics, coding, and image and file analysis. In this study, the GPT-5.5 Instant model available in the ChatGPT interface during the experiment was used. The model name was recorded according to the label visible in the tool interface and documentation at the time of the measurements.

3.2.6. Claude Sonnet 4.6

Claude [15] is a family of large language models developed by Anthropic. The Claude Sonnet 4.6 model was used in this study. The manufacturer describes it as an advanced model for more complex tasks such as coding, long-context reasoning, agent planning, and design.

3.3. Scenarios

The study used three research scenarios. Each scenario corresponded to a specific type of text sample and allowed for the evaluation of a different aspect of the detectors' performance. The testing procedure was the same for all scenarios. The differences concerned the type of sample analysed and the method of interpreting the results.

The first scenario involved detecting human-written texts. The goal of this scenario was to assess whether the detectors correctly identified human-written texts and whether they incorrectly labelled human texts as AI-generated.

The second scenario involved the detection of fully AI-generated texts. This scenario included two sample variants: texts generated by ChatGPT and texts generated by Claude.

The third scenario involved the detection of texts paraphrased by AI. The goal of this scenario was to test how the detectors classified texts originally written by humans but then reworded by the AI model.

In each scenario, a text sample was prepared according to the scenario type, the sample was pasted into the subsequent detectors, the analysis was run in each tool, the raw result returned by the detector was saved,

the result was assigned to a unified category and the result was saved in the final table.

3.4. Standardisation and analysis of results

The analysed tools presented results differently, so before comparing the results, they had to be standardised. Each result was first recorded in its raw form, consistent with the tool's message. It was then assigned to one of three common categories: human text, AI-generated text, or an ambiguous or mixed result.

For percentage scores, auxiliary interpretation thresholds were adopted. Scores between 0% and 40% probability of being generated by AI were marked as human text. Scores above 40% and below 60% were marked as inconclusive or mixed. Scores between 60% and 100% were marked as AI-generated text.

Based on the standardised results, it was determined whether the detector's result was correct, incorrect, or ambiguous. For human-written texts, a correct classification was considered to indicate the human text category. If the detector marked the human text as generated by artificial intelligence, the result was considered a false positive. For texts fully generated by AI, a correct classification was considered to indicate the AI-generated text category. If such text was marked as human text, the result was considered a false negative.

Texts paraphrased using AI were treated as mixed-type samples, as their source content originated from human-written text, but selected passages had been processed by an AI model. For this group, a separate analysis was conducted to determine whether the detector recognised the mixed nature of the sample. A situation in which the paraphrased text was classified as entirely human was considered a failure to detect AI intervention.

For pairwise comparisons of detectors, exact McNemar tests were applied to paired binary outcomes describing whether a given detector classified the same sample correctly or not. Ambiguous or mixed results were treated as incorrect classifications in this analysis, consistently with the accuracy calculation.

4. Results

4.1. Detection of human-written texts

GPTZero, Copyleaks, and Scribbr correctly identified all human-written texts in both language variants. Each tool classified 15 of 15 Polish samples and 15 of 15 English samples as human-written texts. These detectors experienced no AI false positives or unreliable results.

ZeroGPT produced different results. For texts in Polish, the tool correctly identified 5 out of 15 samples as human-written text, whilst it incorrectly classified 6 samples as AI-generated text and classified 4 samples as uncertain results. For texts in English, ZeroGPT achieved a better result, correctly identifying 11 out of 15 samples and marking 4 results as uncertain. Results are shown in Figure 1.

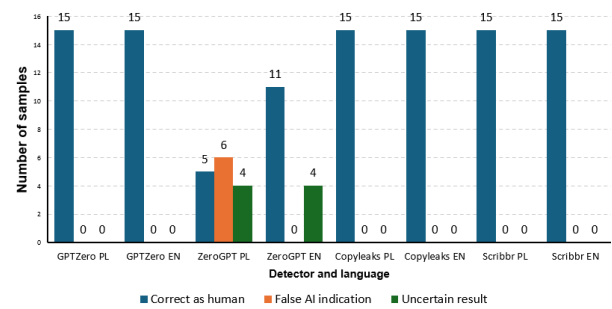


Figure 1: Detection results for human-written texts

4.2. Detection of AI-generated texts

The GPTZero and ZeroGPT tools achieved the highest detection rates for AI-generated text. Both detectors correctly identified 28 out of 30 samples in Polish and all 30 samples in English. This means that, in the case of texts generated entirely by AI, these tools achieved the best results among the detectors analysed.

Copyleaks achieved lower performance, particularly with Polish texts. The tool correctly identified 11 out of 30 Polish samples as AI-generated text, whilst 19 samples were not recognised as having been generated by artificial intelligence. For English texts, the result was significantly better, as Copyleaks correctly identified 23 out of 30 samples.

Scribbr produced the opposite result to Copyleaks. For Polish texts, the tool correctly identified 22 out of 30 samples as AI-generated, whilst for English texts it correctly detected 14 out of 30 samples. Scribbr also produced some inconclusive results: one for Polish texts and four for English texts. The results are shown in Figure 2.

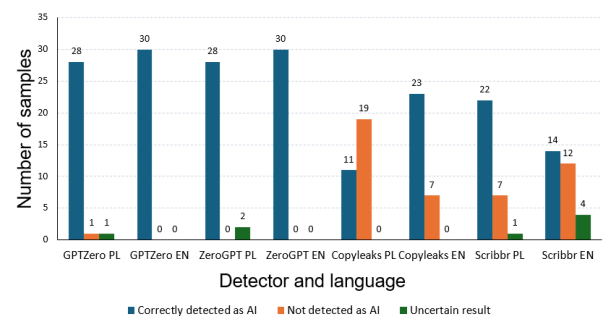


Figure 2: Detection results for AI-generated texts

4.3. Detection of paraphrased texts by artificial intelligence

The greatest differences between the tools were observed in the case of paraphrased texts. GPTZero, Copyleaks and Scribbr mostly classified such samples as human-written texts. GPTZero did not identify any paraphrased text as being entirely AI-generated. For Polish, 29 out of 30 samples were classified as human-written text, with 1 result marked as uncertain. For English, 23 out of 30 samples were classified as human-written text, with 7 results marked as uncertain.

Copyleaks did not detect any paraphrased samples as AI-generated text, nor as mixed or uncertain results. All

samples in Polish and English were classified as human-written text. Scribbr produced similar results, also failing to flag any paraphrased sample as text fully generated by artificial intelligence. For Polish, 28 out of 30 samples were deemed to be human-written, whilst 2 results were uncertain. For English, all samples were classified as human-written.

ZeroGPT achieved the highest detection rate of AI interference in paraphrased texts. For Polish, the tool flagged 19 samples as AI text, 7 samples as human text and 4 results as uncertain. This means that, when taking into account the uncertain results as possible detections of AI interference, the tool detected 23 out of 30 samples. For English, the effectiveness was lower, as ZeroGPT identified 3 samples as AI text, 18 samples as human text, and 9 results as uncertain. Results are shown in Figure 3.

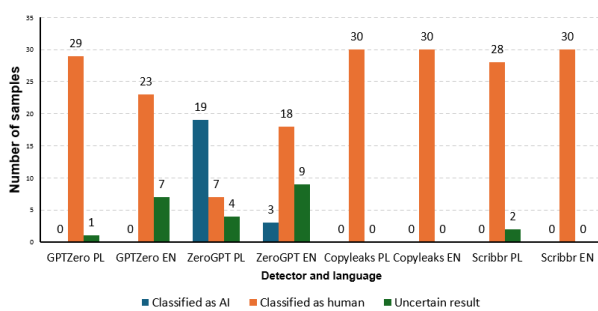


Figure 3: Detection results for AI-paraphrased texts

4.4. Comparison of tool effectiveness

The next part of the analysis involved a general comparison of the effectiveness of the detectors under review. To this end, the results obtained by GPTZero, ZeroGPT, Copyleaks AI Content Detector and Scribbr AI Detector were collated for all the samples analysed. The results are shown in Figure 4.

Table 1: Effectiveness metrics of the analysed AI detectors

Detector	Accuracy	AI precision	AI recall	F1-score
GPTZero	58.7%	100%	48.3%	65.2%
ZeroGPT	64%	93%	66.7%	77.7%
Copyleaks	42.7%	100%	28.3%	44.2%
Scribbr	44%	100%	30%	46.2%

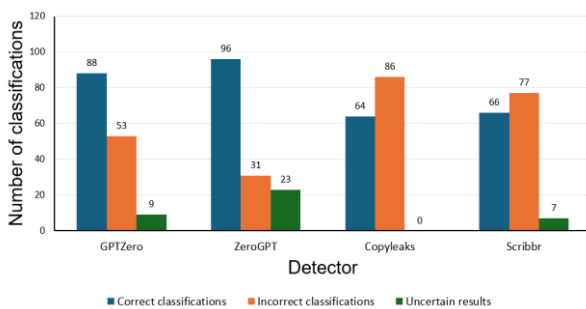


Figure 4: Classification results by detector

Table 1 presents the detailed performance metrics, including accuracy, AI precision, AI recall, and F1-score.

Accuracy was calculated as the number of correct classifications divided by the total number of analysed samples. Ambiguous or mixed results were included in the total number of samples, but they were not counted as correct classifications. Therefore, an ambiguous or mixed result reduced the overall accuracy of a detector, because it did not provide a clear decision. For AI precision, AI recall, and F1-score, the positive class was defined as text classified explicitly as AI-generated.

To estimate the uncertainty of the main accuracy values, 95% confidence intervals were calculated using the Wilson method. The intervals were 50.7%–66.2% for GPTZero, 56.1%–71.2% for ZeroGPT, 35.0%–50.7% for Copyleaks, and 36.3%–52.0% for Scribbr. Since F1-score is derived from precision and recall rather than from a single proportion, its uncertainty was not separately estimated in this study. The obtained confidence intervals indicate that the results should be interpreted as estimates affected by the limited sample size.

ZeroGPT achieved the highest overall accuracy, correctly classifying 96 out of 150 samples. The tool's accuracy was 64.0%, and its F1-score reached 77.7%. This means that, among the detectors analysed, ZeroGPT best combined classification accuracy with the detection of AI-generated text. At the same time, this tool returned the highest number of uncertain results.

The second-best tool in terms of overall accuracy was GPTZero. This detector correctly classified 88 out of 150 samples, which corresponds to an accuracy rate of 58.7%. GPTZero achieved very high precision in detecting AI, but its sensitivity was lower than that of ZeroGPT. This means that the tool rarely misclassified human-written text as AI, but did not detect all text generated or processed by artificial intelligence.

Copyleaks and Scribbr achieved lower scores. Copyleaks correctly classified 64 out of 150 samples, achieving an accuracy of 42.7%, whilst Scribbr correctly classified 66 samples, resulting in an accuracy of 44.0%. Both tools achieved high AI detection precision, but their sensitivity was significantly lower than that of GPTZero and ZeroGPT. In practice, this means that Copyleaks and Scribbr rarely misclassified human-written texts as AI-generated, but often failed to recognise texts generated or paraphrased using artificial intelligence.

The comparison of classification result distributions showed statistically significant differences between the analysed tools, $\chi^2(6) = 68.15$, $p < 0.001$. This indicates that the classification outcomes depended on the detector used. Therefore, hypothesis H2, which assumed that the effectiveness of individual tools differs significantly, was supported by the statistical analysis.

Pairwise exact McNemar tests were additionally performed to compare detectors on the same set of samples. The tests showed no statistically significant difference between GPTZero and ZeroGPT ($p = 0.268$) or between Copyleaks and Scribbr ($p = 0.839$). Statistically significant differences were found between GPTZero and Copyleaks ($p < 0.001$), GPTZero and Scribbr ($p < 0.001$), ZeroGPT and Copyleaks ($p < 0.001$),

and ZeroGPT and Scribbr ($p < 0.001$). These results indicate that GPTZero and ZeroGPT formed the stronger group of detectors, while Copyleaks and Scribbr achieved similarly lower effectiveness.

4.5. Influence of language on detection results

The overall results do not indicate a clear advantage for any single language. For the Polish-language samples, 158 correct classifications were obtained out of 300 results, corresponding to an accuracy of 52.7%. For the English-language samples, 156 correct classifications were obtained out of 300 results, i.e. an accuracy of 52.0%. The difference between the overall results for both languages was therefore small. The results for the individual detectors are shown in Figure 5.

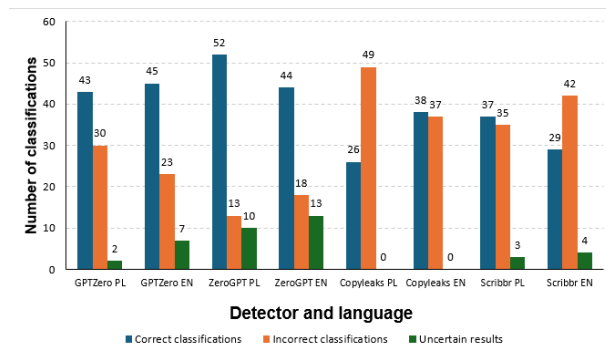


Figure 5: Classification results by detector and language

More significant differences only became apparent when looking at individual tools. GPTZero achieved very similar results for both languages, with an accuracy of 57.3% for Polish and 60.0% for English. ZeroGPT performed better with Polish texts, achieving an accuracy of 69.3%, whilst for English texts it achieved 58.7%. A similar pattern was observed with Scribbr, which achieved 49.3% for Polish and 38.7% for English.

Copyleaks produced a different result. This tool achieved a significantly higher accuracy rate for English texts than for Polish ones. For the Polish language, Copyleaks correctly classified 26 out of 75 samples, which corresponds to an accuracy rate of 34.7%. For English, it correctly classified 38 out of 75 samples, achieving an accuracy of 50.7%. Figure 6 shows a comparison of the accuracy of the detectors tested depending on the language of the analysed samples.

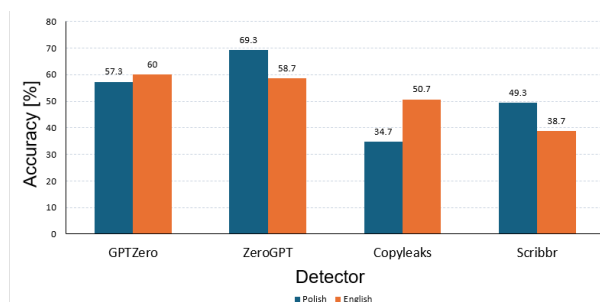


Figure 6: Accuracy of detectors depending on language

The better performance of ZeroGPT and Scribbr on Polish samples may have resulted from several factors. First, AI-generated or AI-paraphrased Polish texts may have contained more regular lexical or syntactic patterns than the corresponding English samples, making them easier for some detectors to identify. Second, the human-written control texts were obtained from sources published before 2017, which reduced the risk of AI contamination but may also have introduced stylistic differences compared with contemporary texts. Third, each detector may rely on different linguistic and statistical features, so the influence of language may vary depending on the tool. Therefore, the results do not support a simple conclusion that detection is generally weaker for Polish. Instead, they suggest that language effects depend on the detector and on the specific characteristics of the analysed samples.

4.6. Influence of the AI model on detection results

Figure 7 shows the detection results broken down by the AI model used to generate the samples and the type of text. The figure shows how many classification results were identified by the detectors as AI text, human text, or an uncertain result.

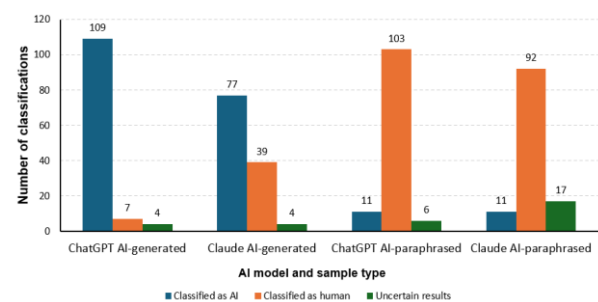


Figure 7: Detection results by AI model and sample type

Texts generated entirely by ChatGPT achieved the highest detection rate. Of the 120 classification results, 109 were identified as AI-generated texts, 7 as human-written texts, and 4 as uncertain results.

A lower detection rate was obtained for texts generated entirely by Claude. In this group, 77 results were classified as AI texts, 39 as human texts, and 4 as uncertain results. The detection rate for texts generated by Claude was 64.2%, meaning that they were less frequently identified by the detectors as machine-generated texts.

In the case of paraphrased texts, the results were significantly lower for both models. For samples paraphrased by both ChatGPT and Claude, the detectability rate was 9.2%. The difference was mainly evident in the number of uncertain results. For texts paraphrased by ChatGPT, 6 uncertain results were recorded, whereas for texts paraphrased by Claude, this figure stood at 17. Table 2 shows the detectability of texts depending on the AI model used to prepare the samples and the type of text.

Table 2: Detectability by AI model and sample type

AI model	Text type	Detectability
ChatGPT	AI-generated text	90.8%
ChatGPT	AI-paraphrased text	9.2%
Claude	AI-generated text	64.2%
Claude	AI-paraphrased text	9.2%

The statistical analysis showed a significant relationship between the AI model used to prepare the samples and the classification result, $\chi^2(2) = 10.66$, $p = 0.005$. This means that the detectors did not classify texts prepared by ChatGPT and Claude in the same way, and the difference was particularly visible for fully AI-generated texts.

5. Conclusions

The study found that the effectiveness of the tools analysed for detecting AI-generated text varied and depended on the type of sample analysed. Texts generated entirely by AI models proved to be the easiest to detect, whilst texts paraphrased using artificial intelligence posed the greatest difficulty for the detectors.

Hypothesis H1 posited that the detectors would perform better at identifying texts generated entirely by AI models than texts written by humans and subsequently paraphrased using artificial intelligence. This hypothesis was confirmed. Texts generated entirely by ChatGPT and Claude were recognised as AI-generated texts significantly more often than paraphrased samples.

Hypothesis H2 concerned differences in effectiveness between the individual tools. The results also confirmed this hypothesis. ZeroGPT achieved the highest overall accuracy, at 64.0%, whilst GPTZero achieved 58.7%. Scribbr and Copyleaks achieved lower results, with accuracy rates of 44.0% and 42.7% respectively. Differences between the tools were also evident in the F1-score and the number of uncertain results. Statistical analysis revealed significant differences between the tools, confirming that their effectiveness was not uniform. Pairwise McNemar tests showed that the differences were significant mainly between the stronger detectors, GPTZero and ZeroGPT, and the weaker detectors, Copyleaks and Scribbr. No significant difference was found between GPTZero and ZeroGPT or between Copyleaks and Scribbr.

Hypothesis H3 posited that detection results depend on the language of the text. The results confirmed this hypothesis only partially. The overall accuracy for Polish and English was very similar, but differences emerged at the level of individual tools. GPTZero achieved similar results for both languages, ZeroGPT and Scribbr performed better on Polish texts, whilst Copyleaks classified English samples more accurately. The overall statistical analysis did not reveal any significant differences between the languages. Therefore, it is not possible to identify a clear advantage of one language over the other.

Another key finding of the study was the comparison of the detection of texts generated by different AI models. Texts generated entirely by ChatGPT were more

frequently identified as artificially generated than those produced by Claude. In the case of paraphrased texts, detection rates were low for both models, standing at 9.2%. This suggests that the model used to prepare the sample may influence the detection result, particularly in the case of texts generated from scratch.

The results obtained suggest that AI detectors can be useful as tools to assist with text analysis, but should not be regarded as a standalone and definitive basis for determining authorship.

A limitation of the study was the number of tools analysed and the use of their free versions. The results may also have depended on the current versions of the detectors, as web-based tools are developed and updated by their manufacturers. The study covered samples of similar length and a specific structure; therefore, the results cannot automatically be generalised to all types of text.

Another limitation was the relatively small size of the dataset. Although 150 samples were sufficient to conduct a controlled comparison of the selected detectors, the number of samples in individual subgroups was limited. This may reduce the statistical power of comparisons between tools, languages, and sample types. Therefore, the numerical results should be interpreted as estimates obtained under the adopted experimental conditions rather than as definitive values for all possible use cases. Future studies should include at least several hundred samples per category to increase the reliability of the comparisons.

The use of human-written texts published before 2017 reduced the risk that the control material contained AI-generated content. However, it may also have introduced a systematic stylistic difference between the human-written samples and contemporary texts. Older texts may differ in vocabulary, sentence structure, and writing conventions, which could have influenced detector decisions. Future research should therefore include verified contemporary human-written texts as an additional control group.

Further research could involve a larger number of detectors, a larger sample set and additional languages. It would also be worthwhile to analyse texts of varying lengths, specialist texts and samples that have undergone varying degrees of human editing.

References

- [1] P. Kumar, Large language models (LLMs): survey, technical frameworks, and future challenges, *Artificial Intelligence Review* 57 (2024) 260, <https://doi.org/10.1007/s10462-024-10888-y>.
- [2] J. Wu, S. Yang, R. Zhan, Y. Yuan, L. S. Chao, D. F. Wong, A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions, *Computational Linguistics* 51(1) (2025) 275–338, https://doi.org/10.1162/coli_a_00549.
- [3] E. N. Crothers, N. Japkowicz, H. L. Viktor, Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods, *IEEE Access* 11 (2023) 70977–71002, <https://doi.org/10.1109/ACCESS.2023.3294090>.

- [4] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, L. Waddington, Testing of detection tools for AI-generated text, *International Journal for Educational Integrity* 19 (2023) 26, <https://doi.org/10.1007/s40979-023-00146-z>.
- [5] M. Perkins, J. Roe, B. H. Vu, D. Postma, D. Hickerson, J. McGaughran, H. Q. Khuat, Simple techniques to bypass GenAI text detectors: implications for inclusive education, *International Journal of Educational Technology in Higher Education* 21 (2024) 53, <https://doi.org/10.1186/s41239-024-00487-w>.
- [6] A. Kartelj, M. Mladenović, S. Vujičić Stanković, Comparison of algorithms for the recognition of ChatGPT paraphrased texts, *Journal of Big Data* 12 (2025) 28, <https://doi.org/10.1186/s40537-025-01082-0>.
- [7] K. Schaaff, T. Schlippe, L. Mindner, Classification of human- and AI-generated texts for different languages and domains, *International Journal of Speech Technology* 27 (2024) 935–956, <https://doi.org/10.1007/s10772-024-10143-3>.
- [8] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Zou, GPT detectors are biased against non-native English writers, *Patterns* 4(7) (2023) 100779, <https://doi.org/10.1016/j.patter.2023.100779>.
- [9] GPTZero, Frequently Asked Questions, <https://gptzero.me/faq>, [19.05.2026].
- [10] ZeroGPT, Frequently Asked Questions, <https://www.zerogpt.com/faq>, [19.05.2026].
- [11] Copyleaks, AI Content Detector, <https://copyleaks.com/ai-content-detector>, [25.05.2026].
- [12] Scribbr, AI Detector, <https://www.scribbr.com/ai-detector>, [27.05.2026].
- [13] OpenAI, What is ChatGPT? FAQ, OpenAI Help Center, <https://help.openai.com/en/articles/12677804-what-is-chatgpt-faq>, [27.05.2026].
- [14] OpenAI, ChatGPT release notes, OpenAI Help Center, <https://help.openai.com/pl-pl/articles/6825453-chatgpt-informacje-o-wydaniach>, [27.05.2026].
- [15] Anthropic, Introducing Claude Sonnet 4.6, <https://www.anthropic.com/news/claude-sonnet-4-6>, [27.05.2026].