RAG-Driven Scholarly Assistant: Automating Research Paper Analysis with Open-Source LLM Benchmarking

Irtefa Waseek[1*], Md Rezaul Karim[1], Md Efatuzzaman Efat[2] and Sumiya Afrose[3]

[1] Department of Statistics and Data Science, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

[2] Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh

[3] Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka-1000, Bangladesh

*Corresponding author: waseekirtefa@gmail.com

Abstract. This work introduces a Retrieval-Augmented Generation (RAG)-based scholarly assistant, for automated reading of papers, which benchmarks several open-source LLMs. The developed system uses a pipeline of document processing, citation and structural analysis, and LLM-based question-answering to produce summaries and insights from academic literature. The benchmarking is done using a range of quantitative metrics majorly BLEU, METEOR, ROUGE scores. Other parameters like Perplexity, factual consistency and computational resource usage are also taken into consideration. The evaluation report is generated by the tool and provide downloadable CSV file. Visual demonstration of the data is also included in the user interface. Our developed toolkit is assessed on five domain-specific research papers (in medicine, literature, economics, computer science and mathematics) ensuring an even comparison across domains. It has been observed that the smaller RAG-based models (DeepSeek-1.5B, 8B), responds faster while exhibiting average higher factual consistency. On the contrary, the larger generative models (Mistral-7B and LLaMA3-8B) provide more detailed answers with higher overlaps. However, it costs higher computation and occasional factually inaccurate outputs. This extensive evaluation bolsters the potential for an open scholarly assistant. Furthermore, it leaves a much clearer impression of domain dependent challenges and strengths as well as a set of directions for future advancements.

Keywords: Retrieval-Augmented Generation (RAG); Large Language Models (LLMs); Automated Literature Review; AI-Driven Document Analysis; OCR; Citation Network Analysis; Benchmarking; NLP.

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_77

1 Introduction

With recent advancements in natural language processing (NLP) and large language model technologies, the capabilities of LLMs have now enabled them to perform well in understanding and generating human-like text, which has invoked interest in using LLM techniques for research assistance[1]. At present, for example, researchers are frequently overwhelmed by information when skimming literatures and AI assistants can help to solve these problems by summarizing papers, extracting main findings or answering domain-specific questions. However a common issue we face with standalone LLMs is the generation of fake or untruthful Fact claims[2]. The need is particularly acute in academic communications, where factual accuracy and good citation practices are key. Retrieval-Augmented Generation Schick and Schütze [3] have introduced the powerful RAG model to prevent hallucination by incorporating LLM outputs in external knowledge sources. In a RAG pipeline, the model retrieves documents or whether snippets (e.g., from relevant paper text or database) that provide evidence to inform on its generated responses. A side-effect of this is that one may get more accurate and contextually relevant answers than with traditional retriever / reader style ranking setups, which fits quite naturally into things like research paper analysis or question-answering.

2 Literature Review

The growth of the academic literature has rendered manual literature review impractical. Amidst the tens of thousands of journals that together publish millions of papers each year [4], even niche subdisciplines are inundated with new research. Hanson et al. refer to this trend as “the squeeze on scientific publishing” with “a drastic increase in burdens for Researcher (writing, reviewing, reading) per unit time”[5]. The explosion of information demands smart automation for evidence synthesis and knowledge organization. Early work that supports literature review included NLP based techniques, such as summarization and topic analysis. Existing extractive summarization methods (typically based on term frequency, or static embeddings) had contributed to compressing content in an easy-to-consume way – but have weak performance on more complicated source documents from the academic domain. For instance, Vishnu et al. (2020) surveyed several NLP-based summarization techniques and pointed out the challenges of dealing with different content (such as figures, tables), maintaining coherence in generated summaries [6]. It grows from contemporary transformer-based architectures and language models. Models are able to take into account contextual and semantic information far more effectively than previously, such that summaries of scientific articles can be made fluent and context-aware. A related development of interest is the retrieval-augmented generation (RAG) framework by Lewis et al. which integrates a parametric neural language model with nonparametric memory (often, a vector-indexed text database) for dynamic retrieval of information combined while generating the out-

put [3]. Lewis et al. found that RAG models fine-tuned for knowledge-heavy tasks perform better than the common sequence-to-sequence models and produce more factual and about-specific-ness outputs. This property applies especially to the literary review where facts are so important and he graduated in sources. In fact, quite a few systems have recently included RAG so that you can query massive document collections (or corpora). Scherbakov et al. (2024) offer an overall coverage of how LLMs can support different review phases. They found that most prototype systems used GPT-based LLMs to perform such tasks as searching databases, screening abstracts, extracting core information and writing sections of the review paper. In particular, their meta study point out that GPT models are overall better than BERT-based models when it comes to information extraction as they have higher average precision and recall [7]. However, very few of the published studies at that time were producing real "auto-generated" reviews – most works aimed to automate parts of a workflow rather than the entire pipeline. This indicates that while parts of literature reviews can be automatized with AI, an end-to-end assistant connecting these parts is still an open research area. Outside of summarization and Q&A, using the bibliographic information from a paper can enhance the literature review process. Such citation networks that trace the links among cited works also serve to ferret out influential studies and thematic clusters in a discipline. Haßler et al. (2024) investigated AI methodologies for automating literature review in education research and suggesting features such as visualizing citation network, map-based display of the co-citation cluster to understand a topic's landscape quickly [8]. Analysis of this nature can help expose, for instance, tightly connected clusters of papers constituting the core literature, compared with scattered papers dealing with peripheral or emerging topics. We take cues from these ideas, and in our system, it works generating references of a paper and (when feasible) linking them to external citations indexes building thus a basic citation graph. This offers readers a sense of what earlier works are most integral to the paper, and how they build on one another. Large Language Models for Scholarly Text The remarkable power of LLMs such as GPT-4 has inspired use of these models on academic text. Yet, reliability and ethical use of these AI, as well as their credibility have been subjects of concern related to research. Hua et al. (2024) point out limitations of ChatGPT, including the risk of generating fake references, lack of knowledge in the domain, and requirement for human supervision [9]. Research on assessing and enhancing the factual correctness of LLM outputs in domain specific tasks is also work-in-progress. Farquhar et al. (2023), for example, suggest the limits of factual consistency of LLM-generated text [10]. To mitigate these short-comings, in our assistant design we leverage the domain-specific models (e.g., DeepSeek tuned to research texts) and by providing source-grounded answers using RAG. Also, in comparing several LLMs directly beside each other for the same task, our effort provides hard evidence of their pros and cons toward academic use cases. This multi-model evaluation follows the recent trend of broader evaluations of language models as advocated by Bommasani et al. (2023) [11] and others, noting that a single measure or model fails to capture performance in full.

In conclusion, literature indicates that a combination of RAG methods, specialized LLMs and supporting analysis tools (e.g., citation mapping) can massively improve the way research papers are reviewed and analyzed. But there has been no unified system

that combines all of these features and systematically benchmarks the many LLM options. Our goal is thus to close this gap by developing a RAG-based research paper assistant that reflects these developments.

3 Methodology and System Architecture

3.1 System Overview:

The architecture of the Research Paper Analysis Assistant takes in a scholarly document (PDF) and generates interactive as well as analytical outputs. A simplified representation of the system architecture is shown in Figure 1. At a high-level, the system takes in a research paper and follows five stages:

PDF Processing: The input academic paper (PDF) is parsed into text while preserving structure (sections, paragraphs, figures/tables) as well as metadata. We do use a PDF extraction module which produces text in a machine-readable format and recognize important structural elements such as section titles, in-text references.

Citation Analysis: The assistant examines citations in the body and correlates them to bibliography. This step enables the system to appreciate citations in their context: for instance, knowing when a particular statement requires referencing, or when an author is referring to previous literature. The system can later use the mapping between in-text citation markers and reference entries to correctly attribute citations in its generated answers, or may also retrieve external article abstracts if necessary.

Structural analysis: In this pass, the system generates a coarse-grained summary as an over-arching structure of the paper. It may produce a draft outline of the paper (e.g., Introduction, Methods, Results, Conclusions) and pinpoint vital features like problem statement, methods used and main results. This prior structure is employed to generate relevant queries at the next stage, and to guide the LLM toward the most salient part of the text.

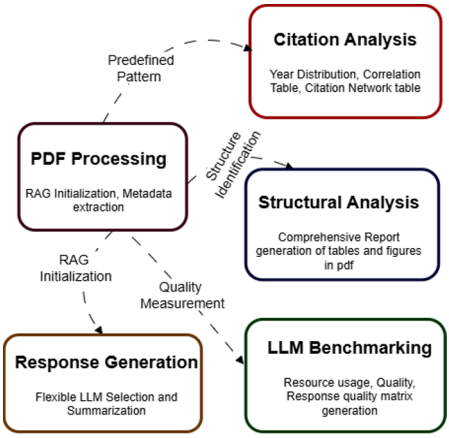

Fig 1. End to end pipeline of the RAG-driven scholarly assistant

LLM Benchmarking: Our main pipeline (MultiLLM-Pipeline) here where a few LLMs are queried to answer questions and/or summarize material related to the topical paper. We stack on top a retriever component relying on the paper's text as knowledge source (e.g., based vector embeddings or keyword lookup) to serve relevant passages to the LLM (this is "augmentation" in RAG). We then guide the LLM to generate responses, such as answers to research questions, section summaries, or citation explanations. In this open-source implementation, we test four generative models here DeepSeek-1.5B, DeepSeek-8B, Mistral-7B and LLaMA3-8B – selected as four points along the scale from smaller to faster models to bigger to more advanced ones. All models are run in the same pipeline, allowing direct performance comparison.

Response: The system retrieves and combines the LLM outputs generating a formatted response to user. This might be an interactive response (with referenced sources from the paper), a summary explanation, or a list of questions/answers.

In addition, the response generation module also logs the content and could conduct a fact-based consistency check by flagging any of such claims that are not supported in their eventuated support. At this point, the assistant can optionally check answers (for instance by fact-checking or verifying that numerical results match the paper).

Here we test the ability of this system to work in 5 different areas by choosing fifty academic articles (10 papers from each domain) which details can be found below provided table 1.

Table 1. Papers selected for multi-domain evaluation with range of published years.

Domain	Paper References	Range of Published Years
Medical	[12-21]	2024-2025
Literature	[22–31]	2013-2025
Economics	[32–41]	2021-2025
Computer Science	[42–51]	2024-2025
Mathematics	[52–61]	2023-2025

The papers act as typical case studies for each of their own fields; the papers are written in various styles and content level. All the papers were processed through all these steps of the pipeline. For each model in LLM evaluating, we posed the same set of prompts consisting in summarizing what the paper's main contributions were and a few key questions about its content. For quantitative testing, we provided reference output for the tasks (counted number of overlaps between known correct summary/answers and system generated ones) in order to allow assessing overlap-based metrics like BLEU and ROUGE. The pipeline captures all output of each model automatically and computes a set of metrics (a suite) which are stored in CSV files for analysis. We instrumented the system so we could measure response time, token use, and system resource usage (CPU and memory) for each model run to provide a detailed profile of performance as a function of cost.

3.2 Evaluation Methodology:

This assessment integrates with a wide range of benchmark metrics to characterize model performance along domains. We consider standard natural language generation quality metrics as well as efficiency metrics related to computational performance. Below we describe each metric:

BLEU Score: BLEU (Bilingual Evaluation Understudy) is a metric that gauges the n-gram overlap between the model output and a reference text. Properly proposed for machine translation, BLEU measures how many generated words appear in the reference translated and takes average of corpus. Greater BLEU is better (closer to the reference); a higher BLEU of course means closer MATCH to the reference(a perfect score would mean that an output perfectly matches with a reference)[62].

METEOR: METEOR scoring computes translation quality through generalized unigram matching and the harmonic mean of precision and recall. It generalizes BLEU to cover synonym and stem matching, and contains a fragmentation penalty that favors those outputs whose word ordering is close to the reference. METEOR's ability to balance recall and meaning match makes it word better with human assessments of quality [63].

ROUGE (ROUGE-1, ROUGE-2, ROUGE-L): Computes recall of n-grams that overlap between the output and the reference; commonly used for summarization. (i.e. ROUGE-1, ROUGE-2), while the longest common subsequence is measured for fluency (ROUGE-L) by considering sequence overlap. Higher ROUGE represents that the model preserved more reference information (which is particularly important for summary-style tasks).[64]

Factual Consistency: This is used to measure the extent to which the model's output remains factually consistent with source information or ground truth. Traditional overlap metrics like BLEU/ROUGE do not guarantee the correctness of content. As such, a fact consistency score (which can be computed using an evaluation mechanism based on a trained verifier model or QA-based approach) is employed to identify hallucination or unsupported statements[65].A value of 1.0 would imply all answers in the output are completely consistent with known facts or the input context.

Perplexity: Perplexity is a measure of uncertainty in the model's way of generating text. It is the exponentiated average negative log-likelihood of the predicted tokens. The lower the perplexity, the more confident and fluent it was in its output; conversely, the higher the perplexity value, the more "surprised" the model became at seeing new text and having to predict words for them [66]. For our benchmark tasks, a lower perplexity is better (as it signifies more fluent, well-modelled language).

N-gram Diversity: We calculate the lexical diversity by counting the percentage of unique n-grams in generated output. This is similar to the Distinct-N metrics described in Li et al. (2016), where for instance Distinct-1 and Distinct-2 are the number of distinct unigrams and bigrams in generated words over total word counts [67]. The higher the n-gram diversity (towards 1.0), this is the less repetitive and more varied text, which is desirable especially for open-end generation.

Coherence: Coherence characterizes the upper-level rationality and structure of the response. It is hard to characterize directly, but it is a critical attribute that human judges

find. In this case, soundness or coherence can also be scored on a 0–1 scale (lower is better), e.g., by the similarity of the output sentences to human ratings, which assign to an output how likely it sounds like text generated automaton via some sort of model-based metric [68]. A lower coherence score suggests that the text is easy to read and coherent.

Response Time: This refers to the time it takes a model to produce the final response from start to finish (in seconds). The average time elapsed per query of each model is reported. Lower response time corresponds to the model being more responsive to input provided by the user, a crucial characteristic in interactive systems.

BERT Score: It measures how closely a model's answer matches the meaning of a reference answer, even if the wording is different. It's like checking if two responses say the same thing in different ways.

Our metrics provide a balance across these two values, and therefore they capture quality (accuracy, fluency, diversity) as well as performance (speed) information. Next, we explore how four models – DeepSeek-1.5B, DeepSeek-8B (domain-tuned small and large), Mistral (open 7B model), LLaMA3-8B (a general-purpose 8B model) and how they perform on these metrics in total or for each domain respectively.

4 Result and Analysis

4.1 Domain Specific performance:

This section gives a comparative assessment of the four language models, namely, DeepSeek-1.5B, DeepSeek-8B, LLaMA-3-8B and Mistral, on five academic topics, including Medical, Literature, Economics, Computer Science, and Mathematics. The evaluation takes into account semantic correspondence, coherence, factual consistency and fluency. In general, there is a significant difference in performance depending on the size of the model and the architecture, and no system can show universal excellence in all measures of performance. A summary of the performance of four models DeepSeek-1.5B, DeepSeek-8B, Mistral, and LLaMA3-8B on five academic domains in terms of BLEU, BERT Score, Coherence, ROUGE-L, factual consistency, and Perplexity is presented in Table 2.

BERT Score, which measures semantic accuracy, is at all times the best with Mistral and LLaMA-3-8B. Mistral has the best semantic scores in Literature (0.793), Economics (0.801), Computer Science (0.804) and Mathematics (0.791). These scores correlate with its low range of perplexity about 15.9 to 17.3, denoting coherent and unperturbed and confidently written text. In the Medical domain (0.712) LLaMA-3-8B is the leader, and in the other domains, the models demonstrate good performance, which proves the ability of both to generate semantically aligned and linguistically natural results.

DeepSeek-8B did well in factual consistency with the highest score in Economics (0.816), Medical (0.782), and Mathematics (0.783). Nevertheless, it has a perplexity value ranging from 22 to 25 which is higher than Mistral or LLaMA-3-8B, indicating that the model prioritizes the retention of facts at the expense of readability and surface-level fluency.

LLaMA-3-8B has also high coherence with domain scores of between 0.658 to 0.795. Its text structuring skill coupled with moderate perplexity of approximately from 18 to 21, makes it a moderate performer in terms of the provision of stability to the overall organization without overwhelming fluency.

DeepSeek-1.5B, in turn, performs poorly in the semantic, coherence, and factual metrics. It has smaller values of perplexity as compared to larger models, but its smaller parameter scale restricts its robustness and cross-domain validity.

Patterns that are tied to domains are also visible: interpretive domains such as Literature tend to be less likely to get high semantic scores, and structured domains such as Economics and Computer Science tend to be more coherent and aligned. Also, the fact that all the models have received the same low BLEU and ROUGE-L scores suggests that they all prefer to use paraphrastic generation over the n-gram-level matching.

Table 2. Domain-Specific Average Performance for 5 domains

Domain	Model	BLEU	BERT Score	Coherence	ROUGE-L	Factual Consistency	Perplexity
Medical	deepseek-1.5b	0.0019	0.678	0.364	0.0826	0.726	19.73
	deepseek-8b	0.0048	0.689	0.382	0.0999	0.782	23.53
	llama3-8b	0.0003	0.712	0.463	0.0753	0.767	18.12
	mistral	0.0010	0.705	0.427	0.0819	0.727	16.65
Literature	deepseek-1.5b	0.0016	0.675	0.403	0.0829	0.776	26.07
	deepseek-8b	0.0016	0.681	0.404	0.0799	0.778	23.07
	llama3-8b	0.0002	0.745	0.409	0.0799	0.730	18.25
	mistral	0.0004	0.793	0.417	0.0820	0.741	15.92
Economics	deepseek-1.5b	0.0009	0.695	0.356	0.0647	0.758	23.98
	deepseek-8b	0.0082	0.720	0.401	0.0966	0.816	22.35
	llama3-8b	0.0002	0.782	0.420	0.0724	0.768	19.52
	mistral	0.0017	0.801	0.449	0.0853	0.771	14.84
Computer Science	deepseek-1.5b	0.0016	0.758	0.362	0.0751	0.692	19.92
	deepseek-8b	0.0036	0.729	0.390	0.0831	0.708	24.84
	llama3-8b	0.0004	0.795	0.418	0.0746	0.658	18.19
	mistral	0.0009	0.804	0.440	0.0734	0.657	17.28
Mathematics	deepseek-1.5b	0.0026	0.653	0.359	0.0861	0.770	23.35
	deepseek-8b	0.0052	0.741	0.408	0.0926	0.782	24.78
	llama3-8b	0.0002	0.796	0.402	0.0756	0.789	20.84
	mistral	0.0012	0.791	0.461	0.0843	0.789	16.28

4.2 **Overall performance:**

In all of these fields, we conclude that both models have their unique strengths and weaknesses. DeepSeek-8B is always an excellent reference-based and factual accuracy tool. It had the best BLEU and ROUGE-L in all domains, and tended to be several times better than the best model, and the highest factual consistency score (reaching 81 percent in economics). These findings highlight the accuracy of DeepSeek-8B to retrieve and combine the relevant information; thus, its responses are very accurate and detailed. The efficiency and fluency trade-off is efficiency and fluency: DeepSeek-8B has the highest response times by far in all domains, and its outputs can be of lower quality - e.g. its perplexity is the highest (worse) in all domains, meaning that its language can be less fluent or more convoluted than the other models. We also noticed DeepSeek-8B to produce excessively long answers, which although improving lexical scores, caused an increase in repetition (reduced N-gram diversity) and mediocre coherence.

Mistral presents a nearly inverse portrait. The smallest of these (7B parameters) but quickest and most reliable in the quality of answers. Mistral is highly efficient with an average response speed that is the highest in domains - 2-3 times higher than DeepSeek-8B - and thus very high. Mistral was the best in terms of the output quality as he always gave the most fluent and logically arranged answers. It has the lowest perplexity and the highest coherence in most domains, and often the highest semantic alignment with references (e.g. highest BERTScore in literature, economics, and medical domains). These scores indicate the answers of Mistral, which are not necessarily full of explicit facts as DeepSeek-8B, but are well-formed and more like a human explanation. The primary weakness of Mistral is a marginally lower factual recall - its factual consistency scores typically scored a few points behind DeepSeek-8B (e.g. 77.1% vs 81.6% in economics). Mistral may not perform as well as the larger retrieval-based model in tasks where detail exhaustiveness or strictness to reference is required. LLaMA3-8B provides a middle-of-the-pack performance that is balanced. It is almost as quick as Mistral (in tens of milliseconds) and generates reasonably coherent responses (at times the most coherent, as in the medical world). Nevertheless, LLaMA3-8B has a tendency to provide short responses, which, presumably, is the cause of its lower BLEU and ROUGE scores, in general. It has a reasonable factual consistency (73-79 per cent depending on domain) and in most cases about a few points lower than DeepSeek-8B. On the whole, LLaMA3-8B is a powerful generalist model, is fast and understandable, yet does not have the accuracy boost of DeepSeek-8B and the fined-tuned fluency of Mistral. The smallest model DeepSeek-1.5B, not surprisingly, is the last on most metrics, but has its value. With that very RAG strategy, DeepSeek-1.5B is able to achieve acceptable factual consistency in certain areas (e.g., 77.6% in literature, practically matching DeepSeek-8B). It means that retrieval augmentation would go a long way in assisting smaller models in generating factually accurate content. This is however limited by the capacity of DeepDeepSeek-1.5B, which serves to score lower on semantic and coherence scores - frequently the lowest BERTScore and coherence (especially in complex fields such as computer science and math). We also noticed that it may be too verbose or repetitive to cover up uncertainty, resulting in reduced N-gram diversity (e.g., 0.920

in literature, the lowest of models). DeepSeek-1.5B is more efficient in speed and resources than DeepSeek-8B and, owing to the overhead of the retrieval process, was not as fast as the other base models in a number of domains.

To conclude, the new comparisons prove that DeepSeek-8B should be selected in case of the priority of factual accuracy and content completeness, and Mistral should be used in case of the priority of fluent and coherent answers in a short period of time. LLaMA3-8B is a fair compromise in terms of speed and acceptable accuracy, and DeepSeek-1.5B is clear evidence that even tiny models can be useful with RAG help, albeit with obvious quality compromises. Such results, which are presented in the revised Table II, explain the significance of the correspondence of model choice to the defined needs of the domain and the metrics of interest to the evaluation process. Both models have their advantages to a scholarly assistant system and using both together or depending on the specific task (e.g. deepseek-8B to answer a technical question in detail and mistral to answer a technical question with a brief explanatory overview) may be an optimal option. The general trends in the performance indicate one significant fact: retrieval-augmentation can greatly enhance the level of performance in factual performance, yet the size of the model and fine-tuning have a significant impact on the level of clarity and efficiency of the responses that are produced. These open-source LLMs may therefore be customized to user needs, and one can focus on accuracy, speed, or readability to analyze research papers automatically.

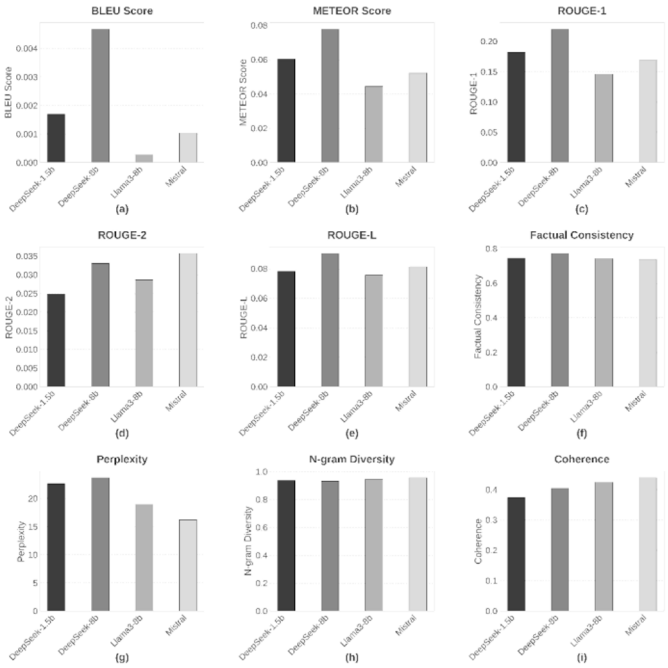

Fig 2: Comparison of performance of four open-source large language models across evaluation metrics

5 Computing System and Code availability:

We performed all experiments on a laptop with AMD Ryzen 5 7640HS CPU, 16 GB RAM and Nvidia GeForce RTX 4050 GPU. The evaluation code repository and scripts are publicly accessible at GitHub (https://tinyurl.com/apv47wan). A demonstration video of the implemented system is available at: https://youtu.be/EgoppOoQ-io .

6 Conclusion

We presented in this paper RAG, an Academia-directed assistant for researchers, which automates LLMs analysis of papers and performed a thorough evaluation over five different academic domains using open-source tools. The new system design also incorporates a PDF parser, citation network and section analyzer in multi-model LLM Q/A engine for generating informed answers and summaries. Our experiments show that the assistant excellently deals with diverse scholarly content, and retrieval expansion enhances factual correctness and domain transferability. We observed that smaller model retrieval systems (DeepSeek series) achieve high precision and efficiency, while larger ones (Mistral-7B, LLaMA3-8B) are richer in detail and fluency indicating a trade-off to be mitigated by future systems. By benchmarking such models on domain-specific tasks, we gain insights into where the strengths and weaknesses are of current open LLMs: for instance, the assistant excels in fact-based and technical Q&A but struggles with more abstract reasoning.

In conclusion, the RAG-driven approach is a promising and powerful strategy for academic AI assistants that utilizes the best of both IR and generation worlds. This work contributes a practical tool (implemented open-source and with an upcoming interactive demo) as well as a conceptual framework for evaluating and developing such assistants. For future work, we plan to incorporate such advanced verification modules, and expand into more domains (including across domain corpora), and continually add the latest LLMs from open source.

7 Acknowledgment

The authors gratefully acknowledge the support of Jahangirnagar University, Dhaka, Bangladesh, and sincerely thank Principal Investigator Prof. Dr. Md. Rezaul Karim for invaluable guidance and encouragement throughout the project.

References

1. 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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J.,

Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners.
2. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 1–38 (2023). https://doi.org/10.1145/3571730.
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, http://arxiv.org/abs/2005.11401, (2021). https://doi.org/10.48550/arXiv.2005.11401.
4. Johnson, J., Douze, M., Jegou, H.: Billion-Scale Similarity Search with GPUs. IEEE Trans. Big Data. 7, 535–547 (2021). https://doi.org/10.1109/TBDATA.2019.2921572.
5. Hanson, M.A., Barreiro, P.G., Crosetto, P., Brockington, D.: The strain on scientific publishing. Quantitative Science Studies. 5, 823–843 (2024). https://doi.org/10.1162/qss_a_00327.
6. Vishnu, J., Devi, K.Y., Sravani, T., Reddy, P.B., Rahul, P.S.: NLP based Machine Learning Approaches for Text Summarization. 13,.
7. Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., Lenert, L.A.: The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. Journal of the American Medical Informatics Association. 32, 1071–1086 (2025). https://doi.org/10.1093/jamia/ocaf063.
8. Haßler, B., Hassan, M., Klune, C., Mansour, H., Friese, L.: Using AI to Automate the Literature Review Process: A Topic Brief. EdTech Hub (2024). https://doi.org/10.53832/edtechhub.1003.
9. Hua, S., Jin, S., Jiang, S.: The Limitations and Ethical Considerations of ChatGPT. Data Intelligence. 6, 201–239 (2024). https://doi.org/10.1162/dint_a_00243.
10. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature. 630, 625–630 (2024). https://doi.org/10.1038/s41586-024-07421-0.
11. Bommasani, R., Liang, P., Lee, T.: Holistic Evaluation of Language Models. Annals of the New York Academy of Sciences. 1525, 140–146 (2023). https://doi.org/10.1111/nyas.15007.
12. Ashwell, G., Russell, A.M., Williamson, A.E., Pope, L.M.: Learning about Inclusion Health in undergraduate medical education: a scoping review. BMJ Open. 15, e092420 (2025). https://doi.org/10.1136/bmjopen-2024-092420.
13. Cheetham, N., Cantle, F., Guise, A., Steves, C.J.: Socio-Economic Diversity of Doctors in the United Kingdom: A Cross-Sectional Study of 10 Years of Labour Force Survey Social Mobility Data, https://www.ssrn.com/abstract=5016429, (2024). https://doi.org/10.2139/ssrn.5016429.
14. Chou, Y.-C., Zhou, Z., Yuille, A.: Embracing Massive Medical Data. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., and Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 24–35. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72378-0_3.

15. Dolat Abadi, P., Zakerimoghadam, M., Abadi, Z.A.D., Rahmanian, M., Riahi, S.M., Khanipour-Kencha, A.: Effects of pre-CABG program on discharge readiness and surgery outcomes for patients undergoing elective CABG surgery: a study protocol for a randomised control trial. BMJ Open. 15, e090256 (2025). https://doi.org/10.1136/bmjopen-2024-090256.
16. Espinel, P., Cho, G., Marshall, N.S., Yee, B.J., Smith, K., D'Rozario, A.L., Ainge-Allen, H.W., Stranks, L., Gauthier, G., Lambert, T., Grunstein, R.R.: Implementation of sleep apnoea testing and treatment services into a cardiometabolic clinic for people living with severe mental illness: a prospective evaluation of a translational programme. BMJ Open. 15, e092034 (2025). https://doi.org/10.1136/bmjopen-2024-092034.
17. Kim, J., Ba, Y., Kim, J.-Y., Youn, B.-Y.: Patient perception of physician attire: a systematic review update. BMJ Open. 15, e100824 (2025). https://doi.org/10.1136/bmjopen-2025-100824.
18. Lindo, A., Alsén, S., Fors, A., Filipsson Nyström, H.: In the shadow of Graves' disease: a qualitative interview study of patients' experiences conducted at a secondary referral centre in Sweden. BMJ Open. 15, e098238 (2025). https://doi.org/10.1136/bmjopen-2024-098238.
19. Rous, B., Clarke, C.A., Hubbell, E., Sasieni, P.: Assessment of the impact of multicancer early detection test screening intervals on late-stage cancer at diagnosis and mortality using a state-transition model. BMJ Open. 15, e086648 (2025). https://doi.org/10.1136/bmjopen-2024-086648.
20. Tan, M., Tang, S., Kan, S., Zhang, H., Wu, B., Ni, Z., Lin, C.C., Ding, J.: Enhancing healthcare providers' advance care planning competence with large language models: protocol for the development of an AI chatbot and its evaluation in a randomised controlled trial. BMJ Open. 15, e099226 (2025). https://doi.org/10.1136/bmjopen-2025-099226.
21. Taxbro, K., Åhman, R., Chew, M.S., Engerström, L.: Impact of socioeconomic status and country of origin on COVID-19 outcomes in Swedish ICUs: a retrospective registry-based cohort study. BMJ Open. 15, e099763 (2025). https://doi.org/10.1136/bmjopen-2025-099763.
22. Cheatle, E.: From Simone de Beauvoir's 'House' to bell hooks' 'Homeplace': Autofiction and Autotheory in Architectural Writing. Architectural Histories. 12, (2024). https://doi.org/10.16995/ah.11690.
23. Dadacz, K.: The Digital Underclass Glitch: Theorising a Digital.
24. Danylyuk, O.: Combat at Gamer's Pace. No Pause nor Reset Button. Body, Space & Technology. 24, (2025). https://doi.org/10.16995/bst.18480.
25. Farantatos, P.: Embodied Memories, Retroactive Traces: Le Corbusier's Travel Sketches in Le Modulor. Architectural Histories. 12, (2024). https://doi.org/10.16995/ah.10728.
26. Gonzalo Encinar, R.: Traces of Expression: AI and Data-Driven Approaches to Dance Archives. Body, Space & Technology. 24, (2025). https://doi.org/10.16995/bst.18264.

27. Insulander, E., Lindstrand, F.: Bringing the past to life: The rhetorical construction of authenticity through digital media in museum exhibitions. Multimodality & Society. 26349795251360535 (2025). https://doi.org/10.1177/26349795251360535.
28. Massoura, K.: The Body that Writes: Gender, Class, and the Abject Female Body in Hannah Kent's Burial Rites (2013).
29. McGuire, M.: #MeToo and the Northern Ireland Troubles: Anna Burns' Milkman. C21 Literature: Journal of 21st-Century Writings. 10, (2023). https://doi.org/10.16995/c21.3397.
30. Rehfeldt, R.A., Chan, S., Katz, B.: The Beethoven Revolution: A Case Study in Selection by Consequence. Perspect Behav Sci. 44, 69–86 (2021). https://doi.org/10.1007/s40614-020-00271-x.
31. Tipping, D.: Illustration as a participatory tool for diverse(queering) discussions on the climate crisis: Exploring the workshop method.
32. Almaharmeh, M.I., Almasarwah, A., Shehadeh, A.: Mandatory IFRS Adoption and Real/Accruals Bases Earnings Management in the UK. JOFRP. 10, 25–39 (2021). https://doi.org/10.35944/jofrp.2021.10.1.002.
33. Björkholm, L., Lehner, O.M.: Nordic green bond issuers' views on the upcoming EU Green Bond Standard. JOFRP. 10, 222–279 (2021). https://doi.org/10.35944/jofrp.2021.10.1.012.
34. Bolt, E.E.T., Cafferkey, K., Townsend, K., Van Der Cingel, M.: Exploring nurses' postturnover experiences in their new employment: A self-determination and job-fit perspective. BRQ Business Research Quarterly. 28, 530–542 (2025). https://doi.org/10.1177/23409444241252665.
35. Gassouma, M.S., Ghroubi, M.: Discriminating between Islamic and Conventional banks in term of cost efficiency with combination of credit risk and interest rate margin in the GCC countries: Does Arab Spring revolution matter? JOFRP. 10, 280–295 (2021). https://doi.org/10.35944/jofrp.2021.10.1.013.
36. Kishor, N.K., Nguyen, N.: Measuring the credit gap: a forecast combination approach. Swiss J Economics Statistics. 161, 2 (2025). https://doi.org/10.1186/s41937-025-00133-w.
37. Momon, Wati, L.N., Sutar: The Role of Political Connections and Family Ownership in Increasing Firm Value. JOFRP. 10, 40–53 (2021). https://doi.org/10.35944/jofrp.2021.10.1.003.
38. Olkhov, V.: Price, Volatility and the Second-Order Economic Theory. JOFRP. 10, 139–165 (2021). https://doi.org/10.35944/jofrp.2021.10.1.009.
39. Oyetade, D., Obalade, A.A., Muzindutsi, P.-F.: Basel capital requirements, portfolio shift and bank lending in Africa. JOFRP. 10, 296–319 (2021). https://doi.org/10.35944/jofrp.2021.10.1.014.
40. Santero-Sánchez, R., Núñez, B.C.: Pursuing equal pay for equal work: Gender diversity in management positions and the gender pay gap throughout the wage distribution. BRQ Business Research Quarterly. 28, 59–73 (2025). https://doi.org/10.1177/23409444221125239.
41. Valero-Gil, J., Suárez-Perales, I., Ferrón-Vílchez, V.: Would you date a liar? The impact of greenwashing on B2B relationships under the managerial trust view.

BRQ Business Research Quarterly. 28, 731–749 (2025). https://doi.org/10.1177/23409444241250360.
42. Belessis, A., Loi, I., Moustakas, K.: Advanced articulated motion prediction. Front. Comput. Sci. 7, 1549693 (2025). https://doi.org/10.3389/fcomp.2025.1549693.
43. Gkolemis, V., Gutmann, M., Pesonen, H.: An Extendable *Python* Implementation of Robust Optimization Monte Carlo. J. Stat. Soft. 110, (2024). https://doi.org/10.18637/jss.v110.i02.
44. Kulikov, A.S., Mikhailin, I., Mokhov, A., Podolskii, V.V.: Complexity of Linear Operators. Theory of Comput. 21, 1–32 (2025). https://doi.org/10.4086/toc.2025.v021a009.
45. Van Der Loo, M.P.J., De Jonge, E.: Data Validation Infrastructure for *R*. J. Stat. Soft. 97, (2021). https://doi.org/10.18637/jss.v097.i10.
46. Barrowman, N., Webster, R.J.: Exploring Data Subsets with **vtree**. J. Stat. Soft. 114, (2025). https://doi.org/10.18637/jss.v114.i04.
47. Diessner, M., Wilson, K.J., Whalley, R.D.: **NUBO** : A Transparent *Python* Package for Bayesian Optimization. J. Stat. Soft. 114, (2025). https://doi.org/10.18637/jss.v114.i01.
48. Kalan, R., Canatalay, P.J., Karsli, E.: Security-Aware Adaptive Video Streaming via Watermarking: Tackling Time-to-First-Byte Delays and QoE Issues in Live Video Delivery Systems. Computers. 14, 404 (2025). https://doi.org/10.3390/computers14100404.
49. Ruíz-Pozo, I., Morales-García, J., Aguirre-Mejía, C.X., Serrano, A.: Using Generative Artificial Intelligence to Improve User Engagement in Content Marketing, https://www.ssrn.com/abstract=4982145, (2024). https://doi.org/10.2139/ssrn.4982145.
50. Allen, S.: Weighted **scoringRules** : Emphasizing Particular Outcomes When Evaluating Probabilistic Forecasts. J. Stat. Soft. 110, (2024). https://doi.org/10.18637/jss.v110.i08.
51. Voboril, F., Peruvemba Ramaswamy, V., Szeider, S.: Generating Streamlining Constraints with Large Language Models. jair. 84, (2025). https://doi.org/10.1613/jair.1.18965.
52. Al-Mekhlafi, S.M.: Numerical simulation of hybrid deterministic-stochastic PDE models for cancer tumor growth. Research in Mathematics. 12, 2567717 (2025). https://doi.org/10.1080/27684830.2025.2567717.
53. Alsahli, G., Karapinar, E., Shahi, P.: On rational contractions in perturbed metric spaces. Research in Mathematics. 12, 2508578 (2025). https://doi.org/10.1080/27684830.2025.2508578.
54. Ameeq, M., Hassan, M.M., Fatima, L., Jawo, E., Alkhaleel, B.A., Hussain, N.: Quartile-based group acceptance sampling plan under the modified power exponential distribution: Applications to failure time and material strength data. Research in Mathematics. 12, 2581408 (2025). https://doi.org/10.1080/27684830.2025.2581408.
55. Anwar, M.N., Khizar, A., Kahungu, M.K.: Empowering mathematics learning: Feedback orientation as a mediator in feedback, achievement motivation, and

achievement. Research in Mathematics. 12, 2578874 (2025). https://doi.org/10.1080/27684830.2025.2578874.
56. Asare, B., Yaa Nchor, E., Dissou Arthur, Y.: Moderating effect of cognitive ability on the nexus between self-regulated learning and problem-solving ability among university students. Research in Mathematics. 12, 2554423 (2025). https://doi.org/10.1080/27684830.2025.2554423.
57. Bihiye, Z.M., Brun, M., Mpimbo, M., Mrema, E.: On interleaving distance as a complete metric for bifiltrations in Rn. Research in Mathematics. 12, 2573572 (2025). https://doi.org/10.1080/27684830.2025.2573572.
58. Bosboom, V., Kraus, M., Schlottbom, M.: A metriplectic formulation of polarized radiative transfer. J. Phys. A: Math. Theor. 56, 345206 (2023). https://doi.org/10.1088/1751-8121/aceae2.
59. Hameed, R., Nosheen, S., Younis, J.: A unified approach for constructing 6 z - point relaxed quaternary subdivision schemes from 4 z - point relaxed binary schemes. Research in Mathematics. 12, 2579377 (2025). https://doi.org/10.1080/27684830.2025.2579377.
60. Panda, S., Nayak, M.M.: Incorporating fuzzy stochastic differential equations into agricultural growth models: A comprehensive overview. Research in Mathematics. 12, 2572194 (2025). https://doi.org/10.1080/27684830.2025.2572194.
61. Sahani, S.K., Oruganti, S.K., Satishkumar, K.: Advanced Mathematical Modeling of Woolen Knitting Dynamics Using Laplace Transform and Fourth-Order Runge–Kutta Method. Journal of Mathematics. 2025, 4985087 (2025). https://doi.org/10.1155/jom/4985087.
62. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. p. 311. Association for Computational Linguistics, Philadelphia, Pennsylvania (2001). https://doi.org/10.3115/1073083.1073135.
63. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.
64. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries.
65. Kryscinski, W., McCann, B., Xiong, C., Socher, R.: Evaluating the Factual Consistency of Abstractive Text Summarization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9332–9346. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.750.
66. Ankner, Z., Blakeney, C., Sreenivasan, K., Marion, M., Leavitt, M.L., Paul, M.: PERPLEXED BY PERPLEXITY: PERPLEXITY-BASED DATA PRUNING WITH SMALL REFERENCE MODELS. (2025).
67. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, http://arxiv.org/abs/1510.03055, (2016). https://doi.org/10.48550/arXiv.1510.03055.
68. Lauriola, I., Campese, S., Moschitti, A.: Analyzing and Improving Coherence of Large Language Models in Question Answering.