Relevance in Reference Recommendation by LLMs: A Meta-Analysis

Domonkos Kostyál, Tanguy Lissenko, and Théo Fuhrmann

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
`domonkos.kostyal01@estudiant.upf.edu`
`tanguy.lissenko01@estudiant.upf.edu`
`theo.fuhrmann01@estudiant.upf.edu`

Abstract. Large Language Models (LLMs) are increasingly used in academic research, including for the task of recommending relevant references. However, there are concerns about the accuracy and relevance of LLM-generated citations. This meta-analysis examines the performance of three state-of-the-art LLMs - Chat-GPT, Claude, and Gemini - recommending relevant scientific papers for 6 research topics with 3 subcategories each in music and sports. Relevance is assessed based on the hallucination rate, publication date, citation count, and topic alignment. The findings reveal that Chat-GPT exhibits the highest citation accuracy (63.3%), but all three models struggle with nuanced citation tasks, sometimes recommending non-existent or irrelevant papers. The results suggest that LLMs can be valuable tools for reference recommendation, but researchers should exercise caution and critically evaluate LLM outputs to ensure accuracy and relevance. Further research should explore the impact of model-specific training data and internet access capabilities on reference recommendation performance.

Keywords: Meta-research · Relevance · Artificial intelligence · Large Language Models · Chat-GPT · Claude · Gemini

1 Introduction

References are the fundamentals of quality research. Therefore, finding relevant papers on a topic is crucial for several key reasons. Most importantly helping to establish credibility, contextualize your contributions, and acknowledge prior works. They also have an essential role in providing ideas for a research task.

Traditional methods of obtaining references include using specialized search tools such as Google Scholar, arXiv, or others. These methods are often tedious and time-consuming for researchers as the databases for a particular domain can be massive, and the selected articles are not always relevant.

In recent years the emergence of AI assistants, especially LLMs have helped people in several aspects of their life. Researchers have also drawn their attention to LLMs to incorporate them in their work for various tasks. One task for which LLMs can be useful is for collecting references for a particular subject, such as

scientific articles, theses, or books. However, there are potential risks in using LLMs for this purpose. The risks may be that the references given are not real, partially correct, irrelevant to the subject, etc.

Therefore, in order to make the best use of LLMs for the reference collection, previous research has investigated and quantified the potential risks of an uninformed use. Past work has focused largely on the accuracy of the bibliographic citations provided by Chat-GPT [4]. Earlier research on early models of Chat-GPT was more pessimistic about the accuracy of the results obtained [6]. Precision was evaluated based on criteria such as title, authors, journal name, etc. However, more optimistic results were obtained when considering more recent, high-performance versions of Chat-GPT [3,13].

Our research compares models that have not been compared in the past: Chat-GPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash. The relevance of the results obtained is judged on the basis of the percentage of hallucination rate (references that do not exist), the date of publication, the journal impact factor (JIF), and the number of citations. We also make a comparison on the basis of different topics, which we subdivide into three categories corresponding to a notion of domain complexity, quantified by the number of publications published on this topic.

We start with a review of the literature, examining what has already been done on the subject. Then we explain our methodology, which is based on a common prompt that we submit to the various models we study. This is followed by an explanation of our results, based on an analysis of the database obtained during the experiment. We conclude with a discussion of our method, results, and future directions.

1.1 Literature Review

Many citation recommendation systems already exist, each employing distinct methods. Färber et al. [8] highlights this diversity by reviewing previously used approaches. Early methods were based on hand-crafted features, but, as he notes, *"In recent years, no novel approaches of this group have been published any more..."*. This shift occurred because deep learning techniques now consistently outperform these traditional models. Färber et al. also discuss methods like topic modeling, machine translation, and neural networks.

Jevin D. West et al. [14] propose a solution using hierarchical clustering. They first rank all papers using a unique algorithm called article-level Eigenfactor, then apply clustering to the results. Some approaches, as noted by Renuka, S. and Raj Kiran [12], leverage Natural Language Processing (NLP) to process and classify articles more comprehensively compared to traditional approaches.

The use of large language models (LLMs) has expanded from everyday life into research fields as well. Nitin et al.'s study [11] explores Chat-GPT's applications across various academic disciplines, examining how LLMs can enhance learning experiences and streamline administrative tasks, while also addressing critical issues related to bias, ethics, and data privacy.

However, as the role of AI in research grows, concerns about over-reliance are also increasing. Buçinca et al. [2] warn that AI's presence in decision-making can lead to *"automation bias"* where researchers accept AI outputs uncritically, potentially bypassing critical literacy skills in favor of AI convenience. This dependence risks diminishing the researcher's role in actively interpreting and validating data findings, which is fundamental to rigorous scientific inquiry.

We identified two significant studies closely related to our research, making it worthwhile to examine their findings in detail. The first research was conducted by Byun et al. [3] and the second is authored by Chelli et al. [5]. Byun compared the accuracy of different GPTs, with the latest, GPT-4. Their method involved providing Chat-GPT with titles and abstracts of existing papers and requesting related works. They tested three different prompts and evaluated them separately. The prompt that yielded the most comparable results in their study was the one in which they asked for only references:

"You are an [NLP or HCI] researcher working on a paper to submit to [EMNLP or CHI]. The paper you are working on is titled: [PAPER TITLE] The abstract for your paper is: [PAPER ABSTRACT] List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format."

With this query, they collected twenty papers from two different venues (HCI, NLP), and their final dataset had 1616 annotated citations. Their total accuracy in the titles was 66% and the precision of authors was 88.64% using GPT-4. However, while these accuracy rates offer valuable insights into the technical capabilities of GPT-4, they do not address whether the recommended citations are genuinely relevant to the specific context and needs of each research topic, a limitation that remains to be investigated. If we compare these to Chelli's results which are in general just a brief overview focusing on medical fields, they collected similar results from 471 citations: The hallucination rate was 28.6% for GPT-4. They also evaluated Bard (the older version of Google Gemini) and their result was 91.4%. To convert their metrics to accuracy we just need to extract them from 1, so respectively 71.4% and 8.6% which is comparable to Byun's result.

We can see that previous research has largely concentrated on metrics tied to information accuracy. Although understanding accuracy is essential, accurate citations that lack relevance remain unhelpful for researchers. In this study, we continue to evaluate accuracy, but we also provide an initial evaluation of the relevance of citations identified by the three aforementioned models.

2   Research methodology

To systematically evaluate the relevance of scientific paper recommendations by LLMs, we designed a consistent prompt schema to be used across all tested models. The prompt schema aimed to simulate an academic inquiry for relevant research articles, following the format:

"I'm writing a research paper on X. To get me started, give me five relevant scientific papers on the topic. Include the authors and published date. Don't include a summary of the article."

Where X is the field of study or a specific topic of interest.

2.1 Fields of study

We selected two broad categories, sports, and music, to introduce diversity and to avoid focusing solely on a single subject area. Within each category, we chose three topics that represent a gradient of specificity (broad, medium, and narrow): For music: Music, Rock Music, and Shoegaze. For sports: Sports, Basketball, and Basketball Shooting Technique (BST).

To ensure relevance in recommendations, we deliberately chose topics that are well-established rather than cutting-edge, which allows us to observe whether the LLMs prioritize recommending foundational or highly regarded works over recent publications that may not yet have proven significance.

In order to objectively assess topic specificity, we collected the number of search results for each topic on Google Scholar and JSTOR. These results indicated the information availability and topic breadth, supporting our categorization from general to highly specific topics. The following figure serves to visually justify the specificity levels and highlights how fewer resources exist as topics become more narrowly focused.
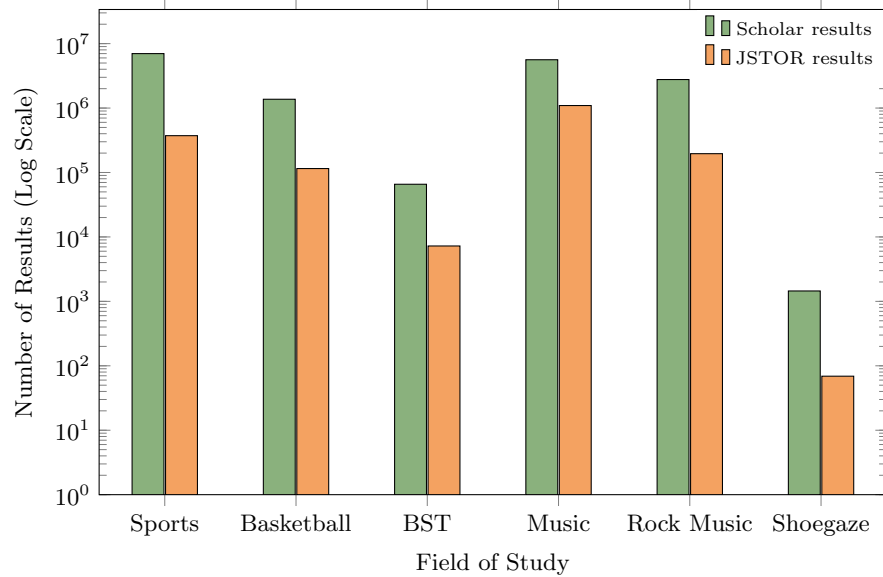

Fig. 1: Number of search results in the Scholar and JSTOR databases.

This setup allowed us to test each LLM's ability to maintain relevance in recommendations across varying degrees of topic specificity.

2.2 Data collection

For each LLM response, we collected the following information to evaluate the relevance of the suggested articles:

- **Verification of Existence:** We first verified if the recommended articles existed. Petiska [10] observed that Chat-GPT often relies on Google Scholar citation counts when suggesting citations. Therefore, we searched Google Scholar for each paper, and if not found there, we attempted to locate it through a regular internet search. For each verified paper, we analyzed the accuracy of the **title**, **authors**, and **publication date**.
- **Publication date:** Recency is important for relevance, as recent works often reflect the latest developments. Modern search engines prioritize new publications, but older influential works can also be highly relevant, especially in well-established fields. Dong et al. [7] emphasize that both recency and reliability affect trust and relevance in information retrieval.
- **Citation Count:** The citation count reflects a paper's relevance and recognition within the scientific community, as a higher number of citations often indicates greater academic impact. According to Bornmann [1], the frequency with which a paper is cited demonstrates its significance and influence. However, it's important to note that more recent papers are less likely to have high citation counts due to their recency, despite potentially being highly relevant in a specific field.
- **Journal Category Relevance and Impact Factor:** The relevance of a journal to the research topic is crucial in determining whether a paper contributes meaningfully to the field. Journals that align with the specific research area ensure that the recommended papers are contextually appropriate. Additionally, the Impact Factor of a journal reflects its overall academic influence and credibility, serving as a measure of the journal's reputation and the quality of the research it publishes. Pagani et al. [9] used a combination of impact factor, publication year, and citation count as crucial features for developing a relevance ranking method, underscoring the importance of these factors in assessing academic relevance.

The collected features provided a framework for quantifying the relevance of each recommendation. Together, they allowed us to assess the models' effectiveness in providing high-quality academic resources for both broad and highly specific topics.

2.3 LLM features

To contextualize the performance of each LLM, we documented relevant characteristics, including the model version, knowledge cutoff date, and internet access

capabilities. These factors may influence the LLMs' recommendation accuracy and relevance, particularly when addressing topics that require up-to-date or lesser-known information.

LLM	Version	Internet Access	Knowledge cutoff
Chat-GPT	4o	Yes	October 2023[1]
Claude	3.5 Sonnet	No	April 2024[2]
Gemini	1.5 Flash	Yes	N/A

Table 1: LLM characteristics.

These model-specific features were recorded to allow for deeper insights into each model's capacity to generate relevant paper recommendations across a spectrum of topic specificity.

3   Results

3.1   Hallucination rate

Once the data has been collected, we commence with an analysis of the recommended articles' accuracy. While similar analysis has been done before, it focused mainly on older versions of LLMs, particularly Chat-GPT. This analysis provides an opportunity to compare with previous research and evaluate whether these models have improved over time.

Our results can be seen in table 2. The following list provides a detailed description of the used labeling for clarification:

- **Correct:** The article is existent and it's referenced correctly.
- **Confused 2 articles:** It confused the names and authors of two different articles. In this case, we decided to keep both articles for the analysis.
- **Wrong author:** These were the situations where the title was correct, but at least one of the authors was incorrectly identified. Usually in these cases, the publication year was also wrong. Despite that, we decided not to focus on the year to simplify the labels.
- **Wrong title:** The title was partially or completely wrong, but the authors and year were correct and belonged to a paper related to the field.
- **Web article:** This is made specifically for Gemini, as it sometimes recommends web pages instead of the requested academic papers.
- **Not Existent:** It's impossible to identify the recommended paper.

To better understand the general accuracy of each model, it is useful to check the accuracy of all titles, whether they exist or not. As shown in Table 3,

---

[1] Source: OpenAI Documentation, available at `https://platform.openai.com`

[2] Source: Anthropic Model Card, available at `https://assets.anthropic.com`
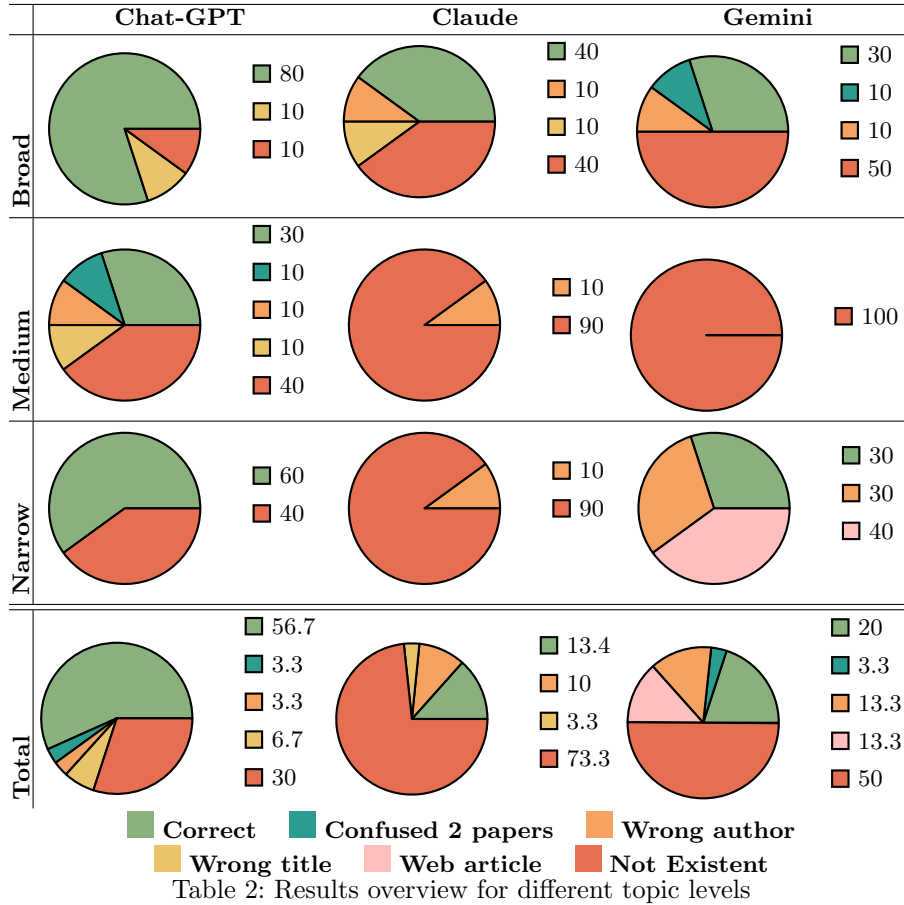
Table 2: Results overview for different topic levels

Chat-GPT outperformed all other models. Interestingly, Gemini demonstrated accuracy comparable to Chat-GPT in the narrow-topic case, despite more than half of its recommendations consisting of web articles. Comparing these results to previous research, both Byun (88.64% with GPT-4) and Chelly (71.4%) reported higher accuracy rates. This discrepancy could be attributed to the limited size of our dataset and the narrow range of topics. However, Gemini showed significant improvement, rising from Chelly's 8.6% (with Bard) to 36.7%. Despite this progress, it achieved only half the performance rate of GPT-4.

Similar trends are observed in author accuracy (Table 4), where Chat-GPT achieved 90.6%, closely matching Byun's results. Meanwhile, Claude and Gemini demonstrated comparable performance, each achieving approximately 60%.

In table 2, we wanted to separate when the model confused two articles or just made up the authors. This phenomenon presents a problem due to the

	Chat-GPT	Claude	Gemini
Broad	80%	60%	50%
Medium	50%	10%	0%
Narrow	60%	10%	60%
Total	63.3%	26.7%	36.7%

Table 3: Title accuracy

	Chat-GPT	Claude	Gemini
Total	90.6%	55.6%	66.8%

Table 4: Author accuracy

inherent nature of LLMs and how they train on datasets. It's model-independent, although it mostly occurred when the topic was more general.

Despite our expectations, there wasn't a significant decrease in accuracy when we narrowed down the topics. Only Claude showed the closest behavior, however it performed the worst probably due to the lack of web access. On the other hand, Gemini also recommended some websites with relevant articles, despite our prompt explicitly requesting research papers. In our case, this was acceptable because we focused on general topics to gain a broader understanding of a field of study.

Surprisingly, the "middle" queries, which were the rock music and basketball topics, performed the worst. Claude hallucinated nearly all of its responses, finding only one relevant source. However, Gemini did even worse, fabricating articles in both fields entirely.

3.2   Publication Years

Figure 2a illustrates the distribution of citation years for bibliographic references retrieved by the three LLMs. Each box plot provides an overview of the spread of publication years. Chat-GPT was assessed using 22 published, non-fiction articles. Claude's evaluation was based on 8 published, non-fiction articles, and Gemini was assessed with 20 published, non-fiction articles. For Chat-GPT, the median publication date is 2015, with the interquartile range spanning from 2008 to 2023. This distribution indicates a potential preference for more recent sources, as the second quartile falls in 2023 and the maximum year reaches 2024. For Claude, the median publication date is 2013, with the interquartile range between 2010.5 and 2014. This suggests a tendency toward older sources, as the second quartile is 2014 and the maximum year is 2015. For Gemini, the median publication date is 2001, with the interquartile range from 2008 to 2020. This distribution points to a broader coverage of sources from various periods. Chat-GPT and Gemini exhibit greater variability in publication years, as indicated by their higher standard deviations around the mean (13.8 and 7.25), whereas Claude shows lower variability with a smaller standard deviation (2.4).

3.3   Number of Citations

Figure 2b illustrates the distribution of citation numbers for bibliographic references retrieved by the three LLMs. Chat-GPT was assessed using 22 published,

non-fiction articles. Claude's evaluation was based on 8 published, non-fiction articles, and Gemini was assessed with 15 published non-fiction articles. Chat-GPT has a median citation number of 59.5 with the interquartile range spanning from 1.0 to 177.5. The standard deviation is 460. Claude has a median citation number of 160 with the interquartile range spanning from 135 to 860.5. The standard deviation is 510.82. Gemini has a median citation number of 18 with the interquartile range spanning from 5 to 1455.5. The standard deviation is 1606.35. From our results, we find that Gemini has the highest citation counts and the highest variability while Chat-GPT has the lowest average and a wider range of citation numbers. Claude's citation counts are more concentrated around the median, with fewer extreme values.

3.4   Journal Impact Factor

Figure 2c illustrates the distribution of journal impact factors for bibliographic references retrieved by the three LLMs. Chat-GPT was assessed using 18 published, non-fiction articles. Claude's evaluation was based on 6 published, non-fiction articles, and Gemini was assessed with 5 published, non-fiction articles. Chat-GPT has a median JIF of 2.9 with the interquartile range spanning from 2.3 to 2.975. The standard deviation is 2.3. Claude has a median JIF of 2.4 with the interquartile range spanning from 1.9 to 2.975. The standard deviation is 3.5. Gemini has a median JIF of 3.7 with the interquartile range spanning from 0.9 to 3.9. The standard deviation is 3.55.

3.5   Journal Categories

The last feature we analyzed was the journal categories using the tables 5 and 6. Starting with Chat-GPT, we can observe a more generalist behavior in the Sports, Shoegaze, and BST topics. However, it introduced journals with more diversity in Basketball, Music, and Rock Music. As for Basketball and Music, not only it explored interesting related categories such as biology, education, and mathematics, but it also identified exact topic matches, recommending "sports sciences" and "music" respectively.

In contrast, Claude had more specialized categories. It showed very relevant and varied categories for the Sport category. Moreover, it identified the matching topic for BST. As for the musical topics, it showed special focus in biology for Music and psychology for Basketball.

Gemini obtained a mix of specialized and generalist categories. It found relevant categories for Sports focusing more on medical fields. It also identified the exact match on the BST category. In the music domain, it was broader in the Music topic and showed more focus on psychology and linguistics in Rock Music and Shoegaze respectively.

3.6   Discussion

Our first finding in this study is that the hallucination rate in newer versions of LLMs remains persistent when providing reference recommendations. However,

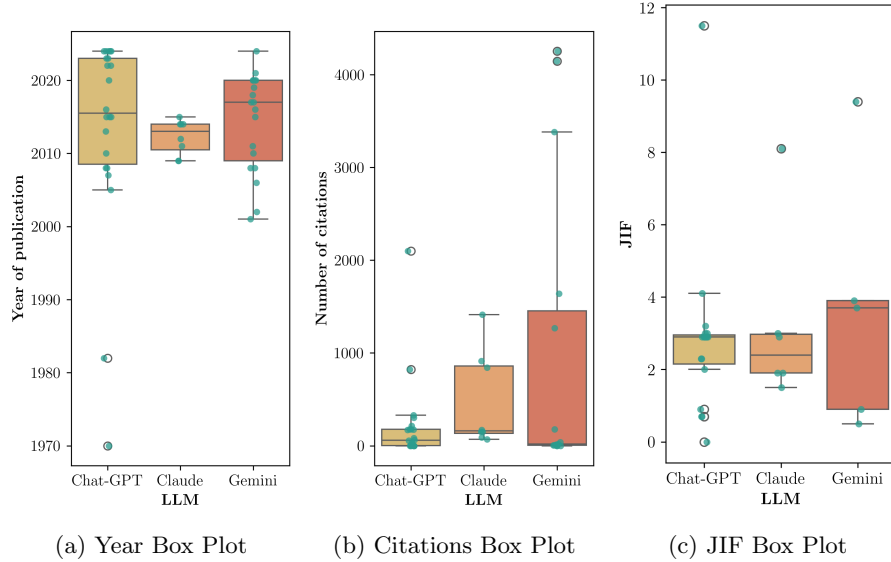(a) Year Box Plot   (b) Citations Box Plot   (c) JIF Box Plot

Fig. 2: Year (a), Citations (b), and JIF (c) box plots. Note: the amount of samples per feature varies due to the different kinds of media collected by the LLMs (i.e. book references didn't belong to any journal, therefore they had no JIF)

Chat-GPT consistently outperformed other models in citation accuracy, particularly in identifying correct titles and authors, with a high accuracy rate of 90.6% in author identification, closely aligning with prior studies, even though OpenAI claims[1] GPT-4o outperforms GPT-4 in most of the benchmarks.

While Gemini showed competitive performance in narrow topics, it also recommended non-academic web articles, highlighting limitations in aligning outputs with specific prompts. These results suggest that, while LLMs are capable of high citation accuracy, they still face challenges with nuanced citation tasks which can lead to confusion for researchers if they heavily rely on them.

Regarding the publication date, Chat-GPT was the most relevant for a user in need of access to recent references. This makes it especially suitable for current and evolving research topics. While Claude is reliable for consistent and older references, the range is not as extensive as in the other models tested here. This makes the tool less ideal for current research needs, probably because of the lack of internet access. Gemini offers a middle ground with its balanced coverage, catering to both recent and older sources, and this adaptability should meet a range of bibliographic needs.

The number of citations showed that Gemini performs the best as it has extremely high variability and potentially sources highly cited references. However,

---

[1] Source: OpenAI GPT-4o release notes, available at `https://openai.com/index/hello-gpt-4o/`

	Basketball	BST	Sports
Chat-GPT	biology, multidisciplinary sciences, biomedical, education, educational research, engineering, sport sciences	business, multidisciplinary sciences	multidisciplinary, multidisciplinary sciences, psychology
Claude	applied, psychology	sport sciences	applied, dietetics, hospitality, leisure, nutrition, psychology, sport, sport sciences, tourism
Gemini		sport sciences	geriatrics, gerontology, gynecology, obstetrics

Table 5: Journal Categories for Sports-Related Topics

	Music	Rock Music	Shoegaze
Chat-GPT	interdisciplinary applications, mathematical methods, mathematics, multidisciplinary sciences, music, social sciences	interdisciplinary applications, mathematical methods, mathematics, social sciences	multidisciplinary, multidisciplinary sciences, psychology
Claude	biochemistry, biology, cell biology, molecular biology, multidisciplinary sciences		
Gemini	multidisciplinary sciences	developmental, psychology	language, linguistics

Table 6: Journal Categories for Music-Related Topics

it often retrieves less-cited articles. This variability can be useful in projects requiring a general perspective but is less suitable for focused research requiring consistently high-impact sources. Nonetheless, Chat-GPT has a wide range in the number of citations it applies, it sources both lower-cited articles and some with more influential references. This makes Chat-GPT versatile for such kinds of research that will make full use of a wide range of citation counts. Claude is more relevant for work that requires reliable, moderately cited references when extreme outliers in the citation counts are less welcome.

As for the JIF, Gemini has the higher mean. Simultaneously, it reports high variability. This means that Gemini has the capability to source articles from journals with a higher average impact; meanwhile, there is some inconsistent reporting as the impact factors range quite widely. Claude has a larger range compared to Chat-GPT, though more limited than Gemini in its reach toward higher-impact journals.

However, these results should be treated with caution, as the number of samples changed for all the analyzed features due to the LLMs' variety of media and performances. The lack of samples in certain features is explained by the hallucination rate.

It's also worth mentioning that we separated the topics into three breadth categories aiming to find performance correlations, but we haven't been able to observe a general pattern in any of the analyzed features regarding relevance.

4   Conclusions

This paper has presented a comparative study related to the relevance and accuracy of reference recommendations made by three large language models: Chat-GPT 4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash. Testing them against a coherent prompt schema for diverse research topics gave insight into their usability in academic work. Chat-GPT was proven to be the most accurate model, with high precision in author identification and title accuracy. It achieved a good balance between the citation count and journal impact factor, and for that reason, it is good at sourcing relevant academic references when recency is an issue. Claude delivered a stable performance regarding citation concentration but struggled with hallucination. While Gemini showed promise in narrow-topic relevance, it also consistently suggested non-academic web content, which is a limitation in aligning with scholarly requirements.

The findings emphasize how one must consider both the issues of accuracy and relevance while using an LLM as a research assistant. These models have indeed improved from previous versions, but performance remains varied depending on topic specificity and even the nature of references needed. Therefore, researchers should use caution and verification of outputs because mere reliance on an LLM may imply the perpetuation of inaccuracies or biases.

The findings emphasize the fact that the studied LLMs, while useful for starting research, are still limited and should not be fully relied upon. Researchers should use them as complementary tools, but also perform manual queries in databases to find relevant recommendations. The models should be used with caution and verification, as their performance is volatile and may not always provide accurate or relevant results.

In this regard, future studies are encouraged to expand the dataset and explore factors such as model-specific training data and internet access capabilities. This could provide an explanation of how such factors contribute to the strengths and weaknesses noted in the reference recommendations. Hybrid systems that couple LLM capabilities with human oversight have the potential to enhance reliability and contextual relevance in academic citation tools.

Code and Data Availability

The collected data and code for the plots can be found in the following GitHub repository: `https://github.com/theofuhrmann/RMProject`

References

1. Lutz Bornmann. Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, 73:775–787, 2017.
2. Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

3. Courtni Byun, Piper Vasicek, and Kevin Seppi. This reference does not exist: An exploration of LLM citation accuracy and relevance. In Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao, editors, *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–39, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
4. ChatGPT by OpenAI. Chatgpt: An advanced ai language model. `https://chatgpt.r4wand.eu.org/`, 2024. Accessed: 2024-11-09.
5. M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J. Raynier, G. Clowez, P. Boileau, and C. Ruetsch-Chelli. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *J Med Internet Res*, 26:e53164, 2024.
6. Terrence Day. A preliminary investigation of fake peer-reviewed citations and references generated by chatgpt. *The Professional Geographer*, 75(6):1024–1027, 2023.
7. Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 11–20, 2010.
8. Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405, 2020.
9. Regina Negri Pagani, João Luiz Kovaleski, and Luis Mauricio Resende. Methodi ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. *Scientometrics*, 105:2109–2135, 2015.
10. Eduard Petiska. Chatgpt cites the most-cited articles and journals, relying solely on google scholar's citation counts. as a result, ai may amplify the matthew effect in environmental science, 2023.
11. Nitin Rane, Abhijeet Tawde, Saurabh Choudhary, and Jayesh Rane. Contribution and performance of chatgpt and other large language models (llm) for scientific and research advancements: a double-edged sword. 5:875–899, 10 2023.
12. S. Renuka, G. S. S. Raj Kiran, and Palakodeti Rohit. An unsupervised content-based article recommendation system using natural language processing. In I. Jeena Jacob, Selvanayaki Kolandapalayam Shanmugam, Selwyn Piramuthu, and Przemyslaw Falkowski-Gilski, editors, *Data Intelligence and Cognitive Informatics*, pages 165–180, Singapore, 2021. Springer Singapore.
13. William Walters and Esther Wilder. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13:article 14045, 09 2023.
14. Jevin D. West, Ian Wesley-Smith, and Carl T. Bergstrom. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, 2(2):113–123, 2016.