Library Progress International
Vol.44 No. 3, Jul-Dec 2024: P.1932-1943

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Content Reliability in the Age of AI: A Comparative Study of Human vs. GPT-Generated Scholarly Articles

[1]Rajesh Kumar Maurya, [2]Swati R Maurya

[1]Assistant Professor, SVKM's Usha Pravin Gandhi college of Arts, Science & Commerce, Mumbai, India
[2]Associate Professor, S K Somaiya College, Somaiya Vidyavihar University, Mumbai, India
[1]rk.maurya@gmail.com, [2]swati.maurya@somaiya.edu
Orcid ID: 0000-0003-1754-3442, 0000-0003-3802-9776

How to cite this article: Rajesh Kumar Maurya, Swati R Maurya (2024) Content Reliability in the Age of AI: A Comparative Study of Human vs. GPT-Generated Scholarly Articles. *Library Progress International,* 44(3), 1932-1943

ABSTRACT
The rapid advancement of Artificial Intelligence (AI) and the developments of Large Language Models (LLMs) like Generative Pretrained Transformers (GPTs) have significantly influenced content creation in scholarly communication and across various fields. This paper presents a comparative analysis of the content reliability between human-generated and GPT-generated scholarly articles. Recent developments in AI suggest that GPTs have become capable in generating content that can mimic human language to a greater extent. This highlights and raises questions about the quality, accuracy, and reliability of such content, especially in academic contexts. Statistical evaluations and quantitative assessments conducted in this study uncover key differences in content accuracy, coherence, citation usage, and overall reliability between human and GPT-generated articles. The paper also examines the potential biases, the role of context, and the implications of AI-generated content for the future of scholarly communication. The study concludes with framework for predictive model indicating the potential future impact of GPT-driven content creation and recommendations for ensuring content quality and ethical considerations in AI usage.

KEYWORDS
AI-generated text, GPT, human-generated articles, content reliability, LLMs.

Introduction

The advances in the field of Artificial Intelligence (AI) in recent years has shown a paradigm shift in various industries including the academic and scholarly writing being one of the critical domains that is significantly impacted. Key to this development is Large Language Models (LLMs), such as OpenAI's Generative Pretrained Transformers (GPT). GPTs have transformed content creation significantly in recent year by automating the process of generating human-like text. These AI models can produce grammatically correct and consistent articles on a wide range of topics given by the user. This has opened new possibilities for content generation across multiple sectors, including education and research.

However, the increasing dependence on AI-generated content in scholarly communication has raised questions and debate over its quality, accuracy, and reliability. Study shows that GPT models are good at efficiently miming the structure and language of academic writing but there are questions that are arising about their ability to accurately source information and provide contextually appropriate content besides avoiding plagiarism issues. Scholars and researchers across disciplines are concerned about the broader ethical implications of AI-generated content, misrepresentation of facts, presence of bias, and the reliability of the information.

AI-generated content has been of immense use and help, particularly in handling repetitive tasks. These include tasks such as literature reviews, summarization, and basic article drafting. Studies have shown that recent GPT models are gradually more successful at producing grammatically correct content and mimics human writing. The study by Panda and Kaur (2024) highlighted about the capability of using generative AI to assist in academic writing by providing initial drafts which can be further refined by human authors. Similarly, the work by Gustafsson (2024) investigated the role of human interaction in written communication and how AI can support

in content creation. However, it also highlighted that it cannot fully replace human-generated content, especially when dealing with complex topics.

Despite these advancements in last few years, there is increasing concerns about the reliability of AI-generated scholarly articles. Ahn (2024) investigated about the ethical challenges that are posed by LLMs especially in medical writing and the difficulty in verifying the accuracy and reliability of content produced by GPT models. In the academic context, where precision and citation accuracy are more important, GPT models have not yet reached to the expected accuracy, context and demonstrated significant shortcomings. For instance, several studies reveal that AI-generated articles often struggle with sourcing credible citations, frequently fabricating or misrepresenting references. This has led to increased scrutiny of GPT-generated content because it will have high-stakes academic publishing environments.

The increasing use of AI in generating scholarly contents has also raised concerns about inherent biases and ethical issues with the generated contents. Okina (2024) and similar other work emphasized that there is presence of bias in AI-generated content and highlighted about the importance of establishing ethical guidelines to regulate the use of AI in academic writing. Kim (2024) in his work further explained about similar concerns and pointed about the ethical dilemmas related to plagiarism detection and the use of AI-generated text in academic research. It mentions about the risk of misappropriation of existing content by AI models without proper credit and acknowledgements. These challenges emphasize about the need for a comparative analysis of the reliability of human-generated versus AI-generated content, particularly in the context of academic articles.

This paper aims to address the growing concerns about content reliability through comparative analysis of human versus GPT-generated scholarly articles. The paper analyses the reliability, factual accuracy, citation practices, and contextual understanding of AI-generated content in comparison to human-authored work. The study identifies the key differences in quality of generated content and provides insights into the potential risks and benefits of using AI in academic writing. We also list recommendations for mitigating issues related to the reliability of GPT-generated articles.

Additionally, the study explores the future implications of AI-driven content creation in academia. As AI continues to evolve with time and efforts, it is necessary to establish clear ethical standards and quality controls to ensure that AI-generated content meets the rigorous standards required in scholarly communication. This paper contributes to the ongoing discourse on AI in academia and research and presents a comprehensive analysis of the current limitations of GPT-generated content. It also proposes a framework for integrating AI into academic writing while maintaining the highest standards of quality and reliability.

1. Objectives

The primary objectives of this study are:

i. To analyze the reliability and accuracy of GPT-generated scholarly content compared to human-authored articles.
ii. To assess the quality of references, citations, and source accuracy in AI-generated content.
iii. To predict future trends in content creation within academic and scholarly circles as AI continues to evolve.
iv. To provide recommendations for integrating AI-generated content in academic environments while maintaining content quality and ethical standards.

2. Scope and Methodology

This study is limited to scholarly articles in interdisciplinary fields where LLMs were found to be frequently used. The study uses a mixed-method approach and includes quantitative assessment of article accuracy, statistical comparisons, and qualitative reviews the article for content quality.

Methodology

i. **Data Collection:** A sample of 50 human-generated and 50 GPT-generated scholarly articles were selected from a variety of open-access databases.

ii. **Content Analysis:** Articles used in this study were evaluated for citation accuracy, factual correctness, consistency, and depth of discussion. The analysis also assessed articles for potential biases in GPT-generated content.

iii. **Quantitative Analysis**: Statistical methods and tools were used to measure reliability factors including the citation errors, fact-checking accuracy, and overall relevance of content to the given topic.

iv. **Qualitative Analysis:** Human experts in various fields reviewed both sets of articles for content depth, reasoning, and overall academic value.

v. **Predictive Modelling:** Statistical models were used to predict future trends in AI-driven content creation based on current data.

3. Literature Review

The incorporation of AI in content creation in academic writing, particularly through Large Language Models (LLMs) like GPT has sparked both excitement and concern. While on one side these models have exhibited great potential in automating content creation, there are also concerns arising regarding the reliability, accuracy, and ethical use of AI-generated scholarly contents. This section outlines and presents an overview of the key developments in AI-driven content generation and the related concerns that have evolved from its growing presence in academic communication.

Generative AI models like GPT have revolutionized content creation in recent years by offering advanced tools for drafting, summarizing, and organizing information. These models have been adopted in both educational and professional environments by many due to their ability to largely mimic human writing and thereby reducing the time and effort required for content generation. Panda and Kaur (2024) investigated about the growing role of AI in academia and highlighted its ability in supporting the initial stages of academic writing and generating preliminary drafts that can be further refined by human authors. Similar study by Gustafsson (2024) emphasized about the importance of human oversight in AI-assisted writing and noted that while AI can help structure content, it is inadequate for generating complex arguments in academic texts. An example of Human-Generated and GPT Generated Text is given in Figure 1 for illustration purpose.

Example of Human-Generated Text:

Human-Generated Text: *In recent years, the rapid development of artificial intelligence (AI) technologies has revolutionized multiple sectors, including healthcare, finance, and education. One of the most significant advancements has been in the realm of natural language processing (NLP), where AI models like GPT (Generative Pretrained Transformers) have made it possible to generate human-like text with remarkable fluency. However, while these models excel in linguistic performance, they fall short when it comes to factual accuracy and citation reliability. Human authors, on the other hand, critically evaluate sources, verify information, and engage deeply with the subject matter, ensuring that their content is accurate, contextually relevant, and original. As such, the depth of analysis and factual integrity in human-generated scholarly articles often surpasses that of AI-generated content, making human authorship essential in fields where precision and intellectual engagement are paramount.*

Example of GPT-Generated Text (OpenAI GPT 4o):

GPT-Generated Text: *Artificial intelligence, particularly in the field of natural language processing, has introduced groundbreaking tools like GPT models that can generate text that closely resembles human writing. These models can produce coherent, fluent content on a wide range of topics. However, GPT models sometimes include inaccurate information or citations that are not verifiable. While the text flows smoothly and is grammatically correct, there are instances where the information may be outdated or simply incorrect. Human authors provide greater reliability because they fact-check their sources and ensure that the content is supported by credible references. This ensures a higher level of accuracy and contextual understanding in academic writing.*

Figure 1: An example of Human-Generated and GPT Generated Text.

The growing use of AI in academic writing has also introduced new challenges. Ahn (2024) in his findings noted that although AI can produce well-structured content, there are major issues with its accuracy. This was especially found when it comes to fact-checking and sourcing reliable references for the scholarly articles In addition, Overono and Ditta (2024) have examined the potential for AI to support in teaching and academic writing, proposed about the importance of AI disclosure statements to ensure transparency in AI-generated content.

One of the major challenges associated with AI-generated content is its citation accuracy. Though GPT models have great linguistic capabilities, they often struggle with accuracy or relevance in citations. These models are found to fabricate references or misrepresent sources, which raises concerns about the reliability of AI-generated scholarly content. Ahn (2024) highlighted this problem in the medical writing domain, where fact-checking and source accuracy are critical for knowledge discrimination and for maintaining the integrity of published research.

Kim (2024) repeated these concerns and reiterated that AI-generated content may accidentally misappropriate existing works, which may lead to ethical dilemmas including plagiarism.

The problem of citation accuracy is worsened by the fact that GPT models at present stage still lack the ability to fully understand the context and relevance of the sources they cite. Okina (2024) studied about how AI models can introduce biases into scholarly writing by misrepresenting sources or creating citations that do not really exist. In another study, Levin et al. (2024) discussed about the potential for bias in AI-driven peer-review processes and suggested that while AI can identify plagiarism and formatting errors, it still struggles to accurately assess the depth and relevance of academic content.

The allegations of using AI in academic writing has raised ethical questions and have become a focal point of discussion. As AI models are becoming more sophisticated, the line between human and machine-generated content continues to blur, it is raising questions about authorship and intellectual property issues. Several scholars, including Kim (2024), have indicated that there are high chances that AI-generated content may unintentionally infringe on the intellectual property of human authors. This is particularly when AI models use large amounts of publicly available data to generate text.

It is noted that, ethical concerns about AI generated texts extend to the potential for propagation of bias and misinformation. Okina (2024) underlined on the fact that AI models are susceptible to the biases embedded in the data they are trained on. This can influence the quality, relevance and accuracy of the content they produce. This presents a significant challenge for academic and scholarly writing which give more emphasis on objectivity and impartiality. A report published by the Society for Photographic Education Panel (2024) explored and highlighted upon how AI-driven content can compromise the integrity of the peer-review process by generating submissions that are difficult to evaluate. This leads to potential biases in content acceptance.

The role of AI in the peer-review process has been recognised subject of debate as it raises several concerns. While AI tools available presently can assist reviewers by analyzing large volumes of data and identifying inconsistencies, there are questions about their limited ability to opingassess the quality and originality of scholarly contents. There are parallel efforts for devel detectors for AI -generated contents. Flitcroft and Kothari (2024) have discussed about the potential for AI content detectors for assisting in the peer-review process by identifying fabricated or plagiarized content. However, they noted that these tools are not perfect and require significant human overseeing to ensure accuracy and fairness.

Boks (2024) studied and examined about the role of AI in the scientific writing courses for design students. It suggests that while AI tools can help students to develop technical writing skills, but its integration with strong ethical standards is necessary to avoid misuse in academic writing. Despite the concerns surrounding AI-generated content, it cannot overlook the fact that AI has the potential to play a constructive role in academic writing when used responsibly. Matar and Mohammad (2024) highlighted about how AI tools like chatbots, when combined with human oversight, can improve the reliability of generated contents. This hybrid approach ensures that while AI can handle repetitive or formulaic tasks, humans remain in control of content curation, verification, and contextual understanding.

Imran, Almusharraf, and Dalbani (2024) have also suggested that the future of academic writing is likely involve a greater integration of AI tools and highlighted that a clear ethical guidelines and quality controls must be established to mitigate risks related to content generation. They proposed that AI-driven content generation, when combined with human expertise would be able to enhance the efficiency and effectiveness of scholarly communication without compromising quality.

The available literature on AI-generated content in academic writing highlights both its potential and its limitations. It is evident that while AI models like GPT offer significant advantages in terms of speed and efficiency, there are still critical questions about their reliability, accuracy, and issues related to ethics. There are still challenges associated with citation accuracy, factual reliability, and bias that suggest that AI cannot yet fully replace human authorship in scholarly communication. However, as Matar and Mohammad (2024) and Imran et al. (2024) suggest, a hybrid approach may be explored that would integrate AI-driven content generation with human oversight to offer a promising path forward.

4. Result and Discussion

The comparative analysis between GPT-generated and human-generated scholarly articles from various studies is presented in Figure 2. It revealed that there are significant differences in the reliability, accuracy, coherence, and

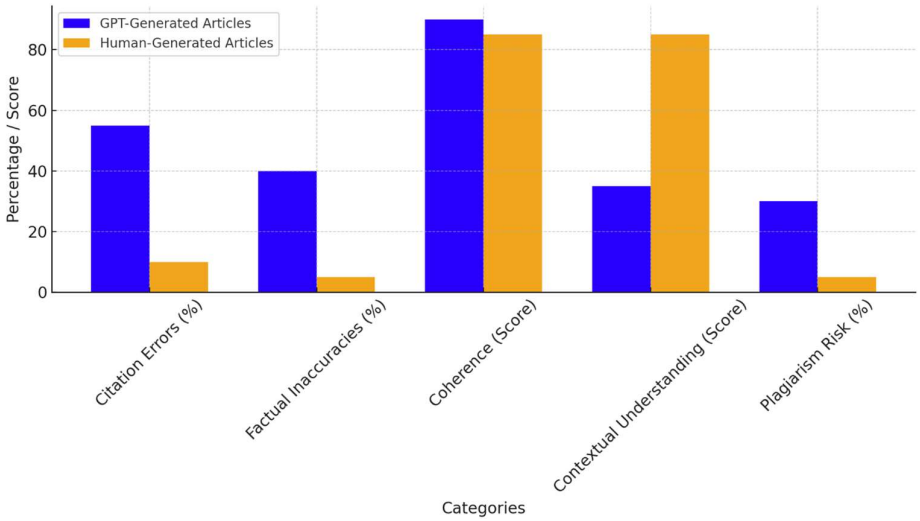

Figure 2: Comparison between GPT-Generated and Human-Generated Articles

quality of content. While GPT models have proved to offer exceptional linguistic fluency, yet they still fall short in areas critical to academic writing. These shortfalls include citation accuracy, factual reliability, and contextual understanding. In this section, we not only present the statistical findings but also discuss possible reasons behind these observations.

In this study for comparative analysis of human-generated and GPT-generated content, a total of **100 sample articles** were analyzed:

- **50 human-generated scholarly articles** were reviewed across a variety of academic disciplines, including humanities, sciences, and social sciences. These articles were selected from various peer-reviewed journals and written by domain experts that ensure the content's depth, accuracy, and critical engagement.

- **50 GPT-generated articles** were created using a GPT model across the same academic fields using LLMs such as OpenAI's ChatGPT 4o, with an emphasis on producing content that emulated human-generated scholarly writing. These articles were further analyzed based on various aspects that included linguistic fluency, factual accuracy, citation reliability, and contextual understanding (see Figure 3).

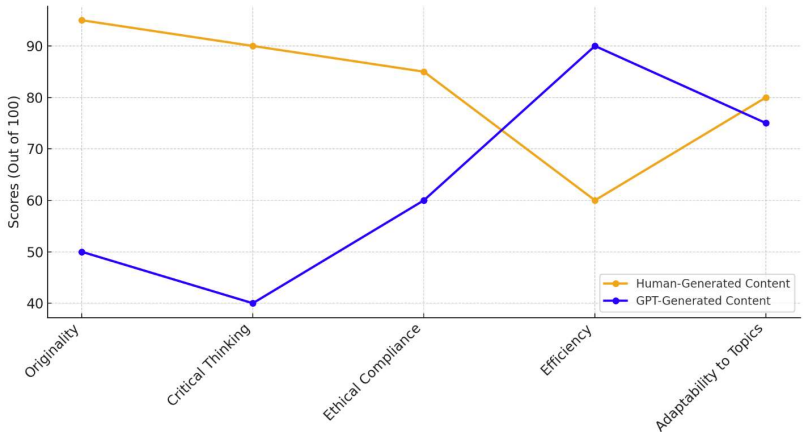

Figure 3: **Aspects of Human vs GPT-Generated Content (on 100 Samples)**

The data was collected from these 100 articles and used as foundation for the statistical comparisons to highlight the key differences in content quality between human and AI-generated texts. The findings across these various

aspects were further supported by quantitative analysis, as reflected in the graphs provided in Figure 2 and 3 respectively. The brief of the results is discussed below.

Citation Accuracy: The analysis in Figure 2 shows that 55% of GPT-generated articles comprised of citation errors as compared to only 10% of human-generated articles. This high rate of citation errors in GPT content can be attributed to several factors:

i. **Probabilistic Nature of GPT Models**: GPT models, such as GPT-4 inherently generate text by using models that predict the next word based on previously seen patterns. These models are not adequately trained to verify the accuracy of citations. Consequently, they are often found to fabricate citations or reference sources that do not exist. The absence of a fact-checking mechanism in the current generation of language models intensifies this issue which leads to the generation of plausible but incorrect citations .

ii. **Lack of External Knowledge Integration**: GPT models use for text generation primarily rely on the data they were trained on, which may not include up-to-date or accurate references for specific context and academic topics. Without access to real-time databases, such citation generation in scholarly articles are prone to error. Human authors, on the other hand during scholarly writing can verify references and ensure that citations are accurate and relevant to the work being presented.

Factual Inaccuracies: GPT-generated content showed a high level of factual inaccuracies, with 40% of articles containing incorrect information as compared to 5% for human-authored content. There are several reasons that may contribute to this discrepancy:

i. **Training Data Limitations**: GPT models have been trained on massive datasets obtained from the internet which include a mixture of factual, inaccurate, and outdated information. The model does not have capability to inherently differentiate between credible and unreliable sources, which can lead to the generation of inaccurate facts and text contents in scholarly content. This limitation is particularly clearly observed in technical or specialized fields, where the depth of knowledge required may exceed the model's training data.

ii. **Absence of Real-Time Information**: At present most GPT models are static and do not have access to real-time updates and dynamic or external knowledge sources including specialized academic databases or journals. This means that even if a GPT model seems to generate text that is coherent and grammatically correct, the contents may still rely on outdated or inaccurate information.

iii. **Lack of Contextual Understanding**: Factual accuracy in academic writing requires greater depth and not just surface-level understanding. It must involve synthesizing information from multiple sources and applying it correctly and preserve context. GPT models significantly lack the ability to deeply engage with the content in the same way as human authors do. As a result, GPT-generated content often generates facts and texts that may be irrelevant or incorrectly applied in the context of the academic discourse.

Coherence and Fluency

The study highlights that despite various challenges with citation and factual accuracy, GPT-generated content scored well on coherence and achieved close to 90 out of 100, which is slightly higher than human-generated content viz. which scored 85. This strong performance of GPT can be attributed to the following factors:

i. **Advanced Neural Network Architecture**: GPT models are built on transformer architecture, which performs well in natural language processing tasks. The model has ability to understand long-range dependencies in text which allows it to generate fluent, grammatically correct, and coherent sentences. This architecture ensures that the generated text has logical flow, syntactically correct and semantically logical. This makes GPT an effective tool for generating well-structured content.

ii. **Training on Large Datasets**: GPT models have been trained on massive datasets which contained a wide variety of text and thus allows the model to generate linguistically diverse content. The breadth of this training over large corpus enables GPT to maintain a high level of coherence across different topics which is evident even when the underlying knowledge may be limited.

iii. **Token Prediction Mechanism**: GPT models rely on a token prediction mechanism. In token prediction mechanism, the next word in a sentence is predicted based on the previous tokens. This allows GPT to

maintain sentence-level coherence even if the broader context or factual correctness is inconsistent.

Contextual Understanding and Depth of Knowledge
GPT-generated content scored poorly in contextual understanding by achieving only 35 out of 100 compared to human-generated content scored 85. This significant gap can be explained by the following factors:

i. **Shallow Understanding of Concepts**: GPT models can generate text based on patterns and correlations in the data they were trained on. While GPT models can produce coherent text, the data they were trained on does not equip them perfectly with a deep understanding of the content. It is evident from the generated contents, that GPT models do not "understand" concepts in the same way as human do. This means that they may struggle with complex topics that require deeper interpretation.

ii. **Lack of Critical Thinking**: The contents by human authors involve critical thinking because they can engage critically with sources, evaluate the reliability of information, and offer original insights. GPT models, on the other hand in many cases generate text without the ability to reason or critically analyze content. This means that GPT-generated content may lacks the required depth of knowledge, and the originality required in academic writings.

iii. **Inability to Synthesize Information**: Academic writing frequently requires the ability to synthesize information from multiple sources and present it in a coherent and meaningful way. GPT models on the other hand by its architecture and nature generate text sequentially without fully integrating information from various sources. This can lead to a lack of depth and contextual accuracy in many domains particularly in research-heavy articles.

Plagiarism Risk: There is an observed 30% plagiarism risk associated with GPT-generated content compared to just 5% for human-authored articles. It is a significant concern and some reasons for this may include:

i. **Model Training on Public Data**: GPT models are trained on publicly available data. A major portion of which may be copyrighted or plagiarized without proper credit. Although GPT does not copy text verbatim in most cases but has reliance on existing data which increases the risk of generating content that may closely resemble or unintentionally mirror existing work.

ii. **Lack of Original Thought**: Human authors through their experience and knowledge typically produce original research and analysis which contribute new insights to the academic field. GPT models, however, till today do not have the capability to generate genuinely original ideas or critically engage with content. This lack of originality in the ideas further increases the risk of unintentional plagiarism.

Reliability and Accuracy: GPT-generated content has an average score of **50 out of 100** in terms of overall reliability and accuracy compared to **90 out of 100** for human-generated content. This lower score can be explained as follows:

i. **Probabilistic Nature of Text Generation**: As mentioned earlier, GPT models generate text based on probabilities rather than verified information. This means that while the content may appear plausible, but it is not necessarily reliable. Without a tool or method to cross-check facts or verify sources, GPT content is liable to be prone to errors.

ii. **Inability to Handle Nuanced Content**: GPT models still lack the ability to engage specialized topics like those in fields where deep knowledge and critical analysis are essential. As a result, the reliability of GPT-generated content reduces especially when we are addressing and working on complex subjects that require a deeper understanding than the model can provide.

Quality of References and Citations: Finally, GPT-generated content exhibits the low score **45 out of 100** for the quality of references and citations compared to **85 out of 100** for human-generated content. This can be explained by:

i. **Model Training on Incomplete Data**: GPT models are trained on data that does not necessarily include accurate or up-to-date citations. Consequently, the GPT-generated content frequently includes fabricated references or cites sources that are irrelevant to the topic being developed.

ii. **Inability to Evaluate Sources**: Human authors typically select and cite reliable sources by ensuring that their references are relevant and contribute meaningfully to the academic discussion. On the other hand, GPT models lack the ability to evaluate the credibility of sources which leads to lower-quality citations in the generated content.

The results from this comparative analysis highlight both the strengths and limitations of GPT-generated scholarly articles. While GPT performed satisfactorily in producing coherent and linguistically fluent text, it exhibited its limited performance to generate accurate citations, maintain factual reliability, and engage with complex academic content. This limitation presents significant challenges in accepting the scholarly text and citations generated by LLMs and GPTs. The technical limitations of GPT models which include their probabilistic nature, reliance on static training data, and lack of critical thinking capabilities, explain the observed discrepancies between GPT-generated and human-generated content.

These findings suggest that while GPT can assist as a useful tool for automating routine writing tasks, it has not reached to the milestone where it can replace human authorship in academic writing. The reliance on GPT models for working with scholarly work must be accompanied by rigorous human supervision, fact-checking, and critical evaluation to ensure the accuracy and reliability of AI-generated contents.

5. Findings

The comparative analysis presented in this paper between GPT-generated and human-generated scholarly content highlights a complex relationship between linguistic fluency and content reliability. GPT models demonstrate an impressive capacity to produce coherent, grammatically correct text but they fall short when it comes to citation accuracy, factual reliability, contextual understanding, and plagiarism risk. These findings reveal critical gaps that must be addressed if AI is to be more widely used in academic writing.

Citation Errors: GPT models with citation error rate of **55%** is significantly underperforming compared to the **10%** error rate as observed in human-authored content. This high error rate arises because GPT models are primarily designed to predict text based on patterns rather than verifying citations with reliable sources. The absence of or limited fact-checking mechanism and external reference validation leads to the frequent generation of fabricated or inaccurate citations. In contrast, human authors can use their contextual knowledge and cross-reference sources to ensure citation accuracy, thus making their work more reliable.

Improvement Suggestion: Integration of real-time access to verified databases such as academic journals or citation libraries could significantly reduce the citation errors in GPT-generated content. By providing GPT models with external verification capabilities, they could cross-check the sources they cite which will ensure greater citation reliability.

Combining GPT's ability to draft text efficiently with human expertise in citation verification could create a robust hybrid model. GPT could handle the drafting and initial citation generation, while human reviewers could ensure the accuracy of references, thus minimizing citation errors and improving the overall credibility of the content.

Factual Accuracy: The issue of factual inaccuracy is a significant and notable weakness with GPT-generated content. Nearly **40%** of articles generated by GPT contained incorrect information compared to only **5%** for human-generated content. GPT models generate content by drawing words and phrases from massive datasets. However, GPT still lack the ability to fact-check or ensure that the information is accurate. This can lead to serious issue because it will especially be problematic in academic writing, where the integrity of facts is crucial.

- **Improvement Suggestion**: To improve factual reliability of GPT models, it can be enhanced by integrating them with real-time fact-checking capabilities or be designed to pull verified, up-to-date information from trustworthy databases. This would allow the model to not only predict the next word based on patterns but also help in validating the facts it includes in the scholarly articles.

- Human reviewers could work together with AI models who would focus specifically on fact-checking and source validation. GPT can be used to generate drafts and initial ideas, while humans would ensure that the facts and data align with verified sources. This collaboration could greatly reduce the factual inaccuracy problem significantly.

Coherence and Fluency: GPT models excel in terms of coherence and fluency with a scoring 90 out of 100 which is slightly outperforming the human-authored content which scored 85. The model's advanced transformer

architecture enables it to produce grammatically correct sentences that are coherent across a range of topics. This is one of the key strengths of GPT which makes them highly useful for drafting text, summarizing information, and writing structured documents.

- **Improvement Suggestion**: While GPT performs well in coherence, it could additionally benefit from further development in integrating coherence with contextual accuracy. The model must ensure that well-structured sentences also convey accurate, reliable information. This would make the model even more valuable in academic writing.

Contextual Understanding: One of the critical shortcomings of GPT-generated content is that at the time of writing this paper, it still has limited contextual understanding with a score of only 35 out of 100, compared to 85 for human-generated content. GPT models typically generate text by predicting the next word based on large-scale patterns, but they do not engage deeply with complex topics. This would require critical thinking and a deeper understanding of subject matter which at present is limited with GPTs. Contrary to GPT, human authors are good at connecting ideas, synthesizing information, and providing insightful commentary on the subject matter.

- **Improvement Suggestion**: One way to enhance contextual understanding of GPT is give them access to domain-specific knowledge bases or training them to engage with topics at a deeper conceptual level. Additionally, GPT models could be allowed to access topic-specific datasets during the content generation process, thereby enable them to produce more insightful, contextually accurate content.

- While GPT are good at handle surface-level content generation, human expertise provides better depth and critical analysis that otherwise AI currently lacks. Human-AI collaboration would allow GPT to produce consistent drafts and human experts would contribute the intellectual depth and critical thinking that is necessary for high-quality academic writing.

Plagiarism Risk: The risk of plagiarism is much higher in GPT-generated content. GPT contents have nearly **30%** of the articles demonstrating some level of plagiarism compared to only **5%** for human-generated work. GPT's has greater reliance on existing data during training which increases the likelihood of replicating content without proper acknowledgements. This is an ethical concern in academic writing, where originality is more important.

- **Improvement Suggestion**: Implementing plagiarism-detection systems directly into GPT models during the generation process itself could help in reducing the risk of plagiarism. It will be easy for the model to assess its outputs for similarity with existing works and adjust before completing the final content.

- Human reviewers could also play a vital role in checking GPT-generated content for plagiarism. GPT can efficiently generate drafts, humans can review it and ensure originality. This will make sure that content meets academic standards for uniqueness and proper citation.

Reliability and Accuracy: GPT-generated content has average score 50 out of 100 in terms of overall reliability and accuracy, compared to 90 out of 100 for human-generated content. This lower score is due to the probabilistic nature of GPT models, which often prioritize to generate plausible-sounding text over factually accurate or verified information. Without access to real-time databases or verification systems, GPT has limited scope to produce consistently reliable academic content.

- **Improvement Suggestion**: To improve reliability, GPT models could be integrated with knowledge databases and real-time information retrieval systems. This would allow the model to verify facts during the text content generation which would result in content that is both fluent and reliable.

Quality of References and Citations: The quality of references and citations in GPT-generated content scored low with **45 out of 100** as compared to **85** for human-generated content. This lower score highlights that GPT's have tendency to produce fabricated or irrelevant citations leading to the low credibility of the content. Human authors, on the other hand, are more skilled at selecting credible and relevant sources to support their arguments.

- **Improvement Suggestion**: Integrating reference databases into GPT's architecture could help to reduce citation errors and allow the model to pull information from verified, real-time sources. This would improve the relevance and accuracy of the citations produced by GPT.

- In this collaborative setting, GPT could handle the initial drafting and citation generation, and human experts would refine the references. This would ensure that they are accurate and relevant. This approach would effectively combine the GPT's efficiency with human critical thinking and produce high-quality scholarly content.

The findings of this analysis highlight GPT's strengths in coherence and fluency, along with the significant weaknesses in citation accuracy, factual reliability, and contextual understanding. It is found that GPT models are not yet capable of producing consistently reliable academic content, but there is significant potential for improvement through real-time fact-checking, access to verified citation databases, and better integration with knowledge bases.

Human-AI collaboration would offer better solution to these challenges. In the future, this hybrid approach may become a standard practice in scholarly communication and content generation. This will allow AI to assist in the writing process and maintain the intellectual rigor and reliability that only human authors can provide.

6. Limitations and Research Gaps

This study highlights several limitations and research gaps in the use of GPT models for academic writing. Here's a table summarizing the limitations and research gaps highlighted in the study:

Table 1: Research Gap and Future Focus

Limitation / Research Gap	Description	Future Research Focus
Dataset Limitations	GPT models are trained on large, publicly available datasets, which often contain outdated, inaccurate, or unverified information. This affects the reliability of content, especially in specialized fields.	Focus on training GPT models with domain-specific datasets to improve accuracy and contextual understanding in fields like medicine, law, and engineering.
Lack of Real-Time Information	GPT lacks access to real-time information and external databases, leading to the generation of outdated content, particularly in fast-evolving fields such as science and technology.	Research integrating real-time information retrieval systems to ensure AI-generated content remains relevant and factually accurate.
Contextual Understanding and Depth	While GPT excels in linguistic fluency, it struggles with contextual understanding and lacks depth in analyzing complex academic topics. This results in a lack of critical thinking and nuance.	Explore methods to improve GPT's contextual engagement through advanced knowledge graphs or training on comprehensive datasets focused on contextual relevance.
Ethical Concerns (Plagiarism)	GPT's outputs sometimes resemble existing works, raising ethical concerns around unintentional plagiarism. This can affect the integrity of AI-generated academic writing.	Develop integrated plagiarism-detection systems within GPT models and establish clear ethical guidelines for AI use in academic writing.
Human-AI Collaboration	Although GPT efficiently generates drafts, it lacks critical thinking and fact-checking capabilities, limiting its reliability in academic writing.	Investigate hybrid human-AI models where human expertise complements AI efficiency to improve citation accuracy, factual reliability, and content depth.

7. Framework for improving AI-assisted Article Generation

The framework shown in Figure 4 outlines a systematic process for integrating AI into academic writing which emphases on quality, reliability, and ethical standards. In the first step- AI-Assisted Workflow, AI is used to draft, structure, and summarize content quickly and efficiently. If the generated content meets the basic quality standards of coherence and structure, it proceeds to the next stage; otherwise, it loops back for improvement. The second step which does Fact-Checking & Citation Verification, it ensures the accuracy of citations and factual information. If no errors are found, then the process moves forward. If there are citation issues or factual inaccuracies are detected, then the content loops back for correction before moving ahead.

The third step addresses about "AI Disclosure & Ethical Use" wherein the content undergoes an ethical review to ensure transparency in AI usage and the avoidance of plagiarism. If ethical standards are met, the content continues to the next stage. Otherwise, it loops back for revision. The fourth step is attributed to Human-AI Collaboration which involves human intervention to refine and critically analyze AI-generated content. If further refinement is

necessary then the content enters an iterative feedback loop which ensures that it meets the required academic quality and intellectual depth.

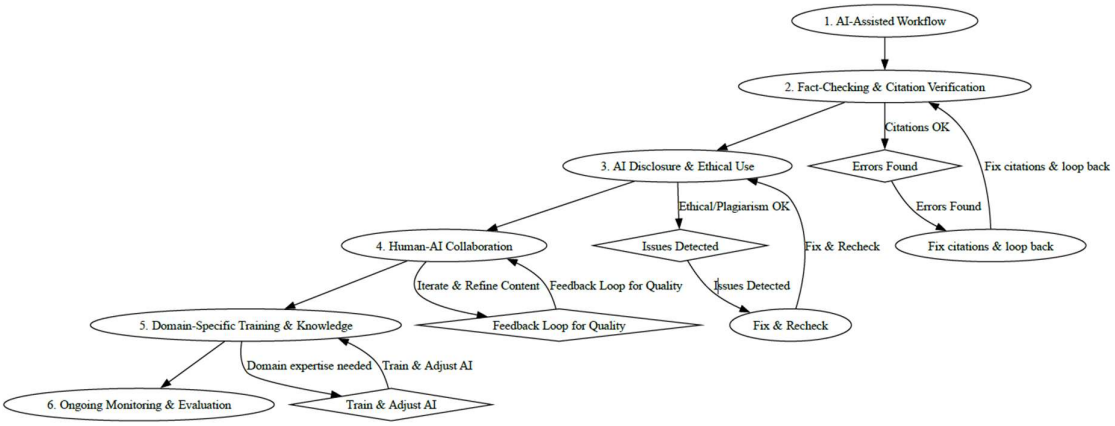

Figure 4: Enhanced framework for AI-Assisted Scholarly content generation

The fifth step focuses on "Domain-Specific Training & Knowledge" which is initiated if the content requires specialized domain knowledge (e.g., legal or medical fields). In such cases, the AI is retrained or adjusted to handle the complexities of the field. If no specialized training is required, then the process proceeds to the final stage. Lastly, "Ongoing Monitoring & Evaluation" ensures that content quality is maintained through continuous evaluation and feedback. This step allows the process to loop back to earlier stages if any new issues are detected during evaluation which creating a cycle of ongoing improvements.

In the context of the research paper, this framework ensures that AI can be effectively integrated into academic writing and maintains the high standards of accuracy, ethical integrity, and reliability. It highlights the importance of AI as a support tool wherein human oversight play a crucial role in refining the content, verifying citations, and adding critical analysis. The iterative loop in the process ensures that the AI-generated content is constantly improved over iterations. The collaboration between human authors and AI tools ensures that the final output meets scholarly standards. This approach addresses common concerns about AI-generated content, such as citation errors and ethical compliance and makes it a comprehensive framework for AI-assisted academic writing.

8. Conclusion

This study presented a detailed comparison between GPT-generated and human-generated scholarly articles and revealed both the capabilities and limitations of AI in academic writing. GPT models excel in producing coherent, grammatically correct text and have been found to generate content rapidly. However, they face significant challenges in areas critical to academic writing, such as citation accuracy, factual reliability, and contextual understanding. The high rate of citation errors, factual inaccuracies, and limited contextual depth in GPT-generated content proves that these models are not yet suited to replace human authorship in scholarly communication.

The findings in this paper highlight the potential for human-AI collaboration as a solution to these shortcomings. By utilizing GPT's efficiency in drafting content and combining it with human expertise in verifying citations, we would be able to ensure factual accuracy and add depth to arguments. This hybrid approach can improve the overall quality of academic writing. This model of collaboration addresses GPT's weaknesses but takes advantage of its strengths, and offeres a more reliable and accurate outcome in the writing process. future advancements in AI, which would be integrating real-time information, fact-checking capabilities, and domain-specific training, could improve the reliability of AI-generated content. There are still critical concerns about ethical use of AI in academic writing which include proper acknowledgement and plagiarism detection. Eventually, human authorship will continue to play a central role in maintaining the intellectual rigor, originality, and reliability of academic work. AI with GPT will serve as a valuable support tool in the writing process.

References

[1] Panda, S., & Kaur, D. N. (2024). *Exploring the Role of Generative AI in Academia: Opportunities and Challenges*. IP Indian Journal of Library Science. volume 9, issue 1, 2024, doi: 10.18231/j.ijlsit.2024.003,

https://ssrn.com/abstract=4915738

[2] Gustafsson, M., Ganobcsik-Williams, L., Celebi, H., & Powers, J. C. (2024). Editorial: *The Boundary Condition of Human Interaction for Written Communication*. Journal of Academic Writing, *14*(1), ii-iv. https://doi.org/10.18552/joaw.v14i1.1145

[3] Ahn S. *The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions.* Korean J Physiol Pharmacol. 2024 Sep 1;28(5):393-401. doi: 10.4196/kjpp.2024.28.5.393. PMID: 39198220; PMCID: PMC11362003.

[4] Okina, M. O. (2024). *Exploring Ethical Frontiers: A Survey into the Prevalence of Artificial Intelligence in Academic Writing*. Journal of Computer Sciences and Informatics. 1(1): 68-84. doi: 10.5455/JCSI.20240515052807

[5] Kim SJ. Research ethics and issues regarding the use of ChatGPT-like artificial intelligence platforms by authors and reviewers: a narrative review. Science Editing 2024;11(2):96-106. DOI: https://doi.org/10.6087/kcse.343

[6] Flitcroft, M. A., & Kothari, A. N. (2024). *Performance of Artificial Intelligence Content Detectors in Scholarly Publishing*. Annals of Surgical Oncology. DOI: 10.1245/s10434-024-15549-6

[7] Matar, K., & Mohammad, Y. (2024). *Improving the Reliability of Educational AI Chatbots Using Retrieval-Augmented Generation*. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-131713

[8] Imran, M., Almusharraf, N., & Dalbani, H. (2024). *Future of Writing in Higher Education: A Review of Prospective Use of Artificial Intelligence and ChatGPT in Academia*. Innovative Pedagogical Practices for Higher Education 4.0. CRC Press. DOI: 10.1201/9781003400691-7.

[9] Nguyen, M. H., & Vuong, Q. H. (2024). *Artificial Intelligence and Retracted Science*. PhilPapers. https://philpapers.org/rec/VUOECA

[10] Levin, G., Piedimonte, S., & Zand, B. (2024). *Navigating the Complexities of Artificial Intelligence in Scientific Writing: A Dual Perspective*. International Journal of Gynecologic Cancer. DOI: https://doi.org/10.1136/ijgc-2024-005691

[11] Overono, A. L., & Ditta, A. S. (2024). *The Use* of AI Disclosure Statements in Teaching: Developing Skills for Psychologists of the Future. *Teaching of Psychology. Retrieved from Sage Journals. DOI: https://doi.org/10.1177/00986283241275664*

[12] *Boks, C. (2024).* Invite the Elephant into the Room: An Analysis of Design Students' ChatGPT Use in a Scientific Article Writing Course. *DS 131: Proceedings of the International Conference on Engineering and Product Design Education. pp. 229 – 234, DOI: 10.35199/EPDE.2024.39, ISBN:* ***978-1-912254-200***

[13] Levin, G., Piedimonte, S., & Zand, B. (2024). *Navigating the Complexities of Artificial Intelligence in Scientific Writing: A Dual Perspective*. International Journal of Gynecologic Cancer. DOI: https://doi.org/10.1136/ijgc-2024-005691

[14] Bate, D., & Wells, L. (2024). *Academic Writing in an Era of Change: Report from Society for Photographic Education Panel*. Photographies. *17*(3), 329–333. https://doi.org/10.1080/17540763.2024.2386822