

Building a Virtual Member of a Community of Practice

Joshua Eckroth[†]
i2k Connect Inc.
Houston, TX, USA
jeckroth@i2kconnect.com

Dayne Freitag
SRI International
San Diego, CA USA
dayne.freitag@sri.com

Jonathan Keefe
Pacific Science & Engineering
San Diego, CA USA
jonathan.keefe@pacific-science.us

Timothy Meyer
Pacific Science & Engineering
San Diego, CA USA
timothy.meyer@pacific-science.us

Karen L. Myers
SRI International
Menlo Park, CA USA
karen.myers@sri.com

Eric Schoen
i2k Connect Inc.
Houston, TX, USA
ejschoen@i2kconnect.com

Pedro Sequeira
SRI International
Menlo Park, CA USA
pedro.sequeira@sri.com

Reid G. Smith
i2k Connect Inc.
Houston, TX, USA
rgsmith@i2kconnect.com

Abstract

We describe a virtual member of a knowledge management Community of Practice (CoP), called ATHENA, that knows an individual, his tasks, his organization, and the community. ATHENA employs an agentic chat capability that combines embeddings with knowledge-based faceted search to provide accurate responses to technical questions along with rationale and citations for efficient validation. ATHENA supports natural, in-the-flow capture of task-related insights to share within a CoP, along with proactive dissemination of information tied to an individual and his current needs. An evaluation involving 75 professionals from the Oil & Gas sector shows that ATHENA dramatically improved outcomes and productivity on a set of well-planning tasks compared to their use of a state-of-the-art RAG baseline. Interestingly, ATHENA also enabled eight non-experts to perform at expert levels.

CCS Concepts

• Information systems->Information systems applications->Enterprise information systems

Keywords: knowledge management; knowledge capture; chat interface; agentic information retrieval; personalization

ACM Reference Format:

Joshua Eckroth, Dayne Freitag, Jonathan Keefe, Timothy Meyer, Karen L. Myers, Eric Schoen, Pedro Sequeira, Reid G. Smith. 2025. Building a

[†]Authors are listed in alphabetical order.

Virtual Member of a Community of Practice. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, November 10-14, 2025, Seoul, Republic of Korea, 8 pages. <https://doi.org/10.1145/3746252.3761585>

1 Introduction

To be successful in today's increasingly complex and information-dense world, workers in many fields must find, evaluate, and integrate knowledge from multiple domains and sources. Knowledge management (KM) tools strive to support workers and organizations with these processes by capturing, organizing, and leveraging information and expertise [1]. When used by individuals and organizations to systematically manage knowledge, KM systems can improve decision making, foster innovation, and enhance productivity across teams by creating two types of connection: "people-to-people" and "people-to-information."

The predominant approach to fostering people-to-people connection is to build a Community of Practice (CoP). CoPs have been the focus for KM initiatives within most organizations [2,3], typically involving online forums for information sharing and expertise finders for locating practitioners with specific knowledge. Technologies like Microsoft SharePoint and search engines remain the primary enabler for people-to-information connection in commercial settings, especially for highly technical disciplines.



This work is licensed under [Creative Commons Attribution International 4.0](https://creativecommons.org/licenses/by/4.0/).

CIKM '25, November 10-14, 2025, Seoul, Republic of Korea.

© 2025 Copyright is held by the owner/author(s).

ACM ISBN 979-8-4007-2040-61/2025/11. <https://doi.org/10.1145/3746252.3761585>

Although AI has been used to enable knowledge management for some time [4], recent advances in the field provide the opportunity to build a new generation of KM technologies that provide better access to knowledge and reduce friction for knowledge capture and sharing.

To this end, we have developed an intelligent assistant called ATHENA that operates as a *virtual member of a CoP*. ATHENA combines three core capabilities embedded within a chat-based user interface that mirrors the natural workflow of person-to-person knowledge sharing (see Figure 1):

- an *agentic approach to information retrieval* that delivers accurate answers rather than documents, along with source material and rationale;

Figure 1. ATHENA overview

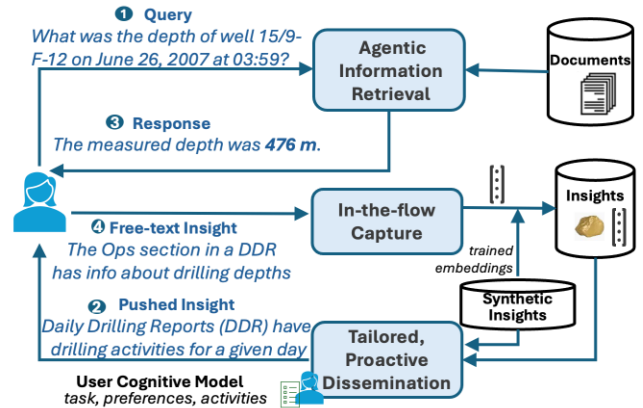
- an *in-the-flow capture of experiential knowledge*, or “insights”, that can be shared within a CoP;
- *proactive dissemination of insights* tailored to an individual’s expertise level and task context via user cognitive modeling and large language models (LLMs).

Much of ATHENA is domain agnostic; however, it was developed in the context of addressing KM challenges for the Oil & Gas sector. We evaluated ATHENA in a realistic setting involving 75 subjects from a professional society devoted to petroleum engineering (the Society of Petroleum Engineers, or SPE), who performed a range of tasks related to well planning. The results show that ATHENA significantly improved outcomes (152% increase in task scores) for experts in the domain compared to their use of a state-of-the-art Retrieval Augmented Generation (RAG) baseline; our technology also increased subject productivity by 283%. Surprisingly, a supplementary evaluation involving 8 subjects showed that ATHENA enabled non-experts to complete tasks as quickly and accurately as experts. The basis for these results was a combination of our agentic retrieval method providing answers that were much more helpful than the baseline and timely dissemination of insights that provided meaningful guidance tailored to a subject’s problem-solving context.

Based on the evaluation results, we have been asked to deploy ATHENA to the SPE’s research portal. In addition, ATHENA’s agentic chat system is in production deployment at a Top-10 Oil & Gas operator and two additional Top-10 operators are in the process of deploying it.

2 Domain: Well Planning

Well planning involves designing and mapping out the wellbore, estimating costs, and identifying the optimal drilling approach to access hydrocarbons efficiently and safely [5]. These tasks require a comprehensive analysis of geological data, regulatory guidelines, and potential environmental impacts. Finding the right information is paramount, as accurate data helps minimize risks, optimize resources, and ensure compliance with safety and environmental standards. Access to reliable information enables



precise decision making, reduces costly errors, and ultimately drives successful project outcomes, positioning companies to maximize productivity while maintaining operational integrity.

We employed a large corpus (17,598 files) containing (a) general material related to the Oil & Gas sector drawn from an extensive petroleum engineering Wiki (PetroWiki, <https://petrowiki.spe.org/PetroWiki>); (b) relevant regulatory matter (US Code of Law, Title 30, <https://uscode.house.gov/browse.xhtml>); and (c) an extensive real-world data set comprising logs, reports, and studies related to drilling operations by Equinor in the Volve oil field (<https://www.equinor.com/energy/volve-data-sharing>).

3 ATHENA User Experience

Current practice for accessing knowledge within a CoP typically involves leveraging a diverse set of tools including file explorers to search documents and data, asynchronous communication channels to request and share information, and dedicated applications for capturing knowledge and completing work. This practice slows the process of finding and sharing knowledge, while also increasing cognitive load by forcing users to constantly switch tasks.

In contrast, our conceptualization of ATHENA as a virtual member of a CoP integrates search, knowledge capture, and sharing functions within a single UI. Users interact with ATHENA via natural language chat messages (see Figure 2). The UI supports chat messages with textual questions, keyword queries and knowledge-based faceted search [6,7] ((A, B) in the figure). For a given answer (C), it provides a clear explanation of the response logic and sources to enable evaluation of the accuracy and completeness of answers (D). Users can view source documents by clicking on their titles in the response (E), which opens the document in an integrated file viewer – removing the need for switching between applications to assess information. To support related questions, we crafted a “document focus” feature (F) that enables a user to directly interrogate a document on the click of an icon by limiting the chat mechanism to respond considering only that document. Once a suitable document has been found,

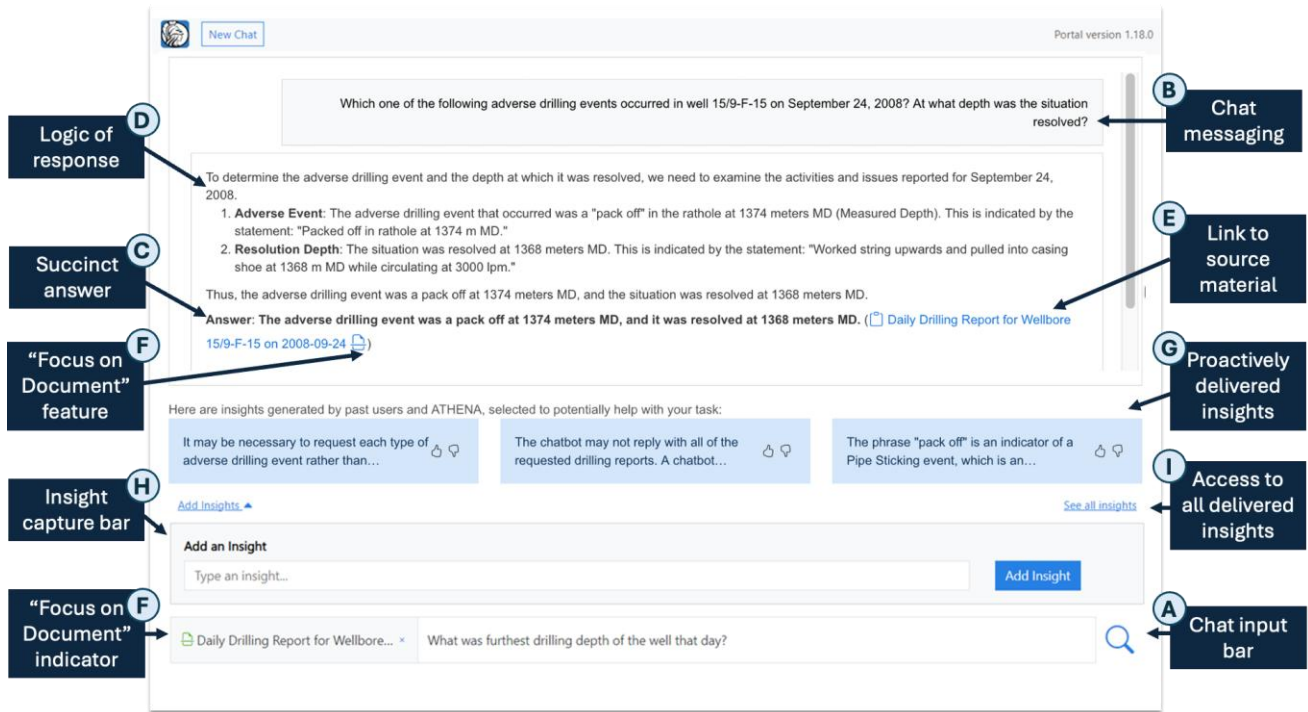


Figure 2. ATHENA user interface supporting chat, capture, and dissemination

this focus reduces the likelihood of hallucinations derived from consideration of irrelevant material.

ATHENA proactively disseminates task-related insights directly within a user’s task flow, in the form of blue text boxes between the chat window and chat input bar (G). These insights are tailored to each user based on their expertise, previously seen insights, and current goals – seeking to deliver the right information at the right time in the right way to maximize its utility [8]. The insights expand and collapse to reduce visual clutter, change color when read, and allow feedback via thumbs up/down buttons to better tailor insights.

ATHENA facilitates the capture of knowledge from users via a lightweight process for submitting insights directly within their task flow. To submit a new insight, users can simply click the ‘Add Insights’ link and type into a text box (H). This process is designed to be quick and easy to encourage sharing with CoP members. The enrichment of captured content to facilitate downstream dissemination is handled by the system (see Sections 5 and 6), rather than requiring additional user input to contextualize the insight. Users can examine insights they have contributed or that they referenced previously (I).

4 Agentic Chat Information Retrieval

Traditional approaches to information retrieval require interacting with a large corpus of documents or data by formulating precise search queries or navigating extensive, hierarchical folder structures within a KM system. This approach increases user cognitive load and creates the challenge of forming

a shared mental model with a search function or file explorer centered around keywords rather than knowledge. Users must spend a significant amount of time learning the system’s content and structure, generating and testing keywords, and filtering results to identify promising data for further evaluation.

As a virtual member of a CoP, ATHENA provides ready access to a vast repository of knowledge, including insights captured from community members. ATHENA offers the convenience of surfacing that knowledge without specialized search skills or machine-centered mental models thanks to its chat interface. This approach enables users to receive relevant answers to their knowledge query in plain text, along with the information needed to evaluate the answer – an ideal workflow akin to discussing a well-sourced document with CoP peers. In this section, we explain how we architected ATHENA’s interpretation of a user message to achieve this goal.

ATHENA’s chat system is a refined RAG architecture [9]. RAG systems typically convert a user message into either a set of keywords used for search or a vector embedding generated by a language model. A search is then performed across the corpus to identify relevant documents or “chunks” (partial documents). Search may involve keyword matching, vector nearest neighbor search using the embeddings of the user input and documents/chunks, or both (called hybrid search). The top-matching documents/chunks from the search are then fed with the original user request to an LLM for filtering and answer generation.

In early experimentation, we found that queries for our domain involving technical concepts such as dates and times, domain-specific document types, and the names of wells, yielded poor results. For example, if a user’s request is for a DDR (daily drilling report) for well 15/9 F-15 for September 2, 2009, then both keyword search and embedding search, and even their combination, often fail to find the right documents or chunks. First, the keyword in the request, “DDR,” is never used in these reports, since it is a shorthand. Second, “15/9 F-15” is often tokenized incorrectly by keyword search, causing irrelevant documents to be retrieved. Third, the date “September 2, 2009” might not appear in that form in the document (or its filename or path), instead it might appear as “2009-09-02” or other formats. In terms of embedded search, the embedding vector for a request that mentions, say September 2, 2009, will be close to the embedding vector for a request that mentions September 10, 2008. Overall, we find that keyword, embedded, and hybrid search (1) fail to understand synonyms or acronyms, (2) fail to recognize well names accurately, and (3) lack the ability to prioritize important parts of the query such as the well name and date to ensure exactly those documents are returned and not others for similarly named wells or similar dates.

To address these issues, we developed an agentic approach [10] in which different agents target specialized parts of the user’s request, performing search for relevant documents and/or chunks that address that aspect of the request. Our agents specialize on Document Classification (e.g., report types), Locations (including wells), Date & Time, Keywords, and Embedding Search. Some agents use an LLM to help identify relevant parts of the query, e.g., the Date & Time agent uses GPT-4o to identify dates and times. Each agent scores its results using a specialized relevance match from its search. Scores are combined via harmonic mean with the top matches then sent to an LLM along with the user’s request to generate the final response, using chain-of-thought prompting [11] to generate references and rationale (see Figure 2).

5 Knowledge Capture

Current practice for capturing on-the-job experiential knowledge often involves dedicated tools that take users “out of the flow”, forcing undesirable context shifts that detract from task performance [12]. Alternatively, lessons learned are documented as part of an after-action review long after a task has been completed, when critical context may be lost and key insights forgotten. To enable subsequent retrieval of such material, contributors must either manually tag documents using predefined ontologies to support indexing or navigate to appropriate folders within their corporate repositories to position content where it is more likely to be discovered by CoP members who need it in the future [13].

In contrast, our approach to capture (see Figure 2) keeps users “in the flow” by having them enter insights as they complete tasks via a co-located text input box. This design makes it easy for users to contribute knowledge and to focus on the information they choose to document. It also avoids the need to learn predefined languages for tagging material or mastery of complex tools.

While this minimalist approach facilitates capture, it increases the challenge for subsequent dissemination to CoP members engaged in similar tasks and contexts. Standard approaches to indexing textual material, such as keyword indexing or vector stores [14] that are commonly used in RAG [15], fail an important test: they do not account for a gap in form and content between the description of a task and any relevant supporting knowledge. This means they will fail to identify useful insights to share that are semantically distinct from the description of the task performed by the user.

As these considerations imply, we seek to optimize the retrieval of captured knowledge given a statement of the current task. To this end, we embed tasks and insights into a shared semantic space and use proximity in the space for relevance assessment. We then retrieve any embedded insights proximal to this point, ranking them by distance from the task embedding. By embedding tasks and insights separately, we provide the system a means to overcome any discrepancies between “task language” and the corresponding “insight language.” Then, the embeddings are optimized to narrow the distance between tasks and the insights relevant to them, using a set of known task-insight relevancy pairings.

Training robust text embeddings requires large volumes of data. To overcome the cold start problem of developing a new system, we generated synthetic task-insight pairings. Specifically, relying on ChatGPT’s prior exposure to language from the Oil & Gas sector, we: (1) presented it with documents from the domain; (2) asked it to generate insights by explaining any specialized language or the purpose of the various document elements; (3) presented it with the insights generated in Step 2 and asked it to invent tasks that might rely on a given insight. The resulting collection of task-insight pairings was used to train our embedding model. This approach to overcoming data insufficiencies for training machine learning models is called data augmentation [16], and generative LLMs, with their ability to produce synthetic data at scale, have opened new possibilities in this area [17]. To our knowledge, we are the first to simulate knowledge elicitation for rapid deployment of KM systems in new domains.

To assist with dissemination, a textual rationale is generated for each captured insight that describes the type of tasks for which the insight is relevant. The rationale is computed from the insight’s content and the current task description using an LLM via few-shot prompting [18], where we provide examples of insight, task, and rationale tuples to the LLM to guide the generation process.

6 Tailored Dissemination of Insights

Task-focused proactive assistance, first addressed in the Lumiere project [19], remains an open challenge. Effective push of helpful suggestions requires understanding of the user’s current task, intent, expertise level, and recent activities [20]. Current approaches to personalizing content often rely on static user profiles built on simple models of expertise and rigid approaches to tailoring knowledge. These limitations can lead to critical

failures to deliver or tailor data, leaving CoP knowledge unused or misunderstood.

Our approach to proactive dissemination of insights involves using an LLM (we used GPT 3.5-turbo) to flexibly select and tailor content to an individual’s expertise and problem-solving context. To assess a user’s expertise online, we maintain a User Cognitive Model (UCM). The UCM includes a Long-Term Memory (LTM) that stores available information about a user’s background, experience, and expertise. In our evaluation (Section 7), the LTM contains users’ responses to an eleven-question, multiple-choice demographics survey designed to assess subject technical background, years of experience in the field, and familiarity with granular concepts related to well planning. The UCM also includes a Short-Term Memory component with information about a user’s current problem-solving context, namely a textual description of the current task, the queries performed for that task, and the documents consulted. UCMs are updated continuously to reflect user activity detected while performing a task.

We architected ATHENA to integrate with task performance systems in which there is access to a textual description of a task and a triggering event that identifies when a user begins a task. Upon receipt of such a triggering event, ATHENA considers whether any insights in its store are relevant (i.e., *selection*) and how they can be tailored to a given user (i.e., *personalization*). For the evaluation described in Section 7, our store of available insights contains both human- and machine-generated content. Human-generated insights originated from two sources: 16 were from a subject-matter expert (SME) who assisted with the formulation of our evaluation tasks, and 74 were from subjects who participated in a separate evaluation of our knowledge capture mechanism. The machine-generated insights comprised 442 general insights for our domain that were used as a test set for training our insight embedding function (derived from our corpus, as described in the previous section), and 442 “distractor” insights generated using ChatGPT that captured general information about well planning. We used this large, diverse set to test the ability of our selection mechanism to identify insights pertinent to a particular user given their context.

Insight selection works as follows: for a given task, insights are first ranked using our embedding function. We then select from among the top insights for further processing, weighting the selection based on source to favor SME-provided insights (10) given their likelihood of greater relevance compared to the other types (5 each). An LLM then selects from that list of 20 insights the 3 that would best help the user perform the current task, given a description of the user’s expertise and experience and insights’ rationale. Users’ descriptions are created via a template listing the insights they have already seen during the task and their answers to the survey.

Selected insights are then further tailored to the user’s modeled expertise. Offline, an LLM is used to identify relevant domain concepts from each insight’s textual content. We then have an LLM generate a short description for each concept at novice, intermediate, and expertise levels with the goal of providing relevant supporting information to such users. We

assume that users classified as experts for some concepts do not require additional information on them. At dissemination time, we again use an LLM to classify a user into one of the expertise levels for each concept based on their responses to the demographics survey in the UCM. We then append the pre-computed concept descriptions for the selected insights according to the user expertise classifications for each concept.

Compared to other work on LLM-based personalization, our approach focuses on user expertise with respect to different in-domain concepts rather than generating user profiles solely from interaction history and embedding them with the LLM inputs for contextualization, e.g., [21,22,23]. Like [24], we integrate user information from the UCM and summarize it for contextualization but our summarization is done online rather than offline to enable more dynamic user modeling.

7 Evaluation

7.1 Experiment Design

We evaluated ATHENA through a multi-phase study in which subjects completed a set of well-planning tasks.

Subjects in the first two phases of the study were volunteers from the Society of Petroleum Engineers (SPE) – the premier technical society for the discipline. By leveraging real-world tasks and experts in the field, we aimed to get an ecologically valid measure of ATHENA’s impact as it would be deployed.

In Phase 1 of the study, 8 SPE volunteers performed two complexity-paired tasks using a baseline system and ATHENA in a counter-balanced order. The baseline was a standard RAG system built on Claude Haiku 3.0 with a chat interface to support information retrieval from our corpus; it was created by a third party engaged by our research funder to compare our work to the state of the art. In Phase 2 of the study, 66 SPE volunteers performed three randomly selected tasks using the ATHENA system. This phase enabled us to collect more data to leverage in between-subject analyses, providing a fuller view of ATHENA’s value for our target users. One additional SPE member who was familiar with the evaluation tasks and both systems served as a “gold standard” user, to provide an estimate of optimal task completion time.

We had planned originally to have all SPE subjects participate in within-subjects testing. After pilot testing, our SPE liaison raised concerns that the baseline was too frustrating to use and would reflect badly on SPE. To address those concerns, we limited the within-subjects experiment to 8 people.

In Phase 3 of the study, eight subjects with technical backgrounds but no experience in petroleum engineering followed the same protocol as for Phase 1. This phase both assessed the extent to which ATHENA could improve performance for non-experts and enabled comparison of novice versus expert performance.

In all phases, subjects completed testing by logging into an online portal that brought them to the first (or only) system they would use. To start, subjects completed a short questionnaire related to their experience in the Oil & Gas sector and familiarity with task-related topics. Data from this survey was used to

Table 1. Results for SPE within-subjects (N = 8), SPE all subjects (N = 74), and non-SPE within-subjects (N = 8)

		Avg. Grade [0-100]	Failure Rate (grade<50%)	Avg. Task Time over Expert (minutes)	Productivity (grade/time)	SUS
SPE	Baseline – 8 subjects	30.7	0.88	7.3	2.4	24.8
	ATHENA – 8 subjects	77.3	0.13	2.9	9.3	40.4
	Change relative to Baseline	152%	-86%	-60%	283%	63%
	ATHENA – all 74 subjects	78.1	0.04	11.9	6.4	40.9
	Change relative to Baseline	154%	-95%	-6%	172%	65%
Non-SPE	Baseline – 8 subjects	26.0	0.69	14.2	1.3	25.4
	ATHENA – 8 subjects	95.8	0.00	5.9	8.4	43.3
	Change relative to Baseline	268%	100%	-70%	534%	70%

formulate general and topic-specific models of user expertise for our personalized dissemination capability, as described in Section 5. Next, subjects were presented with a short training video outlining how to use the first (or only) system. In the video, subjects were told that the chat systems may be helpful but can also produce inaccurate results and so they should validate information provided by the system by checking sources in the corpus. After completing a practice task, subjects then performed their assigned evaluation tasks. Subjects that used both systems repeated this workflow (except the questionnaire) for the second system.

The evaluation tasks involved: analyzing drilling information for a well, estimating pore pressure, interpreting plans for a possible well, and analyzing adverse drilling events. Each task consisted of three related sub-tasks that required finding and analyzing information from the corpus. Subjects were required to provide responses to encourage them to persevere on tasks that may seem challenging. It was suggested to spend at most ten minutes per subtask but time limits were not enforced. An SME-informed scoring rubric assigned equal credit to each subtask.

7.2 Results

The upper portion of Table 1 shows results for the SPE subjects, covering both the within-subjects cohort ($N=8$) of Phase 1 that used both the ATHENA and Baseline systems; and the full set of results for all SPE participants across Phases 1 and 2 using the ATHENA system ($N=74$). As shown, ATHENA enabled much better performance than the Baseline. Average task grade improved appreciably (152% for within-subjects, 154% for all subjects) and was statistically significant ($p < .001$).

We did not find a significant difference in average task time over expert between ATHENA and Baseline conditions when considering the within-subjects data. However, it should be noted that many subjects simply gave up in the Baseline condition because the tool was not helpful in identifying needed materials.

Productivity (grade scaled by average time spent) provides a more meaningful measure of subject performance. As can be seen from the results, productivity increased appreciably in both settings, almost tripling for the within-subjects cohort. *This result clearly shows that ATHENA enabled a higher level of performance,*

and that the significant improvement in task grade was not the result of a speed-accuracy trade-off.

To complement the performance-based measures of system utility, we collected subjective usability ratings from subjects using the System Usability Scale (SUS [25]). Table 1 shows that subjects rated ATHENA as much more usable than the baseline system for both the within-subjects (63% improvement) and all subjects data (65% improvement). Follow-up t-tests indicated that the improvement in usability scores was significant in the within-subjects data, $t(7) = 2.97$, $p = .02$, $d = 1.91$, and for all data, $t(8.39) = 4.56$, $p = .002$, $d = 1.79$.

The bottom half of Table 1 presents the results for the within-subjects experiment run with the 8 non-SPE subjects. The non-expert subjects struggled with the Baseline system but performed their tasks effectively using ATHENA, with productivity increased by 534%. Paired t-tests demonstrate that subjects not only scored significantly better using ATHENA vs. Baseline, $t(7) = 7.49$, $p < .001$, $d = 4.72$, but also completed their tasks significantly faster, $t(7) = 3.09$, $p = .02$, $d = 0.81$. Additionally, we found an improvement in usability scores for ATHENA vs. Baseline, $t(10.9) = 5.34$, $p < .001$, $d = 2.64$.

When using the Baseline, non-SPE subjects performed worse than SPE subjects as one would expect; *surprisingly, those same non-expert subjects outperformed the SPE subjects on average when using ATHENA*. Follow-up Welch’s t-tests indicate a marginally significant effect of subject group when subjects used ATHENA, $t(8.8) = 2.28$, $p = .05$, $d = 1.14$, but not Baseline, $t(12.9) = 1.52$, $p = .15$, $d = 0.76$, indicating that non-SPE subjects performed better than SPE subjects selectively when using ATHENA. The Phase 3 subjects consisted of three with PhDs in computer science, two software engineers, and two non-engineers. A potential explanation for the finding is that the evaluation tasks had an inherent “information foraging” element to them. Two of the non-SPE subjects had backgrounds in information retrieval and all had some familiarity with LLMs. Still, the results were unexpected. Interestingly, the top and bottom results in the non-SPE condition were not correlated with computing expertise.

7.3 Discussion

Greatly improved chat retrieval was a key reason for the substantial improvements in performance compared to the Baseline. Follow-up analyses showed that the baseline system, although built on state-of-the-art RAG technology, failed to identify relevant documents, often selecting the wrong type of report for the wrong date and the wrong well. We believe this confusion was caused by the exclusive use of embeddings for chunk retrieval within RAG (compared to our hybrid approach) which led to confusion of entities that appear in similar contexts (so-called *entity-related hallucination* [26]).

Many subjects felt ATHENA’s “document focus” feature helped them to interact effectively with the chat capability. LLMs excel at summarizing or extracting information from a specified document but are less effective at finding relevant documents. Document focus enabled users to contextualize chat interactions to the target document, enabling improved quality of results and increased user satisfaction.

Proactive dissemination of insights also played a notable role. Table 2 summarizes the distribution of insights by source. It shows that our dissemination mechanism was effective in selecting task-relevant insights, with no “distractors” (those from ChatGPT) presented to users. In response to the question “*Did you find the suggested Insights to be helpful in completing your tasks?*”, 82% of the 49 respondents responded positively while another 10% categorized them as mixed; the remaining 8% either did not use the insights or responded negatively.

Table 2. Source of disseminated insights

Insight Source	% of Disseminated Insights
Captured from SME	82.5
Captured from subject	15.2
Synthesized from document	2.3
Synthesized from ChatGPT	0.0

8 Deployment

Our work with real problems, users, and data has positioned us to have meaningful impact. ATHENA’s agentic chat system is in production deployment at a Top-10 Oil & Gas operator and two additional Top-10 operators are in the process of deploying the solution. In addition, we are working with the SPE, a society with 132,000 members across 146 countries, to deploy ATHENA technology to help find answers to questions across SPE’s substantial corpus of publications (310,000 items currently); the target date for deployment is Q4 2025.

9 Conclusions

We describe the ATHENA knowledge management framework that leverages recent advances in generative AI and user modeling to provide personalized, situation-aware task assistance. Formulated as a virtual member of a CoP, ATHENA’s chat framework built on agentic information retrieval provides accurate question-answering with rationale in the highly technical domain of petroleum engineering. Insights gleaned by

users while completing tasks are captured in the flow, eliminating the need for costly context shifts. Task-relevant insights are proactively disseminated based on a dynamic cognitive model of the user that encodes expertise and activity to support personalization linked to fine-grained domain concepts.

We conducted a large-scale evaluation of our technology firmly grounded in operational realism: tasks were representative of real-world problems related to well planning; the corpus comprised content used widely in the domain; our subjects were professionals in the Oil and Gas sector. The results showed that ATHENA significantly improved outcomes and productivity for experts compared to their use of a state-of-the-art RAG baseline. Surprisingly, a supplementary evaluation showed that ATHENA enabled non-experts to complete tasks on par with domain experts.

While we believe the work reported here constitutes a valuable step forward in developing effective knowledge management systems for the Oil & Gas sector, we plan to extend it in several ways. A truly effective retrieval tool for our domain should be able to extract and interpret information from common forms of diagrams and charts. Captured insights are often of poor quality, making it essential to provide the means to filter (at a minimum) or improve (ideally) the content to improve its downstream utility. Our dissemination mechanism could be improved through personalization achieved over longer usage periods and by integrating disseminated insights within the chat framework.

Acknowledgments

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. The authors are deeply appreciative of the assistance provided by the Society of Petroleum Engineers and its members.

GenAI Usage Disclosure

Generative AI was used as an enabling technology for our research as follows.

- ChatGPT provided the underlying chat capability in the ATHENA RAG system.
- GPT-4o was used to support recognition of specialized entities (e.g., dates and times) within user queries.
- ChatGPT was used to generate data for training our embeddings representation of insights, and to generate distractor insights to support our evaluation.
- GPT 3.5-turbo was used to generate rationale for captured insights and to help select and tailor insights for dissemination to users.
- Claude Haiku 3.0 provided the underlying chat capability in the Baseline RAG system.

Generative AI was not used in the writing of this paper.

REFERENCES

- [1] C. Jackson Grayson and Carla O'Dell. 1998. *If Only We Knew What We Know: The Transfer of Internal Knowledge and Best Practice*. Simon and Schuster.
- [2] Etienne Wenger. 1998. *Communities of Practice*. Cambridge University Press.
- [3] Verna Allee. 2000. Knowledge networks and communities of practice. *OD practitioner*. Vol. 32 (4), 4–13.
- [4] Reid G. Smith and Adam Farquhar. 2000. The Road Ahead for Knowledge Management: An AI Perspective. *AI Magazine*, 21(4):17–40.
- [5] Steve Devereux. 1998. *Practical Well Planning and Drilling Manual*. Pennwell Books.
- [6] Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.
- [7] Reid G. Smith, Eric J. Schoen, Joshua Eckroth, David Mack Endres, Sebastian Flores, Julia Elliott, and Bruce G. Buchanan. 2021. SPE Research Portal: How SPE Uses Artificial Intelligence to Help You Find Technical Information. *The Way Ahead, Society of Petroleum Engineers*.
- [8] Gerhard Fischer. 2012. Context-aware systems: the right information, at the right time, in the right place, in the right way, to the right person. *Proc. of the International Working Conference on Advanced Visual Interfaces*, pp. 287–294.
- [9] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.
- [10] Deepak B. Acharya, Karthigeyan Kuppan and B. Divya. 2025. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access*, Vol. 13.
- [11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, 24824–24837.
- [12] Carla O'Dell and Cindy Hubert. 2011. *The New Edge in Knowledge Management*, John Wiley & Sons.
- [13] Chris J. Collison, Paul J. Corney, and Patricia L. Eng. 2019. *The KM Cookbook: Stories and strategies for organisations exploring Knowledge Management Standard ISO30401*. Facet.
- [14] 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 preprint arXiv:2401.08281.
- [15] Patrick Lewis, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, pp. 9459–9474.
- [16] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. arXiv preprint arXiv:2010.12309
- [17] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh T. Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. arXiv preprint arXiv:2403.02990
- [18] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting Large Language Models. arXiv preprint arXiv:2305.03653
- [19] Eric J., Horvitz, John S., Breese, David Heckerman, David Hovel, and Koos Rommelse. 1998. The Lumière project: Bayesian user modeling for inferring the goals and needs of software users. In *Proc. of Uncertainty in AI*.
- [20] Neil Yorke-Smith, Shahin Saadati, Karen Myers, and David Morley. 2009. Like an Intuitive and Courteous Butler: A Proactive Personal Agent for Task Management. *Autonomous Agents and Multi Agent Systems*. Vol. 1, pp. 337–344.
- [21] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. arXiv preprint arXiv:2402.13598
- [22] Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024. LLMs+Persona-Plug=Personalized LLMs. arXiv preprint arXiv:2409.11901
- [23] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani. 2024. Lamp: When Large Language Models meet personalization. arXiv preprint arXiv:2304.11406.
- [24] Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Z. Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. arXiv preprint arXiv:2310.20081
- [25] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*.
- [26] 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.