Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

The Readability Paradox: Can We Trust Decisions on AI Detectors?

Henry Sanmi Makinde, Akindeji Ibrahim Makinde, Mutiyat Adeola Usman, Hope Adegoke, Baraka Abiodun Makinde-Isola, Wasiu Lawal, and Ibraheem Temitope Jimoh

Educational Research Methodology Department, University of North Carolina Greensboro, US
Department of Information Systems, Federal University of Technology, Akure, Ondo State, Nigeria
Data Science and Analytics Department, North Carolina A&T State University
Educational Research Methodology Department, University of North Carolina Greensboro
Metallurgical and Material Engineering Department, Federal University of Technology, Akure, Nigeria.
Department of Information and Communication Technology (ICT), Federal University of Technology, Akure, Ondo State, Nigeria
Department of Software Engineering, Federal University of Technology, Akure, Ondo State, Nigeria
Akindeji Ibrahim Makinde/iamakinde@futa.edu.ng

Abstract. Artificial Intelligence (AI) detectors are increasingly used in various domains, including healthcare, finance, criminal justice, and more, to make critical decisions. These systems are designed to identify patterns, anomalies, or specific features within data to assist or automate decision-making processes. However, the trustworthiness of AI detectors is a growing concern, particularly as these systems can exhibit biases, errors, and lack of transparency. This study evaluates the effectiveness of AI-detection tools and analyzes the linguistic characteristics distinguishing AI-generated and human-written academic texts across six disciplines. A total of 200 research papers were examined, 100 generated using large language and 100 peer-reviewed human-authored articles. Five AI detection tools were assessed for accuracy in classification. The study also conducted a comparative readability analysis using five established indices which include Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, and total Word Count. Results indicate that while human-written papers were correctly identified by most tools with over 80% accuracy, AI-generated papers were frequently misclassified, especially after paraphrasing. Also from the result, the AI-generated texts were significantly shorter and exhibited higher syntactic complexity, with lower readability scores across all indices and disciplines. These findings underscore limitations in current detection tools and highlight notable stylistic differences in how AI and humans generate academic content, with implications for academic integrity policies and future AI writing systems.

Keywords. Large Language Models (LLMs); AI detection tools; AI-generated text; Linguistic complexity; Readability metrics

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

1. Introduction

Artificial Intelligence (AI) is a field of computer science that has evolved drastically over the years and have been widely used by several sectors including education, finance, healthcare and manufacturing. AI has significantly impacted individuals, organizations and societies as it offers systematic capabilities of reasoning based on inputs and learning through the differences of expected outcomes when it predicts and adapts to changes in its ecosystems and stimuli [1]. The need to detect AI-generated content grows as AI too grows. In education, relying on AI for assignments can undermine the learning process as this raises concerns about the accuracy, ethics, and academic rigor of such work. Several concerns related to the difficulty in differentiating human versus AI authorship within academic and education communities has renewed the debate on the role of traditional human endeavours [1, 2]. AI models from different sources have often produced plausible but inaccurate or misleading content, raising doubts about their reliability and the risk of spreading false information [3].

To address the issues mentioned above, researchers have developed AI detection tools to protect academic and professional integrity. AI detectors have helped to identify patterns or anomalies and determine if content such as text, images, video, audio, or code were created by AI. In the aspect of text, AI detectors assess grammar, word choice, and language patterns to detect AI-generated material [4]. Studies by [5] and [6] shows that some AI detectors were ineffective in identifying paraphrased texts. Some of the detectors may misclassify human-written articles, which can undermine the credibility of academic publications [3,7].

Based on several usage, chatbots have used Natural Language Processing (NLP) to answer user queries by mapping them to the best response sets in the system. Many chatbots have also incorporated language models and deep learning to give customers real-time feedback [1,8]. The language models use generative and discriminative techniques to predict the likelihood of a word sequence that would be produced in a normal human conversation [1,9,10].

One of the recent generative AI model lauched is the DeepSeek. DeepSeek make use of LLMs based on transformer architectures to generate responses for users [11]. Also, OpenAI's ChatGPT is a well-liked and responsive chatbot that has been trained on hundreds of billions of parameters due to its impressive skills in digital health and medicine [3], pure sciences [12], social sciences [13], education, business, and customer support [14]. Still on the recent advancement in generative AI Model, Google DeepMind created the very sophisticated multimodal AI model Gemini, which includes several LLMs and NLP technologies [15]. DeepMind is also designed to handle and process multiple types of data thereby making it a versatile tool for a wide range of applications [16,17]. Another example of an AI-Powered assistant is the Microsoft Copilot developed by Microsoft to enhance productivity and streamline workflows across various applications. Microsoft Copilot has been embedded in their tools thereby making it a versatile companion for professionals, students, and individuals alike. Meta AI is also another widely used AI chat tool that focused on advancing AI technologies in order to improve user experiences across Meta's platforms [18]. These numerous AI tools offer several advantages however due to some of their limitations like providing misleading information, and issues in the education sector and even professional and creative fields, the need for AI-detectors tools arises to ensure authenticity and prevent misuse of AI-generated content.

Several AI detectors such as ZeroGPT, Quillbot’s AI Detector, GPTZero, Turnitin’s AI Detector, and Copyleaks are widely used to identify content generated by AI, including large language models like ChatGPT and GPT-4. The rise of these tools highlights the growing need to distinguish human-written from AI-generated content in fields like education and content marketing. However, no study has yet fully assessed how well these detectors can tell the difference. This study aims to evaluate the effectiveness of several recent AI content detectors in distinguishing human and AI-generated text.

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

[3] compared the accuracy of mainstream AI content detectors and human reviewers in detecting AI-generated rehabilitation-related articles with or without paraphrasing. They collected 50 rehabilitation articles from four peer-reviewed journals and used ChatGPT to generate 50 similar articles. The AI-generated texts were then rephrased using Wordtune. Six AI detectors Originality.ai, Turnitin, ZeroGPT, GPTZero, Content at Scale, and GPT-2 Output Detector were used to identify AI content in the original, ChatGPT-generated, and AI-rephrased articles. Four human reviewers (two professionals and two students) also tried to distinguish between original and AI-rephrased texts. Originality.ai detected 100% of both ChatGPT-generated and AI-rephrased texts. ZeroGPT detected 96% of ChatGPT-generated and 88% of AI-rephrased articles. Turnitin had a 0% false positive rate on human texts but detected only 30% of AI-rephrased articles. Professors correctly identified at least 96% of AI-rephrased articles but misclassified 12% of human-written texts as AI-generated. Students identified only 76% of AI-rephrased articles. [19] also tested detection tools on natural language and programming code, and found that "detecting ChatGPT-generated code is harder than detecting natural language". They noted biases in tools: some tend to over-predict AI authorship (false positives), while others lean toward labeling content as human-written (false negatives).

This paper explores the accuracy and reliability of AI detectors, the discrepancies in detection rates across different AI detectors and the implications of relying on AI detectors for decision-making in academia and other fields.

2. Background

Generative Artificial Intelligence (GAI) encompasses a wide range of advanced computational technologies such as machine learning, neural networks, natural language processing, data mining, and algorithmic systems. Although there have been conversations about AI in education for many years, the emphasis on how these technologies may help students throughout their academic careers has just recently increased [20].

In higher education, AI tools are widely used in three key areas such as adaptive personal tutors, support for group learning, and virtual learning spaces [21]. Intelligent tutoring systems deliver personalized content at scale and give direct feedback, though they often need human guidance. Studies show that chatbots powered by generative AI improve student interest, drive, and results [22]. When students are asked about it, their responses vary, while some students welcome these tools, others raise concerns about data privacy [23]. Despite the concerns, AI tools offer fresh ways to redesign assessments, support reflective writing, and tailor teaching to student needs.

ChatGPT have been known to provides users with a multitude of information, facilitates study assignments, and responds quickly. Many instructors use ChatGPT to handle tedious tasks, give timely feedback, and use data to create lesson plans which allows them more time to focus on creative teaching and student support. Aside privacy concern raised by students, there are additional hazards associated with ChatGPT, which include misuse, decreased human interaction, and concerns to academic integrity [23]. Experts concur that educational institutions need to develop explicit AI policies that uphold morality and encourage innovative use of generative technologies. Teachers are encouraged to view these technologies as forces for change in instruction and assessment rather than as something to be feared [23]. Several research has shown that information produced by AI may contain prejudice, mistakes, or work that is not up to academic standards [24]. Teachers are therefore urged to create tests that emphasize creativity and critical thinking while avoiding the abuse of AI [24].
Some of the available AI-detection tools include

- ZeroGPT – is a free, web-based AI content detection tool designed to identify whether a piece of text was generated by AI models.
- GPTZero – is designed for educational settings as it provides comprehensive sentence-level analyses of AI-likeness and detects AI-generated work using linguistic traits like word predictability and sentence complexity.

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

- Quillbot - Quillbot is best known as a paraphrasing and grammar refinement tool. It is not in the same category as ZeroGPT and GPTZero, but it plays a complex role in the educational context that can both support learning and obscure authorship origins by rewording AI-generated text. This has implications for academic integrity, especially when used to evade detection systems.
- Turnitin - Turnitin has integrated AI-detection capabilities alongside its traditional plagiarism detection services. Turnitin compares submissions with both internet sources and known AI writing patterns, offering indicators of likely AI authorship.
- Copyleaks - Copyleaks supports multiple languages and offers detailed analytics such as token-level heatmaps and content classification and uses deep learning and NLP techniques to detect both plagiarism and AI-generated content.

2.1 AI and Academic Integrity

As AI continues to redefine the landscape of education, academic institutions must balance innovation with integrity. The rapid development of generative tools has opened new opportunities for creativity, personalization, and efficiency in both teaching and learning. However, it also necessitates the responsible use of detection tools, transparent policies on AI usage, and the fostering of essential skills such as critical thinking, analysis, and ethical reasoning [24]. AI-content detection tools such as ZeroGPT, GPTZero, Quillbot, Turnitin, and Copyleaks offer important safeguards, but they are not foolproof. They should be viewed as part of a larger ecosystem that includes faculty training, ethical awareness, and the redesign of assessments to encourage originality and deeper learning.

2.2 Tracking AI-Generated Plagiarism Using Detection Tools

As AI chatbots and large language models (LLMs) become more common, AI-detection tools are now widely used in academic and professional settings. These tools use algorithms and machine learning techniques to analyze written content and flag traits linked to AI-generated text. Many compare submissions against large data sets to spot patterns that suggest non-human authorship. Popular tools include Turnitin, ZeroGPT, GPTZero, Copyleaks, Writer AI, and Winston AI. Each uses advanced methods to separate human and machine writing. However, their performance remains uneven. Research shows some tools often misclassify AI-generated text as human-written, raising doubts about their accuracy. Others have noted discrepancies in detection results across various tools, particularly when analyzing content produced by different LLM versions [25].

Several evaluations have tested a wide range of AI-detection tools. These include assessments of tools such as Copyleaks, Writer AI, Content@Scale, GPTZero, GPTKit, and others. Reported accuracy rates vary considerably, and many tools have been criticized for their inability to consistently distinguish between human- and AI-authored texts. Larger comparative studies have also confirmed these inconsistencies, although some tools like Copyleaks, Originality.AI, and Turnitin have demonstrated relatively stronger performance [25]. Some researchers have investigated how well these tools can withstand adversarial strategies designed to disguise AI-generated content. Example include paraphrasing AI-generated text using automated tools which can dramatically reduce detection rates. In more extensive tests involving content from LLMs such as Bard, Claude 2, and GPT-4, researchers applied techniques including intentional spelling errors, increased lexical variability, paraphrasing, altering sentence complexity, and mimicking non-native English writing. These techniques had varying levels of success in bypassing detection, with paraphrasing and burstiness often significantly reducing detection accuracy. Among the tools evaluated, Copyleaks and Turnitin were generally the most resilient, while others performed less reliably.

Taken as a whole, existing research shows that current AI-detection tools face considerable challenges. Many fail to consistently identify AI-generated text and exhibit high rates of false positives or false negatives. As both detection tools and generative AI systems rapidly evolve, ongoing research is crucial to assess and check the accuracy and reliability of AI detectors.

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

3. Methodology

3.1. Data Collection

The dataset used in this study comprises two distinct corpora: AI-generated and human-written academic texts. The AI-generated corpus consists of 100 academic-style research papers created using large language models such as ChatGPT, Deepseek, and Kimi.ai. These texts span six major academic disciplines Engineering, Biology, Computer Science, Physics, Social Sciences, and Medicine and were designed to reflect a diversity of writing styles and research topics. In parallel, the human-written corpus includes 100 peer-reviewed research papers authored by scholars and retrieved from reputable academic databases, including Scopus and Web of Science. These human-authored papers were selected to match the AI-generated texts by both academic field and topical relevance, ensuring a balanced and comparable dataset for analysis.

3.2. *AI-Detection Tools and Evaluation Process*

Out of the 200 papers, only 186 research papers were input into 5 AI detectors (ZeroGPT, Quillbot's AI Detector, GPTZero, Turnitin's AI Detector and Copyleaks detector) and the detection results were recorded as ("AI-written" or "human-written") for each paper. The percentage of correctly and incorrectly classified papers were calculated for each of the detectors.

3.3. *Data Preprocessing*

All papers were subjected to a uniform preprocessing pipeline prior to analysis which involved removing metadata such as titles, author information, and references to avoid bias. The cleaning steps included sentence segmentation, removal of tables, equations, and figures, and standardization to ASCII encoding to ensure preprocessing steps made the analysis focus solely on the main textual content of each document.

3.4. *Readability and Complexity Metrics*

To assess readability and complexity, five established readability metrics were applied to each paper. The metrics included the Flesch Reading Ease Score, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index (Simple Measure of Gobbledygook), and total Word Count. The indices have wide acceptance in academic readability research because of their ability to provide a comprehensive insight of both surface-level readability and deeper syntactic complexity.

3.5. *Statistical and Visual Analysis*

Statistical analysis was performed to uncover patterns and test for significant differences between the two groups. Descriptive statistics, including mean and standard deviation, were computed for each metric across both AI and human datasets. The visual interpretation, boxplots and bar charts were generated using Python libraries such as matplotlib and seaborn. For inferential analysis, independent samples t-tests were conducted to compare AI and human-generated texts on individual readability indices, while one-way ANOVA tests were used for comparisons across multiple fields. All statistical tests were evaluated at a significance level of $p < .05$.

3.6. *Field Classification*

Each paper was manually categorized into its respective discipline based on thematic content to ensure consistent field-specific comparisons. This classification facilitated targeted analysis and ensured that intra-field comparisons reflected authentic academic subdomain characteristics.

4. Result and Discussion

The Results and Discussion section presents a comparative analysis of AI-generated and human-written academic papers, highlighting detection tool performance and significant differences in readability and linguistic complexity across disciplines.

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

Figure 1 shows the distribution of fields in the data used. Tables 1 and 2 suggest that current AI-detection tools still struggle to identify AI-generated text, while they recognize human-written text with over 80% accuracy. Many studies report these tools detect AI-generated content with only about 50% accuracy or slightly higher. They especially have trouble when ChatGPT rephrases human-written text or mimics a specific style.

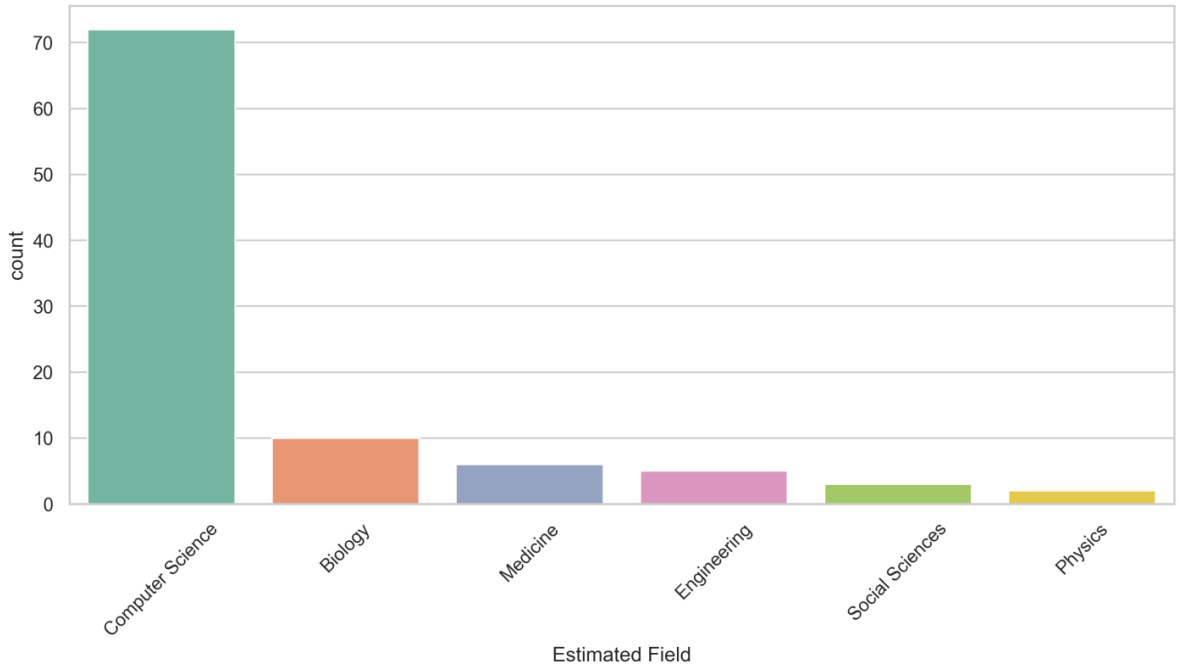

Figure 1: Estimated fields distributions for the papers

Detection tools perform much worse when texts are paraphrased or rewritten. Existing tools still struggle to identify AI-generated text, and detecting ChatGPT-generated code is even harder (Weber-Wulff et al., 2023).

Table 1. Human-written sample detection

Type	ZeroGPT	Quillbot	GPTZero	Turnitin	Copyleaks
Human-written Paper 1	3.42	0	1	0	0
Human-written Paper 2	1.21	0	6	0	0
Human-written Paper 3				0	
Human-written Paper 4	4.29	0	7	0	
Human-written Paper 5	2.34	0	2	0	
Human-written Paper 6	1.02	0	6	0	
Human-written Paper 7	3.06	0	9	0	
Human-written Paper 8	4.05	0	9	0	
Human-written Paper 9				0	0
Human-written Paper 10	2.51	0	3	0	
Human-written Paper 11	77.32	76	56	86	90
Human-written Paper 12	2.11	0	5	0	
Human-written Paper 13	0.53	0	3	0	
Human-written Paper 14	4.14	0	10	0	0
Human-written Paper 15	3.20	0	4	0	
Human-written Paper 16	0.39	0	7	0	

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

Human-written Paper 17	1.39	15	2	0	
Human-written Paper 18	2.94	0	5	0	15
Human-written Paper 19	4.60	0	2	0	0
Human-written Paper 20	3.71	0	5	0	

Table 2. AI-written detection sample results

Type	**ZeroGPT**	**Quillbot**	**GPTZero**	**Turnitin**	**Copy leaks**
AI-written Paper 1	90.6	0	93%	100	100
AI-written Paper 2	91.75	89	81%	27	100
AI-written Paper 3	98.07	0	95%	100	100
AI-written Paper 4	74.54	93	98%	75	100
AI-written Paper 5	95.84	88	79%	100	99
AI-written Paper 6	93.95	96	88%	100	100
AI-written Paper 7	95.29	0	78%	100	99
AI-written Paper 8	97.34	86	86%	100	100
AI-written Paper 9	98.13	89	99%	100	100
AI-written Paper 10	86.79	91	77%	100	99
AI-written Paper 11	93.43	96	78%	100	99
AI-written Paper 12	82.82	89	80%	100	100
AI-written Paper 13	95.82	100	97%	100	99
AI-written Paper 14	87.54	94	95%	100	100
AI-written Paper 15	80.68	0	86%	100	100
AI-written Paper 16	91.86	82	87%	100	100
AI-written Paper 17	98.55	89	98%	100	100
AI-written Paper 18	79.67	83	94%	60	99
AI-written Paper 19	81.74	0	99%	100	100

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

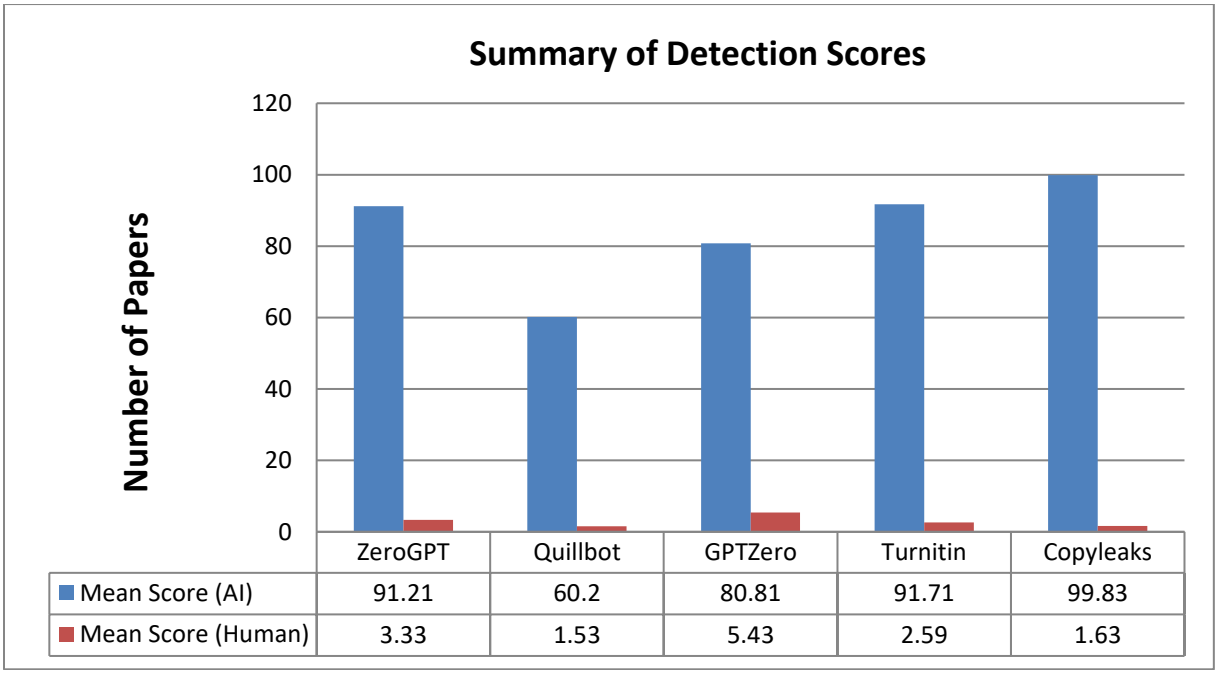

Figure 2: Summary of detection scores

4.1. Overall and Field-wise Readability Comparison between AI and Human Papers

The results presented in the table 3 and figure 3 show a clear distinction between AI-generated and human-written academic papers in terms of readability metrics and structural complexity. On average, human-written papers were substantially longer, with a mean word count of 11,011.14 compared to 1,134.96 for AI-generated texts. The results show that human authors tend to engage more deeply with their subjects, possibly offering greater detail, contextual analysis, and comprehensive literature reviews than AI systems currently do. In terms of readability, human-written papers had significantly higher Flesch Reading Ease scores (mean of 34.90) compared to AI papers (mean of 7.71), indicating that human-written texts are generally easier to read. A low score on this scale for AI-generated texts suggests more complex sentence structures and potentially less accessibility for general audiences. This is further supported by higher average values in all difficulty-oriented indices for AI-generated papers: the Gunning Fog Index (19.98 vs. 15.90), SMOG Index (16.63 vs. 14.57), and Flesch-Kincaid Grade (16.24 vs. 12.67). These findings collectively highlight that AI-generated papers often contain language that is more convoluted or artificially verbose.

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

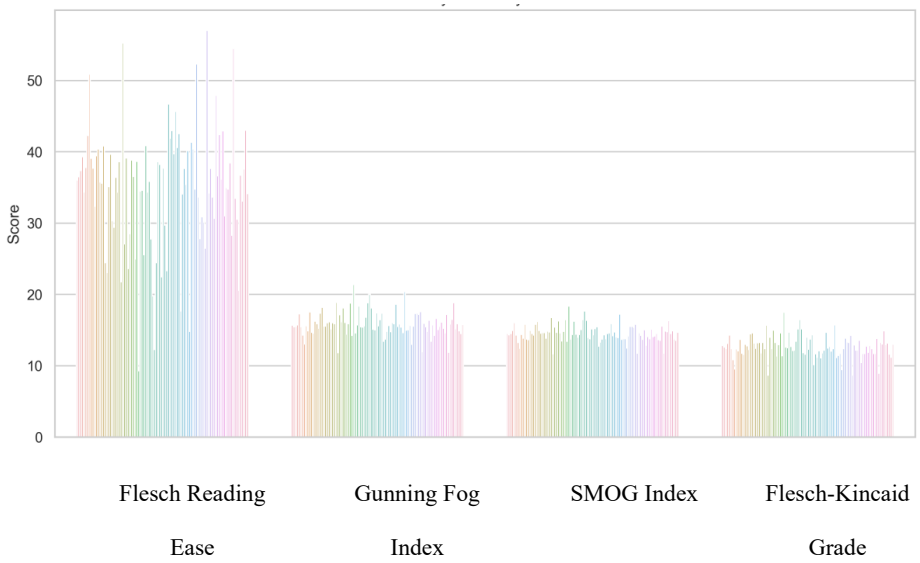

Figure 3: Document Readability Metrics

Table 3. Readability metrics between AI-generated and human-written papers

		AI Paper	Human Paper
Word Count	mean	1,134.96	11,011.14
	std	161.77	6,159.49
Flesch Reading Ease	mean	7.71	34.90
	std	8.95	8.58
Gunning Fog Index	mean	19.98	15.90
	std	1.62	1.67
SMOG Index	mean	16.63	14.57
	std	1.06	1.17
Flesch-Kincaid Grade	mean	16.24	12.67
	std	1.32	1.54

The Field-specific comparisons as shown in table 4 reinforce these trends. Across all disciplines, AI-generated papers scored lower on the Flesch Reading Ease scale, with particularly low readability in Physics (2.68), Social Sciences (5.21), and Computer Science (7.06). Correspondingly, the AI-generated papers also had elevated complexity indices, such as Gunning Fog scores around or above 20 in most fields.

Table 4: Field-Specific Summary

Paper type	**AI Paper**				
	Flesch Reading Ease	Flesch-Kincaid Grade	Gunning Fog Index	SMOG Index	Word Count
Estimated Field					
Biology	12.51	15.79	19.87	16.75	1106.25
Computer Sci.	7.06	16.32	20.01	16.67	1108.98
Engineering	9.62	15.39	18.97	15.55	1139.80
Medicine	14.22	15.21	18.67	15.86	1174.40
Physics	2.68	17.07	21.90	17.65	1244.00
Social Sciences	5.21	16.65	20.17	16.73	1203.65
Paper type	**Human Paper**				
Estimated Field	Flesch Reading Ease	Flesch-Kincaid Grade	Gunning Fog Index	SMOG Index	Word Count
Biology	37.54	11.94	15.20	13.99	10536.10
Computer Sci.	34.52	12.79	16.00	14.65	11440.38
Engineering	36.84	12.57	15.89	14.61	6044.60

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

Medicine	36.59	12.17	15.27	14.11	10070.17
Physics	39.51	11.67	15.23	13.99	15276.00
Social Sciences	25.59	14.27	17.57	15.74	9609.33

4.2. Word Count

The distribution of word counts for the human-written texts, as shown in figure 4, reveals a right-skewed pattern. Most of the academic articles cluster between approximately 5,000 and 15,000 words, with a peak frequency around the 8,000 to 10,000-word range. This indicates that the majority of peer-reviewed papers in the sample tend to conform to typical journal length expectations. The other outliers represent review articles, meta-analyses, or comprehensive theoretical pieces. The overall distribution suggests a heterogeneous corpus with substantial variation in article length, reflecting differences in disciplinary norms, journal formatting requirements, and research complexity.

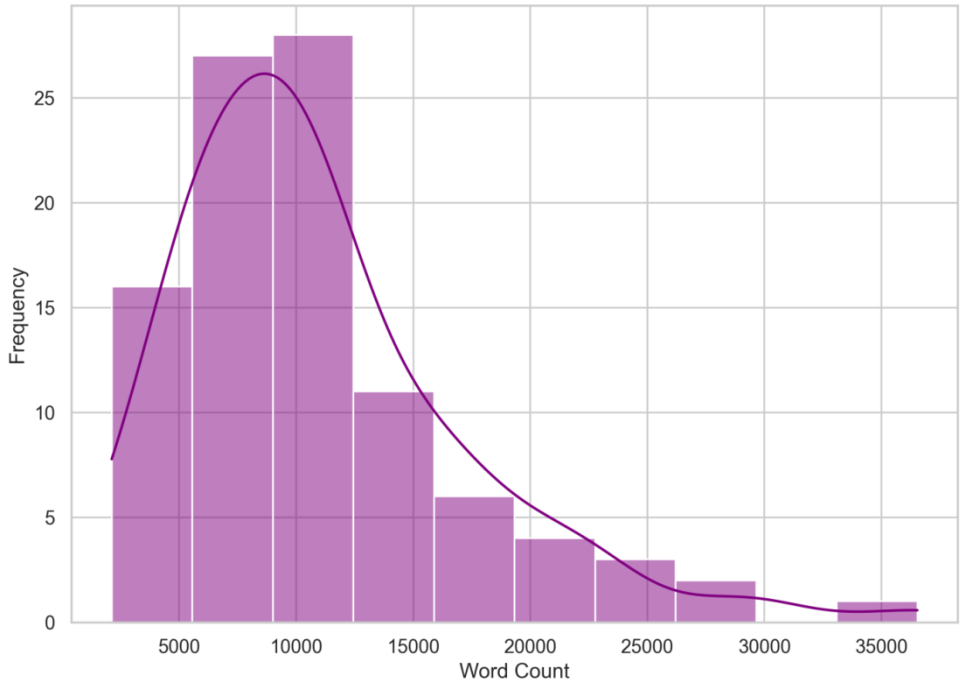

Figure 4: Word Count Distribution

Figure 5 illustrates Word Count by Estimated Field and Paper Type, highlights a substantial and consistent disparity in the length of AI-generated versus human-written academic papers. Across all academic disciplines analyzed, human-written papers exhibit significantly higher word counts than those generated by AI. Human-authored Physics papers have median word counts exceeding 15,000 words, with some extending well above 20,000, while AI-generated texts in the same field mostly shows lower words. The discrepancy suggests that human authors typically produce more detailed discussions, literature reviews, and nuanced argumentation, which AI systems currently struggle to replicate in both depth and volume.

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

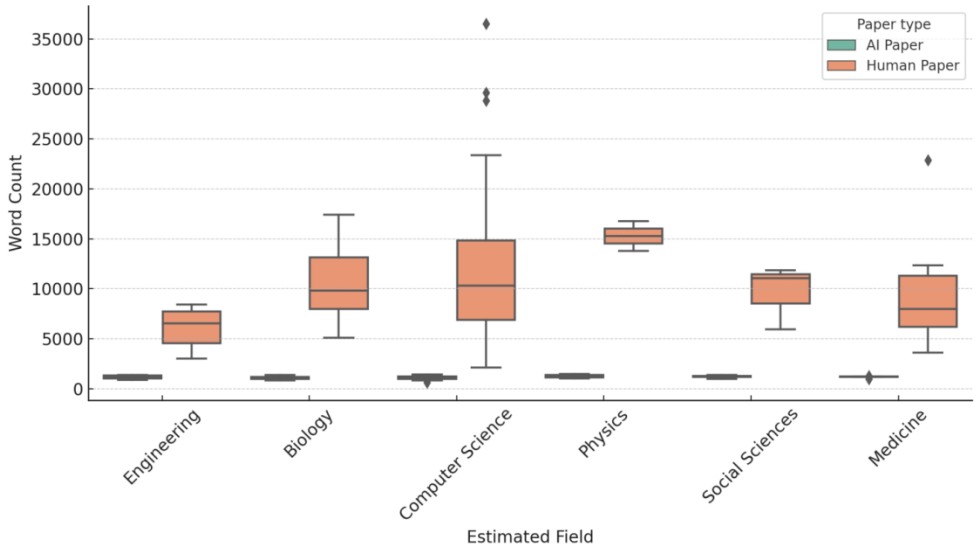

Figure 5: Word count by Estimated field for AI and Human Papers

The boxplot in figure 6 illustrates the Flesch Reading Ease scores across six academic fields, comparing AI-generated and human-written papers. The human-written papers consistently exhibit higher readability scores, indicating they are generally easier to read. These findings suggest that while AI-generated texts may cover academic content effectively, they tend to be more complex or less reader-friendly than their human-authored counterparts.

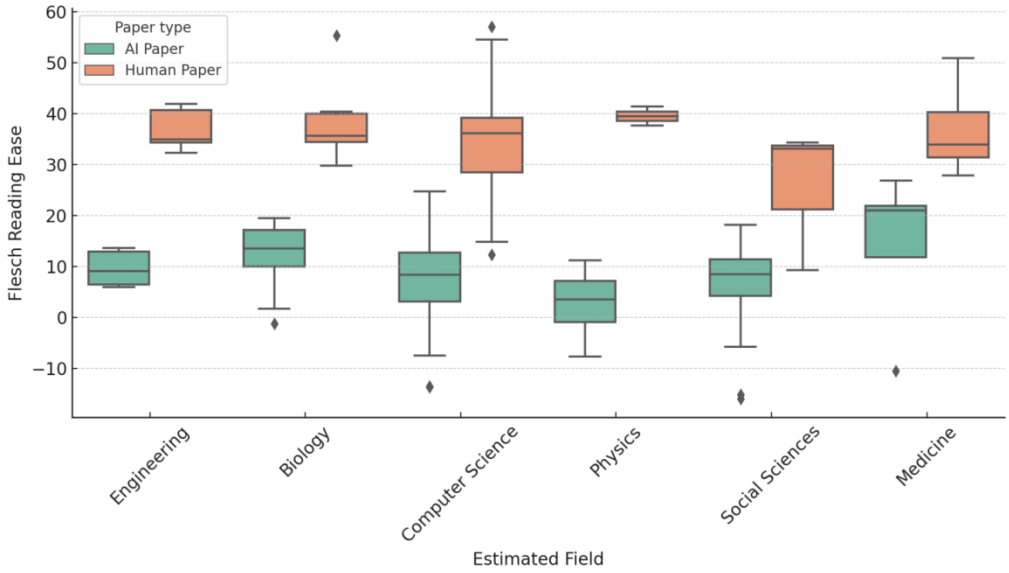

Figure 6: Flesh Reading Ease by Estimated Field and Paper Type

The Gunning Fog Index by Estimated Field and Paper Type boxplot in Figure 7 offers more proof of the linguistic intricacy frequently seen in academic writing produced by AI. Higher scores indicate greater complexity.

Technium
Education and Humanities

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

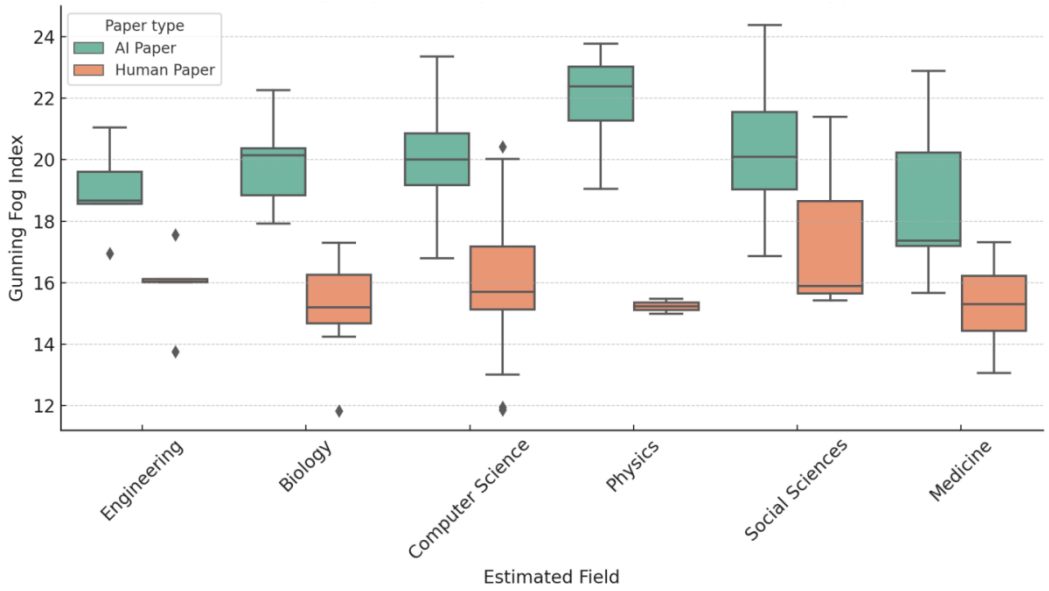

Figure 7: Gunning Fog Index by Estimated Field and Paper Type

Figure 8 shows that AI-generated academic writings are typically more syntactically complex than human-written ones. The result compares SMOG Index scores by field and paper type. In the result, AI-generated articles have higher median SMOG ratings across the majority of fields. The result in physics category shows that the median SMOG score for articles produced by AI is around 18.5, whereas the median score for papers authored by humans is 14. This represents a discrepancy of 4.5 years of necessary schooling and this implies that literature produced by AI in the field can be far more intricate and challenging to understand.

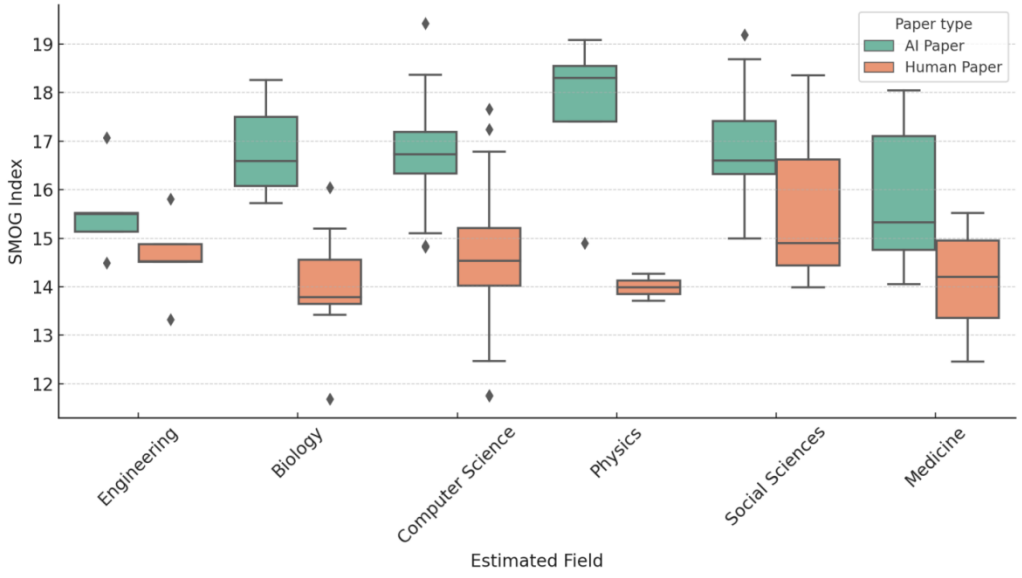

Figure 8: SMOG Index by Estimated Field and Paper Type

The result in figure 9 shows that articles produced by AI have better Flesch-Kincaid Grade Level scores than those authored by humans, indicating more sophisticated vocabulary and sentence structures.

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

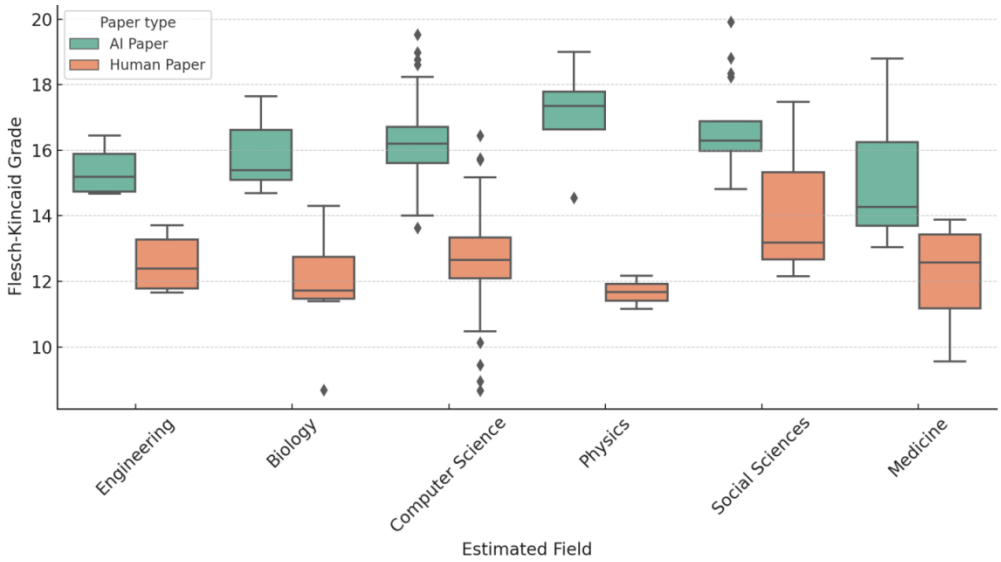

Figure 9: Flesch-Kincaid Grade by Estimated Field and Paper Type

5. Conclusion

The comparative analysis reveals consistent and significant differences between AI-generated and human-written academic papers. The results show that human-written articles are notably longer and easier to read while AI-generated articles display elevated complexity across all readability metrics. This suggest that there are deeper engagement and more accessible language in human-written articles when compare to AI-written articles with verbose or syntactically dense writing. Despite these differences, current AI-detection tools exhibit limited accuracy in reliably identifying AI-generated texts, particularly when such texts are paraphrased or stylized and these findings raise concerns about the robustness of existing detection technologies and emphasize the need for more sophisticated, linguistically aware AI-detection frameworks. As generative AI continues to evolve, both academic institutions and developers must adapt to ensure transparency, accountability, and the preservation of scholarly standards.

References

[1] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M. A. Albashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, S. Chowdhury, T. Crick, S. W. Cunningham, G. H. Davies, R. M. Davison, R. Dé, D. Dennehy, Y. Duan, R. Dubey, R. Dwivedi, J. S. Edwards, C. Flavián, R. Gauld, V. Grover, M.-C. Hu, M. Janssen, P. Jones, I. Junglas, S. Khorana, S. Kraus, K. R. Larsen, P. Latreille, S. Laumer, F. T. Malik, A. Mardani, M. Mariani, S. Mithas, E. Mogaji, J. H. Nord, S. O'Connor, F. Okumus, M. Pagani, N. Pandey, S. Papagiannidis, I. O. Pappas, N. Pathak, J. Pries-Heje, R. Raman, N. P. Rana, S.-V. Rehm, S. Ribeiro-Navarrete, A. Richter, F. Rowe, S. Sarker, B. C. Stahl, M. K. Tiwari, W. van der Aalst, V. Venkatesh, G. Viglia, M. Wade, P. Walton, J. Wirtz, R. Wright: Opinion Paper: "So What if ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *Intern. J. of Information Management*, ISSN 0268-4012, 71, 1–63 (2023).

[2] H. Else: By ChatGPT Fool Scientists. *Nature*, ISSN 0028-0836, 613, 423 (2023).

[3] J. Q. Liu, K. T. Hui, F. Al Zoubi, Z. Z. Zhou, D. Samartzis, C. C. Yu, J. R. Chang, A. Y. Wong: The Great Detectives: Humans Versus AI Detectors in Catching Large Language Model-Generated Medical Writing. *Intern. J. for Educational Integrity*, ISSN 1833-2595, 20 (1), 1–14 (2024).

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

[4] A. Shah, P. Ranka, U. Dedhia, S. Prasad, S. Muni, K. Bhowmick: Detecting and Unmasking AI-Generated Texts Through Explainable Artificial Intelligence Using Stylistic Features. *Intern. J. of Advanced Computer Science and Applications,* ISSN 2158-107X, 14 (10) (2023).

[5] N. Anderson, D. L. Belavy, S. M. Perle, S. Hendricks, L. Hespanhol, E. Verhagen, A. R. Memon: AI Did Not Write This Manuscript, or Did It? Can We Trick the AI Text Detector into Generating Texts? *BMJ Open Sport Exerc. Med*., ISSN 2055-7647, 9 (1), e001568 (2023).

[6] 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. Intern. *J. for Educational Integrity*, ISSN 1833-2595, 19 (1), 1–39 (2023).

[7] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, J. Zou: GPT Detectors Are Biased Against Non-Native English Writers. *Patterns*, ISSN 2666-3899, 4 (7) (2023).

[8] A. K. Kushwaha, A. K. Kar: MarkBot – A Language Model-Driven Chatbot for Interactive Marketing in Post-Modern World. *Information Systems Frontiers*, ISSN 1572-9419, 1–18 (2021).

[9] J. R. Bellegarda: Statistical Language Model Adaptation: Review and Perspectives. *Speech Communication*, ISSN 0167-6393, 42 (1), 93–108 (2004).

[10] A. Vaswani et al.: Attention is All You Need. In: Adv. in Neural Information Processing Systems, 1–15 (2017).

[11] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled: Overview of the Transformer-Based Models for NLP Tasks. *In: 2020 15th Conf. on Computer Science and Information Systems (FedCSIS), IEEE,* 179–183 (2020).

[12] M. L. Tsai, C. W. Ong, C. L. Chen: Exploring the Use of Large Language Models (LLMs) in Chemical Engineering Education: Building Core Course Problem Models with ChatGPT. *Education for Chemical Engineers*, ISSN 1749-7728, 44, 71–95 (2023).

[13] D. Şengür: Using of MATLAB Statistics Toolbox for Data Analysis in Social Sciences with ChatGPT-3 Prompts. *Turkish J. of Science and Technology*, ISSN 1308-9080, 18 (2), 353–361 (2023).

[14] D. Kalla, N. Smith, F. Samaah, S. Kuraku: Study and Analysis of ChatGPT and Its Impact on Different Fields of Study. *Intern. J. of Innovative Science and Research Technology*, ISSN 2456-2165, 8 (3), 827–833 (2023).

[15] M. Farrokhnia, S. K. Banihashem, O. Noroozi, A. Wals: A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research. Innovations in Education and Teaching Intern., ISSN 1470-3297, 61 (3), 460–474 (2024).

[16] M. Imran, N. Almusharraf: Google Gemini as a Next Generation AI Educational Tool: A Review of Emerging Educational Technology. *Smart Learning Environments*, ISSN 2196-7091, 11 (1), 1–8 (2024).

[17] P. Perera, M. Lankathilaka: Preparing to Revolutionize Education with the Multi-Model GenAI Tool Google Gemini? *J. of Advances in Education and Philosophy*, ISSN 2523-2665, 7 (8), 246–253 (2023).

[18] L. Thomas, S. Bhat: An Overview of Facebook's Journey to Meta – A Case Study. Intern. J. of Case Studies in Business, *IT and Education (IJCSBE),* ISSN 2581-6942, 6 (1), 268–287 (2022).

[19] J. Wang, S. Liu, X. Xie, Y. Li: Evaluating AIGC Detectors on Code Content. *arXiv Preprint*, arXiv:2304.05193 (2023).

[20] R. Luckin, W. Holmes, M. Griffiths, L. B. Forcier: Intelligence Unleashed: An Argument for AI in Education. *Pearson Education, London*, UK (2016).

[21] W. Holmes, M. Bialik, C. Fadel: Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. *Center for Curriculum Redesign, Boston*, USA (2019).

[22] S. Wollny, J. Schneider, S. Schimmler, T. Wambsganß: Chatbots in Education: A Systematic Literature Review. *Education and Information Technologies*, ISSN 1360-2357, 26, 3365–3407 (2021).

[23] D. Faggella: AI in Education – Current Applications and Trends. *Emerj Artificial Intelligence*

Technium Education and Humanities
Vol. 11, pp.181-195 (2025)
ISSN: 2821-5079
www.techniumscience.com

Research. https://emerj.com (2023).

[24] I. H. Benarab: Detection of AI-Generated Writing in Students' Assignments: A Comparative Analysis of Some Tools' Reliability. أطراس (Atrās), ISSN 2664-666X, 5 (3), 271–286 (2024).

[25] M. B. Saqib, S. Zia: Evaluation of AI Content Generation Tools for Verification of Academic Integrity in Higher Education. *J. of Applied Research in Higher Education*, ISSN 2050-7003 (2024).