Predicting The Scholarly Impact of Research Papers Using Retrieval-Augmented LLMs

Tamjid Azad[†], **Ibrahim Al Azher**[†], **Sagnik Ray Choudhury**[‡], **Hamed Alhoori**[†]
[†]Northern Illinois University, [‡]University of North Texas
tamjidazad@gmail.com, iazher@niu.edu, sagnik.raychoudhury@unt.edu, alhoori@niu.edu

Abstract

Assessing a research paper's scholarly impact is an important phase in the scientific research process; however, metrics typically take some time after publication to accurately capture the impact. Our study examines how Large Language Models (LLMs) can predict scholarly impact accurately. We utilize Retrieval-Augmented Generation (RAG) to examine the degree to which the LLM performance improves compared to zero-shot prompting. Results show that LLama3-8b with RAG achieved the best overall performance, while Gemma-7b benefited the most from RAG, exhibiting the most significant reduction in Mean Absolute Error (MAE). Our findings suggest that retrieval-augmented LLMs offer a promising approach for early research evaluation. Our code and dataset for this project are publicly available [1] [2].

1 Introduction

Evaluating the impact of a research paper is important to the scientific process, as researchers, funding agencies, and policymakers must make informed decisions (Akella et al., 2021). Typically, impact has been measured using bibliometric indicators, such as citation counts, h-index, i-index, and journal impact factors (Gupta et al., 2023; Waltman, 2016), as well as field-normalized metrics such as Field Citation Ratio (FCR) and Relative Citation Ratio (RCR) (Hutchins et al., 2016; Purkayastha et al., 2019). While each of these metrics provides useful insights into a paper's impact (Gupta et al., 2023), they depend on citation data, which takes time to accumulate.

The delay in assessing the scholarly impact can cause a challenge when making decisions in certain situations. For example, organizations that allocate funding for investments may need to assess the potential of new publications to guide funding, or domains that are evolving quickly may need to identify influential work that is important for directing researchers' attention. While alternative metrics such as altmetrics attempt to capture the engagement of the public immediately through social media and news coverage (Thelwall et al., 2013; Shahzad et al., 2022; Shaikh et al., 2023), they also rely on data after publication and thus cannot provide a true preemptive evaluation.

In this study, we estimate the scholarly impact of research papers by analyzing their content using Large Language Models (LLMs). LLMs have opened new possibilities for evaluating impact (Zhang et al., 2023), allowing researchers to rigorously analyze the research paper is content for more insights (de Winter, 2024; Zhao et al., 2025; Thelwall, 2025). However, despite these models being trained on a vast corpus, their knowledge is fixed at the time of training, so they can't dynamically access external sources during inference (Wang et al., 2024a).

This limitation means that for predicting scholarly impact, LLMs cannot evaluate how a new paper compares to prior related studies or assess its contribution in the context of ongoing research. Since a paper's influence often depends on how original it is compared to prior work and how relevant it is to ongoing research, (James et al., 2023; De Silva et al., 2017), comparing it to other studies is essential for accurately assessing its potential impact.

To address this concern, we use a technique called Retrieval Augmented Generation (RAG), where the retriever collects external sources that are semantically similar to the query being evaluated. These sources are sent to the LLM as context to give a more informed response (Gao et al., 2023). In the context of scientific articles, recent work shows that RAG improves the generation of structured scientific content, such as future work

[1] Code
[2] Dataset

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 124–131
July 31, 2025 ©2025 Association for Computational Linguistics

statements, by grounding predictions in relevant prior research (Azher et al., 2025). RAG could potentially be valuable in the case of impact prediction, where we use prior literature to help LLM reason better. In our study, we will be addressing the following research questions:

RQ1: Does RAG improve the overall performance of LLM compared to zero-shot for scholarly impact prediction?

RQ2: How well do predictions from LLMs generalize across different research disciplines?

RQ3: How often did RAG improve or degrade LLM performance among individual papers?

2 Dataset Collection

We collected research articles published between 2018 and 2022 across five disciplines: Computer Science, Mathematics, Engineering, Physical Sciences, and Psychology. For each discipline and year, we randomly sampled 2,000 articles, extracting their titles, abstracts, and FCR scores from Dimensions.ai[3]. This creates a diverse dataset, sufficient to test the generalizability of prediction models. The FCR adjusts a paper's citation count by comparing it to the average citations of papers in the same field and publication year (Hutchins et al., 2016). Since FCR is an unbounded metric with no upper limit, we used the empirical cumulative distribution function (ECDF) to normalize its values within a $0 - 1$ range for each discipline and publication year. This makes our data more consistent and suitable for the model to learn and analyze for prediction (Kwok et al., 2023).

We then preprocessed the title and abstract columns by converting text to lowercase, discarding special characters, and removing abstracts with fewer than 100 tokens. Based on the ECDF-normalized FCR values, each article was categorized into one of three impact levels: *low* $(0-0.33)$, *medium* $(0.34 - 0.66)$, or *high* $(0.67 - 1)$ impact level. These categorical labels were used to examine the distribution of impact levels within the dataset. To mitigate class imbalance and reduce the risk of model overfitting or bias toward dominant citation patterns, we removed overrepresented classes.

We then evaluated the readability of each paper's abstract using the textstat[4] library. These readability metrics were used as additional input features to quantify the ease or difficulty of text comprehension. Since readability can influence citation patterns, including these metrics enables us to investigate its potential role in the LLM's ability to predict scholarly impact (Ante, 2022; Wang et al., 2022). We used the Flesch Reading Ease (FRE) score[5], where higher values indicate more readable text, and the Gunning Fog Index (GFI)[6], which estimates the years of education required to comprehend the text (DuBay, 2004).

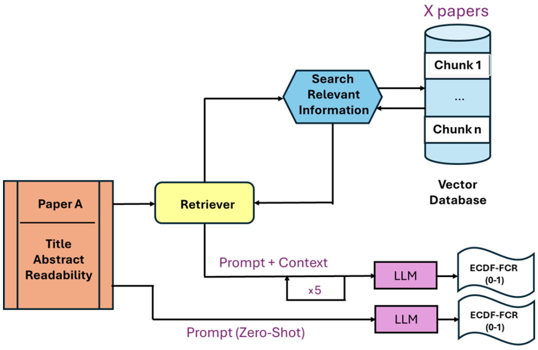

Figure 1: Overview of the RAG and Zero-shot workflow to predict the normalized FCR score.

After preprocessing, the dataset consisted of $6,000$ research articles, each containing its title, abstract, abstract readability scores, and normalized FCR scores. In the experiment, we divided the dataset into a knowledge base containing $5,400$ papers (90% used for retrieval) and a test set of 600 papers (10% used to evaluate the model's predictions).

3 Methodology

Figure 1 illustrates our workflow, which uses RAG to assist the LLM in making its prediction. The experiment used three LLMs (LLama3-8b, Mistral-7b, and Gemma-7b) and a retrieval-augmented setup that combined dense retrieval for contextual grounding and self-consistency to improve prediction reliability. Besides RAG, we used zero-shot as a baseline to assess how much the retriever actually benefited the LLM performance.

3.1 Large Language Models

Zero-Shot Prompting: We use zero-shot as a baseline, where we instruct the LLM to predict the nor-

[3] https://www.dimensions.ai/

[4] https://pypi.org/project/textstat/

[5] https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/

[6] https://readable.com/readability/gunning-fog-index/

malized FCR score (ECDF-FCR) using just the title and abstract of the paper. Previous research has demonstrated the efficacy of zero-shot approaches for tasks such as predicting citation intent, displaying LLM's ability to perform well without additional fine-tuning (Koloveas et al., 2025; Alvarez et al., 2024). As such, zero-shot prompting serves as a benchmark for evaluating our RAG approach (Kumagai et al., 2024).

Self-Consistency: Our implementation of the RAG approach will also utilize self-consistency to further improve the reliability of the predictions. will involve prompting the LLM five times per paper and using the median score as the normalized FCR output. Self-consistency is particularly important in prediction tasks, as it improves the robustness of the model by reducing variance in the response and ensuring that the most consistent response is selected (Nguyen et al., 2024).

3.2 Retrieval-Augmented Generation

Dense Retrieval: Since RAG can be implemented in several ways, we settled on using the dense retrieval approach, which extracts the most comparable documents from a corpus given a query. This is accomplished by representing each document as an embedding and using a search method to efficiently compare pairwise similarities. Unlike keyword retrieval methods, dense retrieval maps documents and queries to a shared embedding space, allowing more semantic matching (Shi et al., 2023).

Facebook AI Similarity Search: FAISS[7] (Facebook AI Similarity Search) is an open-source library designed to find similar items in large datasets, especially when using high-dimensional vectors (Ghadekar et al., 2023; Douze et al., 2024). It supports various search methods, such as L2 distance, cosine similarity, or approximate nearest neighbors (ANN) for vector databases, making it scalable to extensive document collections.

4 Experimental Setup

We evaluated the LLM performance in both the zero-shot (Figure 2, Appendix) and RAG (Figure 3, Appendix) using two types of input sets: **(1)** text only (Title, Abstract) and **(2)** text with readability (Title, Abstract, Flesch Reading Ease, Gunning Fog Index).

We downloaded each model from Ollama[8] to run locally with the default configurations. We implemented a dense retrieval approach using FAISS to identify the most relevant research papers from the knowledge base. The FAISS index was created using the IndexFlatIP method, which is well-suited for cosine similarity search when used with normalized embeddings. Since we used SciBERT to generate the embeddings, we normalized each vector before indexing to ensure that the inner product search approximated the cosine similarity.

To efficiently compute embeddings, we processed research papers in batches of $1,000$, using parallel execution with four workers to speed up the computation. The resulting embeddings were stored directly in FAISS, enabling a flat, brute-force retrieval strategy. During the retrieval phase, the title and abstract of each input paper were encoded using SciBERT and used to query the FAISS index. The retriever will then find five papers that are the most semantically similar to the query containing the test paper and pass them to the LLM, where it will then use those five papers as context when predicting the normalized FCR score.

To evaluate the performance of the LLMs, we measured accuracy and ranking quality using Mean Absolute Error (MAE) and Normalized Discounted Cumulative Gain (NDCG). MAE quantifies accuracy by calculating the average difference between predicted and actual impact scores, with lower values indicating higher accuracy. NDCG assesses how well the model ranks papers by impact, comparing its predicted rankings of FCR scores to their actual rankings, where a value closer to 1 means that it is more accurate at ranking high-impact papers.

5 Results and Discussion

RQ1: Performance of LLMs in Zero-Shot vs. Retrieval-Augmented Generation. The results for zero-shot and RAG predictions are presented in Table 1. In zero-shot, LLama3-8b consistently outperformed Mistral and Gemma-7b in all features, achieving the lowest MAE of 0.222 and the highest NDCG of 0.936 when readability was a part of the input. In contrast, the other models had weaker performance, with Mistral-7b averaging an MAE of 0.304 and an NDCG of 0.918 between the two sets of features, and Gemma-7b receiving 0.309 and 0.916. These findings are consistent with previous research that used LLama3-8b to predict normalized citation counts for newly published articles

[7] https://github.com/facebookresearch/faiss

[8] https://ollama.com/search

Model	Title + Abstract		+ FRE, GFI	
	MAE	NDCG	MAE	NDCG
LLama3-8b	**0.227**	**0.929**	**0.222**	**0.936**
Mistral-7b	0.317	0.923	0.291	0.917
Gemma-7b	0.314	0.910	0.304	0.923

(a) Zero-shot Performance

Model + RAG	Title + Abstract		+ FRE, GFI	
	MAE	NDCG	MAE	NDCG
LLama3-8b + RAG	**0.182**	0.947	**0.195**	**0.953**
Mistral-7b + RAG	0.246	**0.955**	0.260	0.941
Gemma-7b + RAG	0.237	0.940	0.217	0.941

(b) LLM w/ RAG Performance

Table 1: Side-by-side comparison of Zero-shot and RAG performance. Metrics include MAE and NDCG across two input sets: (1) Title + Abstract and (2) Title, Abstract, Flesch Reading Ease (FRE), Gunning Fog Index (GFI).

(Zhao et al., 2025).

After integrating RAG with LLM, the performance of each model for predicting research paper impact improved, although the degree of improvement varied. Gemma-7b had the most substantial gains, reducing its MAE to 0.237 (a 0.077 decrease from zero-shot) with text-only input and 0.217 (a decrease of 0.087) when readability was considered, indicating the model depends on the external context for making its prediction. Mistral-7b also benefited, especially in text-only, where its MAE dropped to 0.246 (a reduction of 0.071). In contrast, LLama3-8b experienced the smallest improvements from RAG, with MAE reductions of 0.045 and 0.027, but still had the lowest MAE out of all models.

RQ2: LLM Prediction Generalizable Across Domains. The influence of RAG on the accuracy of the prediction varied across different domains (Figure 4, Appendix), with some fields benefiting more than others. Computer Science and Engineering showed the most significant improvements across most models, with Gemma-7b showing a reduction in MAE of 0.105 in both fields, the most substantial gain among all domains. Mistral also showed strong improvements, decreasing its MAE by 0.055 in Computer Science and 0.059 in Engineering, while LLama3-8b showed the highest improvement in Engineering only.

RQ3: How Often RAG Improve or Degrade LLM performance. To assess whether the retriever improved or worsened the LLM performance, we compared the absolute prediction error of RAG and zero-shot for each paper across all LLMs. Overall, RAG achieved a performance superior to zero-shot in $57-59\%$ in all cases (Mistral-7B: $1,369$, LLaMA3-8B: $1,376$, Gemma-7B: $1,409$), while zero-shot outperformed RAG in $36-38\%$ of cases (Mistral-7B: 873, LLaMA3-8B: 908, Gemma-7B: 874). This further shows that the context provided by the retriever generally improved the prediction but was not universally effective. These results reveal that while RAG can help, it also introduces noise or conflicting information, a challenge also addressed in Astute RAG, which investigates how to detect and mitigate such retrieval failures in LLMs (Wang et al., 2024b).

6 Conclusion

The evaluation process provides insight into a paper's impact and contribution to the research community. Our study attempts to expedite that process by prompting an LLM with a paper's title, abstract, and abstract readability. To improve the LLM response, we also incorporate RAG, which retrieves relevant papers as context when LLM makes its prediction, offering a faster alternative for assessing impact. While RAG improved the prediction overall, its inconsistent performance in some instances highlights the need to refine the retrieval approach further.

Limitations and Future Works

Our study has some limitations that allow opportunities for further improvement. First, the retrieval mechanism returned irrelevant or low-quality documents, sometimes degrading the prediction. Secondly, the input features for the text are limited to only the title and abstract, which overlooks other sections that could help the LLM. Lastly, because FCR requires at least two years of citation data, the ground truth is unavailable for recently published papers, preventing us from evaluating the performance of newer work. Future work will improve retrieval quality by experimenting with more techniques such as hybrid retrieval or re-ranking methods to match relevant documents better. Furthermore, we will expand the input beyond just the title and abstract, such as the introduction, methodology, and limitations (Azher et al., 2024), so that the model has more content to work with.

References

Akhil Pandey Akella, Hamed Alhoori, Pavan Ravikanth Kondamudi, Cole Freeman, and Haiming Zhou. 2021. Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics*, 15(2):101128.

Carlos Alvarez, Maxwell Bennett, and Lucy Lu Wang. 2024. Zero-shot scientific claim verification using llms and citation text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 269–276.

Lennart Ante. 2022. The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics*, 16(1):101252.

Ibrahim Al Azher, Miftahul Jannat Mokarrama, Zhishuai Guo, Sagnik Ray Choudhury, and Hamed Alhoori. 2025. Futuregen: Llm-rag approach to generate the future work of scientific article. *arXiv:2503.16561*.

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12.

Pali UK De Silva, Candace K Vance, Pali UK De Silva, and Candace K Vance. 2017. Measuring the impact of scientific research. *Scientific scholarly communication: the changing landscape*, pages 101–115.

Joost de Winter. 2024. Can chatgpt be used to predict citation counts, readership, and social media interaction? an exploration among 2222 scientific abstracts. *Scientometrics*, 129(4):2469–2487.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv:2401.08281*.

William H. DuBay. 2004. The principles of readability. Technical Report ED490073, ERIC Clearinghouse.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*, 2.

Premanand P Ghadekar, Sahil Mohite, Omkar More, Praiwal Patil, Shubham Mangrule, et al. 2023. Sentence meaning similarity detector using faiss. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6. IEEE.

Shikha Gupta, Naveen Kumar, and Subhash Bhalla. 2023. Citation metrics and evaluation of journals and conferences. *Journal of Information Science*, page 01655515231151411.

B Ian Hutchins, Xin Yuan, James M Anderson, and George M Santangelo. 2016. Relative citation ratio (rcr): a new metric that uses citation rates to measure influence at the article level. *PLoS biology*, 14(9):e1002541.

Mathew James, Vikas Palakkat, and Gareth JF Jones. 2023. Identifying influential citations in scientific papers. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–4. IEEE.

Paris Koloveas, Serafeim Chatzopoulos, Thanasis Vergoulis, and Christos Tryfonopoulos. 2025. Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. *arXiv:2502.14561*.

Atsutoshi Kumagai, Tomoharu Iwata, and Yasuhiro Fujiwara. 2024. Zero-shot task adaptation with relevant feature information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13283–13291.

Wai Meng Kwok, George Streftaris, and Sarat Chandra Dass. 2023. A novel target value standardization method based on cumulative distribution functions for training artificial neural networks. In *2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 250–255. IEEE.

Alex Nguyen, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2024. When is the consistent prediction likely to be a correct prediction? *arXiv:2407.05778*.

Amrita Purkayastha, Eleonora Palmaro, Holly J Falk-Krzesinski, and Jeroen Baas. 2019. Comparison of two article-level, field-independent citation metrics: Field-weighted citation impact (fwci) and relative citation ratio (rcr). *Journal of informetrics*, 13(2):635–642.

Murtuza Shahzad, Hamed Alhoori, Reva Freedman, and Shaikh Abdul Rahman. 2022. Quantifying the online long-term interest in research. *Journal of Informetrics*, 16(2):101288.

Abdul Rahman Shaikh, Hamed Alhoori, and Maoyuan Sun. 2023. Youtube and science: models for research impact. *Scientometrics*, 128(2):933–955.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, et al. 2023. In-context pretraining: Language modeling beyond document boundaries. *arXiv:2310.10638*.

Mike Thelwall. 2025. In which fields do chatgpt 4o scores align better than citations with research quality? *arXiv preprint arXiv:2504.04464*.

Mike Thelwall, Stefanie Haustein, Vincent Larivière, and Cassidy R Sugimoto. 2013. Do altmetrics work? twitter and ten other social web services. *PloS one*, 8(5):e64841.

Ludo Waltman. 2016. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024a. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv:2411.03350*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2024b. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *Preprint*, arXiv:2410.07176.

Shan Wang, Xiaojun Liu, and Jie Zhou. 2022. Readability is decreasing in language and linguistics. *Scientometrics*, 127(8):4697–4729.

Yang Zhang, Yufei Wang, Kai Wang, Quan Z Sheng, Lina Yao, Adnan Mahmood, Wei Emma Zhang, and Rongying Zhao. 2023. When large language models meet citation: A survey. *arXiv:2309.09727*.

Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. 2025. From words to worth: Newborn article impact prediction with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1183–1191.

Appendix

```
prompt = " You are an expert in evaluating the scholarly impact of research papers.
Given a research paper, predict its normalized FCR score, between 0 and 1, where 0 is
the lowest impact and 1 is the highest impact.

**New Paper:**
Title: the far side of mars two distant marsquakes detected by insight
Abstract: abstract for over three earth years the marsquake service has been analyzing
the data sent back from the seismic experiment for interior structure the seismometer
placed on the surface of mars by nasa insight lander. Although by October 2021, the
mars seismic catalog included 951 events, until recently...
Return only a number. Do not add explanations or text. " ' ' '

Output: 0.304
```

Figure 2: Prompt for Zero-shot with text only feature set.

```
prompt = " You are an expert in evaluating the scholarly impact of research papers.
Given a research paper, predict its normalized FCR score, between 0 and 1, where 0 is
the lowest impact and 1 is the highest impact.

**Context Papers:**
"Title: how to determine the early warning threshold value of meteorological factors on
influenza through big data analysis and machine learning
Abstract: Infectious diseases are a major health challenge for the worldwide population.
Since their rapid spread can cause great distress to the real world, in addition to taking
appropriate measures to curb the spread of infectious diseases..."
Flesch Reading Ease: 26.85
Gunning Fog Index: 15.69
FCR Score: 0.201

"Title: carbon emission of construction materials and reduction strategy take prefabricated
construction in China as an example Abstract: The rapid development of urbanization has
made the building industry a major source of carbon emissions. As the goal of carbon
neutrality becomes clearer, the construction industry faces serious challenges in energy
conservation and emission..."
Flesch Reading Ease: 31.31
Gunning Fog Index: 14.01
FCR Score: 0.149

**New Paper:**
"Title: leveraging user comments for the construction of recycled water infrastructure
evidence from an eyetracking experiment Abstract: Building sufficient recycled water
infrastructure is an effective way to address water shortages and environmental degradation,
playing a strategic role in resource conservation, ecological protection, and sustainable
development. Although recycled water is environmentally..."
Flesch Reading Ease: 5.7
Gunning Fog Index: 23.56
Return only a number. Do not add explanations or text.
"

Output: 0.433
```

Figure 3: Prompt using RAG with full feature set (text and readability).

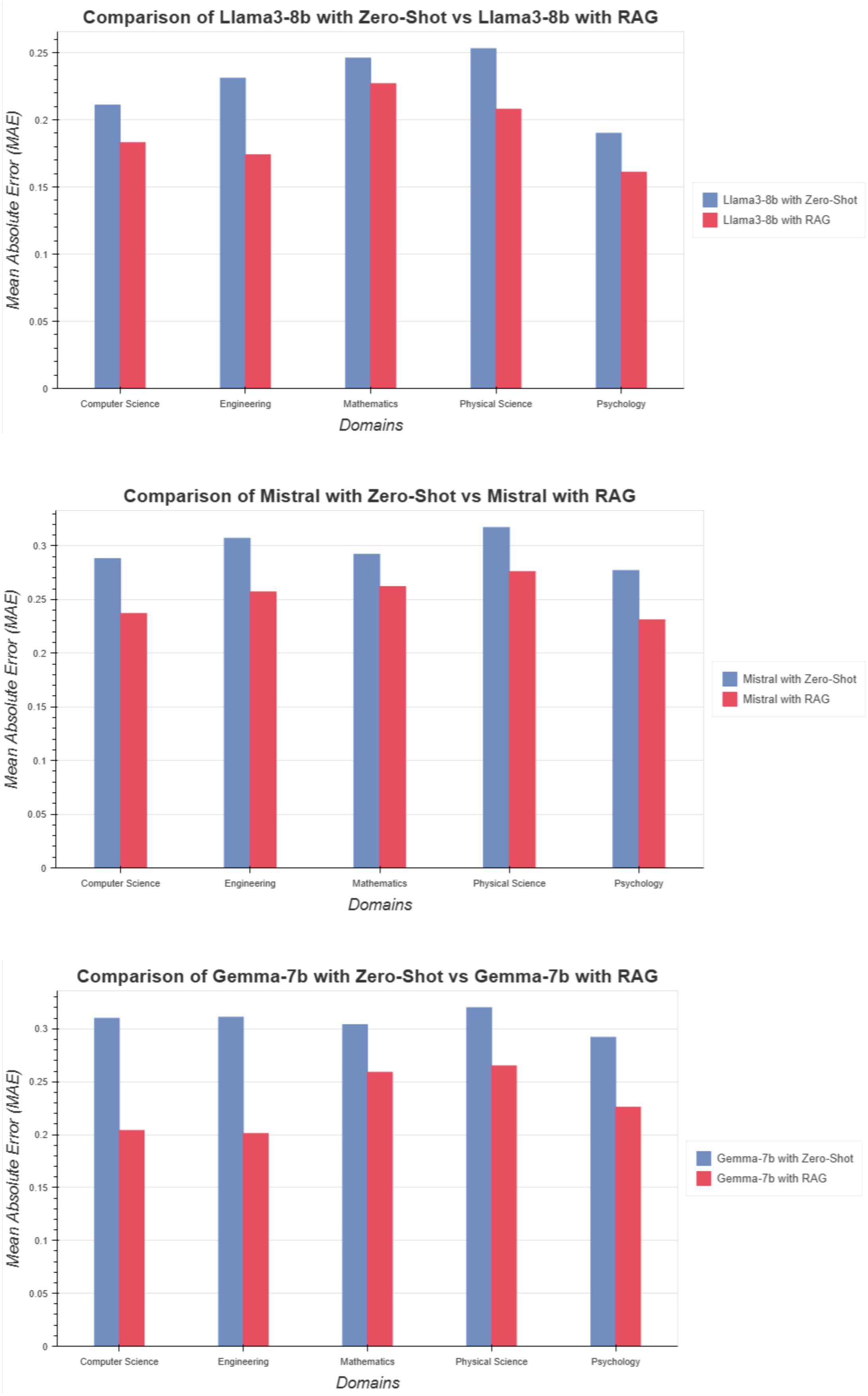

Figure 4: Comparison of MAE across domains for LLama3-8b, Mistral, and Gemma-7b using Zero-shot and RAG approach.