

Keyword search is all you need: Achieving RAG-Level Performance without vector databases using agentic tool use

Shreyas Subramanian^{*1}, Adewale Akinfaderin*, Yanyan Zhang*, Ishan Singh*, Chris Pecora, Mani Khanuja, Sandeep Singh, Maira Ladeira Tanke

Amazon Web Services
410 Terry Avenue North,
Seattle, WA 98109

¹subshrey@amazon.com

Abstract

While Retrieval-Augmented Generation (RAG) has proven effective for generating accurate, context-based responses based on existing knowledge bases, it presents several challenges including retrieval quality dependencies, integration complexity and cost. Recent advances in agentic-RAG and tool-augmented LLM architectures have introduced alternative approaches to information retrieval and processing. We question how much additional value vector databases and semantic search bring to RAG over simple, agentic keyword search in documents for question-answering. In this study, we conducted a systematic comparison between RAG-based systems and tool-augmented LLM agents, specifically evaluating their retrieval mechanisms and response quality when the agent only has access to basic keyword search tools. Our empirical analysis demonstrates that tool-based keyword search implementations within an agentic framework can attain over 90% of the performance metrics compared to traditional RAG systems without using a standing vector database. Our approach is simple to implement, cost effective, and is particularly useful in scenarios requiring frequent updates to knowledge bases.

Introduction

The rapid advancement of Large Language Models (LLMs) has significantly transformed various industries and applications (Hadi et al. 2023), revolutionizing tasks such as text generation, summarization, question-answering systems, and chatbots (Dam et al. 2024; Kumar et al. 2023). These sophisticated models have demonstrated remarkable proficiency in understanding and generating human-like text, leading to widespread adoption in multiple sectors (Hadi et al. 2023). Recent developments in neural architectures and training methodologies (Vaswani et al. 2017; Brown et al. 2020; Liu et al. 2019; Raffel et al. 2020; Xu et al. 2022) have enabled these models to process and generate increasingly complex and contextually relevant responses. However, as LLMs become integral to critical applications, the need for accurate information retrieval and response generation has intensified (Bender et al. 2021; van

Schaik and Pugh 2024). This evolution has prompted a shift from standalone LLM responses to hybrid systems that integrate external knowledge sources, aiming to improve the factual accuracy and relevance of generated content (Es et al. 2023; Zhao et al. 2024a).

Retrieval Augmented Generation (RAG) systems have emerged as a prominent solution, combining LLMs with external databases to ground responses in factual information (Lewis et al. 2020). This approach mitigates issues like hallucinations (instances where models generate plausible but incorrect information) by providing access to up-to-date and domain-specific data. While RAG systems have proven effective, they face challenges in integrating retrieval mechanisms and maintaining knowledge bases. In response, tool-augmented LLM agents have been developed, using search engines and APIs to retrieve information dynamically, offering greater flexibility, particularly when knowledge requires frequent updates (Qu et al. 2024). Despite these advancements, systematic comparisons between RAG and tool-augmented approaches remain limited. Comprehensive benchmarking metrics assessing retrieval accuracy, response quality, latency, and maintenance overhead are essential for understanding their relative trade-offs and guiding the development of efficient, reliable LLM-based systems tailored to specific applications (Gao et al. 2023b; Es et al. 2023).

This study aims to address this gap by conducting a comparison of traditional RAG systems vs. tool-augmented LLM agents. By systematically comparing their retrieval mechanisms and response quality, we seek to provide insights into their respective strengths and limitations. First, we cover a review of related work in the field of LLM-based retrieval systems and their applications. We then detail our methodology, which encompasses the implementation of RAG systems, tool-augmented LLM agentic frameworks, and our approach to evaluation using LLM-as-a-Judge. We then describe the datasets used in this study and our experimental setup. Finally we show how agentic systems without a standing vector database can perform competitively with traditional RAG approaches.

Related Work

Retrieval-Augmented Generation (RAG) has emerged as a crucial approach for improving the accuracy and reliabil-

^{*}These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ity of Large Language Models (LLMs) by combining retrieval mechanisms with generative capabilities (Guu et al. 2020; Yu, Xu, and Akkiraju 2024; Lewis et al. 2020). By grounding responses in retrieved documents from external knowledge bases, RAG significantly reduces hallucinations and improves factual consistency in LLM outputs (Hu et al. 2024b; Xu et al. 2024; Zhao et al. 2024c). Studies have demonstrated that hybrid retrieval methods, incorporating both dense and sparse retrieval techniques, achieve superior performance in document retrieval quality (Hambarde and Proenca 2023; Zhao et al. 2024b). Research by Rakin et al. (2024) shows that RAG systems utilizing dense passage retrieval can reduce model hallucinations significantly compared to base LLMs in fact-based question answering tasks. Additionally, Izacard and Grave (2021) and Wang et al. (2023a) found that the quality of retrieved passages directly correlates with the accuracy of generated responses, with their Fusion-in-Decoder approach demonstrating substantial improvements in response accuracy across various domains. These findings underscore RAG’s significance in enhancing LLM performance through factual grounding and contextual relevance.

Tool-augmented Large Language Models (LLMs) represent an emerging alternative to traditional Retrieval-Augmented Generation (RAG) systems, offering dynamic access to information through integration with external tools such as search engines, APIs, and specialized databases (Inaba et al. 2023; Prince et al. 2024; Parisi, Zhao, and Fiedel 2022). Unlike static knowledge bases, these systems can interact with real-time data sources, allowing them to maintain accuracy in scenarios where information frequently changes (Hong et al. 2024). The architecture of tool-augmented LLMs enables them to execute multi-turn interactions with various tools, significantly expanding their capability to provide contextually relevant and up-to-date responses. This approach proves particularly valuable in domains where information evolves rapidly, such as automated web navigation, automated game playing, database management, or scientific research, where traditional RAG systems might struggle to maintain current information without frequent updates to their knowledge bases (Gur et al. 2023; Wang et al. 2023b; Xu et al. 2023).

The emergence of tool-augmented approaches has demonstrated significant effectiveness in handling complex queries that require current information, representing a notable advancement in LLM capabilities (Xu et al. 2023; Wu et al. 2024). These systems can dynamically select and utilize appropriate tools based on the query context, offering advantages in maintenance requirements compared to traditional retrieval-augmented implementations (Mialon et al. 2023). The reduced need for maintaining extensive vector databases makes tool-augmented LLMs particularly attractive for resource-constrained applications, while their ability to execute real-time searches ensures accuracy in time-sensitive contexts. Furthermore, tool-augmented LLMs demonstrate enhanced adaptability to new scenarios and use cases, as they can leverage existing tools and APIs without requiring extensive retraining or knowledge base updates (Mialon et al. 2023). This flexibility, combined with

their ability to chain multiple tools together for complex reasoning tasks, positions tool-augmented LLMs as a promising direction for developing more versatile and maintainable AI systems (Chen et al. 2024b).

The comparative analysis of RAG and tool-augmented LLM approaches reveals significant gaps in systematic evaluation methodologies, particularly in standardized benchmarking datasets that reflect real-world document formats and retrieval scenarios. While both approaches demonstrate distinct advantages, the absence of comprehensive datasets that include varied document formats (PDFs, web pages, structured databases) hampers robust comparison (Joshi et al. 2017; Kwiatkowski et al. 2019; Pang et al. 2022; Yang et al. 2018). Current benchmarking efforts focus on retrieval accuracy, response quality, latency, and maintenance costs, but often fail to account for the practical challenges of handling diverse document types and formats (Chen et al. 2024a; Gao et al. 2023a). This limitation is particularly notable in evaluating tool-augmented approaches that utilize document processing tools, where standardized testing frameworks are largely absent (Yuan et al. 2024). Effective comparison requires multiple evaluation criteria, including retrieval precision, response accuracy, and user satisfaction, especially in user-facing applications where interaction quality and retrieval depth significantly impact system performance.

Operational considerations further highlight the distinct trade-offs between these approaches. RAG systems typically demand substantial resources for maintaining and updating knowledge bases, particularly challenging in rapidly evolving domains where frequent updates are necessary (Guu et al. 2020; Fan et al. 2024). While tool-augmented LLM agents potentially reduce maintenance overhead by leveraging existing external tools, they introduce dependencies on external services and may face reliability issues when these services are unavailable (Mialon et al. 2023). Vector database maintenance in RAG systems often incurs higher infrastructure costs, especially for organizations requiring frequent data updates (Fan et al. 2024). However, tool-augmented LLM approaches may encounter latency issues and operational risks due to external dependencies (Qu et al. 2024). These trade-offs become particularly crucial in applications where system adaptability and responsiveness are priorities, requiring careful consideration of resource management strategies and reliability requirements. The choice between RAG and tool-augmented approaches ultimately depends on specific use case requirements, available resources, and the balance between maintenance overhead and system reliability.

Methodology

In this study, we built a reference baseline vector database RAG that was used to compare the effectiveness of our agentic keyword-search approach. A high-level diagram of both approaches can be seen in Figure 1. Evaluations used LLM-as-a-Judge metrics computed using the RAGAS python library (Es et al. 2023) using datasets from various domains. Associated code is available in our github repository (see appendix).

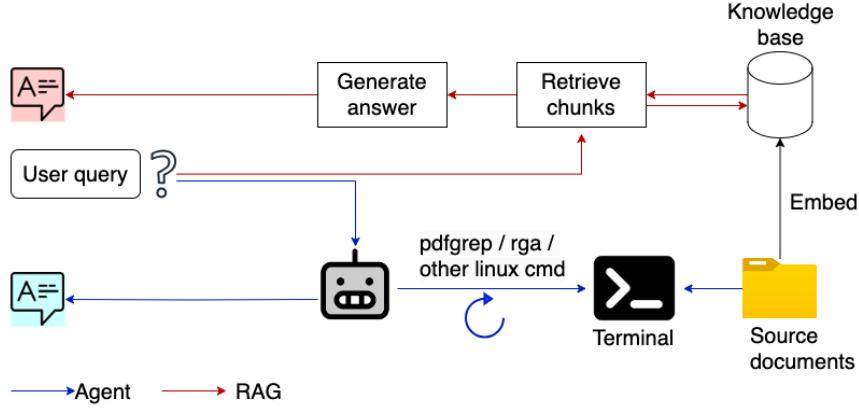


Figure 1: Comparison between RAG (red) and agent-based (blue) pipelines for document QnA

Algorithm 1: Agentic document based Question Answering

Input: User query
Parameter: Folder with source files, Max iterations t_{max}
Output: Final answer

- 1: Let agent iteration $t = 0$.
- 2: Use `pdfmetadata.sh` script to print metadata of all files in the folder
- 3: **while** $t < t_{max}$ **do**
- 4: Observe previous state
- 5: Write `rga` or `pdfgrep` or other linux command
- 6: execute command in linux shell
- 7: **if** additional context found in observation **then**
- 8: Update answer
- 9: **else if** Final answer found **then**
- 10: Stop searching
- 11: **else**
- 12: Continue searching
- 13: **end if**
- 14: **end while**
- 15: **return** Final answer

Datasets

To comprehensively compare traditional vector-based RAG approaches with our proposed agentic search methodology, we curated a diverse corpus of documents that vary in technical complexity, domain specificity, and linguistic structure. These datasets, sourced from Llamahub (Llamahub), were selected as they represent standard RAG benchmarks and ensure reproducibility of our experiments. The datasets include the source pdf, along with questions and associated reference context and answers. The selected datasets represent some of the different challenges commonly encountered in real-world RAG applications:

- **PaulGrahamEssay:** The complete essays from Paul Graham’s personal blog, selected for their complex argumentative structure and interdisciplinary nature (Kamrada 2023).
- **Llama2:** The technical paper describing Meta’s Llama 2 language model architecture and training methodol-

ogy (Touvron et al. 2023).

- **HistoryOfAlexnet:** Covers the development and impact of the AlexNet convolutional neural network (Alom et al. 2018).
- **BlockchainSolana:** Technical documentation and whitepapers related to the Solana blockchain platform (Li et al. 2021).
- **LLM:** A comprehensive survey paper on Large Language Models (Guo et al. 2023).
- **FinanceBench:** A benchmark covering multiple publicly-traded company’s public filings released between 2015 and 2023, including 10Ks, 10Qs, 8Ks, and Earnings Reports (Islam et al. 2023).

Experiment 1: Baseline RAG Implementation

For our baseline comparison, we implemented a standard RAG pipeline using the fully-managed Amazon Bedrock Generative AI platform. An Amazon Bedrock Knowledge Base with Titan Text Embedding Model V2 with 1024 dimensional floating-point embeddings was used along with fixed 300 token chunking strategy with 20% overlap. The source documents from all the datasets were ingested into an associated OpenSearch serverless index. The configuration used for retrieval include max number of chunk retrievals of 5. For response generation, the Anthropic Claude 3 Sonnet model with 200K context window and temperature set to 0.001 was used with a RAG specific prompt detailed in the appendix. Results for the baseline RAG approach were generated for each dataset using the above retrieval and response pipeline with the corresponding dataset questions to be evaluated with ground truth answers and context.

Experiment 2: Agentic Search Framework

Our proposed agentic approach leverages LLMs from the fully-managed Amazon Bedrock platform along with the open-source Langchain framework. The LLM that we leveraged include the Anthropic Claude 3 Sonnet hosted on Amazon Bedrock with a 200K context window. We use the standard ReAct reasoning model (Yao et al. 2022) and set

the temperature to 0.001 for all experiments. The agent orchestration involves initial understanding of the query together and adopting a context-specific search strategy involving the following steps. The agent must begin with an initial metadata analysis of available documents in a folder. Then, the agent can dynamically decide to do broad keyword searches or targeted regex patterns across one or more documents. Using successive context expansion, more keyword searches, and error handling (for e.g. automatic retry) with modified search patterns the agent is able to perform a deep search via a linux shell; this is implemented using the LangChain experimental shell tool. For the agent to understand which documents may be relevant to the question, the agent is invoked with a custom prompt and directed to use several commands within the Linux shell based on the dynamic play between the initial query, search results, and alteration in approach as needed based on errors encountered or the outcomes of previous search iterations. The following are some of the commands the agent can execute:

- *PDF Metadata Tool* that outputs metadata about the directory containing the files, and the individual files.
- *RipGrep-All (rga)* for performing regex-based pattern matching, and multi-keyword search
- *PDFGrep* providing PDF-specific search capabilities, page-range targeting and recursive directory search.

See Algorithm 1. for more details on the agent’s search implementation. In addition, a detailed example agentic workflow can be found in the supplemental material.

Results are generated for each dataset by passing a subset of questions from the dataset to the keyword-search agent producing a candidate answer which can be compared with the ground-truth answer. The text segments selected by the agent from the source document were compared with the ground-truth contexts.

Evaluation Methodology

To compare the baseline RAG approach with the keyword-search agent, the RAGAS evaluation framework Es et al. (2023) was used. This provides a suite of traditional and LLM-as-a-judge based metrics. The primary metrics evaluated include:

- **Faithfulness** which measures factual consistency between the generated answer and the contexts used by the LLM to support its answer.
- **Context Recall** which measures the extent to which all the chunks relevant to answering a query are retrieved.
- **Answer Correctness** which measures the factual accuracy of the generated answer to the ground truth answer.

To assess variance in the agentic approach, metrics were calculated over multiple runs. For the FinanceBench dataset only answer correctness was evaluated and compared for the baseline RAG and agentic approaches due to the complexity of the tables and other complex structures present that poses a challenge for context-chunk based evaluation metrics.

Results

The comparative analysis between the keyword search agent and baseline RAG approaches revealed interesting patterns across three key metrics: faithfulness, context recall, and answer correctness as detailed in Table 1. The agent’s performance relative to RAG was evaluated using attainment scores, which represent the percentage achievement of the agent compared to the RAG baseline. In terms of faithfulness, the agent demonstrated strong performance, achieving an average attainment score of 94.52% across all datasets. The agent performed particularly well on the LLM Survey paper dataset, reaching 99.26% of RAG’s performance, while showing slightly lower but still substantial attainment (88.45%) on the Llama2Paper dataset. Context recall metrics showed more variability, with an average attainment score of 88.05%. The agent achieved near-parity with RAG on the BlockchainSolana and LLM Survey paper datasets (99.62% and 98.71% attainment, respectively), though performance was notably lower for the Llama2Paper dataset (70.56% attainment). For answer correctness, the agent maintained strong performance with an average attainment of 91.48%. Particularly noteworthy were the results on the BlockchainSolana and LLM Survey paper datasets, where the agent achieved virtual parity with RAG (99.97% and 99.51% attainment, respectively). The lowest attainment in this category was observed in the PaulGrahamEssay dataset (79.91%). For a coverage comparison of both approaches on all three metrics, see Figure 2.

Overall, these results suggest that while the keyword search agent generally performed slightly below the RAG baseline, it maintained competitive performance levels, consistently achieving above 88% average attainment across all three metrics without the use of any semantic search via a vector database. We specifically highlight the performance in the BlockchainSolana dataset where context recall and answer correctness are over 99% of the baseline. This indicates that the agent-based approach could serve as a viable alternative to traditional RAG systems, particularly in scenarios where computational efficiency is desired or where vector-database usage is not optimal. Like all retrieval approaches, we observe performance variance across datasets and content types. For example, the PaulGrahamEssay dataset that is characterized by its interdisciplinary writing style, had lower attainment scores and underscores the importance of continued research into improving contextual comprehension beyond keyword matching.

For the complex documents in the FinanceBench dataset, the results show consistent improvement over the traditional RAG baseline, with an average improvement of approximately 6 percentage points. The agent-based approach achieved a mean correctness score of 30.40% ($\sigma = 1.31$), compared to 24.24% for the traditional RAG system as detailed in Table 2. This improvement suggests that our agent’s ability to actively search and interact with complex documents through commands provides more effective information retrieval compared to static chunk-based retrieval methods. The consistency across multiple runs also indicates the stability of our approach.

Dataset Name	Faithfulness			Context Recall			Answer Correctness		
	Agent	RAG	Attain. (%)	Agent	RAG	Attain. (%)	Agent	RAG	Attain. (%)
PaulGrahamEssay	0.8662	0.9056	95.65	0.7527	0.8583	87.70	0.5808	0.7268	79.91
Llama2Paper	0.7252	0.8199	88.45	0.6148	0.8713	70.56	0.5823	0.6661	87.42
HistoryOfAlexnet	0.7280	0.7657	95.08	0.6968	0.8330	83.65	0.6406	0.7073	90.57
BlockchainSolana	0.8122	0.8627	94.15	0.7422	0.7450	99.62	0.5870	0.5872	99.97
LLM Survey paper	0.8061	0.8121	99.26	0.6355	0.6438	98.71	0.5123	0.5148	99.51
Average	94.52 %			88.05 %			91.48 %		

Table 1: Comparison of Agent vs RAG metrics across different datasets, including Attainment Scores (%). Averages are shown in the final row.

System Configuration	Answer Correctness (%)
Traditional RAG	24.24
Agent (3 run Average)	32.71
Agent (Run 4)*	39.64

Table 2: Comparison of Agent vs RAG answer correctness (%) across a subset of the FinanceBench dataset. *Run 4 removed Johnson & Johnson reports from the dataset as the reference files were made unavailable publicly mid-experimentation.

Agentic keyword search vs. Claude computer use

Due to the recent popularity of *Computer Use* capabilities introduced via beta features in Anthropic Claude Sonnet 3.5 v2 model (Anthropic 2024) and more recently with OpenAI’s operator (OpenAI 2025), we compared our method with this new capability that allows direct interaction with computer systems through shell commands and file operations, also offering a way to analyze documents and retrieve information without relying on vector databases (Hu et al. 2024a; Lála et al. 2023). By using the Anthropic Claude Sonnet 3.5 v2 model, we created an agent that follows a structured workflow to answer questions based on access to the raw PDF datasets.

We observe that the agent typically first opens the PDF in Firefox browser, uses keyboard shortcuts (*ctrl + F*) for precise term location, and captures screenshots of the relevant content before answering. We tested the Computer Use agent on seven representative questions from the FinanceBench dataset.(Islam et al. 2023) and manually evaluated results . The agent consistently produced accurate and comprehensive answers. However, the approach faces certain operational challenges, common with our original agent, including occasional failures between API calls. Side-by-side comparisons of our agent and computer use showed effectively similar results (see appendix), though we highlight that our agent approach is far simpler to set up and reproduce.

Conclusion

Our study demonstrated that agentic keyword search approaches can achieve comparable performance to traditional vector database RAG systems in document-based question answering tasks. Our experiments across diverse datasets

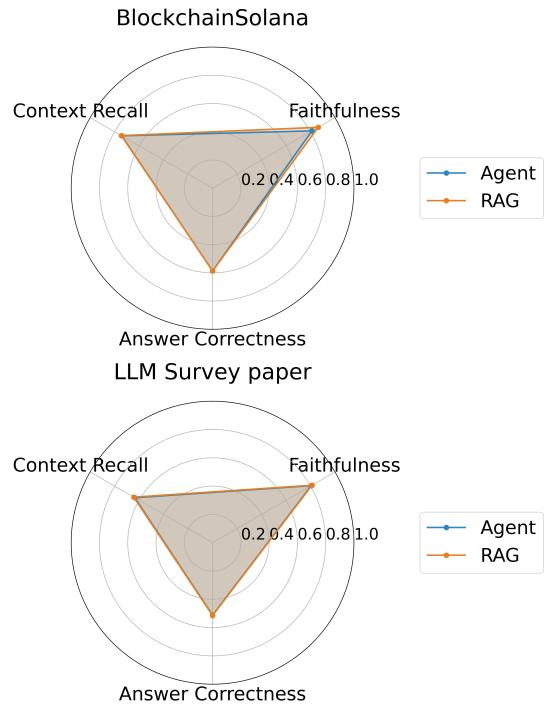


Figure 2: Coverage comparison of Tool-Augmented Agent vs RAG metrics across the BlockchainSolana and LLM Survey Paper datasets

showed that tool-augmented LLM agents using simple keyword search tools can attain over 90% of the performance metrics of vector-based RAG implementations without the need for maintaining standing vector databases. By leveraging existing command-line tools and LLM reasoning capabilities, this method provides a robust alternative to traditional RAG systems, particularly in scenarios where information evolves rapidly or where resource constraints limit the feasibility of maintaining extensive vector databases.

However, several limitations were identified, including performance degradation with large documents, restricted multimedia handling, and context window constraints. The keyword search approach shows limitations in capturing contextual nuances, although partially mitigated through iterative refinement and the agent’s inherent semantic reasoning capabilities. The current implementation also strug-

gles with ambiguous queries and lacks long-term knowledge retention. Furthermore, privacy protections and ethical implications of automated retrieval systems require further investigation, including the implementation of appropriate guardrails such as data access controls and content filtering. Future research will focus on developing more automated, generalizable search strategies that can be useful across document types.

References

Alom, M. Z.; Taha, T. M.; Yakopcic, C.; Westberg, S.; Sidiqe, P.; Nasrin, M. S.; Van Esen, B. C.; Awwal, A. A. S.; and Asari, V. K. 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

Anthropic. 2024. Introducing Computer Use, a New Claude 3.5 Sonnet, and Claude 3.5 Haiku. Accessed: 2024-11-02.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17754–17762.

Chen, Y.; Yoon, J.; Sachan, D.; Wang, Q.; Cohen-Addad, V.; Bateni, M.; Lee, C.-Y.; and Pfister, T. 2024b. Re-Invoke: Tool Invocation Rewriting for Zero-Shot Tool Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4705–4726.

Dam, S. K.; Hong, C. S.; Qiao, Y.; and Zhang, C. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Es, S.; James, J.; Espinosa-Anke, L.; and Schockaert, S. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023a. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; Xiong, D.; et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Gur, I.; Furuta, H.; Huang, A.; Safdari, M.; Matsuo, Y.; Eck, D.; and Faust, A. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hambarde, K. A.; and Proenca, H. 2023. Information retrieval: recent advances and beyond. *IEEE Access*.

Hong, S.; Lin, Y.; Liu, B.; Liu, B.; Wu, B.; Zhang, C.; Wei, C.; Li, D.; Chen, J.; Zhang, J.; et al. 2024. Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679*.

Hu, S.; Ouyang, M.; Gao, D.; and Shou, M. Z. 2024a. The Dawn of GUI Agent: A Preliminary Case Study with Claude 3.5 Computer Use. *arXiv:2411.10323*.

Hu, X.; Ru, D.; Qiu, L.; Guo, Q.; Zhang, T.; Xu, Y.; Luo, Y.; Liu, P.; Zhang, Y.; and Zhang, Z. 2024b. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. *arXiv preprint arXiv:2405.14486*.

Inaba, T.; Kiyomaru, H.; Cheng, F.; and Kurohashi, S. 2023. MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1522–1532.

Islam, P.; Kannappan, A.; Kiela, D.; Qian, R.; Scherrer, N.; and Vidgen, B. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Izacard, G.; and Grave, É. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

Kamradt, G. 2023. LLMTTest-NeedleInAHaystack: Doing Simple Retrieval from LLM Models. <https://github.com/gkamradt/LLMTTest-NeedleInAHaystack>. [Online; accessed 20-March-2024].

Kumar, V.; Srivastava, P.; Dwivedi, A.; Budhiraja, I.; Ghosh, D.; Goyal, V.; and Arora, R. 2023. Large-Language-Models (LLM)-Based AI Chatbots: Architecture, In-Depth Analysis and Their Performance Evaluation. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, 237–249. Springer.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodrigues, S. G.; and White, A. D. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, X.; Wang, X.; Kong, T.; Zheng, J.; and Luo, M. 2021. From bitcoin to solana—innovating blockchain towards enterprise applications. In *International Conference on Blockchain*, 74–100. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Llamahub. 2024. LlamaHub. <https://llamahub.ai/>. Accessed: 2024-11-16.

Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

OpenAI. 2025. Introducing Operator. *OpenAI*.

Pang, R. Y.; Parrish, A.; Joshi, N.; Nangia, N.; Phang, J.; Chen, A.; Padmakumar, V.; Ma, J.; Thompson, J.; He, H.; et al. 2022. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Parisi, A.; Zhao, Y.; and Fiedel, N. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.

Prince, M. H.; Chan, H.; Vriza, A.; Zhou, T.; Sastry, V. K.; Luo, Y.; Dearing, M. T.; Harder, R. J.; Vasudevan, R. K.; and Cherukara, M. J. 2024. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Computational Materials*, 10(1): 251.

Qu, C.; Dai, S.; Wei, X.; Cai, H.; Wang, S.; Yin, D.; Xu, J.; and Wen, J.-R. 2024. Tool Learning with Large Language Models: A Survey. *arXiv preprint arXiv:2405.17935*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rakin, S.; Shibly, M. A.; Hossain, Z. M.; Khan, Z.; and Akbar, M. M. 2024. Leveraging the Domain Adaptation of Retrieval Augmented Generation Models for Question Answering and Reducing Hallucination. *arXiv preprint arXiv:2410.17783*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

van Schaik, T. A.; and Pugh, B. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2832–2836.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Jiayang, C.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023b. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wu, Q.; Liu, W.; Luan, J.; and Wang, B. 2024. ToolPlanner: A Tool Augmented LLM for Multi Granularity Instructions with Path Planning and Feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18315–18339.

Xu, B.; Liu, X.; Shen, H.; Han, Z.; Li, Y.; Yue, M.; Peng, Z.; Liu, Y.; Yao, Z.; and Xu, D. 2023. Gentopia. AI: A Collaborative Platform for Tool-Augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 237–245.

Xu, P.; Patwary, M.; Shoeybi, M.; Puri, R.; Fung, P.; Anandkumar, A.; and Catanzaro, B. 2022. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Xu, Y.; Cai, T.; Jiang, J.; and Song, X. 2024. Face4RAG: Factual Consistency Evaluation for Retrieval Augmented Generation in Chinese. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6083–6094.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yu, T.; Xu, A.; and Akkiraju, R. 2024. In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.

Yuan, S.; Song, K.; Chen, J.; Tan, X.; Shen, Y.; Kan, R.; Li, D.; and Yang, D. 2024. Easytool: Enhancing llm-based agents with concise tool instruction. *arXiv preprint arXiv:2401.06201*.

Zhao, S.; Yang, Y.; Wang, Z.; He, Z.; Qiu, L. K.; and Qiu, L. 2024a. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. *arXiv preprint arXiv:2409.14924*.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2024b. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4): 1–60.

Zhao, Y.; Zhang, J.; Chern, I.; Gao, S.; Liu, P.; He, J.; et al. 2024c. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Appendix

Agent terminal tool instructions

We use the standard zero shot agent template from langchain but provide a detailed description of the following tools it can use via the terminal:

```

1      terminal: Run shell commands on this
2          Linux machine to search
3          information in the "files/"
4          folder. The commands to use are:
5
6 4 ``pdfmetadata.sh```
5 -----
6 YOU MUST first print details of pdf
    files in the files/ folder (ALWAYS
    start with this without any changes)
7 # sh pdfmetadata.sh
8
9 This gives you file level metadata that
    is useful to narrow down the search.
    Then use rga or pdfgrep. Action input
    must start with rga or pdfgrep and
    contain the full command.:
10
11 ````rga`````
12 -----
13 A command line tool to search through
    files via keyword searches and regex
    patterns. All files relevant to this
    task are in the files/ folder.
14
15 - To find a search term in specific file
    (use regex pattern)
16 rga 'searchterm\w*' ./files/filename.pdf
17
18 - To search with multiple keyword
    matches across multiple files:
19 rga 'keyword1|keyword2|keyword3' ./files
    /
20
21 - Use -i for case insensitive search.
22
23
24 ````pdfgrep`````
25 -----
26 Another commandline tool specifically
    for search with PDFs. Useful for
    special cases when:

```

```

27
28 - search in a folder with pdfs, across a
    specific page range for one or more
    keywords (-i is case insensitive, -n
    includes page numbers in output, -r
    is recursive search in files folder,
    and -P is perl compatible regex).
    With pdfgrep you must include '()''
    brackets for the pattern
29 pdfgrep -inrP --page-range 1-4 '((
    keyword1|keyword2)' ./files/
30
31 - Search all .pdf files whose names
    begin with foo recursively in the
    current directory:
32 pdfgrep -r --include "foo*.pdf" pattern
33
34 Other tips:
35 *****
36 - If a complex query fails, try a series
    of simpler queries instead.
37 - ALWAYS try to return larger context
    with -C 5 with both rga and pdfgrep
    to get 2 or more lines around the
    returned keyword match.
38 - remember that you MUST do 'sh
    pdfmetadata.sh' first to understand
    what files you are dealing with and
    then continue your search.
39 - your action should always just be "
    terminal" and action input is the
    full command you want to run in the
    terminal
40 *****

```

Example detailed agent run

The table 3 below shows a partial run of how our agent plans and executes commands to reach the final answer. The entire run can be seen in our supplementary material section of our associated github repository.¹ For easier readability, commands run in the terminal corresponding to the original user question are shown in blue. The agent first uses the 'pdfmetadata' shell script to discover details about files existing in a local folder. Then, it uses 'rga' to look for content related to Hyperledgerfabric. Then, it dives deeper into specific pages using the 'pdfgrep' command to finally get the context required for the final answer.

¹<https://github.com/amazon-science/aws-research-science/tree/main/ShellAgent>

Computer Use Agent Interactions

Table 4 below shows example responses from our agent to various questions about American Express's 2022 financial statements, including the context used to derive each answer. These examples demonstrate our Computer Use agent's capabilities, which employs command-line tools and direct file

Step	Agent observation	operations to extract and verify information from PDF documents without relying on vector databases.
Question	What are the three main components in Hyperledger Fabric and what role does each component play in the system?	Which debt securities are registered to trade on a national securities exchange under American Express' name as of 2022?
Action input	sh pdfmetadata.sh	As of 2022, American Express has only its Common Shares (par value \$0.20 per share) registered under Section 12(b) of the Securities Exchange Act to trade on the New York Stock Exchange under the symbol 'AXP'. No debt securities are registered to trade on any national securities exchange.
Observation	-e File Title Author Subject Keywords Creator Producer CreationDate ModDate Tagged Pages Encrypted PageSize FileSize Optimized PDFVersion ./files/Blockchain-Solana.pdf Markus Richter ...	The company explicitly states under 'Securities registered pursuant to Section 12(b) of the Act' only its common shares. While American Express has significant debt securities outstanding, these are not exchange-listed. The company maintains debt ratings from major agencies (Moody's, S&P, and Fitch) for its unsecured debt programs. Their debt securities are primarily issued through private/institutional channels rather than listed exchanges.
Thought	The pdfmetadata output does not seem to contain information directly relevant to the main components of Hyperledger Fabric. I should search through the PDF files using the rga or pdfgrep tools to look for content related to "Hyperledger Fabric components", "Fabric architecture", or similar phrases.	
Action input	rga -i 'hyperledger fabric components fabric architecture' ./files/BlockchainSolana.pdf:Page 14: reflected in the increase in performance and strength on confidentiality. The Hyperledger Fabric architecture is shown in Fig. 9. ./files/BlockchainSolana.pdf:Page 15: Fig. 9. Hyperledger Fabric Architecture	What are the geographies that American Express primarily operates in as of 2022?
Observation	The search results indicate that the PDF file "BlockchainSolana.pdf" contains information about the Hyperledger Fabric architecture and its main components. I should skim through the relevant pages to identify the three main components and their roles.	American Express operates globally with a strong presence in the United States and significant international operations. As of 2022, approximately 78% of their total revenues net of interest expense were generated from U.S. operations, while 22% came from international activities. Their largest international presence outside the United States includes the United Kingdom, Japan, the European Union, Australia, Canada, and Mexico.
Action input	pdfgrep -inrP -page-range 14-16 '(component architecture)' ./files/BlockchainSolana.pdf -C 5	The company operates in approximately 103 countries and territories through network partnerships. They have a regional structure divided into: United States, EMEA (Europe, Middle East and Africa), APAC (Asia Pacific, Australia and New Zealand), and LACC (Latin America, Canada and the Caribbean). International operations are managed through their International Card Services (ICS) segment.
Observation	(Results truncated here. Please refer to supplementary material.)	(Results truncated here. Please refer to supplementary material.)
Final Answer	The three main components of Hyperledger Fabric are: <ol style="list-style-type: none"> 1. Membership - Provides identification services and allows nodes to access the system through membership services. 2. Blockchain - Provides consensus services for the distributed ledger. 3. Chaincode - Refers to the programs or smart contracts that execute on the blockchain network. 	

Table 3: Partial agent run showing how the original question is answered using keyword search tools.

Table 4: Computer Use Agent Query Examples - Part 1