Scientific Report Title

Author Name

May 31, 2025

Abstract

This is the abstract section. It should provide a concise summary of the research, including the main objectives, methods, results, and conclusions. The abstract should be self-contained and typically not exceed 250 words.

1 Introduction

The introduction should provide the background and context for your research. It should:

- Research Problem Presented, and Claims in the Original Paper
- Retrieval Augmented Generation
-
- Explain the significance of the work

Large Language Models (LLMs) have the potential to help scientists with retrieving, synthesizing, and summarizing the scientific literature CITE. However, issues such as hallucinations CITE, lack of detail CITE, and underdeveloped retrieval and reasoning benchmarks, hamper the direct use of LLMs in scientific research.

The field is rapidly developing, with new models such as Google's Gemini 2.5 CITE and OpenAI's o3 CITE being able to 'reason', and excel at coding, maths, and langaage benchmarks. This also highlights the developments of new cutting-edge benchmarks for scientific performance in areas such as scientific discovery CITE, analysis CITE, reasnoning CITE, programming CITE, CITE, and mathematics CITE.

Alongside the development of the fundamental models themselves, techniques such as Retreival Augemented Generation (RAG) and the use of Multi-Agent Systems (MAS) allowed better use of LLMs in scientific research.

1.1 Retrieval Augmented Generation

RAG address issues found within foundational generative models by combining two components: a retrieval component and a generative component. For a given prompt, the retrieval levarages dense vector respresentations to summarise and find relevant information from large data sources. These representations are then given to the generative component (LLM), which generates responses grounded in the retrieved knowledge. This helps prevent hallucinations, as the generative process is grounded in (assumed) factually-correct, external knowledge. This is why RAG has potential applications in a wide number of fields (open-domain question answering, scientific discovery, medical diagnoses, ...), where factual accuracy and contextual understanding is crucial CITE.

1.2 Multi-Agent Systems

MAS enhances the capabilities of a single LLM by leveraging the collaboration of LLMs and the fact that LLMs can be fine-tuned to perform specific tasks CITE. Frameworks such as Microsoft's AG2 CITE and Google's LangChain allow the interaction between agents and LLMs to better perform at specific tasks. This interaction between agents, with the ability to 'converse' in AG2's case, is comparable to a team of scientists working together. This marks a major leap forward toward automated scientific discovery, with tools such `cmbagent` CITE and `Robin` CITE being developed.

1.3 PaperQA2

To make scientific discoveries, one must be able to synthesise scientific knowledge. Skarlinski et. al. believe this can be broken down into three vital tasks: scientific question answering, summarising, and detecting contradictions in the literature. In their paper, it is shown that their developed tool, PaperQA2, outperforms the human benchmark in all three areas CITE. The focus of this report is to understand and reproduce the question-answering result.

1.3.1 PaperQA2 Architecture

PaperQA2 is RAG agent that treats retrieval as a multi-step agent task, comprised of multiple tools CITE. The Gather Papers tool transforms the user prompt into a keyword search to find suitable papers. The papers are then parsed into chunks of information, and then ranks these chunks. What makes PaperQA stand out, is the fact that these chunks are then reranked (by relevance) and contextually summarised by an LLM. This allowed irrelevant chunks to be excluded, and is 'critical' to RAG CITE. The framework also includes a Citation Traversal tool, exploiting citation graphs to provide additional relevant sources.

1.3.2 Key Result

The paper claims that PaperQA's performance beats the human benchmark, specifically for the precision of questions answered, achieving a precision of 73.8%. The benchmark metrics are accuracy and precision.

$$Accuracy = \frac{CorrectQuestions}{AllQuestions} \tag{1}$$

$$Precision = \frac{CorrectQuestions}{AnsweredQuestions} \tag{2}$$

The process of evaluating the performance of PaperQA2 was to ask questions, where the answers were found in newly published papers and could not be inferred from either the title, abstract, or conclusion.

2 Methods

2.1 Data Collection

The data used to train and evaluate PaperQA2 was found on HuggingFace. This contained information, such as the question proposed, true answer, distractor values, the orignal paper DOI, and metadata. Due to licensing issues, the actual papers themselves were not directly available, and so they had to be manually retrieved from their DOIs. The evaluation papers were made up of both the training and test set.

Once the papers were collated, some initial investigation of the papers themselves were conducted. Figure CITE shows the years in which the papers were published. The criteria for papers was that they were published after the GPT Training Cutoff in September 2021. However, a number of the papers did not meet this criteria, and were subsequently removed from the evaluation papers.

HuggingFace data Retrieval Incorrect Papers in the original Paper

2.2 Evaluation Methodology

The field of LLMS, multi-agent AI, and RAG are new and ever-changing. This highlights a need for a unified system to benchmark the performance of these tools. The UK's AI Security Institute (AISI) released an open-source framework called

Present your findings in a clear and organized manner.

2.3 Statistical Analysis

Include relevant statistical results.

2.4 Figures and Tables

Example of a table:

Table 1: Sample Table

Category	Value 1	Value 2
A	1.23	4.56
B	7.89	0.12

Example of a figure reference:

Figure 1: Sample Figure

3 Discussion

Interpret your results and discuss their implications. This section should:

- Interpret the results
- Compare with previous studies
- Discuss limitations
- Suggest future work

4 Conclusion

Summarize the main findings and their significance.

Acknowledgments

Acknowledge any funding, technical support, or other contributions.

Acknowledgments

Acknowledge any funding, technical support, or other contributions.

References