Agentic RAG Redefining Retrieval-Augmented Generation for Adaptive Intelligence

Meethun Panda
Cornell University

Agentic RAG Redefining Retrieval-Augmented Generation for Adaptive Intelligence

Meethun Panda

Associate Partner, Enterprise Technology, Data & AI at BAIN & COMPANY, Dubai, UAE
Jumeirah, Dubai

---***---

ABSTRACT: This paper presents Agentic RAG, a new framework of retrieval augmented generation complemented with autonomous learning. Retrieval Augmented Generation (RAG) is a novel paradigm that unifies the advantages of generative and retrieval systems to improve the quality and the relevance of responses in Natural Language Processing (NLP) tasks. We discuss key innovations, highlight practical applications, discuss challenges, and highlight future research directions, all informed through the necessity for increased retrieval accuracy, scalability, and ethical AI integration.

KEYWORDS: AI, Agentic, Augmented, RAG.

I.INTRODUCTION

The RAG system exploits available external knowledge from a retrieval component (e.g., a database or search engine) as an extension to the raw knowledge passed to the generative model. This integration empowers the system to draw abundant external knowledge with a degree of dynamism, capable of dealing with multiple situation, and various complex questions. RAG systems work in their traditional form where relevant documents or pieces of information are retrieved from external sources and then consumed by a generative model to produce contextual responses.

In domains where the richness of external data can enhance the system's performance, this mechanism has emerged as a promising solution within areas such as those of question answering, dialogue systems, and information retrieval tasks. Although, traditional RAG systems are typically bound by its dependency on the predetermined retrieval process but have difficulties in leveraging new experiences to continuously learn from.

By incorporating agentic AI principles, the RAG framework the way in which retrieval augmented systems operate becomes significantly different. Agentic AIs are systems that can do some of tasks but also perform actions on their own, with goal and with adapting to the future dynamically changing environment. With RAG embedded with agentic behaviour, the system is able to retrieve, generate and adapt in a desired manner without constant human intervention by being driven by specific objectives.

Fig. 1 RAG and Agentic RAG (Dell Technologies Info Hub, 2024)

Incorporating autonomy into the system results in it transitioning from a reactive tool to a more proactive agent, one that learns from its experience and optimizes for performance in real time. An agentic RAG system is not bound to a predetermined retrieval mechanism, but can improve its query refinement, selection of most relevant data and generation of responses that are suited to particular goals and objectives, while growing its knowledge and capability.

The goal of Agentic RAG is to build a system that is not only reactive, but also adaptive and goal oriented. These systems are built to continually improve by learning from their interactions with the environment, and outputs. The goal-oriented aspect is important in that it allows the system to focus the work, select the priority as well to plan the actions and adapt changes in behaviour among others as the system receives the new information.

Such is the power of Agentic RAG that it is especially useful in dynamic and ever-changing fields where the context or the requirement can vary so often. For example, in healthcare, a system that autonomously retrieves medical literature, suggests diagnostics, and adjusts its advice according to feedback from doctors or patients, but ensures that its advice is relevant and always up to date. An agentic RAG could retrieve case law, generate summarisation of case rulings and judicial decisions, and clarify their understanding of legal precedents based on the evolution of the rulings and judicial decisions in a legal system.

An autonomous investment research platform in finance could, based on new market data and historical trends, continuously learn its strategy from past predictions and outcomes, and could themselves adapt their strategies to the new information. What is needed is to create intelligent systems that can learn, adapt, and make efficient decisions in complex, real world situations.

II.COMPONENTS

Agentic RAG (Retrieval-Augmented Generation) is a simple yet effective framework of advanced retrieval techniques combined with autonomous decision making that produces adaptive, goal-oriented intelligence. The use of these components, motivated by reinforcement learning and contextual awareness, makes it possible for the system not only to seek information, but to use it to make intelligent decisions and adapt its behaviour as part of continuous feedback loops. In this section we explore the main parts of Agentic RAG, a powerful tool in creating AI.

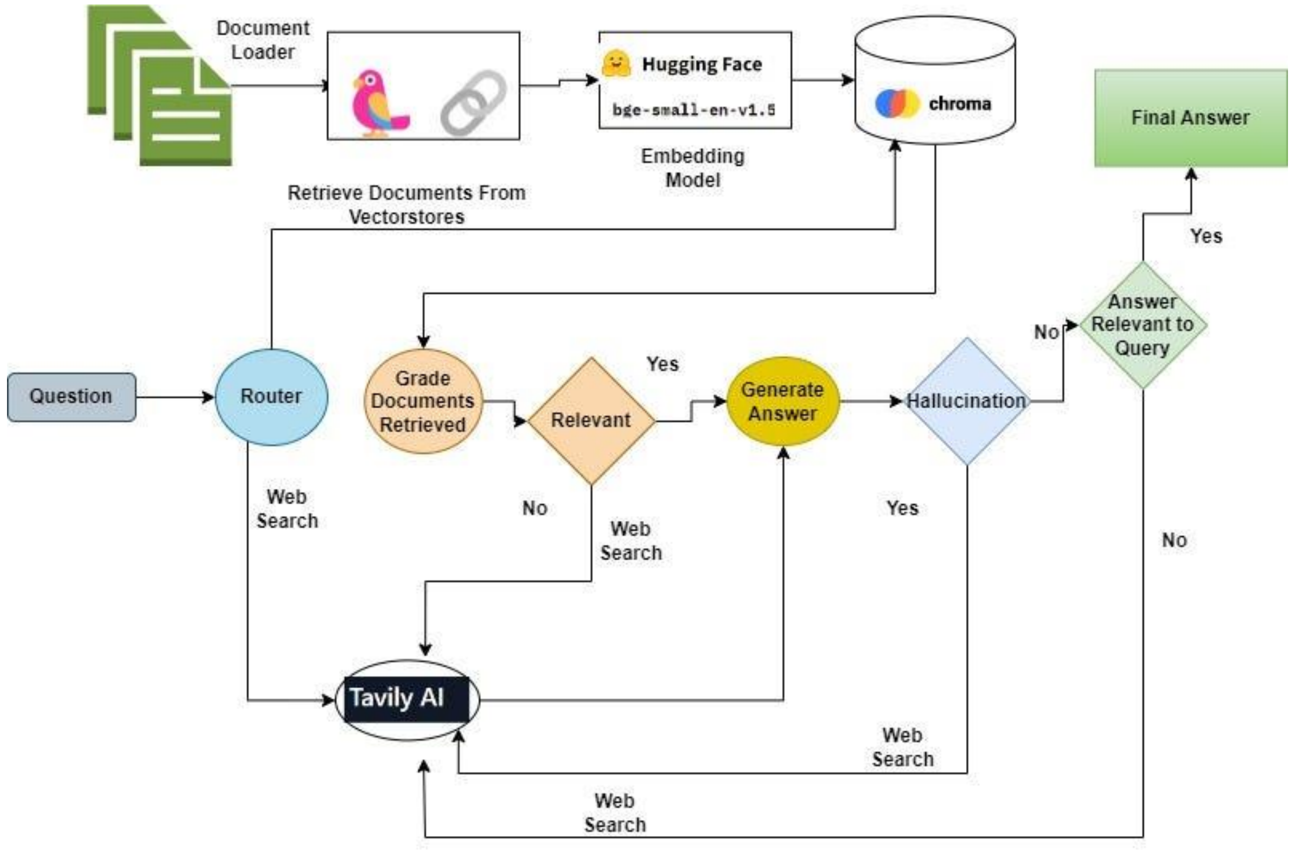


Fig. 2 RAG Flowchart (Medium, 2024)

Retrieval Mechanisms

RL strongly enhances a dynamic retrieval mechanism at the heart of Agentic RAG. A traditional retrieval system takes advantage of defined queries and stable indexation on the new information, however in the case of real time systems these solutions are very limited. On the other hand, instead of choosing an offline policy that does not store the user feedback, Agentic RAG solves the retrieval process using RL to optimally update the retrieval process in the continuous environment.

Where it's possible something similar to the supportive curriculum is applied, making it possible for the system to learn from experience, and refine its approach to retrieve relevant information from the external environment as well as from the actions of that system, without any need for manual intervention. The system can dynamically tweak the parameters governing the retrieval process to produce more effective queries, to favour some type of information over others, to learn to filter out unhelpful or inferior material.

RL driven retrieval mechanism runs in cyclical form. Initially, the system uses simple retrieval strategies; however, as it encounters its environment and receives rewards or punishment for returned information, it learns to change its strategy. As inputs change the system continues to converge towards retrieving the most pertinent data.

For applications which need to retrieve recent information or responses tailored to the context of the query, this adaptive retrieval process is highly valued. Consider healthcare domain, where a lot of medical knowledge changes continuously, the system should be learning and learning continuously how to retrieve the latest research, or case studies, if one knows about a certain diagnosis. In the legal domain the system should be able to search of relevant case law not just from any point in time, but also from within the context of the latest interpretations and rulings of the law.

Decision-Making Loops

After relevant data retrieval, Agentic RAG uses context aware decision-making loop to decide how to process and use the information. The system's goals and priorities, which are continually refined under the context of each query and the emergent mission goals of the task at hand, guide these decision-making loops. The context-aware loops give the involvement of the system in evaluating the harvested data, determining the relevance status, making the decision of creating the best responses that are consistent with the stated mission.

With context agentic RAG allows your system to make decisions in a wide range of scenarios. For example, for legal research, the system must learn if we want to know the general sense of a legal principle or the way this principle is applied in the particular case. For instance, in a healthcare scenario the system might have to choose whether to fulfil the most recent medical guidelines or think through a bigger picture of the patient's medical history and symptoms. The system does not make these decisions blindly, they are dynamically changed by the context and outcomes to achieve always actions consistent with the most relevant and most immediate goals.

Agentic RAG integrates context aware decision making to choose amongst its responses in such a way as to bias outcomes, to achieve long term goals. Feedback loops evaluated continuously by the system as to the effectiveness of their decisions and adjusting them as applicable to gain the best service for themselves. The system's performance and the appropriateness of the produced content in given context provide information to inform these feedback loops for more delicate understanding about how to act in further interactions.

Architecture of RAG

An analysis of an Agentic RAG system architecture is laid out to seamlessly combine the retrieval and generation components and to provide decision making skill with reinforcement learning and context awareness abilities built into the architecture. It bases on numerous interconnected layers, allowing dynamic retrieval, decision making and iterative learning.

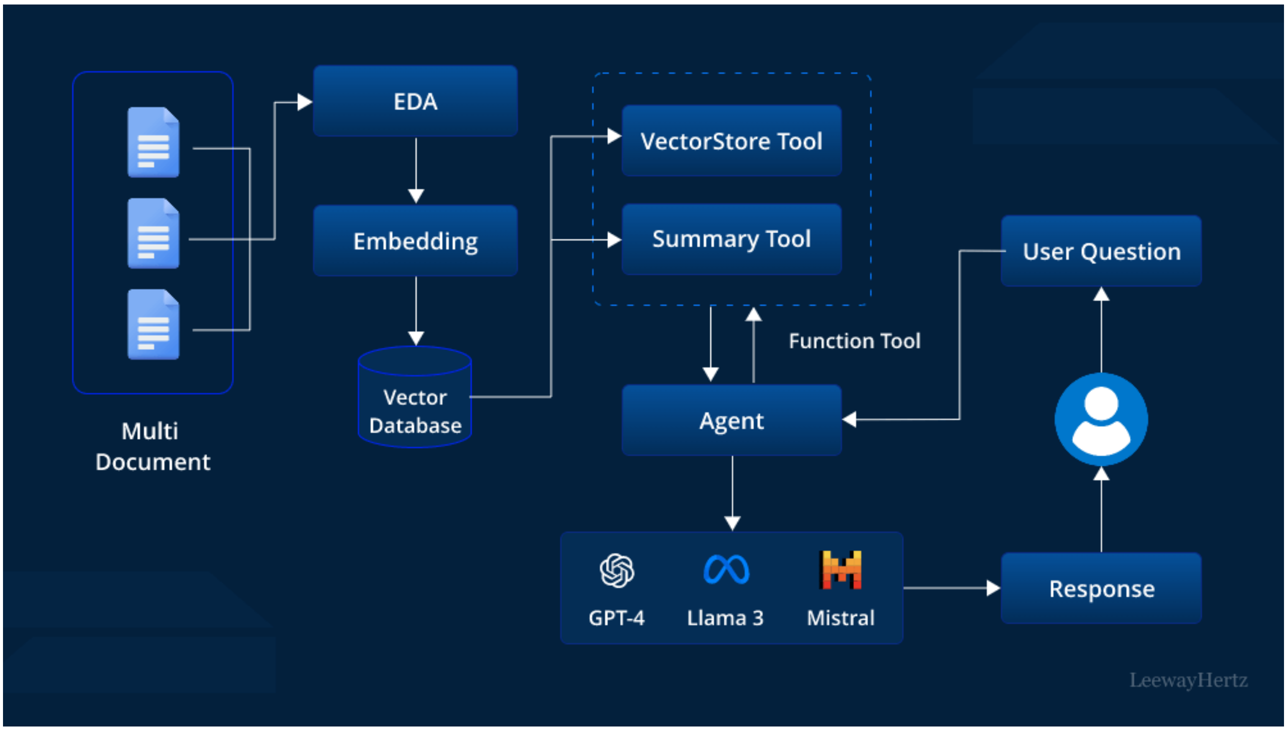


Fig. 3 RAG Architecture (LeewayHertz, n.d.)

The first layer of the retrieval system employs traditional search techniques and self-guided query creation as it interfaces with an external knowledge base or corpus of data. The RL-driven mechanism is useful to this layer as it continuously adapts query strategy according to retrieval performance. The retrieved data then feeds into the next layer where the making of decision takes place.

After retrieving data, the system's decision-making layer then evaluates retrieved data with the task at hand. It uses this information, then prioritizes it according to its relevance to the system's goals. In this case, the system gets involved with the context aware loops, where the system continues to refine its decision making and make further decisions from the feedback of previous actions. At this stage the system has shown the agentic principle, in autonomous fashion, deciding how to use the retrieved information to produce contextually relevant and goal-oriented responses.

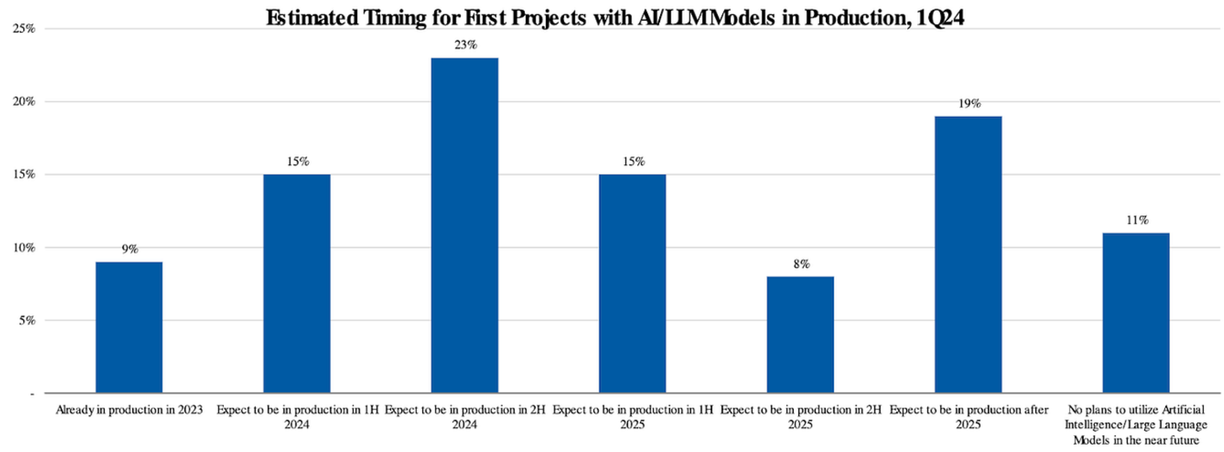
The generation layer takes the retrieved data and uses its context aware decisions to generate a response. This can be text generation, decision making suggestions or actions on the outside of the system's environment. Importantly, this layer is iterative and feedback from generated content is fed into future, future retrieval and decision-making strategies. It gives a system a natural cyclical nature, something which will ensure that the system can continuously adjust itself, refine its strategies, and continuously get better at responding.

These components come together to make an Agentic RAG architecture that can retrieve, generate, and adapt in real time independently. This architecture is flexible enough such that it can be adapted based on the required domain or application specifically and it includes all of the required components for iterative learning and continuous improvement. The net effect is a model that can not only produce high quality responses, but learns from its environment to adapt its behaviour in more promising ways no matter the state of the world.

III.INNOVATIONS

Self-Directed Query

The most important element of Agentic RAG is the change from fixed retrieval to the author generation of the self. Traditional RAG systems are dependent on predefined formulating mechanisms for queries which can be formulated from pre-defined inputs data and external knowledge bases. These systems use fixed strategies to retrieve information that is most likely relevant.



Source: AlphaWise, Morgan Stanley Research. n=100 (US and EU data)

Fig. 4 RAG Production Graph (Medium, 2024)

static retrieval exhibits such a limitation that it often cannot support fine grained or dynamic contexts. However, unlike present RAG systems, Agentic RAG allows the system to autonomously create queries in real-time. It is driven by this transition powered by reinforcement learning whereby the system gets empowered to individually evaluate the relevance of retrieved data and revise the query generated in response to feedback.

Because Agentic RAG systems are self-directed, they can continually fine tune the retrieval process by self-generating query. Through interaction with the environment, the system elaborates its query generating strategies which are more aligned to specific objectives and contexts. This innovation is just critical for applications where the information is constantly updated or refined in response to new inputs.

If we take the example of the field, in healthcare one such system might have to come up with queries such as querying on the most recent medical research about a patient's condition and medical history. Through the shift away from static query formulation, Agentic RAG systems become better enabled to deliver more personalized, more timely and more accurate information.

Feedback Loop

for iterative learning, the beginning of Agentic RAG contains feedback loop mechanisms. With traditional systems, when the data is fetched and the response is created, there is no mechanism to improve the process. The adaptive nature of Agentic RAG is implemented through continuous feedback loops which are central features of the system.

These loops enable the system to evaluate retrieved and generated content with respect to desired goals and outcomes. On this feedback, the system's performance is refined simultaneously on the retrieval and generation processes.

The iterative learning model is a dynamic learning model where the system is not static but every interaction with the system environment, the system evolves. The feedback mechanism then tells the system what out of date, redundant data it brought back, so that the system can correct future search and data selection types.

This enables the system to learn from the action experience of its previous actions and to adapt its decision-making over time, so that its action becomes better aligned with the goal. This particular process has a lot of merit in dynamic fields such as finance or the law, because the system must be capable of adjusting to new developments, trends or legal precedents at regular times.

Goal Prioritization

Another cornerstone of the innovations within Agentic RAG are in goal prioritization and contextual adaptation. In traditional retrieval systems, when there is little or no consideration given to the context or the long-term goals for a given query, the information retrieval system is perceived to be working well. Agentic RAG takes the next step and is more sophisticated, in that it does not just pull up useful data, it also negotiates how to plan its actions cognitively given the context and overall goals that need to be accomplished.

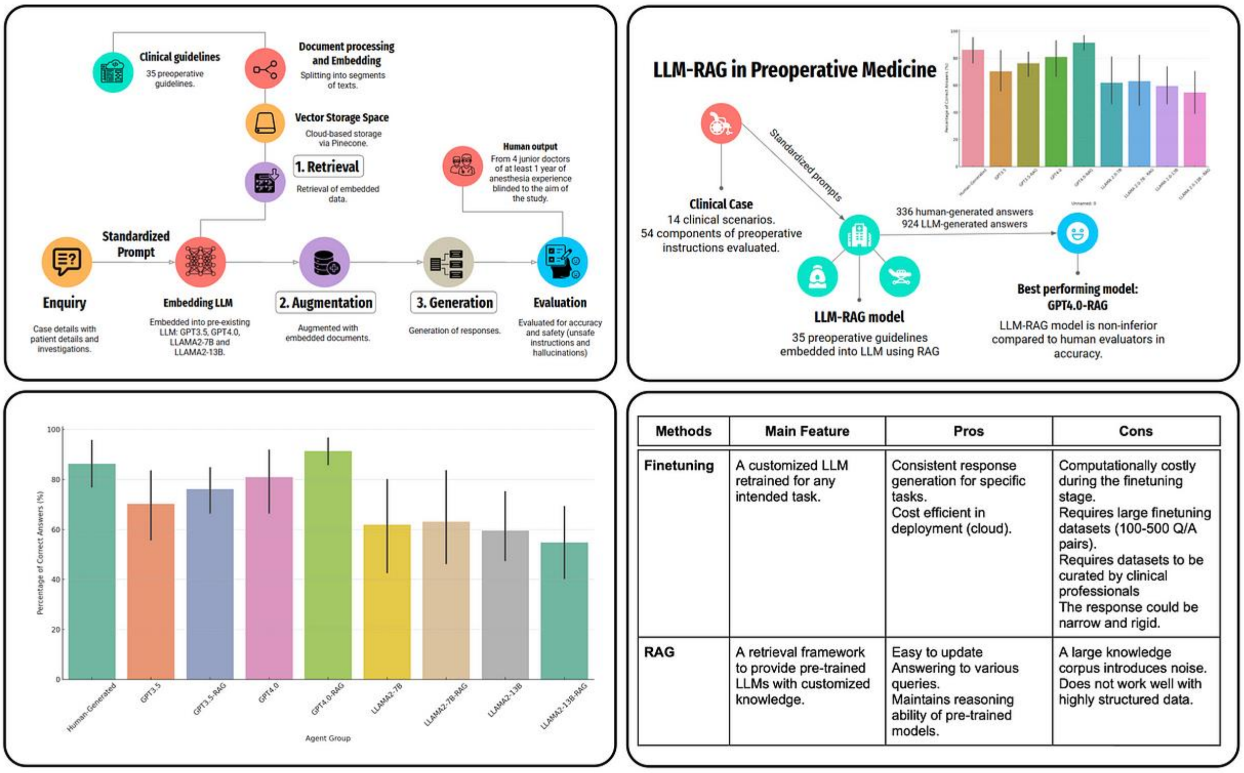
It allows the system to have priorities, to weigh trade-offs, and figure out how to do it in a dynamic manner to always meet their end goals better. By ensuring the system can adapt to changing conditions, or ever evolving requirements, the system can maintain its suitability for the intended purpose. To take an example, in a legal research application, Agentic RAG system would be able to prioritise retrieval of case law relevant to current litigation or specific jurisdiction.

This is further refined by the goal prioritization mechanism such that the system can decide which pieces of information are most important to the task at hand. Agentic RAG systems can integrate goal prioritization with contextual adaptation to tailor their own behavior for particular applications requiring real time medical decisions, financial analysis, and providing personalized legal advice. As a consequence, we obtain more intelligent, adaptive, and goal-oriented systems that are constantly tuned in reaction to their interactions and objectives.

IV.PRACTICAL APPLICATIONS

Healthcare

Adaptive intelligence in medical diagnosis, which we propose can be enabled with agentic RAG, has the potential to revolutionize healthcare. Current diagnostic systems rely solely on traditional static based algorithms or pre-defined decision trees, imposing a constraint that may not be applicable to sophisticated medical problems. On the other hand, Agentic RAG systems are capable of extracting the most relevant medical research, patient histories, and clinical data, in real-time, and adaptively to a range of diagnostic needs.


Clinical guidelines
35 preoperative guidelines.
Document processing and Embedding
Splitting into segments of texts.
Vector Storage Space
Cloud-based storage via Pinecone.
1. Retrieval
Retrieval of embedded data.
Human output
From 4 junior doctors of at least 1 year of anesthesia experience blinded to the aim of the study.
Standardized Prompt
Enquiry
Case details with patient details and investigations.
Embedding LLM
Embedded into pre-existing LLM: GPT3.5, GPT4.0, LLAMA2-7B and LLAMA2-13B.
2. Augmentation
Augmented with embedded documents.
3. Generation
Generation of responses.
Evaluation
Evaluated for accuracy and safety (unsafe instructions and hallucinations)
LLM-RAG in Preoperative Medicine
Standardized prompts
Clinical Case
14 clinical scenarios.
54 components of preoperative instructions evaluated.
336 human-generated answers
924 LLM-generated answers
LLM-RAG model
35 preoperative guidelines embedded into LLM using RAG
Best performing model:
GPT4.0-RAG
LLM-RAG model is non-inferior compared to human evaluators in accuracy.
100
80
60
40
20
0
Percentage of Correct Answers (%)
Human-Generated
GPT3.5
GPT3.5-RAG
GPT4.0
GPT4.0-RAG
LLAMA2-7B
LLAMA2-7B-RAG
LLAMA2-13B
LLAMA2-13B-RAG
Agent Group

Methods	Main Feature	Pros	Cons
Finetuning	A customized LLM retrained for any intended task.	Consistent response generation for specific tasks. Cost efficient in deployment (cloud).	Computationally costly during the finetuning stage. Requires large finetuning datasets (100-500 Q/A pairs). Requires datasets to be curated by clinical professionals The response could be narrow and rigid.
RAG	A retrieval framework to provide pre-trained LLMs with customized knowledge.	Easy to update Answering to various queries. Maintains reasoning ability of pre-trained models.	A large knowledge corpus introduces noise. Does not work well with highly structured data.

Fig. 5 RAG in Medicine (Medium, 2024)

Agentic RAG systems continuously learn from patient data, treatment outcomes and external medical knowledge to refine their diagnostic recommendations, in order to remain current with the state of the art and tailored to individual cases. This dynamical adaptability demonstrates the utility of the system for physicians to improve patient care outcomes and to help them make better decision.

Legal

Since working in this field, we have seen that Agentic RAG can also help in making the process of case law summarization faster and more effective by performing automatically retrieving sound legal precedents, and synthesizing them into concise, context specific summaries. Traditional legal research is time consuming and frequently requires an in-depth knowledge of the complex case law.

Through embedding dynamic retrieval mechanisms and self-directed query generation, Agentic RAG systems can identify quickly the most relevant legal cases toward a query. The systems are able to offer accurate summaries, and adapt their responses to legal practitioners' special requirements. In fact, lawyers can save precious time, improve their research effectiveness, and make better decisions.

Finance

Practically, Agentic RAG can be used in the finance industry, where the intelligent platforms for autonomous investment research development are well suited. Traditional investment research is very reliant on human analysts who scour huge amounts of data and get insights from it. Agentic RAG enables investment platforms with the ability to autonomously retrieve market data via scraping, retrieve and analyse macroeconomic data and company reports, and dynamically generate queries to identify market trends and opportunities.

The system's ability to tackle goals—such as to maximize returns or minimize risks—enables it to adjust to changing market conditions and generates real time, personalized investment recommendations. Agentic RAG Drive platforms learn continuously from market shifts and past investment outcomes for better, more efficient ideas on the investment strategies.

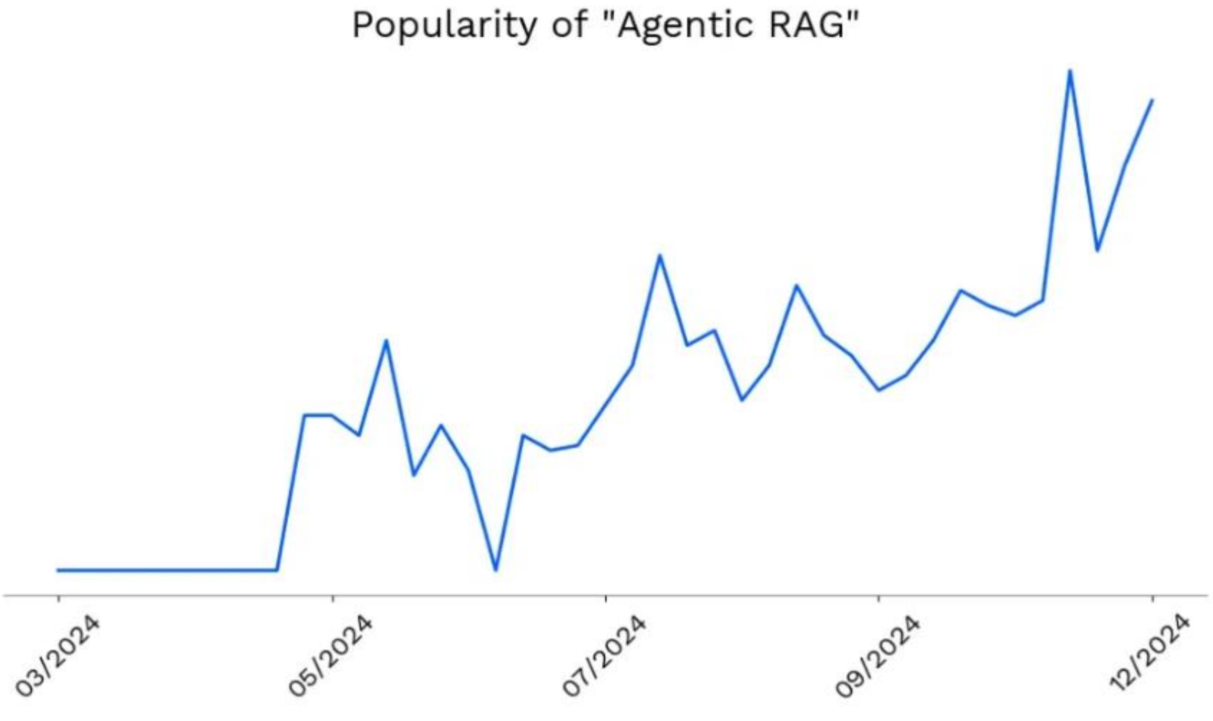


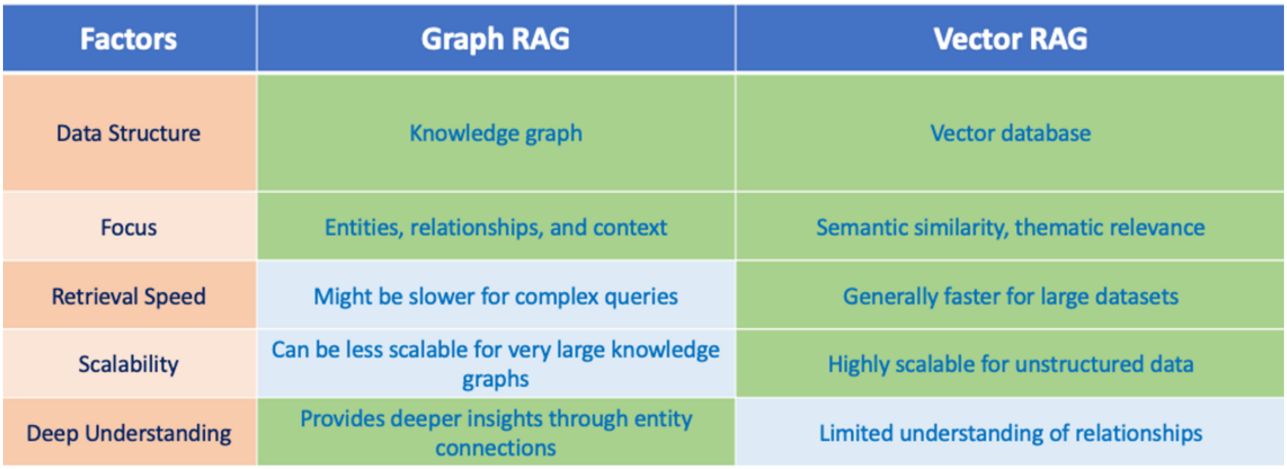
Fig. 6 Agentic RAG Popularity (Research AIMultiple, 2025)

V.CHALLENGES

An important difficulty associated with the adoption of Agentic RAG involves a trade-off between system latency and retrieval accuracy. Since the accuracy of the information to be retrieved is essential to provide good inputs for decision making in Agentic RAG systems, the retrieval and generation processes must be in real time. However, quickly retrieving the most relevant data from vast and complex external knowledge sources is a very big problem.

However, the entire system could be undermined if the retrieval mechanism doesn't efficiently handle what would be the most pertinent information or generates irrelevant queries. The system might need to run multiple retrievals to get its desired data. Latency added in that is also not favourable for real time applications such as healthcare diagnosis or trade financing where you have to get the response instantly. However, it is this key hurdle in the widespread adoption of Agentic RAG systems.

One of the biggest challenges encountered is the fact that systems are limited in their scalability. With the increasing scope of applications of Agentic RAG, the systems must also be able to handle increasing, diverse inputs across different domains. As the volume of data to process increases, the scaling will bottleneck into faster retrieval and generation processing due to scalability issues in the retrieval and generation processes, slowing response times down or decreasing performance.



Factors	Graph RAG	Vector RAG
Data Structure	Knowledge graph	Vector database
Focus	Entities, relationships, and context	Semantic similarity, thematic relevance
Retrieval Speed	Might be slower for complex queries	Generally faster for large datasets
Scalability	Can be less scalable for very large knowledge graphs	Highly scalable for unstructured data
Deep Understanding	Provides deeper insights through entity connections	Limited understanding of relationships

Fig. 7 Difference between Graph RG and Vector RG (Artificial Intelligence in Plain English

VI.FUTURE DIRECTIONS

Future opportunities for investigating Agentic RAG will need the incorporation of multi modal data to generate richer output. These applications need responses that are richer, contextually aware, and can be addressed through these systems as we use a wide spectrum of data sources (text, images, audio). Moreover, studying the more advanced feedback loops would improve continuous learning, so that systems can deal better with dynamic environments. They could research ways to improve the feedback to make it quick, and still remain accurate. Finally, the fields of cross union application and integration of ethical AI principles must be explored, such that these adaptive systems embody in their design, fairness, transparency, and accountability in multiple markets.

VII.CONCLUSION

Combining retrieval augmented generation with agentic, autonomous, and adaptive intelligence is transformative, and retrieval augmented generation (RAG) combined with agentic RAG is one such approach. Agentic RAG systems have the potential to aid decision making in a wide range of domains, including healthcare and finance, by leveraging dynamic retrieval mechanisms, self-directed query generation and iterative learning. Although still facing challenges such as retrieval accuracy and scalability, the improvements in reach and continuous improvement in the integration of multi modal data promise major improvements in adaptive, goal-oriented AI systems.

REFERENCES

[1] Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv preprint arXiv:2501.09136*. https://doi.org/10.48550/arXiv.2501.09136

[2] Bousetouane, F. (2025). Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents. *arXiv preprint arXiv:2501.00881*. https://doi.org/10.48550/arXiv.2501.00881

[3] Bousetouane, F. (2025). Physical AI Agents: Integrating Cognitive Intelligence with Real-World Action. *arXiv preprint arXiv:2501.08944*. https://doi.org/10.48550/arXiv.2501.08944

[4] Finsås, M., & Maksim, J. (2024). *Optimizing RAG Systems for Technical Support with LLM-based Relevance Feedback and Multi-Agent Patterns* (Master's thesis, NTNU). https://hdl.handle.net/11250/3160478

[5] Li, X. (2025, January). A Review of Prominent Paradigms for LLM-Based Agents: Tool Use, Planning (Including RAG), and Feedback Learning. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 9760-9779). https://aclanthology.org/2025.coling-main.652/

[6] Mitra, A., Del Corro, L., Zheng, G., Mahajan, S., Rouhana, D., Codas, A., ... & Awadallah, A. (2024). Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*. https://doi.org/10.48550/arXiv.2407.03502

[7] Li, X. (2024). A Review of Prominent Paradigms for LLM-Based Agents: Tool Use (Including RAG), Planning, and Feedback Learning. *arXiv preprint arXiv:2406.05804*. https://doi.org/10.48550/arXiv.2406.05804

[8] Ravuru, C., Sakhinana, S. S., & Runkana, V. (2024). Agentic retrieval-augmented generation for time series analysis. *arXiv preprint arXiv:2408.14484*. https://doi.org/10.48550/arXiv.2408.14484

[9] Lee, M. C., Zhu, Q., Mavromatis, C., Han, Z., Adeshina, S., Ioannidis, V. N., ... & Faloutsos, C. Agent-G: An Agentic Framework for Graph Retrieval Augmented Generation. https://openreview.net/forum?id=g2C947jjjQ

[10] An, Z., Ding, X., Fu, Y. C., Chu, C. C., Li, Y., & Du, W. (2024). Golden-Retriever: High-Fidelity Agentic Retrieval Augmented Generation for Industrial Knowledge Base. *arXiv preprint arXiv:2408.00798*. https://doi.org/10.48550/arXiv.2408.00798

[11] Das, R., Maheswari, K., Siddiqui, S., Arora, N., Paul, A., Nanshi, J., ... & Sengupta, D. (2024). Improved precision oncology question-answering using agentic LLM. *medRxiv*, 2024-09. https://doi.org/10.1101/2024.09.20.24314076

[12] Khanda, R. (2024). Agentic AI-Driven Technical Troubleshooting for Enterprise Systems: A Novel Weighted Retrieval-Augmented Generation Paradigm. *arXiv preprint arXiv:2412.12006*. https://doi.org/10.48550/arXiv.2412.12006

[13] Jang, J., & Li, W. S. (2024, December). AU-RAG: Agent-based Universal Retrieval Augmented Generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (pp. 2-11). https://doi.org/10.1145/3673791.3698416

[14] Alam, H. M. T., Srivastav, D., Kadir, M. A., & Sonntag, D. (2024). Towards Interpretable Radiology Report Generation via Concept Bottlenecks using a Multi-Agentic RAG. *arXiv preprint arXiv:2412.16086*. https://doi.org/10.48550/arXiv.2412.16086

[15] Li, X., Dong, G., Jin, J., Zhang, Y., Zhou, Y., Zhu, Y., ... & Dou, Z. (2025). Search-o1: Agentic Search-Enhanced Large Reasoning Models. *arXiv preprint arXiv:2501.05366*. https://doi.org/10.48550/arXiv.2501.05366