JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 01/06 pp105−121
Volume 5, Number 2, June 2024
DOI: 10.23919/JSC.2024.0011

Impact of Large Language Models on Scholarly Publication Titles and Abstracts: A Comparative Analysis

Phoey Lee Teh* and Chukwudi Festus Uwasomba

Abstract: Artificial Intelligence (AI) tools become essential across industries, distinguishing AI-generated from human-authored text is increasingly challenging. This study assesses the coherence of AI-generated titles and corresponding abstracts in anticipation of rising AI-assisted document production. Our main goal is to examine the correlation between original and AI-generated titles, emphasizing semantic depth and similarity measures, particularly in the context of Large Language Models (LLMs). We argue that LLMs have transformed research focus, dissemination, and citation patterns across five selected knowledge areas: Business Administration and Management (BAM), Computer Science and Information Technology (CS), Engineering and Material Science (EMS), Medicine and Healthcare (MH), and Psychology and Behavioral Sciences (PBS). We collected 15 000 titles and abstracts, narrowing the selection to 2000 through a rigorous multi-stage screening process adhering to our study's criteria. Result shows that there is insufficient evidence to suggest that LLM outperforms human authors in article title generation or articles from the LLM era demonstrates a marked difference in semantic richness and readability compared to those from the pre-LLM. Instead, it asserts that LLM is a valuable tool and can assist researchers in generating titles. With LLM's assistance, the researcher ensures that the content is reflective of the finalized abstract and core research themes, potentially increasing the impact and accessibility and readability of the academic work.

Key words: semantic alignment; cosine similarity; patterns; large language models; technology impact

1 Introduction

The integration of Artificial Intelligence (AI) technology has become a commonplace occurrence across various industries, including the field of education. AI has significantly accelerated the pace of research and is now an integral part of our daily operations[1].

Artificial intelligence has proven to be a valuable asset to humanity, particularly in the realm of medical procedures[2]. The upcoming generation is fortunate to have the advantages of AI in their work, but there are also valid concerns about potential job displacement[3].

Regardless of the outcomes, AI will undoubtedly continue to evolve and advance, offering new opportunities for employment a voxel-based method.

At this juncture, in anticipation of this transition, we seek to establish a baseline and assess the transformation resulting from AI implementation, which in this project, we operate under the assumption that papers predating 2015 are less likely to involve AI assistance, marking a shift when AI began playing a more substantial role after 2015, a time when AI-enhanced work was not prevalent, and compare it to the works following the introduction of AI. The point before the AI-enhanced world, often referred to as the pre-AI era, is typically considered to be the time before the widespread integration and use of AI technologies

• Phoey Lee Teh is with the Department of Cyber and Computing, Wrexham University, Wrexham, LL11 2AW, UK. E-mail: phoey.teh@wrexham.ac.uk.
• Chukwudi Festus Uwasomba is with the School of Computing and Computing, STEM Faculty, The Open University, Milton Keynes, MK7 6AA, UK. E-mail: chukwudi.uwasomba@open.ac.uk.
* To whom correspondence should be addressed.
Manuscript received: 2023-12-12 ; revised: 2024-05-07; accepted: 2024-05-31

in various aspects of our lives. This era can be loosely defined as the period before the mid-2010s, with 2015 serving as a reference point for when AI technologies began to gain significant traction and influence[4]. Conversely, the point after the AI-enhanced world would be the period marked by the extensive use and impact of AI in our daily lives, which continues to evolve and advance beyond 2015.

To establish a baseline for academic writing in publications before and after the onset of the AI era, our goal is to explore the differences between titles produced by authors and those generated by Large Language Models (LLMs) in key domains of computer science—specifically, deep learning, cybersecurity, data science, and software engineering. We aim to scrutinize the relevance, coherence, fluency, informativeness, and clarity of statements within publications from the pre-AI era. The objective is to assess whether the title accurately reflects the content claimed in the paper. Our investigation seeks to identify any relationship between titles before the AI era, enabling us to provide recommendations to future authors on whether or not they should utilize LLMs to generate titles from the abstracts pasted onto OpenAI.

This endeavour could enable us to not only confirm but also navigate any notable distinctions between text generated before the pre-AI era and in the forthcoming post-AI era. Researchers have attempted to distinguish between user-generated text and AI-generated text[5, 6], and explored how natural language processing and AI are used for enterprise management in the era of industry[7]. They are also studying whether text is user-generated or robot-generated for several important reasons, for instance, to precipitate and decide on authenticity in ethical and legal implications such as deep fake analysis[8], or other information verification processes[9]. However, there still lacks focus on the attempt to identify the difference between the versions of AI generated text.

It is important to know whether a published paper was produced before or after the advent of AI for several reasons. Firstly, for a reader to read a paper that is published, we should establish a very accurate contextual understanding; besides the ethical considerations, papers that are produced after the introduction of AI might incorporate machine learning techniques, natural language processing, or other AI-related methods, which can affect the quality and reliability of the research. Hence, understanding the use of, or the level of used AI in assessing the research is important.

In this study, we aim to thoroughly explore crucial aspects relevant to this context, specifically focusing on the clarity and coherence between the title and abstract. To achieve this goal, our study involves an investigation into the correlation between the title and citation, emphasizing noticeable variations in semantic depth and citation patterns. This analysis is particularly important when comparing with the pre-LLM era, allowing us to highlight the influence of LLMs. Additionally, we seek to foster collaboration by inviting perspectives and insights from researchers and scholars in the field to enrich the comprehensiveness of our findings.

This research also gives us an indication of the advancement of technology-change across time in identifying the disparities between the different generations. Also, to determine the disparities in information between the article's title and the content intended to be conveyed in the research paper. Hence, the objective of our study is as follows:

(1) Compare the semantic similarity score between the authors' given title and GPT-4's generated title for high and low cited articles before and during the LLM era.

(2) Compare the readability score between the authors' given title and GPT-4's generated title for low cited articles before and during the LLM era.

(3) Determine if the variation in these scores are statistically significant.

2 Literature Review

When it comes to analysing context, multiple aspects of natural language processing are utilized for assessment. This involves evaluating the coherence, logical flow, informativeness, and clarity of the content. In the realm of research publications, each article is meticulously composed to facilitate clear understanding and effectively convey research findings to the target audience, providing valuable insights and knowledge. Therefore, clarity in communication, particularly in the context of intricate research, is vital for simplifying complex ideas and facilitating reader understanding[10].

There was once a notion suggesting that if students

were able to obtain correct answers through AI, the questions might be overly simplistic[11]. While AI can supply responses to a wide range of questions, do all of these answers truly address the question? When a researcher compiles a paper and seeks to communicate their findings, the foremost objective is to guarantee readability and comprehension. Often, the title is rephrased to make it more appealing or captivating. This is critical for ensuring the accessibility of their research findings, particularly in a world where readers often have diverse multidisciplinary backgrounds. However, do all the "titles" effectively convey the intended message within the entire article? Ensuring comprehension has become vital to engage readers effectively in the realm of research and to foster a persuasive argumentative approach, which is essential in research.

Given the advancements in AI, which include versatile tools like ChatGPT, it is difficult to imagine researchers not utilizing them for article composition or relying on their assistance in the writing process. Recently, tools have been developed and researched to distinguish between text produced with the help of LLM and text created without the assistance of it. However, as recently pinpointed by Ref. [12], ChatGPT has outperformed humans in terms of emotional context. Thus, differentiating between human and AI-generated text using emotional context, or as the determining factor to compare the disparities, remains uncertain.

To use ChatGPT, or not to use ChatGPT[13], demonstrates that there are no existing methods that can effectively detect ChatGPT-generated content. They have tested a large, benchmarked dataset spanning a variety of subjects to assess text from different contexts resulting in a 90% true negative rate of detection. The study also states that most detectors are prone to classifying any text as human-written even though the text was ChatGPT-generated, with those functions as an intelligent tool adept at comprehending and producing human-like text. The core focus of this paper is on the coherence between the generated title and its relevance to the content discussed in the published article. Our aim is to determine if AI-generated titles closely match the abstract (i.e., the content) of the overall information that is meant to be conveyed.

Relevance in text analytics refers to the degree to which a particular piece of text or document is related to a specific topic, query, or context[14]. It is a measure of how closely the content matches the information or subject of interest. Normally, a publication title creates an initial expectation for potential readers, this includes if the content is relevant to the title, or, it can confuse or mislead readers, leading to distrust or dissatisfaction. Ensuring relevance means that the research or discussion remains focused on the core topic. This focus is essential for clarity and depth. Irrelevant content can lead to poor content matches to what it claims in the results or in the discussion and hence would decrease user satisfaction. Checking the credibility also refers to checking the consistency in relevance throughout a document, and it enhances the credibility of the author and the document itself. Readers are more likely to trust and reference documents that stay on topic and deliver on their title's promise.

According to Ref. [15], a scholarly article's title is crucial, guiding readers by signaling its subject and scope. An effective title encapsulates the research essence, setting accurate expectations. Title precision is vital for relevance, not just academic convention. The abstract summarizes the research, maintaining cohesion with the title. Alignment between title and abstract is pivotal, ensuring continuation of promised premise. Misalignment could mislead readers, impacting relevancy and citation potential. A well-aligned abstract serves as a bridge, previewing the article's content and value. Its relevancy is judged by accurately reflecting core content and assisting readers in deciding on further reading.

To delve deeper into how ChatGPT operates, here are several key points to consider. ChatGPT maintains an organized writing style, it provides a very structural response, enabling the presentation of ideas coherently. It typically starts by summarizing the main points of the questions and then provides a comprehensive response. This is often followed by a concluding summary[16]. In contrast to human interaction, human communication can inherently encompass emotions or viewpoints[17]. Besides that, ChatGPT's tendency to offer detailed responses is shaped by Reinforcement Learning with Human Feedback (RLHF)[18]. It usually

provides thorough explanations unless specifically asked for brevity, in other words, ChatGPT tends to give long and detailed answers because of the way it learns from human feedback. It is crucial to grasp the diversity in presentation methods. Occasionally, the manner of writing can influence how it is perceived by humans, which includes the utilization of textual diversity[19]. When discussing sensitive subjects, ChatGPT remains neutral. This quality is especially useful for addressing politically charged issues, as it avoids bias or harmful content; it can handle sensitive topics[20]. In a study that compares the impact of anticipating a chat with a chatbot or a human, research findings indicate that humans typically have expectations of communication quality while also displaying empathy[21]. There are also studies that pinpointed that ChatGPT may generate incorrect details, particularly when responding to queries beyond its expertise[22], as a robot might occasionally fabricate information to answer the prompt or request by a human. Regardless of whether a robot can generate repetitive phrases, it holds potential due to the inclusion of statements generated based on the assumption of the "well-formedness" of ontologies, which may not always hold true in the context of linked open data[23]. Also, ChatGPT's sentences can adhere to specific templates, leading to predictable language patterns, and its responses might lack accuracy in the context, leading to logically or contextually irrelevant information[24]. Even if ChatGPT does not fully grasp the context, it attempts to fill in gaps using patterns from its training data.

The core principles of quality assessment in academic publishing emphasize rigor, reliability, and transparency. Metrics like citation impact and peer review, along with AI integration, could reshape benchmarking practices, notably with LLMs for title generation. Recognizing these shifts is crucial. It is also vital to address challenges and controversies in scholarly publishing benchmarking, given the field's dynamic nature and ongoing efforts to improve assessment methods.

2.1 Metrics used in measuring similarity

There are several ways to investigate textual coherence or accessibility. Coherence, in this context, pertains to the extent to which the information and insights derived from a given text or set of texts demonstrate logical and meaningful connections when analysed collectively[25]. Coherence is typically assessed qualitatively, and it is not quantified with numerical values like cosine similarity. Coherence plays a pivotal role in text analytics, as it serves as an indicator of the quality and efficacy of the analysis in yielding meaningful outcomes. When a topic and its abstract are coherent, they contribute to a clear representation of the intended information within an article, ensuring clarity. Clarity signifies the quality of being easily comprehensible, devoid of confusion or ambiguity. It stands as a crucial component of effective communication, ensuring that the message is conveyed in a direct and understandable manner. Clarity holds significance in various communication forms, including writing, public speaking, technical documentation, and user interfaces.

Readability measures the ease with which a text can be read and comprehended. It typically considers factors like sentence length, word complexity, and overall text structure. Readability metrics, such as the Flesch-Kincaid Grade Level or the Gunning Fog Index, assess how accessible a text is to its intended audience. The Flesch-Kincaid method or Flesch reading ease is a method that is based on two factors: average sentence length and average number of syllables per word[26].

When exploring methods for assessing similarity in textual content, in our case title and abstract, two forms of measurement exist: qualitative and quantitative. Qualitative measurements encompass aspects such as coherence, relevance, and informativeness. In contrast, our focus lies on quantitative measurements, which can be gauged by generating similarity scores.

One of the common approaches to quantify relevance is by measuring Term Frequency-Inverse Document Frequency (TF-IDF), a widely used formula. TF-IDF assesses how important a term (word or phrase) is to a specific document within a collection (corpus) with the formula as "TF- IDF = (Term Frequency) × (Inverse Document Frequency)" where Term Frequency (TF) measures how frequently a term appears in a document. It is calculated as the number of times a term appears in the document divided by the total number of terms in the document. It reflects the importance of a term within a specific document. TF(t, d) = (Number of times term t appears in document d)/(Total number of

terms in document *d*)[27]. It is useful in linking news pages that are relevant or news events that occur in succession[28]. Despite that, we have no intention to measure just by frequency of terms, therefore, it is not a suitable metric to determine in our case.

Another commonly employed technique is sentiment analysis; often it is utilized to delve into the intricate details of sentiments expressed in comments[29, 30]. These investigations center on domains or subjects and utilize keywords to articulate the focus of their research goals. For example, they may examine opinion reviews from TripAdvisor or sentiments toward specific commercial products or services. It is crucial to note that our goal does not involve reviewing comments; rather, we seek to validate whether the title aligns with the abstract and accurately reflects the content.

Perplexity is used to compare different language models or evaluate their performance on specific tasks. Lower perplexity generally indicates better language understanding and fluency of a model on a given text. Perplexity is not a metric to compare human and AI generated texts but is used to score how well an LLM can predict text itself. It also requires access to a pretrained language model and requires extensive computational resources, therefore, is not appropriate for the model used here[31, 32].

Cosine similarity is a quantitative method that calculates the cosine of the angle between these vectors. The resulting value ranges from − 1 (completely dissimilar) to 1 (perfectly similar), with 0 indicating no similarity. Cosine similarity is commonly used to compare the similarity between documents or to find similar documents in a corpus. It is a fundamental tool for tasks like document retrieval, information retrieval, and text clustering[31]. Cosine similarity has found wide application in numerous research studies. It is used in diverse contexts: On gamification-based e-learning platforms, it helps detect similar essay scoring[33]; It facilitates string matching capabilities[34] in the realm of plagiarism detection in thesis documents, it plays a crucial role[32]; It contributes to document summarization efforts[35]. For tasks like image captioning and similarity ranking, it proves valuable[36]. Even in the development of movie recommendation systems, cosine similarity is employed[37]. In these studies, researchers have consistently found that cosine similarity is a useful tool that provides a quick and effective means of understanding text matching.

2.2 Prompt engineering in LLMs

Prompt engineering, an emergent field within AI, primarily focuses on optimising the input (or "prompt") provided to AI models, especially LLMs, to enhance the quality and relevance of their outputs. A significant study by Ref. [38] highlighted that this concept transcends mere input manipulation, delving into the realm of understanding model behaviour and response patterns. This understanding is crucial for applications ranging from content creation to problem-solving. Furthermore, Brown et al.[39] demonstrated through GPT-3, one of the most advanced LLMs, that the way a prompt is structured can significantly influence the model's output, making prompt engineering a critical skill in AI. Peters et al.[40] expanded on this by examining the nimpact of prompt engineering on AI ethics, emphasizing how carefully crafted prompts can mitigate biases inherent in AI systems. However, Poola[41] cautioned against over-reliance on prompt engineering, arguing that it could lead to overfitting in AI responses and a reduction in model generalisability. Within the context of this research, we argue that prompt engineering signifies a critical intersection between human creativity and machine intelligence. It is a testament to the idea that while AI can process and generate vast amounts of information, the direction and quality of its output are still heavily influenced by human input.

3 Methodology

3.1 Research model

The research model for our study is designed to evaluate the impact of LLMs on the dissemination of scholarly articles within the computer science domain. Our model is established on the assumption that advancements in AI technologies, particularly post-2015 with the emergence of LLMs, have significantly influenced the research landscape. Authors are now utilising LLM to generate their titles, which was not possible pre-LLM time.

3.2 Conceptual framework and research variable

The study is structured around a comparative analysis framework that examines articles from two distinct eras: the pre-LLM era (2004 to 2014) and the LLM era

(2015 to 2024). We posit that LLMs have played a transformative role in research focus, dissemination, and citation patterns within the 5 selected knowledge areas covered, which include Business Administration and Management (BAM), Computer Science and Information Technology (CS), Engineering and Material Science (EMS), Medicine and Healthcare (MH), and Psychology and Behavioral Sciences (PBS). The era of publication (pre-LLM vs. LLM) and citation counts (high vs. low) serve as independent variables. While, the semantic content of titles and abstracts, as well as the readability of titles in relation to reflecting the content of the research done serve as the dependent variables. Given these, we hypothesize that: LLM is capable of outperforming human authors in article title generation and articles from the LLM era will demonstrate a marked difference in semantic richness and readability trajectory compared to those from the pre-LLM.

3.3 Data identification and selection process

In conducting our research, we applied a methodical approach to selecting and analysing titles and abstract of articles. Utilising a data extraction technique, we collated a dataset of 15 000 titles and abstract of articles from the field of BAM, CS, EMS, MH, and PBS as mentioned earlier. This expansive selection was narrowed down to 2000 journal papers through a multi-stage screening process that ensures adherence to our study's stringent criteria (Fig. 1).

3.3.1 Datasets gathering

We extracted titles and abstract from three eminent databases such as Google Scholar, Scopus, and OpenAlex using a proprietary or specialized system referred to as " Publish or perish version 8"[42]. This tool allowed us to specify the timescale and number of extractions for the datasets. So, with an even extraction of 2500 datasets (representing 500 per selected field) from each era (pre-LLM and LLM era) per source, we gathered a total of 15 000 titles and abstract. Thereafter the extraction process, we merged the datasets according to their respective field.

3.3.2 Datasets filtering and screening

In view of the aim of the research, the first level of our filtering process is the removal of books and other extracted titles and abstract of articles that are not peer-reviewed. Each article underwent a rigorous screening process to ensure it met our high standards. The

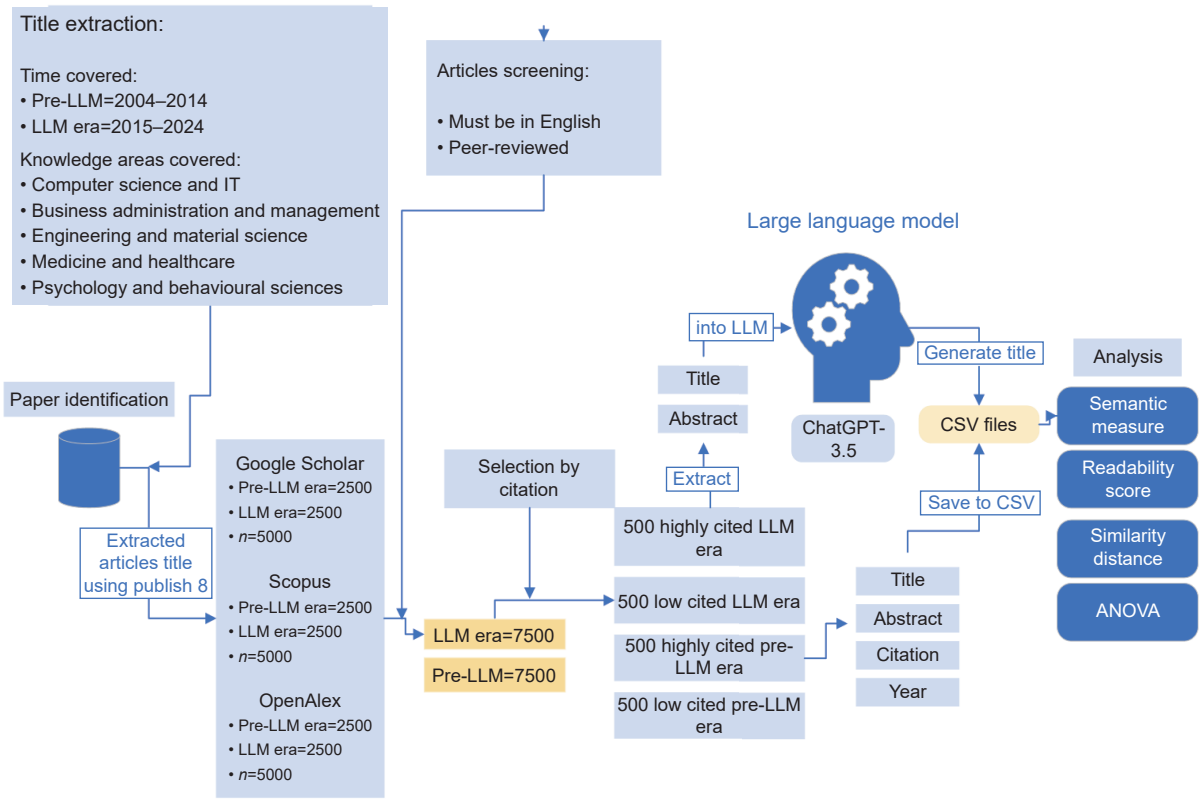

Fig. 1 Data identification and selection process.

research had to be articulated in English, subjected to the peer-review process, and focused on cutting-edge domains as mentioned earlier. These criteria were established under the premise that advancements in AI, particularly LLMs, would have a pronounced impact on these areas. This is because, these domains are recognised for their receptiveness to AI technology, and a higher likelihood of integrating AI into research by researcher. Further refinement was carried out based on citation metrics, a reliable indicator of a paper's influence and recognition within the scientific community[43, 44]. It is important to note that some recent studies have highlighted that citations, while commonly used in academia, may not consistently reflect a paper's true influence or recognition in the scientific community[45, 46]. Nonetheless, in the context of our research, we maintain that citation metrics should be regarded as a viable indicator[47].

3.3.3 Datasets selection

In our datasets selection process, considering the different durations available for citations to accumulate, we implemented a Citation Index (CI), calculated as the number of citations an article has received divided by its age.

$$\mathrm{CI} = C/A,$$

where CI represents the citation index, C represents the number of citations received by the article, and A represents the age of the article (measured in years). Subsequently, we arranged the datasets within each domain based on their CI to delineate highly cited and low cited articles. We selected the top 100 titles and abstract of highly cited papers and the bottom 100 out of 1500 from each dataset as the low cited papers. This process involved categorizing articles into four quadrants: 500 highly cited and 500 low cited papers for both pre-LLM and LLM eras, yielding a total of 2000 titles and abstract of articles utilized in this study. Also, in the case where the extract abstract in incomplete, we manually search the database to complete them. Our meticulous identification and selection methodology aim to curate datasets that offer a balanced representation of influential and emerging research. The collected titles and abstracts, spanning both pre-LLM and LLM periods, serve as the foundation for our analysis, facilitating a comparative examination of the academic landscape before and after the introduction of LLMs.

3.4 Analysis tool and technique

3.4.1 Computational environment

The analyses were conducted using Python programming language within a Jupyter Notebook environment. This platform was chosen for its interactivity, which facilitates an iterative analysis process, and its compatibility with a range of libraries essential for data analysis. The datasets, code, and its detailed commentary have been shared on GitHub for transparency and reproducibility: https://github.com/cfestus/LLM4Titles. Microsoft Excel was also utilised for its robust data aggregation capabilities as all extracted datasets were stored in a CSV file. ChatGPT (GPT-3.5) web version was employed to generate all the titles analysed in this study using a prompt engineering approach. The choice of this version of ChatGPT is on the assumption that most researchers would likely use this, since it is free.

3.4.2 Prompt engineering approach

In our study, we employed the role-playing technique of prompt engineering. This innovative method was utilised to effectively guide the model's output. The approach involved assigning a specific role to the model, providing a detailed context, and framing both a question and a preferred response format. The model was assigned the role of a "Professor" with extensive experience in academic publishing, as exemplified in the following prompt:

"Assume the role of a Professor in Computer Science, specializing in Knowledge Engineering, with 25 years of experience in writing for journals and conference proceedings."

A carefully tailored context was provided. We included the "Title" and "Abstract" of the research paper to train the model's responses in relevant information. We then posed a question to the model, asking it to:

"Analyzing the Title and Abstract provided below, Crafting the best and Optimal Title for the Publication."

We chose not to limit the model's response format in our experiments, aiming to give the model the freedom to generate the best possible title, regardless of length or style.

3.4.3 Similarity measurement

We used the cosine similarity score[48], a metric used to

determine how similar two documents are irrespective of their size, as the primary method of comparison. A TF-IDF vectorizer is employed from the sklearn library to convert textual data into numerical representations. This vectorizer is fit-transformed on the "Abstract", "AuthorsTitle", and "GPTTitle" columns, enabling the extraction of meaningful features from the text. Cosine similarity scores are computed between the abstract and the "AuthorsTitle" as well as "GPT4Title" columns using the cosine_similarity function from the sklearn.metrics.pairwise module. This process measures the semantic similarity between the abstract and titles, aiding in the identification of more related contents. We then employed descriptive statistics such as the mean and standard deviation, to provide insights into the central tendency and dispersion of similarity scores across the datasets. The scores were further examined through quartile-based metrics; the 25th percentile (Q1), the median (50th percentile), and the 75th percentile (Q3) to understand the distribution of similarity scores.

3.4.4 Readability measurement

We utilized the Flesch-Kincaid Grade Level readability formula to assess the readability of titles (both original and generated) within a dataset. Readability assessment is crucial for understanding the complexity of textual content, which can impact accessibility and comprehension. The Flesch_Kincaid_Grade function from the textstat library is applied to each title, providing a numerical assessment of its readability complexity. The scores are added as new columns ("AuthorsTitle_FK_Grade" and "GPTTitle_FK_Grade"). This integration allows for easy comparison and analysis of readability levels between original and generated titles.

3.4.5 Similarity distance

We calculated the Levenshtein distance between the original titles ("AuthorsTitle") and titles generated by GPT ("GPTTitle"). The Levenshtein distance is a metric used to quantify the difference between two sequences, in this case, strings representing titles. This is calculated using the Levenshtein.distance function from the Levenshtein module. This function measures the minimum number of single-character edits required to change one string into the other. The distances are added as a new column ("Levenshtein distance"). This column contains normalized distance values, indicating the similarity between each pair of titles. A value of 1 represents identical titles, while lower values indicate greater dissimilarity.

3.4.6 Statistical analysis

We conducted an Analysis of Variance (ANOVA) on the semantic similarity scores for authors and ChatGPT, as well as the readability scores, using a significance level (α) of 0.05. This statistical test allowed us to assess whether the observed changes in these metrics were statistically significant.

4 Result

4.1 Similarity measure

Tables 1 and 2 present the descriptive statistics of titles' semantic similarity scores for highly cited articles in the LLM era and the pre-LLM era. The mean across various titles ranged from 0.32 to 0.38 in the LLM era compared to the pre-LLM which ranged from 0.32 to 0.43, with MH having the highest mean score. The computed mean scores of authors titles' semantic similarity scores in LLM era are slightly higher (0.37)

Table 1 Descriptive statistics for highly cited titles in LLM era.

Similarity	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title semantic similarity	BAM	0.38	0.17	0.00	0.26	0.35	0.47	0.83
	CS	0.37	0.13	0.03	0.29	0.37	0.45	0.69
	EMS	0.36	0.14	0.00	0.28	0.36	0.43	0.72
	MH	0.37	0.18	0.00	0.25	0.41	0.48	0.70
	PBS	0.35	0.14	0.00	0.24	0.34	0.44	0.67
LLM title semantic similarity	BAM	0.32	0.15	0.00	0.21	0.33	0.40	0.72
	CS	0.33	0.13	0.01	0.25	0.34	0.42	0.67
	EMS	0.32	0.12	0.00	0.23	0.32	0.40	0.65
	MH	0.31	0.16	0.00	0.19	0.31	0.45	0.57
	PBS	0.32	0.13	0.00	0.23	0.31	0.41	0.66

Table 2 Descriptive statistics for highly cited titles during pre-LLM.

Similarity	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title semantic similarity	BAM	0.40	0.18	0.00	0.27	0.39	0.52	0.93
	CS	0.41	0.16	0.01	0.31	0.44	0.50	0.75
	EMS	0.36	0.14	0.00	0.26	0.35	0.44	0.71
	MH	0.43	0.14	0.00	0.33	0.43	0.53	0.69
	PBS	0.36	0.17	0.00	0.25	0.37	0.50	0.74
LLM title semantic similarity	BAM	0.34	0.17	0.00	0.22	0.33	0.46	0.73
	CS	0.36	0.15	0.01	0.26	0.36	0.46	0.71
	EMS	0.32	0.14	0.00	0.23	0.33	0.41	0.75
	MH	0.35	0.14	0.00	0.27	0.36	0.47	0.73
	PBS	0.32	0.14	0.00	0.24	0.33	0.42	0.65

compared to LLM title semantic scores (0.32), this trajectory is also maintained in pre-LLM where the computed mean score are 0.39 and 0.34 for LLM era and the pre-LLM, respectively. In addition, the standard deviation across all titles ranged from 0.12 to 0.17 in the LLM era compared to the pre-LLM which ranged from 0.14 to 0.18.

Also, the results in Tables 1 and 2 indicated that both eras exhibited similar minimum values of 0.00, with a varied maximum valve of 0.83 and 0.93 for LLM era and pre-LLM, respectively. Authors' titles have a marginally higher maximum score compared to the generated titles. The quartile ranges (25%, 50%, and 75%) show consistent patterns across disciplines between the two eras, with slightly lower values observed in the LLM era, with MH of LLM title scoring a significant low score of 0.19. The 75th percentile for authors title in pre-LLM, ranging from 0.44 to 0.53 is higher than all other categories on the same scale.

Tables 3 and 4 present the descriptive statistics of semantic similarity scores for titles of low-cited articles in both the LLM era and the pre-LLM era. Across various titles, the mean ranged from 0.36 to 0.44 in the LLM era, whereas in the pre-LLM era, it ranged from 0.35 to 0.40, with PBS (authors title, LLM era) exhibiting the highest mean score. The mean scores of authors' titles' semantic similarity in the LLM era (0.41) were slightly higher compared to LLM-generated title semantic scores (0.39), a trend that persisted to a lesser extent in the pre-LLM era with mean scores of 0.38 and 0.37, respectively. Moreover, the standard deviation across all titles ranged from 0.13 to 0.16 in the LLM era, similar to the range observed in the pre-LLM era.

Furthermore, the findings in Tables 3 and 4 revealed that both eras demonstrated comparable minimum values of 0.00, while demonstrating varied maximum values of 1.00 (in LLM-generated titles) and 0.78 for LLM era and pre-LLM, respectively. LLM titles (in LLM era) shows a higher maximum score compared to the Authors titles. Furthermore, the quartile ranges (25%, 50%, and 75%) demonstrated consistent patterns across disciplines in both eras, albeit with slightly lower

Table 3 Descriptive statistics for low cited titles during LLM era.

Similarity	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title semantic similarity	BAM	0.36	0.14	0.04	0.32	0.41	0.47	0.76
	CS	0.39	0.16	0.00	0.30	0.39	0.51	0.76
	EMS	0.40	0.14	0.00	0.33	0.41	0.51	0.75
	MH	0.41	0.15	0.00	0.29	0.42	0.52	0.72
	PBS	0.40	0.15	0.00	0.34	0.45	0.54	0.82
LLM title semantic similarity	BAM	0.40	0.13	0.02	0.26	0.38	0.44	0.68
	CS	0.40	0.14	0.16	0.30	0.38	0.48	0.74
	EMS	0.41	0.14	0.10	0.32	0.38	0.48	1.00
	MH	0.41	0.14	0.11	0.32	0.41	1.48	1.00
	PBS	0.44	0.14	0.00	0.30	0.43	0.49	0.71

Table 4 Descriptive statistics for low cited titles during pre-LLM.

Similarity	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title semantic similarity	BAM	0.38	0.15	0.02	0.31	0.41	0.52	0.74
	CS	0.37	0.16	0.03	0.25	0.38	0.47	0.78
	EMS	0.36	0.13	0.08	0.29	0.39	0.46	0.70
	MH	0.35	0.15	0.00	0.29	0.39	0.50	0.67
	PBS	0.40	0.16	0.00	0.31	0.40	0.50	0.77
LLM title semantic similarity	BAM	0.40	0.14	0.01	0.30	0.38	0.48	0.63
	CS	0.37	0.14	0.68	0.28	0.37	0.45	0.73
	EMS	0.38	0.13	0.62	0.27	0.37	0.45	0.70
	MH	0.38	0.16	0.00	0.26	0.37	0.46	0.69
	PBS	0.40	0.15	0.01	0.30	0.42	0.50	0.73

values observed in the pre-LLM. Particularly noteworthy is the case of CS of author title, which scored lower at 0.25. In addition, the 75th percentile for the generated titles in LLM era, ranging from 0.44 to 1.48, surpassed all other categories on the same scale.

4.2 Readability

Tables 5 and 6 illustrate the descriptive statistics of readability scores for titles of highly cited articles in both the LLM era and the pre-LLM. From Tables 5 and 6, the mean readability scores across all titles ranged from 11.60 to 15.31 in the LLM era, contrasting with the pre-LLM era, where the range extended from 12.42 to 16.30, with BAM and PBS boasting the highest mean scores. The computed mean scores of generated titles' readability in the LLM era (13.90) exceeded those of the original titles' readability (12.80), a trend that persisted in the pre-LLM era, with mean scores of 15.38 and 13.51 for the LLM era and the pre-LLM, respectively. Moreover, the standard deviation across

Table 5 Descriptive statistics for highly cited titles during LLM era across five domains.

Readability	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title readability	BAM	13.22	4.37	0.9	11.1	13.7	16.2	24.5
	CS	11.60	4.90	−1.9	8.7	11.7	15.1	22.1
	EMS	13.53	5.54	0.9	9.6	12.9	16.3	32.4
	MH	13.24	5.94	0.5	10.3	13.1	15.6	44.2
	PBS	12.40	4.60	0.9	9.6	12.3	15.1	24.5
LLM title readability	BAM	13.35	5.09	−15.7	10.7	13.9	16.3	24.1
	CS	14.40	4.90	1.3	11.5	14.3	17.0	27.7
	EMS	15.31	4.34	6.4	12.3	14.7	18.2	26.5
	MH	12.62	3.95	5.2	9.9	12.3	15.1	22.5
	PBS	13.80	4.50	5.2	10.3	13.5	16.8	28.8

Table 6 Descriptive statistics for highly cited titles during pre-LLM across five domains.

Readability	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title readability	BAM	13.87	5.55	−1.20	11.10	14.30	17.80	28.10
	CS	12.42	4.58	1.30	9.60	12.70	14.78	26.50
	EMS	12.58	6.19	−2.30	8.00	12.30	16.60	29.20
	MH	14.21	4.76	3.70	11.80	14.30	17.80	30.40
	PBS	14.47	6.89	2.10	10.70	13.90	17.40	55.60
LLM title readability	BAM	16.30	4.81	5.60	13.40	16.05	19.20	28.10
	CS	14.46	3.64	7.20	11.50	14.70	16.30	24.10
	EMS	15.40	4.61	4.80	11.90	15.10	18.20	36.30
	MH	14.46	3.86	4.80	12.30	14.30	16.30	25.70
	PBS	16.30	4.58	8.00	12.70	15.90	19.40	28.80

all titles ranged from 3.95 to 5.95 in the LLM era, while in the pre-LLM, it ranged from 3.64 to 6.89.

The insights from Table 5 unveil that the original titles exhibit a minimum value of −1.9, observed in the CS category during the LLM era, contrasting with the minimum value of −2.30 recorded in EMS during the pre-LLM era (Table 6). There are varied maximum values, with EMS in original titles reaching 32.40 for the LLM era and PBS in original titles reaching 56.60 for the pre-LLM. Remarkably, the original titles in the pre-LLM era display a higher maximum readability score compared to the generated titles. The quartile ranges (25%, 50%, and 75%) display consistent patterns across disciplines in both eras, albeit with slightly lower values observed in the pre-LLM. Particularly is the case of EMS in the original title, which scored lower at 8.00. Furthermore, the 75th percentile for the generated titles in the LLM era, ranging from 16.30 to 19.40, surpasses all other categories on the same scale, with an average of 17.88.

Tables 7 and 8 display the descriptive statistics regarding the readability scores for titles of low-cited articles in both the LLM era and the pre-LLM. According to Tables 7 and 8, the mean readability scores across all titles varied from 13.29 (PBS, original title) to 17.40 (EMS, generated title) in the LLM era (Table 7). This contrasts with the pre-LLM era, where the range extended from 12.41 (MH, original title) to 18.00, with EMS (generated title) also showing the highest mean scores (Table 8). The computed mean scores of generated titles' readability in the LLM era (16.20) were higher than those of the original titles' readability (14.66), a trend that persisted in the pre-LLM era, with mean scores of 16.50 and 14.20 for the LLM era and the pre-LLM, respectively. Furthermore, the standard deviation across all titles ranged from 3.73 to 5.66 in the LLM era, while in the pre-LLM, it ranged from 3.30 to 6.59.

The findings depicted in Table 7 reveal that the generated titles show a minimum value of −15.70, observed in the PBS category during the LLM era, distinct with the minimum value of −2.30 recorded in article titles within PBS during the pre-LLM era (Table 8). There are varied maximum values, with MH in

Table 7 Descriptive statistics for low cited titles during LLM era across five domains.

Readability	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title readability	BAM	14.47	4.67	2.10	11.90	14.70	16.70	31.20
	CS	14.43	4.66	2.90	11.10	14.70	18.20	30.80
	EMS	16.28	4.76	6.00	13.10	16.20	19.00	38.30
	MH	14.83	5.66	1.30	12.30	14.70	17.40	44.30
	PBS	13.29	4.43	3.30	10.00	12.90	16.60	24.10
LLM title readability	BAM	16.79	3.74	9.50	14.28	16.60	19.10	26.80
	CS	15.60	3.73	6.80	12.70	15.50	17.40	25.70
	EMS	17.40	3.97	7.20	15.10	17.40	20.00	34.00
	MH	15.36	4.05	5.00	12.70	15.40	18.20	25.30
	PBS	15.68	5.24	−15.70	13.10	15.80	18.20	28.00

Table 8 Descriptive statistics for low cited titles during pre-LLM across five domains.

Readability	Knowledge area	Mean	Standard deviation	Min	25% datasets	50% datasets (Median)	75% datasets	Max
Authors' title readability	BAM	14.25	5.02	−1.20	11.40	14.45	17.50	24.10
	CS	13.45	4.76	0.90	10.30	13.10	16.30	26.50
	EMS	16.62	4.35	6.00	13.90	16.20	18.80	30.80
	MH	12.41	4.64	0.50	10.30	12.70	15.05	24.50
	PBS	14.26	6.59	−2.30	10.30	13.90	17.80	55.60
LLM title readability	BAM	16.50	4.16	6.00	13.40	16.20	19.00	30.00
	CS	16.47	3.89	3.70	14.30	16.20	18.60	27.30
	EMS	18.00	3.30	10.30	15.50	17.80	20.35	27.20
	MH	14.95	4.21	5.60	12.10	15.00	17.80	26.50
	PBS	16.58	3.68	8.00	13.90	16.60	19.00	28.10

original titles reaching 44.30 for the LLM era and PBS in original titles reaching 56.60 for the pre-LLM. The original titles in the pre-LLM exhibit a higher maximum readability score compared to the generated titles. The quartile ranges (25%, 50%, and 75%) present consistent patterns across disciplines in both eras, with slightly lower values observed in the pre-LLM. Particularly are the case of CS, MH, and PBS in the original title, which scored lower at 10.30. The 75th percentile for the generated titles in the pre-LLM, ranging from 17.80 to 20.35, surpasses all other categories on the same scale, with an average of 18.58.

4.3 Similarity distance

Tables 9 and 10 provide insights into the percentage distribution of Levenshtein distances between highly cited authors' titles and LLM-generated titles for highly cited articles during the LLM era and the pre-LLM era. In the LLM era (Table 9), the results reveal that, on average, 89.9% of scores fall within the moderate distance (0.21–0.30) to very low distance (0.41–1.00) ranges between highly cited authors' titles and LLM-generated titles. Particularly interesting is the distribution of distance scores in the PBS domain, where there is a significant concentration of scores from high similarity to low similarity compared to other domains. In PBS, the distance scores are concentrated within the moderate (0.11–0.20) to low (0.31–0.40) ranges. Only 2.9% of the scores fall under the category of very high similarity (0.00–0.10).

In the pre-LLM era (Table 10), the results reveal that, on average, 94.3% of scores fall within the moderate distance (0.21–0.30) to very low distance (0.41–1.00) ranges between highly cited authors' titles and LLM-generated titles. Notably, EMS demonstrates higher percentages (56.88%) in the very low similarity (0.41–1.00) range in the pre-LLM era compared to the LLM era (34.58%, Table 9). The prevalence of very high to high similarity scores was more pronounced in the LLM era, with an average percentage distribution score of 10.1%, contrasting with 5.71% in the pre-LLM era. The distance scores primarily concentrate between moderate to very low similarity, with only 1.76% of the scores falling under the category of very high similarity (0.00–0.10) ranges.

Tables 11 and 12 provide findings of the percentage distribution of Levenshtein distances between low cited authors' titles and LLM-generated titles for during the LLM era and the pre-LLM era. In the LLM era (Table 11), the results indicate moderate (0.21–0.30) to very low (0.41–1.00) percentages of distance between titles,

Table 9 Levenshtein distance distribution for highly cited titles during the LLM era. (%)

Range	BAM	CS	EMS	MH	PBS
0.00–0.10	0.94	0.00	1.87	0.00	0.09
0.11–0.20	4.72	6.48	8.41	6.36	21.65
0.21–0.30	27.36	29.63	24.30	22.73	30.93
0.31–0.40	31.13	35.19	30.84	39.09	30.93
0.41–1.00	35.85	28.70	34.58	31.82	16.40

Note: Very high similarity: 0.00–0.10. High similarity: 0.11–0.20. Moderate similarity: 0.21–0.30. Low similarity: 0.31–0.40. Very low similarity: 0.41–1.00.

Table 10 Levenshtein distance distribution for highly cited titles during the pre-LLM. (%)

Range	BAM	CS	EMS	MH	PBS
0.00–0.10	0.93	0.00	0.00	0.00	0.86
0.11–0.20	7.18	3.33	2.56	5.05	8.62
0.21–0.30	25.00	20.00	13.04	23.24	22.42
0.31–0.40	29.51	35.83	27.52	34.34	30.17
0.41–1.00	37.38	40.84	56.88	37.37	37.93

Note: Very high similarity: 0.00–0.10. High similarity: 0.11–0.20. Moderate similarity: 0.21–0.30. Low similarity: 0.31–0.40. Very low similarity: 0.41–1.00.

Table 11 Levenshtein distance distribution for low cited titles during the LLM era. (%)

Range	BAM	CS	EMS	MH	PBS
0.00–0.10	0.00	0.85	0.00	1.86	0.93
0.11–0.20	2.68	10.26	2.66	9.35	11.11
0.21–0.30	24.11	21.37	20.35	25.23	35.19
0.31–0.40	41.07	27.35	31.86	30.84	26.85
0.41–1.00	32.14	40.17	45.13	32.72	25.92

Note: Very high similarity: 0.00–0.10. High similarity: 0.11–0.20. Moderate similarity: 0.21–0.30. Low similarity: 0.31–0.40. Very low similarity: 0.41–1.00.

Table 12 Levenshtein distance distribution for low cited titles during the pre-LLM. (%)

Range	BAM	CS	EMS	MH	PBS
0.00–0.10	0.00	1.82	0.00	3.48	0.00
0.11–0.20	4.72	10.91	2.75	5.22	4.57
0.21–0.30	20.75	34.55	24.77	25.22	22.94
0.31–0.40	35.85	28.18	41.28	32.17	27.52
0.41–1.00	38.68	24.54	31.20	33.91	44.97

Note: Very high similarity: 0.00–0.10. High similarity: 0.11–0.20. Moderate similarity: 0.21–0.30. Low similarity: 0.31–0.40. Very low similarity: 0.41–1.00.

with a mean score of 92.0%. Particularly is the distribution of distance scores in CS and PBS, where there is a significant concentration of scores ranging from high similarity to low similarity compared to other domains, with the distance scores mainly concentrated between moderate to very low. Only 3.6% of the scores fall under the category of very high similarity (0.00–0.10), which is slightly higher compared to that of highly cited titles during the LLM era.

In the pre-LLM era (Table 12), there are higher proportions of low (0.31–0.40) to very low similarity (0.41–1.00), averaging at 93.3%. PBS demonstrates elevated proportions (44.97%) of very low similarity (0.41–1.00) in the pre-LLM era compared to the LLM era. The prevalence of high similarity scores was more prominent in the LLM era, with an average score of 7.2%, contrasting with 5.6% in pre-LLM. The distance scores mainly concentrate between moderate to very low similarity, with only 1.0% of the scores falling under the category of very high similarity (0.00–0.10).

4.4 Statistical test

Table 13 presents ANOVA tests comparing the readability scores of authors' titles and LLM-generated titles between the LLM era and the pre-LLM era, for both highly cited and low cited articles. In the highly cited LLM era, the ANOVA tests reveal a significant difference ($p < 0.05$) in the readability scores of LLM-generated titles, a trend mirrored in the highly cited pre-LLM where the p-value also falls below the specified alpha value. For the titles of low cited articles in the LLM era, significant differences are observed between authors' titles and LLM-generated titles ($p < 0.05$).

The ANOVA tests comparing the semantic similarity scores of authors' titles and LLM-generated titles between the LLM era and the pre-LLM era, for both highly cited and low cited articles are presented in Table 14. The ANOVA tests reveal a significant difference ($p < 0.05$) in the scores of authors' titles.

5 Discussion and Conclusion

This study aimed to explore the semantic alignment between titles-abstracts produced by human authors and those generated by LLMs in five different research domains: BAM, CS, EMS, MH, and PBS. It analyzed semantic alignment, readability, and similarity scores, subjecting them to statistical tests to detect any chance differences.

In highly cited articles during the LLM era, both authors and LLMs displayed comparable mean similarity scores, with slight quartile variations. This suggests LLMs perform similarly to human authors in generating titles aligning closely with abstracts. The nuanced quartile differences indicate consistent performance across different dataset percentiles, affirming LLMs' ability to produce high-quality titles. In pre-LLM, original titles had slightly higher similarity scores than LLMs, indicating consistent alignment with abstracts. This reflects LLMs' advancement in generating closely aligned titles.

Table 13 ANOVA tests between the readability scores.

Period	Authors' title readability			LLM title readability		
	F	*p*-valve	*F* crit	*F*	*p*-valve	*F* crit
Highly cited LLM era	2.3383	0.0544	0.2899	5.0632	0.0005	2.3899
Highly cited pre-LLM	2.3827	0.0506	2.3899	4.0067	0.0033	2.3899
Low cited LLM era	4.9708	0.0006	2.3899	4.6365	0.0011	2.3899
Low cited pre-LLM	8.9821	5.2183	2.3899	7.6590	5.4131	2.3899

Note: Significant at p-value < 0.05.

Table 14 ANOVA tests between the semantic similarity scores.

Period	Authors' title semantic similarity			LLM title semantic similarity		
	F	*p*-valve	*F* crit	*F*	*p*-valve	*F* crit
Highly cited LLM era	0.4415	0.7234	2.6274	0.1372	0.9685	2.3899
Highly cited pre-LLM	4.0670	0.0030	2.3899	1.4815	0.2066	2.3899
Low cited LLM era	1.0365	0.3878	2.3899	1.2909	0.2725	2.3899
Low cited pre-LLM	1.0201	0.3963	2.3899	2.1099	0.0784	2.3899

Note: Significant at p-value < 0.05.

Before LLMs, human authors consistently crafted titles reflecting abstract essence. The closeness in mean similarity scores underscores LLMs' ability to align with reference abstracts, supporting human-authored titles. This highlights LLMs' effectiveness in understanding academic content and suggests historical AI language generation progression. In low cited articles, LLM titles slightly outperformed authors in both LLM era and pre-LLM, suggesting better alignment with abstracts. This implies LLMs capture key concepts more succinctly than human authors, providing relevant representations.

In highly cited scholarly publications during both LLM era and pre-LLM, there is a trend showing slightly higher readability in LLM-generated titles, evidenced by higher mean scores compared to original titles. Original titles tend to be simpler to read, with variations in readability across disciplines affected by LLMs. Consistent quartile patterns suggest that LLMs may enhance overall readability, especially in LLM era with notably higher percentiles. For low-cited articles, LLM-generated titles consistently had higher readability scores in both eras, indicating a consistent trend. Variations in standard deviation suggest differences in readability score variability between eras. These findings suggest that LLMs have impacted the readability of titles for low-cited articles, potentially enhancing accessibility and comprehension. Differences between original and generated titles highlight the benefits of using LLMs in scholarly communication to improve readability and engagement.

For highly cited articles during LLM era and pre-LLM, the Levenshtein distance between titles indicates moderate to low similarity levels in LLM era, with minimal very high similarity scores. This suggests moderate title similarity during this era. The pre-LLM shows a higher percentage of low to very low similarity, indicating greater title diversity. LLM era exhibits a higher prevalence of high similarity scores, suggesting more uniform titles by LLMs, while pre-LLM suggests wider title variations. For low-cited articles, LLM era demonstrates a relatively high mean score of 92.0%, indicating significant title dissimilarity and potential variation in similarity levels. The pre-LLM shows an even higher average score of 93.3%, implying greater dissimilarity and potential lack of uniformity in title composition. These differences in similarity patterns between eras highlight the influence of historical context and evolving research practices on scholarly communication title composition.

The ANOVA tests comparing the readability scores of authors' titles and LLM-generated titles between the LLM era and the pre-LLM era, for both highly cited and low-cited articles provides crucial insights. The results indicate significant differences ($p < 0.05$) in the readability scores of LLM-generated titles compared to authors' titles in both highly cited LLM era and pre-LLM, suggesting a notable impact of LLMs on title readability. The null hypothesis is rejected in favor of the alternative hypothesis, indicating a statistically significant difference between the two groups. For low-cited articles in the LLM era, significant differences are observed between authors' titles and LLM-generated titles, indicating a consistent trend. In the ANOVA tests comparing semantic similarity measure, a significant difference ($p < 0.05$) in the scores of authors' titles is revealed in highly cited pre-LLM, underscoring potential shifts in semantic composition influenced by the advent of LLMs. The null hypothesis is similarly rejected in this case, highlighting the statistical significance of the observed differences. These findings highlight the substantial influence of LLMs on the readability and semantic similarity of scholarly titles across different titles and eras, suggesting important implications for scholarly communication and research practices.

This study contends that there is insufficient evidence to suggest that LLM outperforms human authors in article title generation or articles from the LLM era demonstrating a marked difference in semantic richness and readability compared to those from the pre-LLM. Instead, it asserts that LLM is a valuable tool and can assist researchers in generating titles. The findings resonate with the existing literature on AI-generated text and its implications for academic writing, aligning with the works of Refs. [6, 49] and who concluded that technology significantly improve students' writing competence hence reducing the workload of professors in supporting scientific writing. The observed similarities in semantic similarity scores between titles authored by humans and those generated by LLM are in accordance with the findings of Refs. [18, 50], who concluded that LLM feedback is useful in the manuscript improvement process and can

sometimes bring up novel points not covered by humans. Furthermore, the readability analyses conducted in this study align with previous research emphasizing the importance of complementing or shaping user-generated text with AI-generated text, particularly in academic contexts, as highlighted by Ref. [51], who demonstrated the effectiveness of LLM in accurately measuring the typicality of text documents in specific concepts, offering a significant improvement over previous state-of-the-art methods that required extensive model training. This alignment with existing literature strengthens the robustness of our study's contributions to the understanding of the interplay between human-authored and LLM-generated titles in academic publications. This research contributes to the understanding of LLM-generated text in academic writing, shedding light on the capabilities of LLMs in generating titles across citation categories. The findings hold significance for researchers, authors, and publishers, recommending the integration of LLM into academic writing processes to enhance efficiency and creativity. Authors could leverage LLM's advanced language capabilities to draft and refine abstracts, a crucial component that succinctly encapsulates the essence of their research. By generating initial drafts or suggesting improvements, LLM could streamline the abstract writing process, ensuring clarity and coherence while maintaining the author's original intent. This approach could save considerable time, allowing researchers to focus on refining and validating the content. Moreover, writing titles last, with LLM's assistance, ensures that they are reflective of the finalised abstract and core research themes, potentially increasing the impact and accessibility of the academic work. This paradigm shift in academic writing, with AI as a collaborative tool, could lead to more effective communication of complex ideas and findings, benefiting the broader scientific community.

The limitations of this research centred on the focus of comparing the semantic match of titles and abstracts of articles within the research areas we assumed might be influenced by the implementation of AI. This approach does not provide a comprehensive assessment of the articles. Therefore, it is advisable to replicate the study in different research contexts, for instance, the use of research methodology sections of articles instead of "abstract" for a more thorough examination. In future research, one avenue to explore involves empowering LLMs to generate an ontology that functions as a reference point for all journal titles. This ontology can also facilitate the linking of other pertinent words, aligning with individual literature reviews of related works. This capability is poised to streamline the integration of highly relevant information within the research landscape. Such an ontology, akin to news feeds, offers a valuable reference point for researchers navigating the expansive domain of academic literature. The cohesion achieved through the consistent use of closely related words across projects further enriches the interconnectedness of this knowledge network.

References

[1] H. Zohny, J. McMillan, and M. King, Ethics of generative AI, *Journal of Medical Ethics*, vol. 49, no. 2, pp. 79–80, 2023.

[2] M. Z. M. Hurmuz, S. M. Jansen-Kosterink, I. Flierman, S. del Signore, G. Zia, S. del Signore, and B. Fard, Are social robots the solution for shortages in rehabilitation care? Assessing the acceptance of nurses and patients of a social robot, *Comput. Hum. Behav. Artif. Hum.*, vol. 1, no. 2, p. 100017, 2023.

[3] C. R. Yeh, C. J. Wong, W. V. Chang, and C. S. Lai, Labor displacement in artificial intelligence era: A systematic literature review, *Taiwan Journal of East Asian Studies*, vol. 17, no. 2, pp. 25–75, 2020.

[4] Z. Tekic and J. Füller, Managing innovation in the era of AI, *Technol. Soc.*, vol. 73, p. 102254, 2023.

[5] S. Salter, P. L. Teh, and R. Hebblewhite, Human-created and AI-generated text: What's left to uncover? in *Intelligent Computing*, K. Arai, ed. Cham. Switzerland: Springer, 2024, pp. 74–80.

[6] M. Jakesch, M. French, X. Ma, J. T. Hancock, and M. Naaman, AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness, in *Proc. 2019 CHI Conf. Human Factors in Computing Systems*, Glasgow, UK, 2019, pp. 1−13.

[7] P. M. Mah, I. Skalna, and J. Muzam, Natural language processing and artificial intelligence for enterprise management in the era of industry 4.0, *Appl. Sci.*, vol. 12, no. 18, p. 9207, 2022.

[8] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation, *Eng. Appl. Artif. Intell.*, vol. 110, p. 104673, 2022.

[9] K. Somoray and D. J. Miller, Providing detection strategies to improve human detection of deepfakes: An experimental study, *Comput. Hum. Behav.*, vol. 149, p. 107917, 2023.

[10] N. Serki and S. Bolkan, The effect of clarity on learning: Impacting motivation through cognitive load, *Commun. Educ.*, vol. 73, no. 1, pp. 29–45, 2024.

[11] B. L. Frye, Should using an AI text generator to produce academic writing be plagiarism? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4292283, 2022.
[12] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, ChatGPT outperforms humans in emotional awareness evaluations, *Front. Psychol.*, vol. 14, p. 1199058, 2023.
[13] A. Pegoraro, K. Kumari, H. Fereidooni, and A. R. Sadeghi, To ChatGPT, or not to ChatGPT: That is the question! arXiv preprint arXiv: 2304.01487, 2023.
[14] N. Rodrigues, M. Burch, L. Di Silvestro, and D. Weiskopf, A visual analytics approach for word relevances in multiple texts, in *Proc. 21st Int. Conf. Information Visualisation* (*IV*), London, UK, 2017, pp. 1–7.
[15] M. S. Tullu, Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key, *Saudi J. Anaesth.*, vol. 13, no. Suppl1, pp. S12–S17, 2019.
[16] H. H. Thorp, ChatGPT is fun, but not an author, *Science*, vol. 379, no. 6630, p. 313, 2023.
[17] T. Holtgraves, Implicit communication of emotions via written text messages, *Computers in Human Behavior Reports*, vol. 7, p. 100219, 2022.
[18] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, et al., Can large language models provide useful feedback on research papers? A large-scale empirical analysis, arXiv preprint arXiv: 2310.01783, 2023.
[19] P. L. Teh, P. Rayson, I. Pak, S. Piao, J. S. Y. Ho, A. Moore, and Y. Cheah, Textual variations affect human judgements of sentiment values, *Electron. Commer Res. Appl.*, vol. 53, p. 101149, 2022.
[20] S. Jiang, R. E. Robertson, and C. Wilson, Reasoning about political bias in content moderation, in *Proc. 34th AAAI Conference on Artificial Intelligence* (*AAAI-20*), New York, NY, USA, 2020, pp. 13669–13672.
[21] Q. Zhou, B. Li, L. Han, and M. Jou, Talking to a bot or a wall? How chatbots vs. human agents affect anticipated communication quality, *Computers in Human Behavior*, vol. 143, p. 107674, 2023.
[22] M. Khalil and E. Er, Will ChatGPT get you caught? Rethinking of plagiarism detection, in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, eds. Cham, Switzerland: Springer, 2023, pp. 475–487.
[23] K. Cao, C. Chen, S. Baltes, C. Treude, and X. Chen, Automated query reformulation for efficient search based on query logs from stack overflow, in *Proc. 43rd Int. Conf. Software Engineering*, Madrid, Spain, 2021, pp. 1273–1285.
[24] B. D. Lund and T. Wang, Chatting about ChatGPT: How may AI and GPT impact academia and libraries? https://ssrn.com/abstract=4333415, 2023.
[25] S. Syed and M. Spruit, Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation, in *Proc. IEEE Int. Conf. Data Science and Advanced Analytics* (*DSAA*), Tokyo, Japan, 2017, pp. 165–174.
[26] V. P. E. Reyes, Exploring the use of the phoneme frequency scale method in determining word difficulty levels and readability scores, in *Proc. 2019 7th Int. Conf. Information and Education Technology*, Aizu-Wakamatsu, Japan, 2019, pp. 284–288.
[27] S. Albitar, S. Fournier, and B. Espinasse, An effective TF/IDF-based text-to-text semantic similarity measure for text classification, in *Web Information Systems Engineering—WISE 2014*, B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, eds. Cham, Switzerland: Springer, 2014, pp. 105–114.
[28] X. Ao, X. Yu, D. Liu, and H. Tian, News keywords extraction algorithm based on TextRank and classified TF-IDF, in *Proc. Int. Wireless Communications and Mobile Computing* (*IWCMC*), Limassol, Cyprus, 2020, pp. 1364–1369.
[29] Y. Li and B. Shen, Research on sentiment analysis of microblogging based on LSA and TF-IDF, in *Proc. 3rd IEEE Int. Conf. Computer and Communications* (*ICCC*), Chengdu, China, 2017, pp. 2584–2588.
[30] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection, arXiv preprint arXiv: 2301.07597, 2023.
[31] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, AI vs. human—Differentiation analysis of scientific content generation, arXiv preprint arXiv: 2301.10416, 2023.
[32] R. Saptono, H. Prasetyo, and A. Irawan, Combination of cosine similarity method and conditional probability for plagiarism detection in the thesis documents vector space model, *Journal of Telecommunication*, vol. 10, nos. 2–4, pp. 139–143, 2018.
[33] M. Ali Fauzi, D. C. Utomo, B. D. Setiawan, and E. S. Pramukantoro, Automatic essay scoring system using N-gram and cosine similarity for gamification based E-learning, in *Proc. Int. Conf. Advances in Image Processing*, Bangkok, Thailand, 2017, pp. 151–155.
[34] S. Tata and J. M. Patel, Estimating the selectivity of TF-IDF based cosine similarity predicates, *SIGMOD Rec.*, vol. 36, no. 4, pp. 75–80, 2007.
[35] S. T. Yuan and J. Sun, Ontology-based structured cosine similarity in speech document summarization, in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence* (*WI'04*), Beijing, China, 2004, pp. 508–513.
[36] K. Jain, S. Gandhi, S. Singhal, and S. Rajput, Semantic image captioning using cosine similarity ranking with semantic search, in *Proc. 2023 15th Int. Conf. Contemporary Computing*, Noida, India, 2023, pp. 220–223.
[37] S. Akter, A. E. Siam, K. M. M. Monir, M. H. K. Mehedi, and A. A. Rasel, Bengali movie recommendation system using K nearest neighbor and cosine similarity, in *Proc. 2023 9th Int. Conf. Computer Technology Applications*, Vienna, Austria, 2023, pp. 25–29.
[38] L. S. Lo, The art and science of prompt engineering: A new literacy in the information age, *Internet Ref. Serv. Q.*, vol. 27, no. 4, pp. 203–210, 2023.
[39] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.
[40] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, Responsible AI—Two frameworks for ethical design practice, *IEEE Trans. Technol. Soc.*, vol. 1 , no. 1 , pp.

34–47, 2020.
[41] I. Poola, Overcoming ChatGPTs inaccuracies with pre-trained AI prompt engineering sequencing process, *International Journal of Technology and Emerging Sciences*, vol. 3, no. 3, pp. 16–19. 2023.
[42] A. W. Harzing, *The Publish or Perish Book: Your Guide to Effective and Responsible Citation Analysis*. Melbourne, Australia: Tarma Software Research Pty Limited, 2010.
[43] H. Huang, D. Zhu, and X. Wang, Evaluating scientific impact of publications: Combining citation polarity and purpose, *Scientometrics*, vol. 127, no. 9, pp. 5257–5281, 2022.
[44] B. I. Hutchins, X. Yuan, J. M. Anderson, and G. M. Santangelo, Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level, *PLoS Biol.*, vol. 14, no. 9, p. e1002541, 2016.
[45] D. W. Aksnes, L. Langfeldt, and P. Wouters, Citations, citation indicators, and research quality: An overview of basic concepts and theories, *SAGE Open*, vol. 9, no. 1, p. 2158244019829575, 2019.
[46] M. R. Dougherty and Z. Horne, Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences, *R. Soc. Open Sci.*, vol. 9, no. 8, p. 220334, 2022.
[47] A. P. Akella, H. Alhoori, P. R. Kondamudi, C. Freeman, and H. Zhou, Early indicators of scientific impact: Predicting citations with altmetrics, *Journal of Informatics*, vol. 15, no. 2, p. 101128, 2021.
[48] F. Rahutomo, T. Kitasuka, and M. Aritsugi, Semantic cosine similarity, in *Proc. 7th International Student Conference on Advanced Science and Technology ICAST*, Seoul, Republic of Korea, 2012, p. 1.
[49] T. Schmohl, A. Watanabe, N. Fröhlich, and D. Herzberg, How artificial intelligence can improve the academic writing of students, https://conference.pixel-online.net/library_scheda.php?id_abs=4769, 2020.
[50] X. Yu, Y. Qi, K. Chen, G. Chen, X. Yang, P. Zhu, W. Zhang, and N. Yu, GPT paternity test: GPT generated text detection with GPT genetic inheritance, arXiv preprint arXiv: 2305.12519, 2023.
[51] G. Le Mens, B. Kovács, M. T. Hannan, and G. Pros, Uncovering the semantics of concepts using GPT-4, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 120, no. 49, p. e2309350120, 2023.

Chukwudi Festus Uwasomba received the BSc degree in information technology and business information systems (First Class Honours) from Middlesex University, UK in 2015, the MSc degree in business information systems management (Distinction) from Middlesex University, UK in 2016, and the PhD degree in computing from Sunway University, Malaysia in 2023, specializing in natural language processing and ontologies during his PhD. He was awarded the Prestigious Sunway University PhD Studentship. He is a research associate at The Open University, UK. He led the Polifonia CHILD pilot project, a European Union (EU)-funded initiative under Horizon 2020, pioneering innovative methods for curating documentary evidence in humanity databases. He also played a key role in SPICE (another EU project), where he led the development of an intelligent agent for privacy preservation. Currently, he leads research in the Agile Research Network, employing both qualitative and quantitative research approaches to understand the transformative effects of agile practices and methodologies on organisational culture. He has served as a program committee member for several reputable conferences and has authored and co-authored numerous publications. His research primarily focuses on natural language processing, advanced computational techniques, and enhanced agility.

Phoey Lee Teh received the PhD degree in management information system from Universiti Putra Malaysia, Malaysia in 2011, and the master degree in information technology from University Science Malaysia, Malaysia in 2003. She is an academic at Wrexham University, UK. She was a full professor expertise social computing at Sunway University, Malaysia. She has been a PC chair, reviewer, track chair, and post-publication chair at several conferences. She has authored and co-authored up to 80 research articles, book chapters, and conference papers of which 60 of them are Scopus-indexed. She has also published a book. She holds a PGCert in Academic Practice and is honored as a senior fellow of Advance Higher Education (HE) and senior member of IEEE.