INTERNET REFERENCE SERVICES QUARTERLY
https://doi.org/10.1080/10875301.2024.2426793

Routledge
Taylor & Francis Group

"ChatGPT 4.0 Ghosted Us While Conducting Literature Search:" Modeling the Chatbot's Generated Non-Existent References Using Regression Analysis

Dharel P. Acut[a], Nolasco K. Malabago[b], Elesar V. Malicoban[c], Narcisan S. Galamiton[d] and Manuel B. Garcia[e,f]

[a]College of Education, Cebu Technological University, Cebu City, Philippines; [b]Graduate School, Cebu Technological University, Cebu City, Philippines; [c]College of Education, Mindanao State University - Iligan Institute of Technology, Iligan City, Philippines; [d]College of Computer, Information and Communication Technology, Cebu Technological University-Main Campus, Cebu City, Philippines; [e]College of Education, University of the Philippines Diliman, Quezon City, Philippines; [f]Educational Innovation and Technology Hub, FEU Institute of Technology, Manila, Philippines

ABSTRACT

The integration of AI technologies like ChatGPT has transformed academic research, yet substantial gaps exist in understanding the implications of AI-generated non-existent references in literature searches. While prior studies have predominantly focused on medical and geography fields using descriptive statistics, a systematic investigation into ChatGPT 4.0's effectiveness in generating accurate references within the realm of science and technology education remains unexplored, highlighting a significant dearth of research in this critical area. This study, therefore, investigates the reliability of AI-generated references in academic writing utilizing ChatGPT 4.0. Employing a non-experimental correlational design, the research examines the impact of prompt specificity on citation accuracy across various types of prompts, including general, specific, methodological, review, and interdisciplinary prompts. The findings indicate that specific, review, and interdisciplinary prompts correlate positively with accurate references, while general prompts frequently result in non-existent references. Visualizations, including a confusion matrix and precision-recall curve, illustrate the model's performance. Ultimately, the study underscores the necessity of well-structured prompts to enhance reference quality and cautions against AI-induced hallucinations that produce non-existent references, which can significantly undermine research credibility.

KEYWORDS

AI hallucination; artificial intelligence; ChatGPT 4.0; fake references; logistic regression; prompts; research integrity

Introduction

In the realm of academic research, the integration of advanced technologies such as artificial intelligence (AI) has revolutionized the way scholars and practitioners access, analyze, and disseminate information (Dwivedi et al., 2023;

CONTACT Dharel P. Acut sirdharel.acut@gmail.com

Garcia, 2024a; Ofosu-Ampong, 2024). ChatGPT 4.0, an AI language model developed by OpenAI, has become a popular tool for generating content, including aiding researchers in conducting literature searches (OpenAI, 2024). However, despite its advanced capabilities, there are concerns regarding the accuracy and reliability of the information it provides, particularly in the context of generating references and citations (Elkhatat, 2023; Mai et al., 2024; Ray, 2023).

Accurate referencing is a cornerstone of academic integrity, ensuring the credibility and verifiability of scholarly work (Lo, C.K., 2023). Instances of fabricated or erroneous references can undermine the quality of research, leading to misinformation and academic misconduct (Aïmeur et al., 2023). The phenomenon of AI-generated fake references is an emerging issue that demands thorough investigation, especially as more researchers rely on tools like ChatGPT for academic writing and literature searches (Livberber & Ayvaz, 2023; Picazo-Sanchez & Ortiz-Martin, 2024).

Existing literature highlights the potential and pitfalls of using AI in academic research, particularly in generating references (Khalifa & Albadawy, 2024; Rogayan, 2024; Wagner et al., 2022). Studies have explored the capabilities of AI in generating coherent and contextually appropriate text, improving the efficiency of literature reviews, and aiding in information synthesis. Previous research predominantly involved ChatGPT 3.5, which, despite its advancements, exhibited limitations in maintaining accuracy and reliability in reference generation (Giray, 2023b). These studies often lacked the rigorous application of inferential statistical methods necessary for a comprehensive understanding of the issue. In contrast, ChatGPT 4.0 introduces significant improvements in language processing and contextual understanding, theoretically reducing the likelihood of generating non-existent references. Thus, this study explores variables such as query specificity, prompt structure, and contextual factors to uncover patterns and predictors of inaccurate reference generation. The findings will provide valuable insights into the limitations of AI tools in academic research and offer recommendations for mitigating the risks associated with their use.

Literature review

AI in academic research and scholarly communication

AI has increasingly become a tool for aiding academic research, offering capabilities that streamline tasks such as literature reviews, reference management, and even content generation (Malik et al., 2023). Studies like those by Collins et al. (2021) and Xu et al. (2021) emphasize the role of AI in enhancing the efficiency of research processes by quickly synthesizing

information from large datasets. AI language models such as ChatGPT have been instrumental in helping researchers generate coherent, contextually relevant text. Researchers have found that AI can assist in automating parts of the academic writing process, such as drafting abstracts or creating summaries (Khalifa & Albadawy, 2024).

However, despite the growing role of AI in academia, concerns remain about its reliability in generating accurate scholarly information. Many scholars argue that while AI-generated content can reduce time spent on repetitive tasks, it often lacks the depth and critical thinking required in academic writing (Zhai et al., 2024). This limitation calls for caution when using AI tools in scholarly communication, especially in generating references, which are foundational to academic integrity. Moreover, researchers have noted that AI's capabilities are still evolving, and models like ChatGPT must be used judiciously in academic work (Yu, 2024). In summary, while AI tools offer potential for enhancing academic writing and research efficiency, their use in scholarly communication raises questions about quality, accuracy, and credibility. As AI continues to evolve, its role in research will likely expand, but users must remain aware of its limitations, especially in generating verifiable references.

Challenges of AI-generated references

One of the most pressing challenges in using AI tools like ChatGPT for academic research is the generation of non-existent or fabricated references. Walters and Wilder (2023) investigated this issue, finding that ChatGPT-generated references often included inaccuracies, missing details, or completely fabricated citations. Their study revealed that while AI-generated content might appear polished on the surface, the underlying data it draws from may lack verifiable sources, potentially leading to significant errors in scholarly work. These fabricated references can undermine the credibility of academic output and cause issues for researchers who rely heavily on AI for literature searches.

Moreover, the proliferation of AI-generated non-existent references poses risks to academic integrity, as it can lead to misinformation in research and publications. Studies have shown that the rate of AI-generated false citations varies depending on the complexity and specificity of the prompt (Giray, 2023b). The more general the prompt, the higher the likelihood of non-existent references being generated. In contrast, more specific prompts may reduce this risk but do not entirely eliminate it. Researchers have called for systematic approaches to verify the references generated by AI, stressing the need for users to cross-check all references using reliable databases such as PubMed, Scopus, Google Scholar, CrossRef, and Semantic Scholar (Alyasiri et al., 2024). Given the growing concerns about

AI-generated references, scholars argue for greater scrutiny and verification in AI-assisted academic research (Khalifa & Albadawy, 2024). This includes developing tools or mechanisms to flag non-existent references automatically, a critical step toward maintaining the reliability and trustworthiness of AI in academic environments (Miao et al., 2023).

Prompt engineering and its impact on AI performance

Prompt engineering, the practice of carefully crafting queries to AI models to generate desired outputs, has gained attention as a key factor influencing the quality of AI-generated content, including academic references (Walter, 2024). Studies suggest that prompt specificity plays a crucial role in determining the accuracy and relevance of AI outputs (Giray, 2023a). For instance, a general prompt, such as "Provide references on the impact of technology in education," leaves significant room for interpretation, which can lead to a broad range of outputs. This lack of precision might result in a mixture of reliable and non-existent references, as the AI may struggle to filter its vast dataset to meet ambiguous or unspecific requirements. As ChatGPT operates by predicting the next word based on probability and prior training, vague instructions make it more likely to generate content that may include fabricated references to fill in gaps where specific data is lacking.

In contrast, a highly specific prompt like "Provide references for experimental studies on gamification in higher education from 2018 to 2023" narrows the focus significantly, guiding the AI toward a more targeted and verifiable set of responses. By including particular parameters—such as study type (experimental), topic (gamification in higher education), and time frame (2018 to 2023)—the prompt helps filter relevant information and reduces the model's inclination to generate non-existent references. This is because specificity helps the model focus on a well-defined data subset, improving the quality and relevance of the output (Sivarajkumar et al., 2024). Moreover, research by Kalyan (2024) supports the notion that specific prompts yield more reliable outputs from language models, as they provide clear boundaries within which the model operates.

Recent research highlights the role of prompt engineering in optimizing AI performance for academic tasks. Knoth et al. (2024) demonstrated that refining prompts to be highly specific not only enhances content quality but also reduces the risk of generating incorrect or non-existent references. In the context of this study, prompt types were categorized into general, specific, methodological, review, and interdisciplinary to analyze how each influences the quality of references produced by ChatGPT. These categories align with user intent and the AI's ability to process varying degrees of contextual complexity. However, the relationship between prompt design

and AI output is not foolproof. Even with highly specific prompts, models like ChatGPT can still generate false or misleading citations due to gaps in their training data. As shown by Ekin (2023) and Giray (2023a), factors such as domain specificity, clarity of the input, and inherent limitations of the model influence the accuracy of the references produced.

Research gaps and the need for regression analysis

Despite the growing reliance on AI tools like ChatGPT for academic tasks, significant research gaps exist in understanding the extent and implications of AI-generated non-existent references in literature searches. While previous studies have explored the accuracy of AI-generated references, these efforts have largely focused on medical and geography-related fields and have predominantly relied on descriptive statistics (Alyasiri et al., 2024; Alter et al., 2024; Walters & Wilder, 2023). Crucially, no research to date has systematically examined ChatGPT 4.0's performance in generating references within the context of science and technology education. Given the interdisciplinary nature of this field—combining pedagogical strategies, student outcomes, and technological integration—this domain presents unique challenges that make it vital to assess the AI's ability to produce accurate citations.

To address these gaps, this study applies logistic regression analysis to model the factors influencing the generation of both existent and non-existent references. This statistical approach is critical for understanding how various prompt types (e.g., general, specific, methodological, review, interdisciplinary) affect reference accuracy. Unlike descriptive methods, logistic regression allows for a more nuanced examination of the relationships between input variables and the AI's performance, offering predictive insights into the likelihood of generating non-existent references. By modeling binary outcomes such as citation existence, the study provides a rigorous framework for identifying the factors most likely to lead to reference fabrication.

Thus, this research fills a gap in the literature by investigating ChatGPT 4.0's reliability in generating academic references within the field of science and technology education. Through a systematic analysis of prompt types and their influence on reference accuracy, the study contributes to the ongoing discourse on AI's role in academia and highlights the need for critical validation of AI-generated content. Specifically, the research addresses two key questions:

1. How frequently does ChatGPT 4.0 generate non-existent references in terms of types of prompts?
2. How do different types of prompts affect the likelihood of generating non-existent references?

Methods

To quantify the relationship between types of prompts and the likelihood of generating non-existent citations, this study utilized a quantitative research approach with a non-experimental correlational design (Chiang et al., 2015) in July 2024. This approach involves numerical data and statistical analysis, focusing on observing and analyzing existing data to identify correlations rather than manipulating independent variables or randomly assigning predictors. The independent variables in this study include various types of prompts. The dependent variable is the binary outcome indicating whether the rate of non-existent references exceeds a specified threshold.

Development and testing of prompts

The authors devised and categorized prompts into five distinct types: general, specific, methodological, review, and interdisciplinary, each with unique characteristics and justifications based on the experiences in immersing with ChatGPT and grounded in existing research strategies and guidelines (Ekin, 2023; Giray, 2023a; Lo, L.S., 2023). General prompts are broad and open-ended, often resulting in varied responses. Less structured prompts can lead to diverse but sometimes less accurate outputs. Specific prompts, in contrast, are narrowly focused, guiding the AI toward precise information. Clarity in prompts reduces ambiguity and improves accuracy. Methodological prompts direct the AI to concentrate on research methods and enhance the quality of references by providing a clear framework for the AI to follow. Review prompts request summaries or critiques of existing literature, leveraging the AI's ability to synthesize information. Review-focused queries yield comprehensive overviews. Finally, interdisciplinary prompts draw on multiple fields of study, encouraging the AI to integrate diverse perspectives. Such prompts can enrich the depth and breadth of generated content. This categorization allows for a nuanced analysis of how different types of prompts influence the accuracy of AI-generated references (Table 1).

The prompt types were chosen to mimic real-world research needs in science and technology education, reflecting a range of common academic inquiries. Each type was tested in pilot sessions with ChatGPT 4.0 to observe its responses and assess alignment with expected outputs. Key elements of the design process included the context (education/social sciences), input data (specific reference requests), and output indicators (number of non-existent references). We mapped each prompt type to ensure it aligned with user intent, model understanding, domain specificity, clarity, and an effort to minimize bias. The prompts were

Table 1. Categories of prompt types utilized in this study.

Prompt type	Prompt	Characteristic
General	Please provide 50 references (APA format) for recent articles on the impact of technology in science education.	Asks for references on a broad topic without specifying any particular aspect or detail.
Specific	Please provide 50 references (APA format) for the studies published in the last five years on the effectiveness of inquiry-based learning on student learning outcomes.	Narrows the request to studies on a particular pedagogical approach, inquiry-based learning.
Methodological	Please provide 50 references (APA format) for recent experimental studies on the impact of gamification on student learning outcomes.	Specifies the type of study design (experimental) required.
Review	Please provide 50 references (APA format) for systematic reviews on the effectiveness of blended learning in higher education.	Requests references for systematic reviews.
Interdisciplinary	Please provide 50 references (APA format) for studies exploring the psychological effects of educational technology on student motivation.	Intersects multiple fields of study, educational technology and psychology.

designed with clear user intent (e.g., requesting APA-formatted references) and crafted to minimize ambiguity, ensuring ChatGPT could interpret the requests accurately (Ekin, 2023). We incorporated domain specificity by focusing on science education and varied the prompts' levels of specificity to assess the model's ability to handle different degrees of complexity (Giray, 2023a). Constraints, such as timeframes (e.g., recent studies within the last five years), were deliberately included to test their influence on the generation of non-existent references. To minimize bias, all prompts followed a consistent structure, requesting the same number of references (50) and adhering to APA formatting. This uniformity helped reduce variability in responses due to differences in prompt construction.

In the initial testing phase, we started with smaller reference quantities (e.g., 5 or 10 references per prompt). However, this resulted in a high rate of fabricated references, particularly for general prompts, where the AI struggled to provide valid citations even with fewer requests. The generalized nature of these prompts limited the AI's ability to draw from relevant sources, leading to a higher proportion of non-existent references. Based on these initial results, we increased the reference count to 50 per prompt. This decision was driven by the need to create a substantial dataset for regression analysis, which requires a large number of data points for meaningful statistical insights. A higher reference count provided a clearer picture of the AI's ability to generate diverse and relevant citations, while also highlighting potential gaps in its outputs. This approach balanced the need for comprehensiveness with the practical feasibility of conducting a thorough analysis, offering insights into the types of prompts that generated more reliable references and helping refine overall prompt design.

Data collection and analysis

Several methodical steps were conducted to ensure comprehensive and accurate examination of the AI-generated references. Initially, references generated by ChatGPT 4.0 were encoded into an Excel spreadsheet for systematic cross-verification against existing research databases, including Google Scholar, Education Resources Information Center (ERIC), and Semantic Scholar (Ewald et al., 2022). Each reference was checked for its existence and accuracy within these databases, allowing the identification of non-existent references. For instance, we input the title of each reference into Google Scholar, which is a widely used and comprehensive source of academic citations. If no match was found, we then cross-checked the reference in ERIC for education-related topics, which aligns with the focus of our research. Semantic Scholar was also used for broader academic sources across various disciplines. In cases where the reference could not be located in any of these databases, we manually verified the author names and publication dates individually, ensuring that even partial matches were considered before classifying a reference as non-existent.

By "partial matches," we refer to situations where certain elements of the reference—such as the author list, journal name, volume/issue number, or publication date—differed from the generated citation. For instance, some references had inconsistencies in the number or names of authors, where ChatGPT might have listed only a subset of the actual authors or even entirely different authors. Similarly, the journal names might have been incorrect or slightly altered, making the articles difficult to locate. We also encountered incorrect volume and issue numbers that did not correspond to the actual article, as well as inaccurate publication dates. We approached these partial matches with caution, resolving minor inconsistencies before classifying a reference as non-existent. For example, if the title and author list were correct but the publication date or volume number was incorrect, we considered the reference a valid match after manually correcting the discrepancies.

When interacting with ChatGPT, we employed several strategies to ensure consistency, reproducibility, and minimize bias throughout the process. First, each interaction followed a standardized protocol in which we used predefined, clearly articulated prompts across all interactions. This allowed us to reduce variability in the phrasing of prompts, which could have influenced the model's output. Every prompt was carefully constructed and followed a uniform structure, ensuring that variations in how we asked questions did not inadvertently affect the reference generation. To further reduce bias, we implemented multiple test sessions at different times of the day and across several days in July 2024. This approach helped ensure that temporary fluctuations in the model's output

due to server load or updates did not influence the results. Additionally, each reference generation process was replicated multiple times, and we compared the outputs for consistency. In cases where ChatGPT generated differing sets of references for the same prompt, we recorded these variations to account for potential inconsistencies in the model's responses. To increase reproducibility, all interactions with ChatGPT were conducted using the same browser (Google Chrome) and the same computer, ensuring a controlled and consistent environment throughout the experiment. This approach helped eliminate variability that could result from differences in hardware or software, such as using different computers or browsers, which could affect how the AI processed the prompts. Maintaining these consistent conditions could be essential to ensure reproducibility and reliability in the results, as external factors like system performance and network stability remained constant across all trials. We also ensured that no additional external factors, such as changes in prompt structure or data inputs, influenced the outputs.

The non-existent references were then categorized based on the types of prompts used: general, specific, methodological, review, and interdisciplinary. This categorization enabled the understanding of the distribution of inaccuracies across different prompt types and facilitated further analysis. The frequency of non-existent references for each prompt category was calculated, followed by correlation analysis to explore potential relationships between the types of prompts and the frequency of non-existent references. This step helped identify whether certain prompt types were more prone to generating inaccuracies. Appendix A lists selected non-existent references generated by ChatGPT in this study.

Additionally, the one-hot encoded categories for prompt types play a crucial role in elucidating the relationship between specific prompt characteristics and reference accuracy. One-hot encoding is a technique used to convert categorical data into a numerical format that is suitable for statistical analysis and machine learning models (Dahouda & Joe, 2021). By transforming each category of the prompt types into a separate binary variable (where each variable indicates the presence or absence of a specific category), this method preserves the distinct nature of each category without imposing any ordinal relationship between them. This transformation is essential because it allows the categorical data to be accurately analyzed alongside numerical data, facilitating a more precise and meaningful correlation analysis. As a result, one-hot encoding enables researchers to explore how different types of prompts may influence the accuracy of references generated, thereby providing deeper insights into the effectiveness of various prompt characteristics in relation to the study's outcomes.

To further investigate the relationship between the predictor variables (types of prompts) and the binary outcome (high vs. low non-existent

reference rate), logistic regression analysis was applied. This statistical method allowed the assessment of the likelihood of a high non-existent reference rate based on the type of prompt used, providing deeper insights into the predictive power of each prompt category, while controlling for potential confounding factors and accommodating the binary nature of the outcome variable. Additional inferential statistical tests, including the chi-square test and correlation analysis, were conducted to validate the findings. The chi-square test was used to determine the significance of the association between prompt types and the occurrence of non-existent references, while correlation analysis examined the strength and direction of these associations.

All computations were performed using IBM SPSS Statistics, and visualizations were generated using Google Colab, facilitating efficient data processing and graphical representation of the results. The visualizations included frequency distribution charts, correlation matrices, and logistic regression plots, providing clear and interpretable insights into the data.

Results

In July 2024, ChatGPT 4.0 was utilized to conduct a literature search in the context of science and technology education research. This domain aligns closely with our expertise, allowing us to critically assess the accuracy of AI-generated references. During this period, the chatbot was utilized for ideation and reference generation, systematically examining its outputs to assess the accuracy and reliability of the 250 references it produced. The results of this investigation are presented below, detailing the cross-verification process, frequency analysis, and subsequent statistical evaluations.

Cross-verification and frequency analysis

The general prompt type shows a stark disparity, with only 10% of the references being existent and 90% non-existent, compared to the expected 22.4 existent and 27.6 non-existent references (Table 2). This indicates a substantial gap between what was expected and what was generated, suggesting that ChatGPT 4.0 struggles significantly with general prompts. The high rate of non-existent references in this category raises concerns about the reliability of using ChatGPT for general literature searches, emphasizing the need for users to independently verify references when using such broad prompts (Giray, 2023b).

In contrast, the review and interdisciplinary prompt types demonstrate significantly better performance, with existent reference rates of 76% and 84%, respectively. These figures are notably higher than the expected 22.4

Table 2. Distribution of existing and expected references by prompt type.

Prompt type	Existent references	Non-existent references	Existent references rate	Non-existent references rate	Expected existent references	Expected non-existent references
General	5	45	10.0	90.0	22.4	27.6
Specific	16	34	32.0	68.0	22.4	27.6
Methodological	11	39	22.0	78.0	22.4	27.6
Review	38	12	76.0	24.0	22.4	27.6
Interdisciplinary	42	8	84.0	16.0	22.4	27.6

existent references, indicating that ChatGPT 4.0 is more accurate in generating references for these types of prompts. This improved performance can be attributed to the structured and comprehensive nature of review and interdisciplinary prompts, which often require the AI to integrate and synthesize information from multiple sources. This synthesis may leverage more established and widely recognized references, enhancing the AI's ability to provide accurate citations. The higher accuracy in these prompt types suggests that ChatGPT 4.0 performs better when handling specialized or well-defined topics, where comprehensive information is more readily available. However, the presence of non-existent references, though lower, still underscores the importance of careful verification by users to ensure the credibility of AI-generated content.

The specific and methodological prompt types exhibit intermediate performance, with reference rates of 32% and 22%, respectively. Although these rates are higher than those in the General category, they still fall short of the expected reference rate of 10%. The term "expected count" refers to the rate we anticipate based on certain criteria or benchmarks, while "observed count" is the actual rate measured in the data. In this case, the observed reference rates are significantly higher than the expected rate of 10%, indicating that while performance has improved compared to the general category, there is still substantial room for enhancement. Additionally, the high rates of non-existent references—68% for specific prompts and 78% for methodological prompts—highlight the need for users to meticulously verify the references generated by ChatGPT 4.0 to ensure their accuracy (Alshami et al., 2023).

The Chi-Square test results were crucial in determining whether the differences between observed and expected frequencies were statistically significant. The computed Chi-Square statistic of 89.059 far exceeded the critical value of 9.488 for 4 degrees of freedom at the 0.05 significance level. This indicates a significant discrepancy between observed and expected reference counts, suggesting that the variations are unlikely due to random chance. The p-value of < 0.000 further confirms that the null hypothesis, asserting no difference between observed and expected frequencies, can be confidently rejected. The Chi-Square test also relates the association between the type of prompt and existent reference. Understanding

such association provides each user the ability to recognize the type of prompt to hasten finding references at high valid results (Singhal & Rana, 2015).

These results underscore a significant inconsistency in the types of references generated by ChatGPT 4.0. The tool demonstrates better performance for specialized and review-type prompts but exhibits notable reliability issues with general and methodological prompts. The challenges with general and methodological prompts indicate that the AI struggles with broader or less defined areas where specific, reliable sources are less readily available. This variability highlights the need for ongoing improvements in AI models to enhance accuracy and reliability across all prompt types. Consequently, users should exercise caution and verify AI-generated references to uphold the integrity of their academic and research work (Davis & Lee, 2023; Dwivedi et al., 2023).

Correlation analysis

The correlation heatmap visualizes the relationships between different types of references and reference rates across various prompt types. One of the first observations from the heatmap is the inverse relationship between existent and non-existent references, indicated by strong negative correlations. This aligns with research indicating that increased rigor in certain prompt types (Giray, 2023a), such as review and interdisciplinary studies, often results in fewer non-existent references due to thorough cross-referencing and verification practices (Figure 1).

The heatmap also shows strong positive correlations between prompt types like review and interdisciplinary with existent reference rates. For instance, review prompt shows a high positive correlation with existent reference rate ($r = 0.53$) and a high negative correlation with non-existent reference rate ($r = -0.53$), which suggests that review-type prompts are more likely to produce references that are existent and verifiable. Interdisciplinary prompts also exhibit strong correlations: positively with existent references rate ($r = 0.66$) and negatively with non-existent references rate ($r = -0.66$). This correlation suggests that interdisciplinary prompts tend to yield more existent references compared to other prompt types. This could be due to the broader and more integrated approach taken in interdisciplinary research, which may involve more comprehensive literature reviews and higher verification standards.

Conversely, prompt types such as general, specific, and methodological exhibit higher non-existent reference rates. The general prompt type has a negative correlation with existent reference rate ($r = -0.69$) and a positive correlation with non-existent reference rate ($r = 0.69$). This could be attributed to the broad nature of general prompts, which might not demand

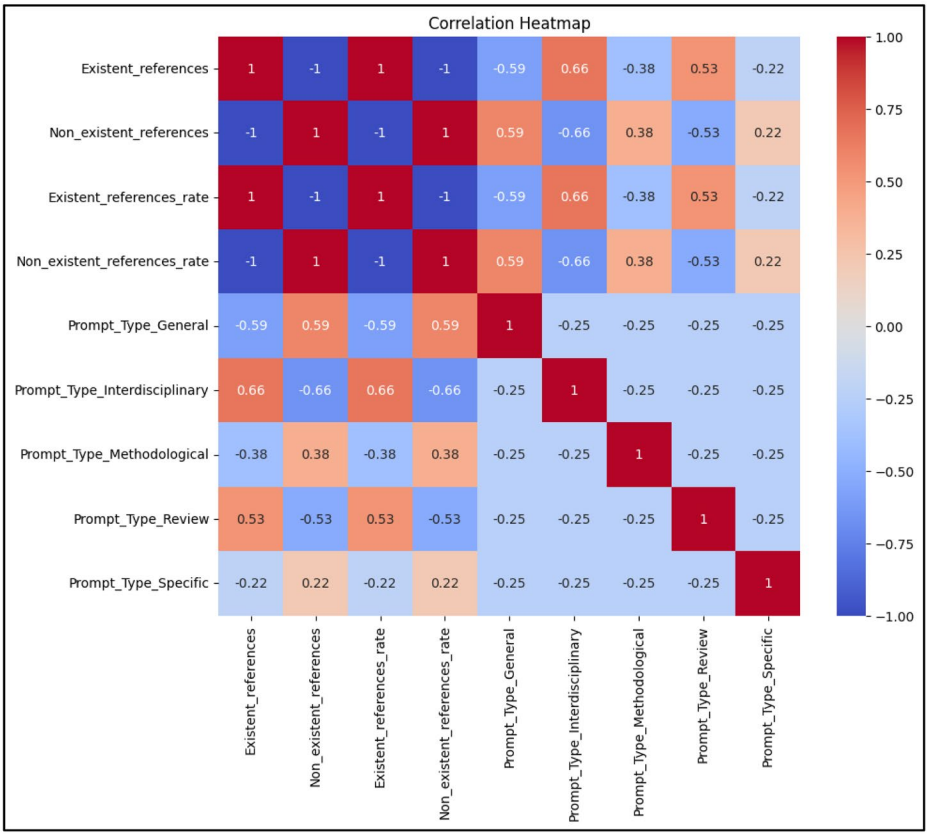

Figure 1. Heat map of reference generation correlations.

as rigorous a validation process as more specialized prompts. Specific prompt has a negative correlation with existent reference rate ($r = -0.22$) and a positive correlation with non-existent references rate ($r = 0.22$). This could be attributed to highly specific prompts that may lead to a narrower focus that the AI model cannot adequately address due to limitations in its training data. Methodological prompt, on the other hand, has a slight negative correlation with existent reference rate ($r = -0.38$) and a slight positive correlation with non-existent reference rate ($r = 0.38$). These prompts might focus on process over content, sometimes leading to oversight in cross-referencing. Studies have shown that prompts requiring less stringent validation criteria can lead to an increased likelihood of non-existent references (Ekin, 2023; Kochanek et al., 2024).

Overall, the heatmap provides valuable insights into how different prompt types influence the presence of existent and non-existent references, highlighting distinct patterns and relationships that might not be immediately apparent through traditional analysis. This visualization emphasizes the need for tailored validation methods depending on the prompt type to enhance reference accuracy in academic and research settings, as certain

prompts may be more prone to generating inaccurate references. Such findings underscore the importance of customizing reference-checking protocols to align with the specific characteristics of each prompt type, ensuring a higher standard of reliability in research outputs (Davis & Lee, 2023).

Logistic regression analysis

The linear regression analysis of ChatGPT 4.0's reference generation yielded interesting insights, though the overall model did not explain much variance. The R-squared value of 0.000 suggests that the expected references (both existent and non-existent) are not strong predictors of the actual number of existent references generated. Despite this, the coefficients for both expected existent references (0.3971, $p = 0.039$) and expected non-existent references (0.4893, $p = 0.039$) were statistically significant, indicating that these variables do have a measurable impact on reference generation accuracy.

This mixed outcome implies that while ChatGPT 4.0's reference accuracy might be influenced by expected reference patterns, the variance is largely unexplained by these factors alone. The significant coefficients suggest that certain prompt types may push the model toward generating more or fewer accurate references, but the overall low explanatory power could be attributed to the small dataset or potential multicollinearity issues, as highlighted by the design matrix singularity warning. This calls for a larger and more diverse dataset to improve model reliability and reduce potential overfitting, providing a clearer picture of how ChatGPT handles different prompt types in reference generation.

However, the confusion matrix heatmap shows a perfect classification with no misclassifications. Each cell in the matrix represents the count of true positive, true negative, false positive, and false negative predictions. In our case, the matrix shows a count of 1 for both true positives and true negatives, with counts of 0 for false positives and false negatives. This indicates that the model has correctly classified all instances in the test set. While this is a positive outcome, it is important to note that the limited size of the test set (only 2 instances) can make the model's performance appear better than it might be on a larger, more varied dataset. Overfitting is a potential risk when dealing with such limited datasets (Charilaou & Battat, 2022).

The classification report further supports the confusion matrix findings, showing perfect precision, recall, and F1-scores for both classes (0 and 1). Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives among all actual positives. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns.

Achieving perfect scores in these metrics suggests that the model has performed exceptionally well on the given test set, indicating its effectiveness in distinguishing between real and non-existent references generated by the chatbot (Hicks et al., 2022). While these results are promising, it is important to ensure the model's robustness through further validation and testing on additional datasets to confirm its reliability across different scenarios (Faber & Fonseca, 2014).

The precision-recall curve offers another perspective on model performance, especially for imbalanced datasets (Richardson et al., 2024). In this case, the curve is an ideal, straight line at the top right corner of the plot, reflecting perfect precision and recall at all thresholds. This visualization confirms the findings of the confusion matrix and classification report, showing that the model maintains high performance regardless of the decision threshold. Precision-recall curves are particularly useful for evaluating models when the classes are imbalanced, as they focus on the performance of the positive class (Barros et al., 2019). Here, the curve's ideal shape indicates that the model has no trouble distinguishing between the classes in the test set, which, in this context, likely refers to accurately detecting non-existent references (Figure 2).

Discussion

Accurate and reliable research is the cornerstone of scholarly work, as it ensures that findings are credible, replicable, and contribute meaningfully to the body of knowledge in a given field (Shaheen et al., 2023). The integrity of research depends not only on the methodology and analysis but also on the reliability of the sources used to support arguments and conclusions (Acut & Antonio, 2023). References serve as the foundation upon which researchers build their work, allowing them to ground their findings in established knowledge, acknowledge prior contributions, and provide readers with a roadmap to verify and further explore the subject matter (Divecha et al., 2023). Without accurate references, research becomes speculative, and the trustworthiness of the findings is compromised, which undermines the academic rigor essential to advancing science, education, and even professional practice. Recently, there has been a growing use of ChatGPT in scholarly writing. The academic community quickly raised concerns regarding the accuracy and validity of AI-generated content (Garcia, 2024b). Therefore, there is a necessity to critically evaluate the reliability of ChatGPT, especially when using its content for scientific research.

According to our results, ChatGPT 4.0 frequently generates non-existent references. This finding is consistent with prior studies in the field of health research, where similar issues have been observed (Wu & Dang, 2023;

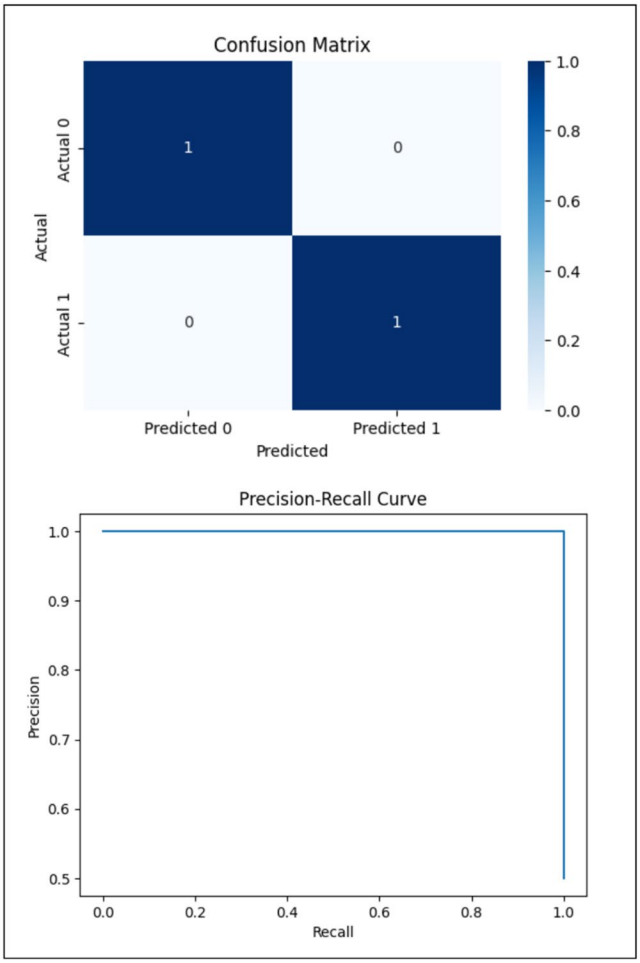

Figure 2. Normalized confusion matrix and precision-recall curve.

Wagner & Ertl-Wagner, 2024). These comparable results suggest that the issue of non-existent references is prevalent across various fields and persists even in the latest version of ChatGPT. While there is evidence that ChatGPT outperforms other large language models in generating references (Dhane et al., 2024), it remains prone to fabricating citations. Given that content generation and writing assistance are commonly cited as key potentials of ChatGPT (Baig & Yadegaridehkordi, 2024), the frequent generation of non-existent references raises serious concerns about its reliability in academic and professional contexts. Similar to other publications that have accidentally included ChatGPT-generated content (e.g., see the first sentence in the retracted paper by Zhang et al., 2024), these non-existent citations can easily go unnoticed and be included in scholarly work. The risk is especially high for users who may assume that AI-generated content is inherently trustworthy without performing the necessary

verification. This trend underscores the critical need for caution when using AI tools in academic research and highlights the importance of maintaining rigorous standards for reference checking to ensure the integrity of scholarly work.

One feature of this paper is the evaluation of generated citations based on prompt types. Our analysis revealed that the likelihood of ChatGPT generating non-existent references significantly increases with the use of broad, open-ended prompts. One plausible explanation is that when prompts lack specificity, the AI attempts to synthesize plausible responses based on incomplete or generalized patterns from its training data (Dwivedi et al., 2023). This scenario often leads to the creation of references that do not exist but appear credible, as they fit the context of the query (Knoth et al., 2024). The model suggests that broad prompts yield less reliable outputs due to the AI's probabilistic nature, where it fills gaps by generating fabricated references that seem relevant but have no basis in actual literature. The difference between broad and specific prompts in generating non-existent references can be explained through the lens of language generation models. Broad prompts leave more interpretive leeway, prompting ChatGPT to extrapolate widely from its training data (Brown et al., 2020). This results in a higher chance of ghosted references as the AI ventures into areas with less concrete data. Conversely, specific prompts guide the AI more narrowly, reducing ambiguity and helping it retrieve information more accurately (Giray, 2023a). The lack of constraint in broad prompts increases the cognitive load on the model, often leading it to "hallucinate" references when precise data is unavailable or insufficient. AI hallucination is a well-documented phenomenon where AI systems, particularly language models, generate information that appears factual but is not grounded in reality (Farquhar et al., 2024; Kouzelis & Spantidi, 2024; Reddy et al., 2024).

The results of the linear regression analysis reveal interesting insights into the performance of ChatGPT 4.0 in generating academic references, though the overall model demonstrated limited explanatory power. While the expected references (both existent and non-existent) were significant predictors of reference accuracy, the model failed to explain much of the variance in the actual number of existent references generated. This suggests that reference generation is influenced by factors beyond the expected reference patterns. One possible explanation is that prompt specificity and context play a more important role in determining reference accuracy. Prior studies, such as those by Walters and Wilder (2023) and Dwivedi et al. (2023), have similarly found that AI tools like ChatGPT perform better when given precise, well-defined prompts, whereas vague or general prompts tend to result in more errors. These findings indicate that

ChatGPT's reference generation relies heavily on the quality of the input it receives, further emphasizing the need for careful prompt design.

Moreover, the significant coefficients indicate that certain prompt types, such as review and interdisciplinary, tend to generate more accurate references, while general or methodological prompts are more prone to producing non-existent references. This aligns with previous research by Alyasiri et al. (2024) and Giray (2023a), which highlighted that AI models tend to struggle with abstract or less structured prompts. The low explanatory power of the model could also be attributed to the limited dataset used, underscoring the need for larger, more diverse datasets to improve model reliability and reduce the risk of overfitting. Overall, the study underscores the necessity of continued improvements to AI models to enhance their accuracy and reliability, particularly in academic settings, where the integrity of AI-generated references is crucial. These findings suggest that users must remain cautious and verify AI-generated content, particularly when dealing with general or methodological topics.

Practical implications

Our findings have significant implications for researchers, educators, and libraries using AI tools like ChatGPT for literature searches and content generation. While AI can enhance the speed and breadth of initial literature exploration, the generation of non-existent references presents a serious challenge to research integrity. Users must remain vigilant, as over-reliance on AI without proper verification could result in the inclusion of fabricated references in academic work. For educators, these findings highlight the need to incorporate critical thinking and AI literacy into academic curricula to ensure that students can identify and cross-check sources generated by AI. The broader impact of non-existent references on academic integrity is also profound, especially if users fail to recognize these ghosted citations. Such references, when incorporated into academic papers, could undermine the credibility of the research, leading to flawed conclusions or invalid citations (Rivkin, 2020). If unchecked, the proliferation of fabricated references could erode trust in AI-assisted academic work and compromise the peer-review process. More troubling is the potential for these inaccuracies to spread through citation chains, with one unverified reference being cited by others, leading to a cascade of misinformation within scholarly communities.

In addition, the findings of this study hold significant relevance for libraries, particularly in the areas of academic reference services, AI-driven literature searches, and citation management. As libraries incorporate AI tools like ChatGPT into their services, understanding the limitations and strengths of these tools is essential (Nehra & Bansode, 2024). This study

highlights that while AI can assist in generating references, its accuracy is highly dependent on the quality and specificity of prompts. For librarians, this insight offers an opportunity to enhance user education by providing targeted guidance on how to craft effective prompts that result in more reliable AI-generated references. In relation to academic reference services, libraries could integrate these insights by offering workshops or instructional materials focused on the nuances of AI-based tools for literature searches. Librarians could demonstrate how general or poorly defined prompts may lead to inaccurate or non-existent references, thereby emphasizing the importance of prompt specificity in ensuring the integrity of search results. Additionally, AI-driven citation management tools may be optimized by developing algorithms or systems that flag potentially non-existent references, encouraging users to verify AI-generated content before including it in their academic work (Jhajj et al., 2024; Rogayan, 2024). By doing so, libraries can not only help users make the most of AI tools but also uphold the rigor and reliability of academic research outputs. This approach positions libraries as key facilitators in the responsible integration of AI into scholarly workflows.

To minimize the occurrence of non-existent references, ChatGPT users should adopt a comprehensive multi-step approach. First, using more specific and narrowly focused prompts is essential. When users provide clear, concise, and targeted prompts, it reduces the likelihood of ChatGPT generating fabricated citations, as the AI is guided by more explicit parameters. This means avoiding overly broad or open-ended questions and instead focusing on particular aspects of a topic. For example, asking about specific studies or well-known theories in a given field will yield more accurate and relevant information compared to vague prompts that leave too much interpretive leeway. Cross-verification of AI-generated references should also be standard practice. Users must check every reference generated by ChatGPT against reputable academic databases such as Google Scholar, PubMed, Scopus, or other resources. This manual verification process ensures that each reference exists and is correctly cited. For added reliability, directly accessing the cited papers' Digital Object Identifier (DOI) or specific publication sources can further authenticate references. However, given the current limitations of ChatGPT in reliably generating accurate citations, it may be prudent to refrain from using the tool for literature searches entirely until more robust solutions or improvements are developed (Haman & Školník, 2024).

Limitations of the study and future works

While the current results are promising, they should be interpreted with caution due to several limitations. First, the analysis is based on a test

set that may not fully represent the complexity of real-world scenarios. The model's performance in this context may not generalize effectively to larger or more diverse datasets.

A key limitation lies in the choice of prompt structure. The uniform structure across all prompt types, designed to minimize variability, may have introduced biases, particularly when handling niche or highly specialized topics. The decision to request 50 references per prompt, while aimed at generating a substantial dataset for analysis, may have inadvertently led to an increased number of fabricated citations, especially in fields where fewer legitimate sources are available. For example, highly specialized subjects or emerging areas of research may not have the volume of studies to meet such a high reference count, thereby prompting the AI to generate non-existent citations after exhausting available sources. A more flexible approach, adjusting the number of requested references based on topic breadth and depth, could mitigate this issue in future studies.

The verification process also has limitations. Although we employed a multi-platform strategy using Google Scholar, ERIC, and Semantic Scholar, these databases may not cover all relevant academic sources. Furthermore, there were inconsistencies in verifying citations, including errors in authorship, journal names, or publication details. Some references, though partially correct, had discrepancies such as missing authors, erroneous journal names, or incorrect volume/issue numbers and publication dates. These inconsistencies complicated the verification process, and the lack of DOIs for most generated citations added to the challenge of confirming their validity.

Potential model inconsistencies also pose a challenge. The AI may respond differently to similar prompts depending on slight variations in phrasing or context, which could affect the reliability of its outputs. This introduces another layer of complexity, as even minor changes in prompt wording could influence the model's ability to generate accurate references. Future research should explore several key indicators to address these limitations. Prompt complexity and ambiguity may affect the model's ability to generate accurate references, while topic familiarity should be considered, as AI models may perform differently based on how closely the subject matter aligns with their training data. Confidence scores in the AI's outputs could also be examined, as higher confidence might correlate with better accuracy, while lower confidence may indicate a need for closer scrutiny. Additionally, different reference types (e.g., journal articles vs. books) may have varying error rates.

The impact of repeated prompts on the AI's performance should also be analyzed, as generating citations on similar topics multiple times might either improve or degrade accuracy. Broader and more varied datasets, as

well as the role of iterative prompt reformulations, could help refine the AI's outputs. Future investigations into these factors will lead to a more nuanced understanding of the model's strengths and weaknesses, contributing to its reliable performance across diverse applications.

Conclusion

This study provides valuable insights into the challenges and potential of AI tools like ChatGPT 4.0 for reference generation in academic research. Despite the significant coefficients observed in the logistic regression model, its ability to explain the variance in actual references was limited. This underscores the critical role of prompt specificity and complexity in achieving accurate AI-generated references. The study reveals that while reputable journals are abundant in science and technology education, a uniform prompt structure and high reference quantity may lead to an increased number of fabricated citations, particularly in emerging areas where sources are limited. Additionally, despite employing robust cross-verification processes through major databases like Google Scholar, ERIC, and Semantic Scholar, citation inconsistencies and the absence of DOIs presented challenges in validating references. Future research should address these limitations by exploring variations in prompt complexity, topic familiarity, and model confidence, and by employing broader datasets and iterative prompt reformulations. Such measures will enhance the reliability and generalizability of AI-generated references, contributing to a more nuanced understanding of the model's performance across diverse contexts. Overall, this study highlights the need for trustworthy references in scholarly work, as the generation of non-existent citations can undermine research integrity and lead to problems such as being "ghosted" in subsequent literature searches and the necessity for cautious and critical use of AI in literature searches and underscores the importance of verifying references to maintain research credibility and integrity.

Acknowledgements

Heartfelt gratitude is extended to the peer reviewers and the editor for their insightful feedback and constructive criticism, which significantly enhanced the quality of this manuscript. Their thorough evaluations and expert guidance helped clarify key concepts and refine our arguments. We appreciate their commitment to fostering scholarly dialogue, which has been instrumental in bringing this work to fruition.

Author's contributions

The contributions of the authors to this research are as follows: DPA was responsible for conceptualization, gap identification, literature search, coding, data analysis, visualization,

and drafting the original manuscript. MBC contributed to editing the manuscript, enriching the discussion section, analyzing and interpreting the data, verifying statistical results, and reviewing the manuscript. NKM, EVM, and NSG handled the cross-verification of references, statistical modeling, data analysis, visualization, and the review and editing of the manuscript. All authors have read and approved the final version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Use of AI tools declaration

The authors declare that ChatGPT, an AI language model developed by OpenAI, was utilized for generating references based on provided prompts. However, the ideation, analysis, and writing of this work are entirely the authors' efforts.

Data availability and material statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon request.

References

Acut, D., & Antonio, R. (2023). Effectiveness of Science-Technology-Society (STS) approach on students' learning outcomes in science education: Evidence from a meta-analysis. *Journal of Technology and Science Education*, *13*(3), 718–739. doi:10.3926/jotse.2151

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, *13*(1), 30. doi:10.1007/s13278-023-01028-5

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T. (2023). Harnessing the Power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, *11*(7), 351. doi:10.3390/systems11070351

Alter, I. L., Chan, K., Lechien, J., & Rameau, A. (2024). An introduction to machine learning and generative artificial intelligence for otolaryngologists-head and neck surgeons: a narrative review. *European Archives of Oto-Rhino-Laryngology*, *281*(5), 2723–2731. doi:10.1007/s00405-024-08512-4 38393353

Alyasiri, O. M., Salman, A. M., Akhtom, D., & Salisu, S. (2024). ChatGPT revisited: Using ChatGPT-4 for finding references and editing language in medical scientific articles. *Journal of Stomatology, Oral and Maxillofacial Surgery*, *125*(5S2), 101842. doi:10.1016/j.jormas.2024.101842

Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in the higher education: A systematic literature review and research challenges. *International Journal of Educational Research*, *127*, 102411. doi:10.1016/j.ijer.2024.102411

Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, *9*(4), 275. doi:10.3390/educsci9040275

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan,

T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (pp. 1877–1901). Curran Associates Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Charilaou, P., & Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, *28*(5), 605–607. doi:10.3748/wjg.v28.i5.605

Chiang, I. C. A., Jhangiani, R. S., & Price, P. C. (2015). *Overview of nonexperimental research*. Pressbooks. https://opentextbc.ca/researchmethods/chapter/overview-of-nonexperimental-research/

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383. doi:10.1016/j.ijinfomgt.2021.102383

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381–114391. doi:10.1109/ACCESS.2021.3104357

Davis, R. O., & Lee, Y. J. (2023). Prompt: ChatGPT, create my course, please!. *Education Sciences*, *14*(1), 24. doi:10.3390/educsci14010024

Dhane, A. S., Sarode, S., Sarode, G., & Singh, S. (2024). A reality check on chatbot-generated references in global health research. *Oral Oncology Reports*, *10*, 100246. doi:10.1016/j.oor.2024.100246

Divecha, C. A., Tullu, M. S., & Karande, S. (2023). The art of referencing: Well begun is half done!. *Journal of Postgraduate Medicine*, *69*(1), 1–6. doi:10.4103/jpgm.jpgm_908_22

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. doi:10.1016/j.ijinfomgt.2023.102642

Ekin, S. (2023). Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *TechRxiv*, doi:10.36227/techrxiv.22683919

Elkhatat, A. M. (2023). Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *International Journal for Educational Integrity*, *19*(1). doi:10.1007/s40979-023-00137-0

Ewald, H., Klerings, I., Wagner, G., Heise, T. L., Stratil, J. M., Lhachimi, S. K., Hemkens, L. G., Gartlehner, G., Armijo-Olivo, S., & Nussbaumer-Streit, B. (2022). Searching two or more databases decreased the risk of missing relevant studies: A metaresearch study. *Journal of Clinical Epidemiology*, *149*, 154–164. doi:10.1016/j.jclinepi.2022.05.022

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. doi:10.1590/2176-9451.19.4.027-029.ebo

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. doi:10.1038/s41586-024-07421-0

Garcia, M. B. (2024a). Addressing the mental health implications of ChatGPT dependency: The need for comprehensive policy development. *Asian Journal of Psychiatry*, *98*, 104140. doi:10.1016/j.ajp.2024.104140

Garcia, M. B. (2024b). Using AI tools in writing peer review reports: Should academic journals embrace the use of ChatGPT? *Annals of Biomedical Engineering*, *52*(2), 139–140. doi:10.1007/s10439-023-03299-7

Giray, L. (2023a). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. doi:10.1007/s10439-023-03272-4

Giray, L. (2023b). ChatGPT references unveiled: Distinguishing the reliable from the fake. *Internet Reference Services Quarterly*, *28*(1), 9–18. doi:10.1080/10875301.2023.2265369

Haman, M., & Školník, M. (2024). Using ChatGPT to conduct a literature review. *Accountability in Research*, *31*(8), 1244–1246. doi:10.1080/08989621.2023.2185514

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. doi:10.1038/s41598-022-09954-8

Jhajj, K. S., Jindal, P., & Kaur, K. (2024). Use of artificial intelligence tools for research by medical students: A narrative review. *Cureus*, *16*(3), e55367. doi:10.7759/cureus.55367

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. doi:10.1016/j.nlp.2023.100048

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, *5*, 100145. doi:10.1016/j.cmpbup.2024.100145

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225. doi:10.1016/j.caeai.2024.100225

Kochanek, M., Cichecki, I., Kaszyca, O., Szydło, D., Madej, M., Jędrzejewski, D., Kazienko, P., & Kocoń, J. (2024). Improving training dataset balance with ChatGPT prompt engineering. *Electronics*, *13*(12), 2255. doi:10.3390/electronics13122255

Kouzelis, L. R., & Spantidi, O. (2024). Enhancing historical extended reality experiences: Prompt engineering strategies for AI-generated dialogue. *Applied Sciences*, *14*(15), 6405. doi:10.3390/app14156405

Livberber, T., & Ayvaz, S. (2023). The impact of artificial intelligence in academia: Views of Turkish academics on ChatGPT. *Heliyon*, *9*(9), e19688. doi:10.1016/j.heliyon.2023.e19688

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. doi:10.3390/educsci13040410

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. doi:10.1016/j.acalib.2023.102720

Mai, D. T. T., Van Da, C., & Van Hanh, N. (2024). The use of ChatGPT in teaching and learning: A systematic review through SWOT analysis approach. *Frontiers in Education*, *9*. doi:10.3389/feduc.2024.1328769

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, A., & Marzuki, N. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, *5*, 100296. doi:10.1016/j.ijedro.2023.100296

Miao, J., Thongprayoon, C., Suppadungsuk, S., Valencia, O., A G., Qureshi, F., & Cheungpasitporn, W. (2023). Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: A narrative review. *Clinics and Practice*, *14*(1), 89–105. doi:10.3390/clinpract14010008

Nehra, S. S., & Bansode, S. Y. (2024). Exploring the prospects and perils of integrating artificial intelligence and ChatGPT in academic and research libraries (ARL): Challenges and Opportunity. *Journal of Web Librarianship*, 1–22. doi:10.1080/19322909.2024.2390413

Ofosu-Ampong, K. (2024). Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. *Telematics and Informatics Reports*, *14*, 100127. doi:10.1016/j.teler.2024.100127

OpenAI. OpenAI. (2024). *GPT-4.0 and more tools to ChatGPT free*. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/

Picazo-Sanchez, P., & Ortiz-Martin, L. (2024). Analysing the impact of ChatGPT in research. *Applied Intelligence*, *54*(5), 4172–4188. doi:10.1007/s10489-024-05298-0

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. doi:10.1016/j.iotcps.2023.04.003

Reddy, G. P., Kumar, Y. V. P., & Prakash, K. P. (2024). Hallucinations in large language models (LLMs). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)* (pp. 1–6). doi:10.1109/eStream61684.2024.10542617

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns (New York, N.Y.)*, *5*(6), 100994. doi:10.1016/j.patter.2024.100994

Rivkin, A. (2020). Manuscript referencing errors and their impact on shaping current evidence. *American Journal of Pharmaceutical Education*, *84*(7), ajpe7846. doi:10.5688/ajpe7846

Rogayan, D. V. (2024). "ChatGPT assists me in my reference list:" Exploring the chatbot's potential as citation formatting tool. *Internet Reference Services Quarterly*, *28*(3), 305–314. doi:10.1080/10875301.2024.2351021

Shaheen, N., Shaheen, A., Ramadan, A., Hefnawy, M. T., Ramadan, A., Ibrahim, I. A., Hassanein, M. E., Ashour, M. E., & Flouty, O. (2023). Appraising systematic reviews: A comprehensive guide to ensuring validity and reliability. *Frontiers in Research Metrics and Analytics*, *8*, 1268045. doi:10.3389/frma.2023.1268045

Singhal, R., & Rana, R. (2015). Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences*, *1*(1), 69. doi:10.4103/2395-5414.157577

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, *12*, e55318. doi:10.2196/55318

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. doi:10.1177/02683962211048201

Wagner, M. W., & Ertl-Wagner, B. B. (2024). Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, *75*(1), 69–73. doi:10.1177/08465371231171125

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*(1). doi:10.1186/s41239-024-00448-3

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, *13*(1), 14045. doi:10.1038/s41598-023-41032-5

Wu, R. T., & Dang, R. R. (2023). ChatGPT in head and neck scientific writing: A precautionary anecdote. *American Journal of Otolaryngology*, *44*(6), 103980. doi:10.1016/j.amjoto.2023.103980

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., … Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *Innovation (Cambridge (Mass.))*, *2*(4), 100179. doi:10.1016/j.xinn.2021.100179

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*, *10*(2), e24289. doi:10.1016/j.heliyon.2024.e24289

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1). doi:10.1186/s40561-024-00316-7

Zhang, M., Wu, L., Yang, T., Zhu, B., & Liu, Y. (2024). The three-dimensional porous mesh structure of Cu-based metal-organic-framework - Aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries. *Surfaces and Interfaces*, *46*, 104081. doi:10.1016/j.surfin.2024.104081

Appendix A. Selected fake references generated by ChatGPT 4.0

Prompt type	Selected fake reference generated by ChatGPT 4.0	Cross-verification			
		Google Scholar	ERIC	Semantic Scholar	Remarks
General	Abazi-Bexheti, L., Abazi-Alili, H., & Rushiti, S. (2020). The impact of using digital tools on student engagement and achievement in science education. Journal of Educational Technology & Society, 23(2), 23-32.	Non-existent	Non-existent	Non-existent	Fabricated
General	Ali, R., & Alharbi, H. (2021). Enhancing science education through augmented reality: A case study. International Journal of Science and Mathematics Education, 19(1), 73-87.	Non-existent	Non-existent	Non-existent	Fabricated
General	Blanchard, M. R., LePrevost, C. E., Tolin, A. D., & Gutierrez, K. S. (2021). Investigating the impact of mobile devices on student engagement in a science classroom. Computers & Education, 167, 104175.	Non-existent	Non-existent	Non-existent	Fabricated
General	Bos, B., & Tarus, J. K. (2020). Investigating the effects of digital storytelling on student learning in science. Journal of Science Education and Technology, 29(6), 781-791.	Non-existent	Non-existent	Non-existent	Fabricated
General	Buabeng-Andoh, C. (2020). Exploring the relationship between teachers' technology use and students' science achievement. International Journal of Science and Mathematics Education, 18(8), 1603-1622.	Non-existent	Non-existent	Non-existent	Fabricated
Specific	Adesoji, F. A., & Idika, O. P. (2020). Effects of inquiry-based learning on senior secondary school students' achievement in chemistry. Journal of Science Education and Technology, 29(4), 577-588.	Non-existent	Non-existent	Non-existent	Fabricated
Specific	Aljarrah, A., & Fakhouri, H. (2021). Inquiry-based learning and its effect on students' academic achievement in science. International Journal of Science Education, 43(3), 451-466.	Non-existent	Non-existent	Non-existent	Fabricated

(*Continued*)

Continued.

Prompt type	Selected fake reference generated by ChatGPT 4.0	Cross-verification			
		Google Scholar	ERIC	Semantic Scholar	Remarks
Specific	Anwar, S., & Hussain, Z. (2020). The impact of inquiry-based learning on students' cognitive and affective outcomes in physics education. Journal of Science Education and Technology, 29(6), 822-833.	Non-existent	Non-existent	Non-existent	Fabricated
Specific	Avsec, S., & Kocijancic, S. (2021). The impact of inquiry-based learning on students' knowledge and attitudes in STEM education. International Journal of Science and Mathematics Education, 19(2), 319-336.	Non-existent	Non-existent	Non-existent	Fabricated
Specific	Chen, Y., & Lo, S. K. (2021). Inquiry-based learning and its effects on students' science process skills and attitudes. Journal of Science Education and Technology, 30(1), 44-56.	Non-existent	Non-existent	Non-existent	Fabricated
Methodological	Abdollahpour, Z., & Asadzadeh, S. (2020). The impact of gamification on EFL learners' motivation and academic achievement. Journal of Educational Technology & Society, 23(3), 105-115.	Non-existent	Non-existent	Non-existent	Fabricated
Methodological	Bautista, G., & Graf, S. (2021). The impact of gamification on online learning environments: A systematic literature review. International Journal of Learning, Teaching and Educational Research, 20(5), 1-22.	Non-existent	Non-existent	Non-existent	Fabricated
Methodological	Boulet, G., & Bédard, D. (2021). The effect of gamified learning environments on student achievement: A meta-analysis. Educational Research Review, 33, 100389.	Non-existent	Non-existent	Non-existent	Fabricated
Methodological	Chen, C., & Xie, H. (2020). Exploring the effects of gamification on student motivation and performance in higher education. Journal of Educational Technology Development and Exchange, 13(2), 11-30.	Non-existent	Non-existent	Non-existent	Fabricated
Methodological	Clark, L. A., & Welderufael, M. (2021). The impact of gamification on student engagement and academic performance in STEM disciplines. Computers & Education, 163, 104097.	Non-existent	Non-existent	Non-existent	Fabricated
Review	Altun, A. (2019). Examining the impact of blended learning on learning outcomes: A meta-analysis. Educational Research Review, 27, 133-144.	Non-existent	Non-existent	Non-existent	Fabricated

(*Continued*)

Continued.

Prompt type	Selected fake reference generated by ChatGPT 4.0	Cross-verification			
		Google Scholar	ERIC	Semantic Scholar	Remarks
Review	Boelens, R., De Wever, B., & Voet, M. (2017). Four key design principles for blended learning: A systematic literature review. Educational Technology Research and Development, 65(4), 1-18.	Non-existent	Non-existent	Non-existent	Fabricated
Review	Finn, A., & Bucceri, M. (2017). A case study of blended learning for flexible and collaborative learning in a large first-year undergraduate course. Journal of Higher Education Policy and Management, 39(6), 586-597.	Non-existent	Non-existent	Non-existent	Fabricated
Review	Garrison, D. R., & Kanuka, H. (2018). Blended learning environments: Perspectives on leveraging technology for high-touch learning experiences. Internet and Higher Education, 11(3-4), 21-30.	Non-existent	Non-existent	Non-existent	Fabricated
Review	Owston, R. D., & York, D. N. (2018). The effect of blended learning on student performance at course and program levels. Education and Information Technologies, 23(2), 633-644.	Non-existent	Non-existent	Non-existent	Fabricated
Interdisciplinary	Alias, N. A., & Zainuddin, A. M. (2018). Innovation for better teaching and learning: Adopting the flipped classroom approach. European Journal of Social Sciences Education and Research, 12(1), 110-116.	Non-existent	Non-existent	Non-existent	Fabricated
Interdisciplinary	Clinefelter, D. L., & Aslanian, C. B. (2017). Online college students 2017: Comprehensive data on demands and preferences. Learning House.	Non-existent	Non-existent	Non-existent	Fabricated
Interdisciplinary	Fokides, E., Atsikpasi, P., & Kourtis-Kazoullis, V. (2019). Factors affecting students' motivation and academic performance in e-learning: The case of a university in Greece. European Journal of Educational Research, 8(2), 457-473.	Non-existent	Non-existent	Non-existent	Fabricated
Interdisciplinary	Jeno, L. M., & Grytnes, J. A. (2018). Motivation and learning environment: A longitudinal study of the relationship between student motivation and learning environment. Educational Psychology, 38(6), 759-780.	Non-existent	Non-existent	Non-existent	Fabricated
Interdisciplinary	McGill, T. J., & Klobas, J. E. (2018). The role of spreadsheet knowledge in technology acceptance. Computers & Education, 50(4), 1415-1426.	Non-existent	Non-existent	Non-existent	Fabricated