



Society of Petroleum Engineers

SPE-228117-MS

Building an Intelligent Research Assistant – RAG Isn't All You Need

J. Boden, J. Eckroth, and S. Gipson, i2k Connect, Houston, Texas, USA; K. Myers, SRI International, Menlo Park, California, USA; R. Smith and E. Schoen, i2k Connect, Houston, Texas, USA; J. Keefe, Pacific Science and Engineering, San Diego, California, USA; P. Sequeira, SRI International, Menlo Park, California, USA

Copyright 2025, Society of Petroleum Engineers DOI [10.2118/228117-MS](https://doi.org/10.2118/228117-MS)

This paper was prepared for presentation at the SPE Annual Technical Conference and Exhibition held in Houston, Texas, USA, 20 - 22 October 2025.

This paper was selected for presentation by an SPE program committee following review of information contained in an abstract submitted by the author(s). Contents of the paper have not been reviewed by the Society of Petroleum Engineers and are subject to correction by the author(s). The material does not necessarily reflect any position of the Society of Petroleum Engineers, its officers, or members. Electronic reproduction, distribution, or storage of any part of this paper without the written consent of the Society of Petroleum Engineers is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of SPE copyright.

Abstract

Based on experimentation, this paper outlines the efficiency and accuracy improvements that can be achieved using a cutting-edge generative AI chat enabled research assistant to facilitate well planning exercises. The technology innovations required to achieve those gains are also outlined.

With the goal of measuring these improvements, during the summers of 2024 and 2025, the research team tested an experimental, domain aware agentic research assistant which could answer a question, cite the sources and show the researcher the original material it had used to determine the answer.

Seventy-five (75) volunteers from thirty (30) countries participated in a simple well planning exercise answering a series of questions requiring analysis and interpolation of information contained in Equinor's Volve data set which consists of approximately 30,000 documents of mostly unstructured data (PDFs, etc.). To assist in answering the questions, the system had access to a good portion of those documents including well design documents, well logs, geological and stratigraphical data and operational reports (e.g. daily drilling reports) as well as SPE's PetroWiki.

The system was benchmarked against a more standard but still state of the art RAG (Retrieval Augmented Generation) solution. By all measures, the experimental system outperformed the state-of-the-art RAG system, improving the accuracy of the answers by 152% and reducing the time required to answer by 86%. This paper explores the efficiency gained when a researcher, especially one with less industry experience, has access to a capable virtual research assistant which, in turn, has access to a broad set of documents covering the domain. The paper describes the limits of a state-of-the-art RAG system, the impact of building domain awareness into a novel system and the technology areas that make a difference in that system's usefulness when presented with the volumes of material common in the industry.

While the technology continued to be improved in general, the next round of research focused on compiling answers across a large set of documents and finding information within plots and graphs. Similar testing, this time with testers who did not have oil and gas experience, showed that subjects achieved 86% correct answers with the improved solution vs 27% with an updated baseline RAG system.

Finally, we performed ablation testing to understand the efficacy of various search approaches in achieving these results for various types of questions.

Introduction

The work described in this paper is the result of three-year project funded by DARPA called KMASS (Knowledge Management at Scale and Speed). KMASS "... aims to research, develop, integrate, evaluate, and demonstrate underlying technology that will enable effective use of documented knowledge, acquisition of new knowledge as part of regular workflows, and application of useful knowledge when and where it is required and with necessary granularity" [1].

The release of the ChatGPT application that used the GPT 3.5 model in November 2022 was disruptive not only to the AI industry [2] but resulted in the recasting of the KMASS program towards more fundamental research on the use Large Language Models (LLMs) in knowledge management (KM). As a result, the focus of the project shifted from using AI in a search application, to point the user to the documents most likely to contain the answers, to using a greatly simplified chat interface, which provided both the actual answer and where that answer could be found within the documents.

Although there have been some studies on the use of LLMs for work linked to various roles including consulting [6] or software engineers [7], those studies have focused on creative rather than analytical tasks. Few, if any, efforts have sought to evaluate LLM-based technologies for knowledge-intensive work on realistic tasks as we have done during this project.

This paper focuses on the results and learnings post the KMASS program shift to LLMs and a conversational chat user experience.

Experimental Design and Results

The program involved three rounds of human testing, each separated by approximately a year, with research and development undertaken to improve the technology between each round of testing.

The testing consisted of a series of research tasks focused on the type of due diligence which would be required for a real world well planning exercise. The testers were not asked to fully plan a well, since such an exercise would be as much a test of the individual's well planning skills as a test of the system's ability to speed up information gathering and synthesis. Thus, evaluation was focused on the due diligence portion of the planning process. Specifically, the tasks involved finding information in the Volve dataset [3], a collection of all the subsurface and production data for the Volve field which was released by Equinor for research, study and development purposes. The dataset consists of approximately 30,000 files containing logs, reports and other documents. PetroWiki [4] was also added so people could look up relevant background information without having to leave the system.

All tests were conducted in a double blind and randomized fashion. Each test was also benchmarked against a state-of-the-art system developed by a large university research lab with experience in AI. The work to build this system was funded by DARPA and outside the control of the team to ensure an unbiased baseline for comparison. This system had access to the same Volve and PetroWiki content as the system under test.

Testing Phase One

The first round of testing compared ATHENA, a smart research portal similar to SPE's research portal (<https://search.spe.org>) [6], to a baseline 'SharePoint-like' search experience. [Figure 1](#): ATHENA Portal Interface for First Phase of Testing shows what this interface looked like. During the testing, the user searched for the relevant documents in the search tab, read the document to find the answer then entered that answer in the answer sheet. The Phase One approach is not the focus of this paper; however, we note that advanced AI based search tools and filtering on the ATHENA portal did lead to a modest improvement in task speed and accuracy among the approximately 130 volunteer testers.

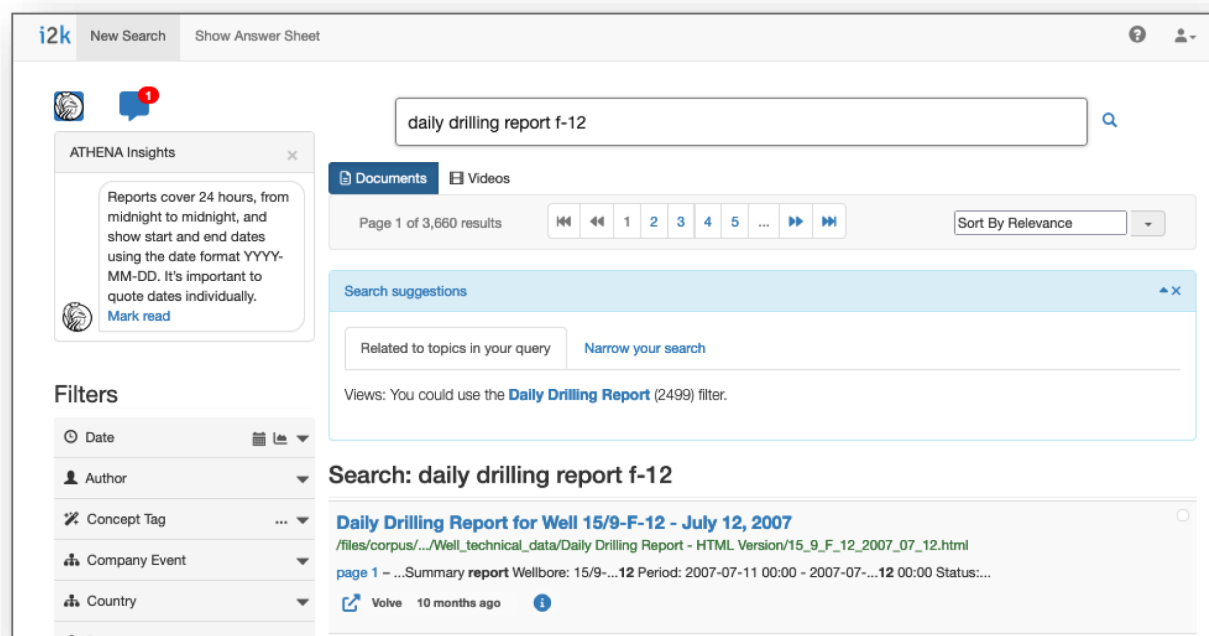


Figure 1—ATHENA Portal Interface for First Phase of Testing

Testing Phase Two

As the KMASS program shifted to focus on the use of Large Language Models, ATHENA's user interface changed to a conversational chat interface as shown in Figure 2: ATHENA Conversational Chat Interface Circa Mid-2024.

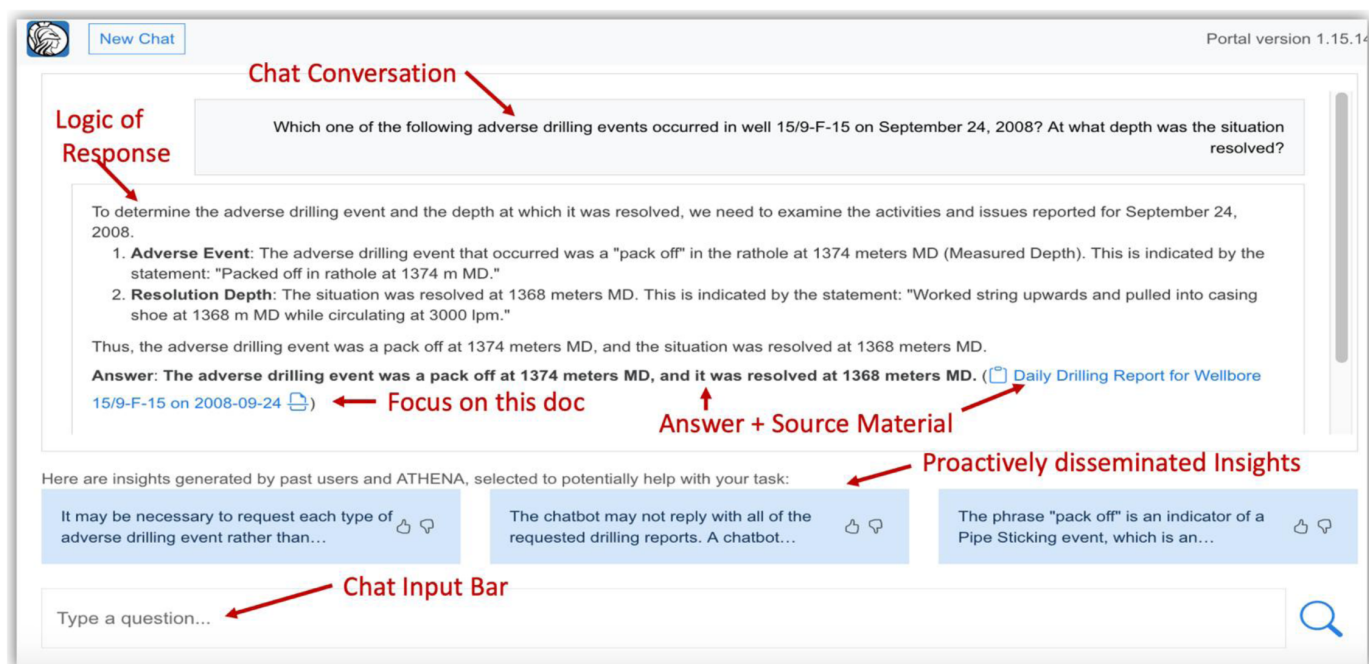


Figure 2—ATHENA Conversational Chat Interface Circa Mid-2024

For the second round of testing, the volunteers performed data gathering tasks in the following areas:

Task 1: Analyze drilling info for a well.

- Measured depth at a particular time
- Maximum depth
- Drilling activities at designated times

Task 2. Analyze data for a planned well.

- Timing of operations
- Data acquisition strategy
- Well path analysis

Task 3: Analyze adverse drilling events.

- Drilling interruptions
- Testing conducted
- Adverse event at a given time

Task 4. Pore pressure analysis.

- Pressure estimation for a particular well
- Id key parameters used for an analysis method
- Determine values in a specific analysis

The testers completed a random selection of half these tasks using the baseline system and half using the ATHENA system.

Figure 3: Test Scores and Time Taken to Complete Task, shows both the percent of correct results achieved and the time taken for the testers to answer the random selection of tasks on both the ATHENA and baseline systems. It should be noted that the times taken on the baseline system might be misleading since testers often gave up when they struggled to find the answer. Net, the testers were able to achieve significantly higher grades using the ATHENA system than the baseline.

The testers, on average, completed 30.7% of the tasks correct when they used the baseline system while they averaged 77.3% correct when using the ATHENA system, a 152% improvement in results over the baseline. Similarly, failure rates dropped by 86% when using the ATHENA system (13% failure on ATHENA vs 88% on the baseline).

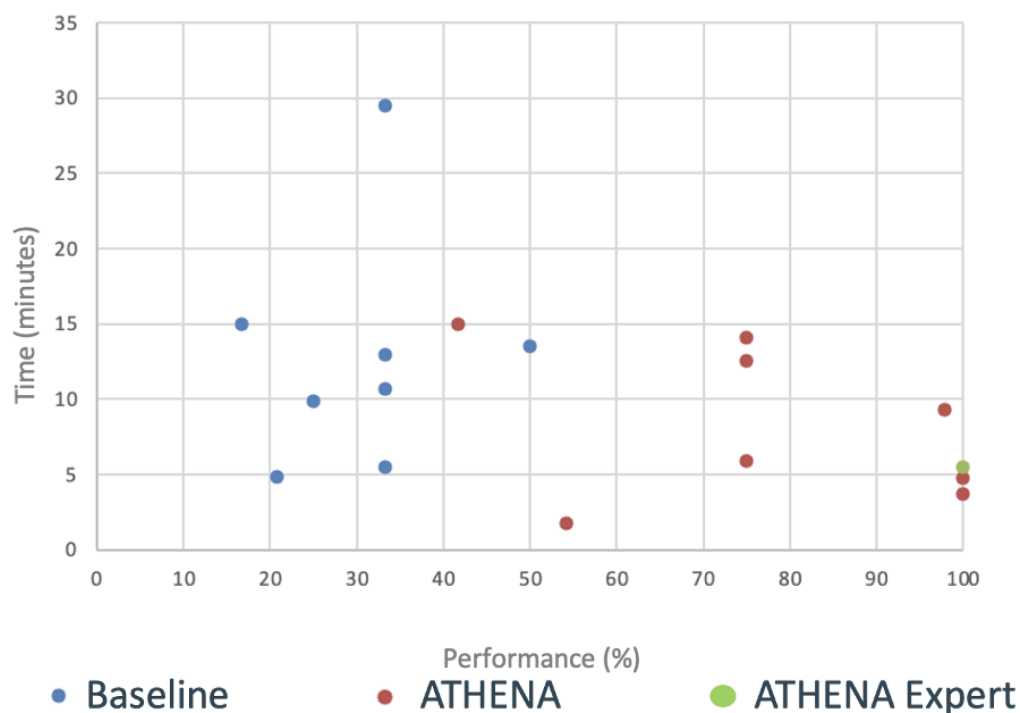


Figure 3—Test Scores and Time Taken to Complete Task

The testing was extended to an additional 74 testers who had a petroleum engineering background but who didn't have the deep domain expertise of the previous group. Again, the tasks were performed in random order but only on the ATHENA system. Figure 4: Performance vs Time for all Testers, shows the scores and times taken for all testers. Even with less domain expertise, testers using the ATHENA system generally outperformed the results on the baseline system.

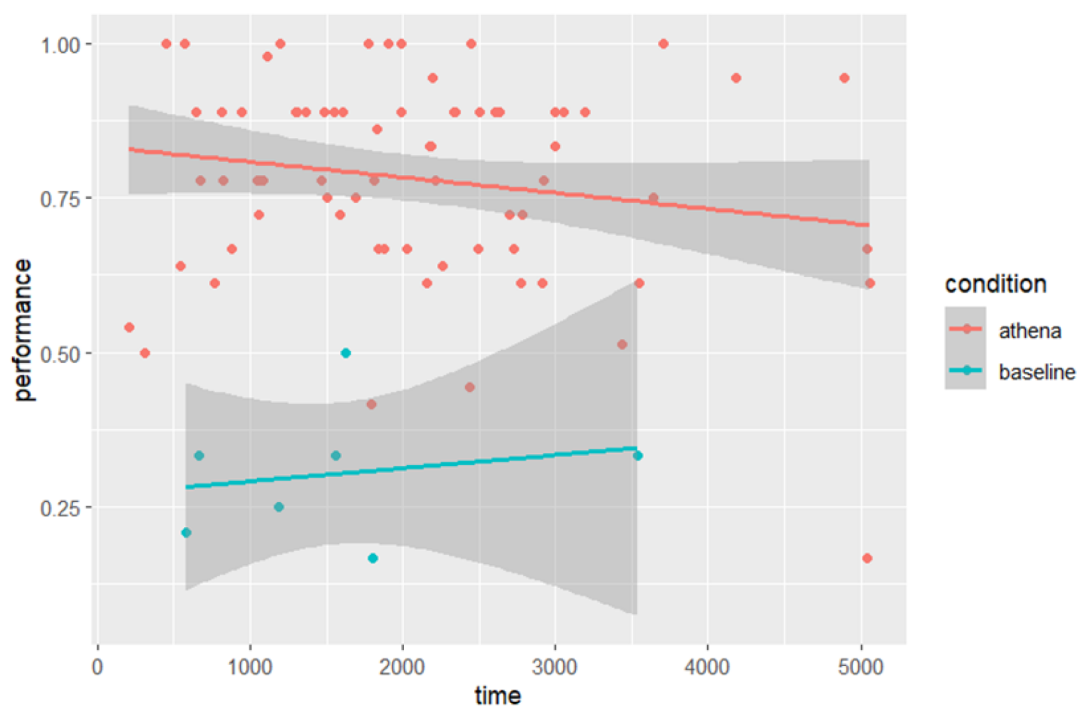


Figure 4—Performance vs Time for all Testers

Testing Phase Three

For the third round of testing both the ATHENA system and the baseline system were enhanced. We added tasks that required the tester to aggregate information for many documents (for instance creating a table of lost circulation events or a table of MWD tools used by casing diameter) and sub tasks where the answer was only available in charts or graphs.

As was done for the Phase Two baseline vs ATHENA testing, the testers were assigned tasks randomly to be completed on a random selection of either the baseline or ATHENA system.

The baseline system performed much better than in the Phase Two testing on the "finding a needle in a haystack" tasks in which specific information had to be found within a single document. For such tasks, it performed approximately on par with the ATHENA system. On the tasks which required aggregation of information over many documents and those requiring multi-modal understanding, ATHENA clearly outperformed the baseline system. This is shown in the results the testers achieved in [Figure 5: Tester's Grade vs Average Task Time](#). The testers, on average, only completed 27.2% of the tasks successfully on the baseline system whereas they achieved an 85.8% success rate using ATHENA. That translated to a failure rate (below 50% of tasks completed correctly) of 62% on the baseline but 0% on ATHENA.

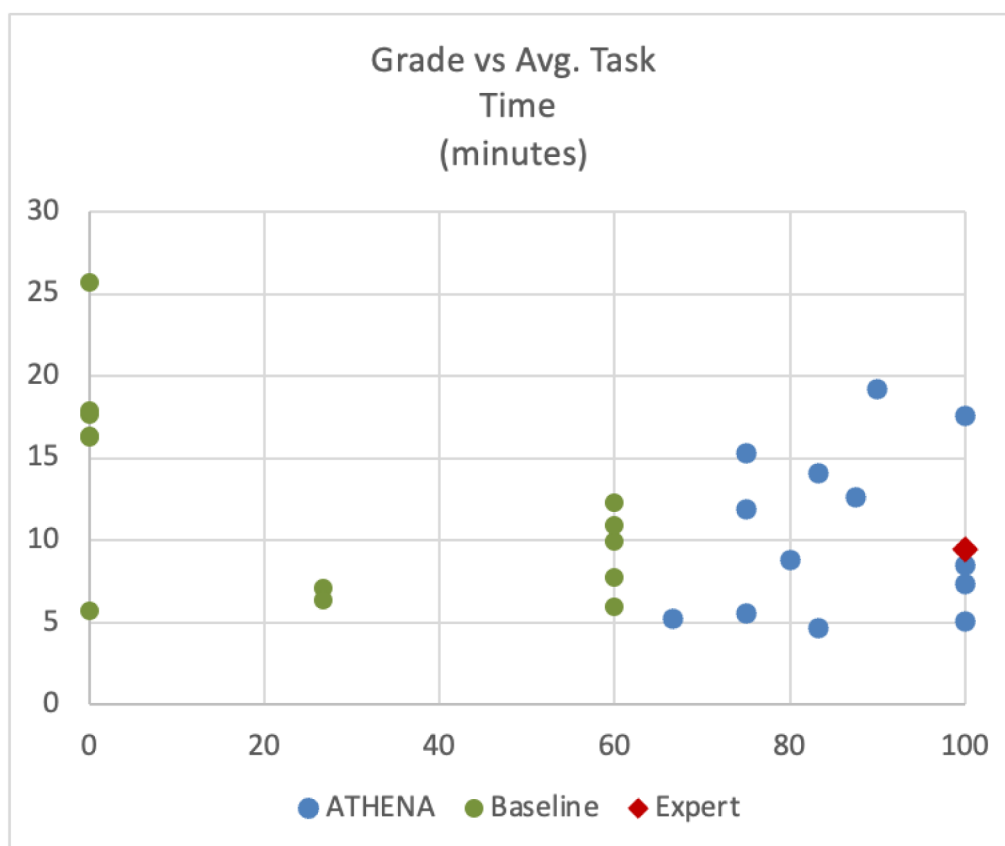


Figure 5—Tester's Grade vs Average Task Time

Discussion and Analysis

Early testing showed that finding the correct source content was the key reason ATHENA's chat outperformed the baseline. Further analysis showed that while built on state-of-the-art RAG technology, the baseline system frequently failed to identify the relevant documents. It often selected the wrong type of report, the wrong date or the incorrect well. We believe this was caused by its exclusive use of embeddings for retrieval compared to the hybrid approach used by ATHENA. This confusion of entities which appear in similar contexts is known as entity-related hallucinations [8]. Further evidence of this theory is that once

the baseline system incorporated key word extraction for dates and well names it performed on par for tasks which required handling of dates or focus on a specific well.

Phase Two Differentiators

Both the baseline and ATHENA systems used very capable LLMs and both used embeddings for search, an approach that was consistent with the state-of-the-art RAG systems at the time.

In addition to dense vector search, the ATHENA system pre-processed the Volve and PetroWiki content to classify it using industry aware classifiers. A single document is allowed to be about multiple things both within a specific taxonomy and across taxonomies. For example, classifiers detect well names and thus fields and basins so these can be used later to find any references to a particular well or all wells within a field, etc. Classifiers detect the type of document such as daily drilling reports (DDR), logs, etc. Summary reports or PowerPoint summaries will often include sections or slides from multiple document types so a single document which contains multiple types of material is classified into all those categories so the content can be found later. Adverse events such as lost circulation, stuck pipes, etc. are also detected in advance using classifiers. The ATHENA system used what became known as agents to handle search strategies (such as for dates) and to be able to filter content to find the answer to the user's questions. This pre-processing using domain specific classifiers seemed to be the differentiator that allowed ATHENA to outperform the baseline system.

LLMs excel at extracting information from and summarizing specific documents but are less effective at finding relevant documents, especially as the number of documents grows or the individual documents get larger. Many subjects felt ATHENA's "document focus" feature was helpful since it allowed them to focus LLM interactions to the target document or documents, improving the control they had on the scope and thus the quality of results.

Phase Three Differentiators

By Phase Three testing, both systems were using an agentic approach with multiple task specific LLMs. The baseline system maintained the use of embeddings but adopted a keyword extraction method to handle dates and well names, which greatly improved its performance on several tasks. The ATHENA system switched from an embeddings similarity search to a combination of keyword search and filtering more similar to that used in Phase One, but with a major difference: the Phase Three system does many searches in parallel, ranks and combines the results into a single list that is processed for appropriate content to provide to the LLM to generate an answer. This technique works both for 'needle in a haystack' searches and for situations where the answer is spread over many (potentially hundreds) of documents. The research team have named this technique "Search, Selection, Retrieval Augmented Generation" (SS-RAG).

To explain the impact of these choices, let's first look at [Figure 6](#) which shows the mean reciprocal rank results across the task set for ATHENA's search compared to the results achieved with selected embedding models independently of the introduction of the multiple searches. The ATHENA search, closely modeled on the search used for the Phase One, identified the correct material more reliably than using embedding similarity (cosine distance).

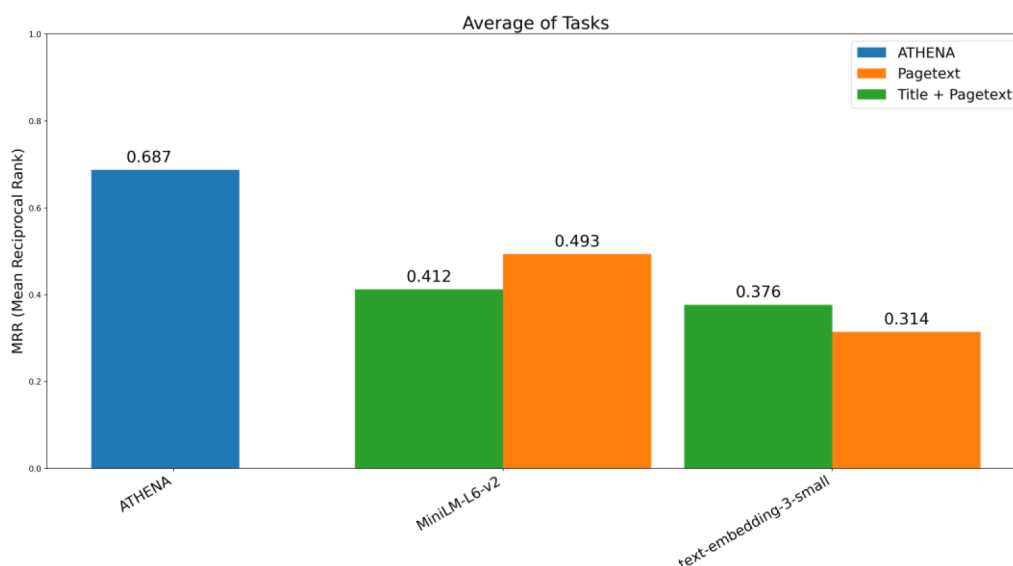


Figure 6—MRR for ATHENA Search vs Embedding Based Search

Ablation testing was performed to understand the individual contribution of components, in this case types of search, to the overall system performance. The results of ablation testing with and without the multi-search capability across several techniques for choosing the search parameters are shown in Figure 7. The multi-search technique used by ATHENA has a clear advantage over a single search when it comes to finding the right content.

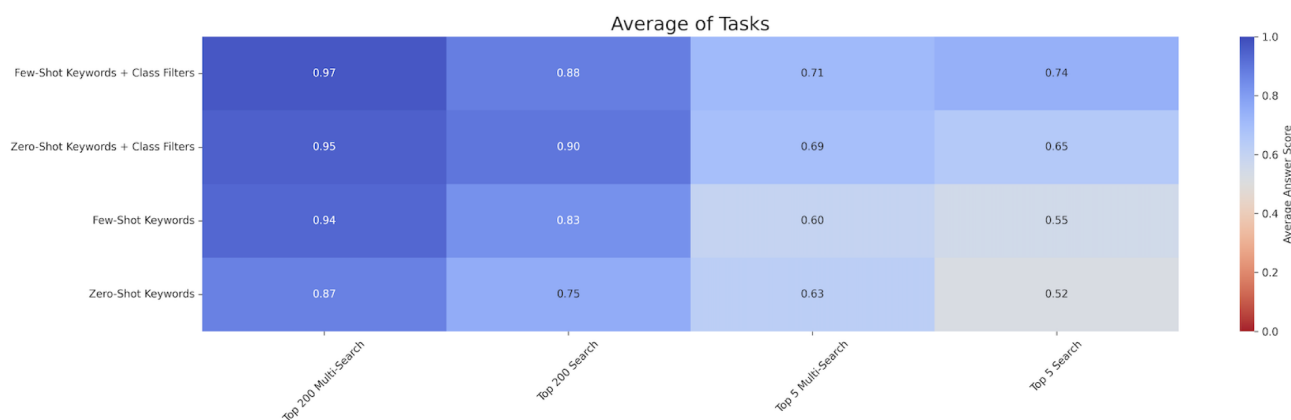


Figure 7—Ablation Testing Summary Comparing Multi-Search vs Single Search

In Phase Three, ATHENA also includes an LLM fine-tuned on the type of charts and plots used in the industry so it can answer questions where the information is only available in charts. This gives it a clear advantage over the baseline for the sub tasks which requires reading a chart.

Conclusions

Adding appropriate domain aware components to an agentic RAG implementation greatly improves the solution's performance. This was shown with ATHENA in Phase Two testing and used to greatly improve the baseline system performance for Phase Three testing.

Real world diligence in the oil and gas industry requires not only the ability to find a single piece of information in a vast corpus of documents but also the ability to find all the answers across that corpus.

The research team created a novel approach, called SS-RAG, which greatly improves the performance of a question answering solution for both these use cases.

Acknowledgement

The team would like to extend its deep gratitude to the volunteers who gave generously of their time to test the systems.

The work reported in this paper was supported by U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA), Contract No. W911NF22C0048. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. KMASS: Knowledge Management at Scale and Speed – <https://www.darpa.mil/research/programs/knowledge-management-at-scale-and-speed>
2. Wikipedia description of ChatGPT and its impact on AI: <https://en.wikipedia.org/wiki/ChatGPT>
3. Equinor. (2018). Volve field data (Equinor Open Data License). Equinor. <https://www.equinor.com/energy/volve-data-sharing>
4. PetroWiki, at the time of testing, was a Wiki run by the Society of Petroleum Engineers. It is now included in OnePetro, SPE's library and can be found at <https://onepetro.org/pages/petrowiki>
5. Eric Schoen, Reid Smith, John Boden. (2018). AI Supports Information Discovery and Analysis in an SPE Research Portal, <https://doi.org/10.2118/191758-MS>
6. Fabrizio Dell'Acqua et al 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. 2023. *Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013*, <http://dx.doi.org/10.2139/ssrn.4573321>.
7. McKinsey Digital. 2024. Unleashing developer productivity with generative AI. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai>. Accessed March 22, 2024
8. Jianbin Shen, Junyu Xuan, and Jie Liang. 2023. Mitigating Intrinsic Named Entity-Related Hallucinations of Abstractive Text Summarization. *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15807–15824.