DLINE JOURNALS

Journal of Digital Information Management

ISSN: 0972-7272

JDIM 2025: 23 (3)

Adaptive and Enhanced Retrieval Augmented Generation (RAG) Systems: A Summarised Survey

Pit Pichappan
Digital Information Research Labs, Chennai 600017. India
pichappan@dirf.org

ABSTRACT

This brief review examines adaptive and enhanced Retrieval-Augmented Generation (RAG) systems, focusing on overcoming the limitations of standard RAG models, such as inefficiency, excessive resource utilisation, and rigid retrieve-then-generate workflows. It highlights advancements like Dynamic RAG and Parametric RAG, which enable context-aware retrieval and parameter-level knowledge integration. The paper emphasises adaptive mechanisms that intelligently decide when and how much to retrieve, improving efficiency and relevance. It also discusses query rewriting, verification, and multimodal extensions to enhance precision. Furthermore, the integration of forecasting latent retrieval methods is introduced, where deep latent dynamics models extract predictable, interpretable components from limited time series data, improving forecasting accuracy. Applications in education, edge computing, and domain-specific contexts are explored, showing reduced hallucinations and better scalability. The review outlines the evolution of RAG toward intelligent, responsive systems using reinforcement learning and hybrid architectures. We identified several potential future research directions, including agentic workflows, multi-modal adaptation, and domain-specific RAG models, which promise more reliable, scalable, and context-aware generative AI applications across various fields.

Subject Categories and Descriptors: [**H.3.3 Information Search and Retrieval]:** Retrieval models; [**I.2.7 Natural Language Processing**] Text analysis

General Terms: RAG, AI-based retrieval, RAG, GAI, Retrieval Augmented Generation

Keywords: Adaptive Retrieval-augmented generation, RAG, Generative AI, Standard RAG, Forecasting Latent Retrieval

Received: 24 June 2025, Revised 4 July 2025, Accepted 20 July 2025

Review Metrics: 0/6, Review Score: 4.9, Inter-reviewer consistency: 79.4%

DOI: https://doi.org/10.6025/jdim/2025/23/3/190-195

1. Introduction

Users often face challenges in formulating effective queries for retrieval tasks, and this issue persists despite several studies addressing it. To aid in query writing and rewriting, the system automatically refines user queries, thereby enhancing retrieval effectiveness. One is interactive retrieval, where users engage with the system to refine their queries, ensuring proper representation. Furthermore, embeddings-based retrieval becomes possible with the aid of learned representations.

Generative AI retrieval is an innovative disruption in information retrieval, leveraging large language models (LLMs) to create a ranked list of document identifiers for a given query. It enhances the retrieval process efficiency by replacing the large external index with complex parameters. These frameworks detect the relationship between queries and document identifiers, which have many inherent issues in finding the relevance between queries and documents. Several improved models have been proposed to address this issue and ensure high precision in results. [1,2,3]

In resource-limited edge environments, the RAG model is known to reduce LLM hallucinations and integrate external knowledge into LLMs without requiring fine-tuning; however, its retrieval process can introduce significant latency. [4,5,6] Among the many inclusive features and characteristics, the decisive component in RAGs remains unexplored.

Enhanced large language models utilise a modular hybrid retrieval and fine-tuned generation approach for adaptive and personalised applications across a wide range of domains. Enhanced with real-time, engaging, context-based learning content, it increases engagement and scalability, thereby reducing the workload of educators. It accounts for issues of accuracy, multimodality, and ethics, which is a significant change in AI education [7,8,9].

2. Standard RAGs

RAGs use external knowledge for query reformulation and enhancement. Typically, RAG systems follow a fixed retrieve-then-generate workflow and rely on in-context knowledge integration. This approach, however, is inefficient for complex tasks that require the integration of multiple chains of reasoning for flexible information retrieval and deeper integration of external knowledge. The two rapidly expanding areas of research include the models of Dynamic RAG and Parametric RAG. Dynamic RAG investigates how LLMs can determine during generation what and when to fetch, allowing them to adjust in real-time to changing information needs. Gao, et al. [23] identified three paradigms of RAG, namely, the Naive RAG, Advanced RAG, and Modular RAG. On the other hand, Parametric RAG aims to refine how retrieved knowledge is integrated, shifting from input-level to parameter-level knowledge injection for enhanced efficiency and effectiveness. [8] The objective of RAG is to enhance the understanding and generative capabilities of vision models by combining internal model knowledge with reliable external knowledge sources, rather than relying solely on internal model knowledge. [14]

In the early stages of RAG production, standard models emerged, which were later found to possess inherent limitations, including the excessive use of computing resources and time, as well as unwarranted reliance on external knowledge for common queries. When queries are effective, the standard models unnecessarily modify them, which may lead to improper concept representation in queries. Moreover, these standard models fail to reflect the domain-specific query requirements.

While studies indicate an improvement in LLM reasoning through RAG augmented with Knowledge Graphs, the deployment of said solution faces challenges due to ever-changing user demands and environments. To address the stated issues, Tang et al proposed a RAG framework modified through the use of Multi-objective Multi-Armed Bandit techniques that integrate various retrieval methods. Such a framework can evolve and maintain balance between performance, and responsiveness in non-stationary contexts. [9]

3. Adaptive and Enhanced RAG models

The Adaptive RAG (Retrieval-augmented generation) is an advanced version of all types of RAG systems, wherein the system disruptively decides when and how much information needs to be retrieved based on the query from the context or the confidence level of the internal model, rather than resorting to the usual static retrieval approach for every single query.

In essence, the underlying philosophy is to make retrieval intelligent and aware of context so that it can be more efficient in providing the information, precise in delivering relevant information, and tuned to precisely what is needed for each generation task or shortage.

To understand the adaptive and enhanced RAG models, we generate a workflow with proper descriptions.

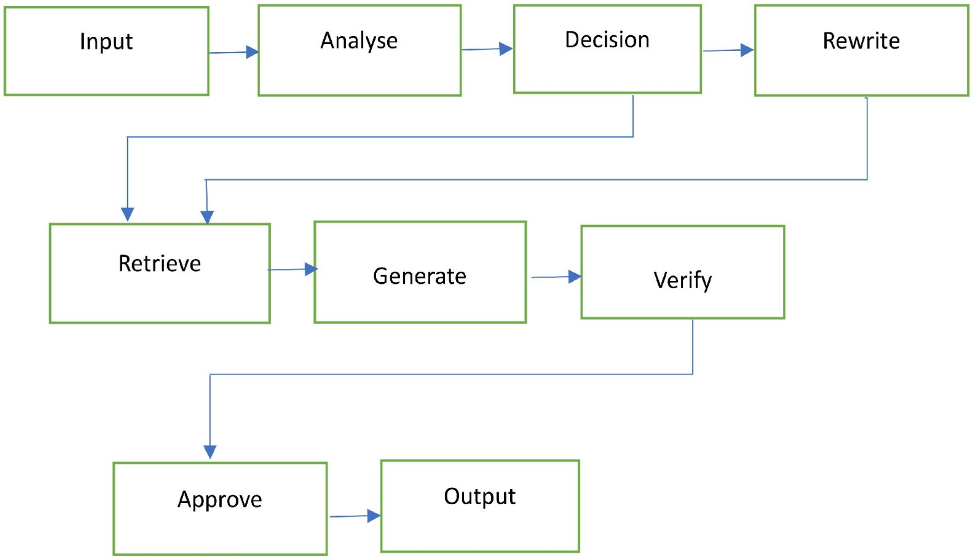

Figure 1. Illustration of the workflow of the Adaptive RAG model

We present an illustration to explain how the RAG models work.

3.1 Illustration of how Enhanced and Adaptive RAG Models Work

Ineffective queries in Information Retrieval are characterised by too broad expressions, ambiguity, lack of context, vagueness, lack of specificity, poor precision, and unclear intent.

Input**:** "recent studies about climate change effects on electric cars"

Decision**:** 'Recent studies' is ambiguous. The intersection of two domains is inconsistent. Boolean operators are not properly fixed, leading to structural complexity.

Query Rewrite**:** "Studies on the Effect of low carbon emissions of electric cars in the last two years."

Retrieve: Pulls a few recent research papers or summaries, particularly generated in the last two years.

Generate**:** Draft summary.

Verify: Detects mention of "low carbon emissions of electric cars" → also detects unclear and imprecise→ Elimination of results on it- Lead to final retrieval.

Approve: Matching of retrieved content with query.

Final Output**:** Well-grounded, detailed summary with conceptual, semantic and technical description.

4. Assessment of RAG Models

Assessment metrics for the Adaptive and improved RAGs are necessary to confirm their efficiency and enhance the models. The existing evaluation techniques of RAG, including independent evaluation and end-to-end evaluation [12], EM for question-answering assignments [13], UniEval and E-F1 for summarisation tasks [14], and BLEU for machine translation [15], are analysed to determine their appropriateness for evaluating Adaptive RAG.

In an early evaluation task, the RAG models were found to produce language that is more precise, varied, and accurate compared to a leading parametric-only seq2seq baseline. [16]

5. Forecasting Latent Retrieval

Recently, Forecasting Latent Retrieval models have been proposed, where retrieval mechanisms that utilise latent representations have the potential to improve predictive models. However, these models are not yet proposed with clear frameworks, and hence, they are related to the research threads that align with this concept. The deep latent dynamics models infer meaningful, predictable latent factors from limited temporal data. By emphasising predictability, sufficiency, and identifiability, these models decompose time series into independent, interpretable components, enhancing forecasting accuracy and long-term efficiency. [10]

The forecasting framework for short, high-dimensional time series forecasting combines the low-rank temporal matrix factorisation approach with the optimal model selection via cross-validation on the latent factors. The forecasting of the components outperforms univariate methods, thereby enhancing performance, especially with limited data. [11] The current research on latent forecasting relies on additional knowledge [17], [18], [19], [20] and modelling relations of latent representations [21], [22].

6. Conclusions and Future Directions

The enhanced and adaptive RAG models offer promising approaches to learn retrieval as a strategy with the assistance of reinforcement learning. In the adaptive models, the additional/bespoke modifications follow the feedback from user interactions. The Multi-modal adaptation, including the option to pull images or tables, is being investigated, which may yield an interesting outcome. For the agentic workflows, RAG serves as one of the several tools. Once these models attain maturity, the research on RAG has the potential to evolve into domain-specific models, where specific fields, such as medicine and business, have unique requirements.

Rather than focusing on what is to be retrieved, when, and how much, retrieval-augmented generation serves a greater purpose and leveraging adaptable retrieval mechanisms to function effectively in specific contexts. Such an approach is pivotal to intelligent and responsive knowledge-augmented generation, marking the shift away from strict input-output pipelines. It is a critical piece in constructing dependable, scalable, and user-friendly applications built on LLM.

References

[1] Lu, P., Dong, X., Zhou, Y., Cheng, L., Yuan, C., Mo, L. (2025). DOGR: Leveraging Document-Oriented Contra/stive Learning in Generative Retrieval, *In*: Proceedings of the AAAI Conference on Artificial Intelligence, 39 (23), 24732-24740

[2] Lee, Dohyeon., Kim, Jongyoon., Kim, Jihyuk., Hwang., Seung-won., Park, Joonsuk. (2025). *tRAG: Term-level retrieval-augmented generation for domain-adaptive retrieval*. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6566–6578).

[3] Zheng, Xu., Weng, Ziqiao., Lyu, Yuanhuiyi., Jiang, Lutao., Xue, Haiwei., Ren, Bin., Paudel, Danda., Sebe, Nicu., Van Gool, Luc., Hu, Xuming. (2025). *Retrieval-augmented generation and understanding in vision: A survey and new outlook*. arXiv:2503.18016v1 [cs.CV] 23 Mar 2025

[4] Qin, Qitao., Luo, Yucong., Lu, Yihang., Chu, Zhibo., Liu, Xiaoman., Meng, Xianwei. (2025). *Towards adaptive memory-based optimization for enhanced retrieval-augmented generation*. arXiv:2504.05312v3 - March 2025.

[5] Ouyang, T., et al. (2025). *AdaRAG: Adaptive optimization for retrieval augmented generation with multilevel retrievers at the edge*. In *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications* (pp. 1–10). doi.org/10.1109/INFOCOM55648.2025.11044685

[6] Li, Siran., Stenzel, Linus., Eickhoff, Carsten., Bahrainian, Seyed Ali. (2025). *Enhancing retrieval-augmented generation: A study of best practices*. arXiv:2501.07391v1 [cs.CL] 13 Jan 2025].

[7] Shan, R. (2025). *LearnRAG: Implementing retrieval-augmented generation for adaptive learning systems*. In: *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 224–229).

[8] Su, Weihang., Ai, Qingyao., Zhan, Jingtao., Dong, Qian., Liu, Yiqun. (2025). *Dynamic and parametric retrieval-augmented generation*. In *Proceedings of SIGIR '25*, Padua, Italy. Association for Computing Machinery.

[9] Tang, Xiaqiang, Li, Jian, Du, Nan, & Xie, Sihong. (2025). *Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs*. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 12). AAAI-25 Technical Tracks 12.

[10] Cheng, Sibo., Prentice, I. Colin, Huang., Yuhan, Huang., Jin, Yufang., Guo., Yi-Ke., Arcucci, Rossella. (2022). *Data-driven surrogate model with latent data assimilation: Application to wildfire forecasting. Journal of Computational Physics, 464*, 111302.

[11] Charotia, Himanshi., Garg, Abhishek., Dhama, Gaurav., Maheshwari, Naman. (2021). *Optimal latent space forecasting for large collections of short time series using temporal matrix factorization: A preprint. arXiv preprint arXiv:2112.08052v1.*] 15 Dec 2021.

[12] Liu, Nelson F., Lin, Kevin, Hewitt, John, Paranjape, Ashwin, Bevilacqua, Michele, Petroni, Fabio, & Liang, Percy. (2023). Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172.

[13] Borgeaud, Sebastian., Mensch, Arthur., Hoffmann, Jordan., Cai, Trevor., Rutherford, Eliza., Millican, Katie., Van Den Driessche, George Bm., Lespiau, Jean-Baptiste., Damoc, Bogdan., Clark, Aidan. (2022). Improving language models by retrieving from trillions of tokens. In International Conference on Machine Learning (pp. 2206–2240). Proceedings of Machine Learning Research (PMLR).

[14] Jiang, Zhengbao., Xu, Frank F., Gao, Luyu., Sun, Zhiqing., Liu, Qian., Dwivedi-Yu., Jane, Yang., Yiming, Callan, Jamie., Neubig, Graham. (2023). Active retrieval augmented generation. arXiv preprint arXiv:2305.06983.

[15] Zhong, Zexuan., Lei, Tao., Chen, Danqi. (2022). Training language models with memory augmentation. arXiv preprint arXiv:2205.12674.

[16] Lewis, Patrick., Perez, Ethan., Piktus, Aleksandra., Petroni, Fabio., Karpukhin, Vladimir., Goyal, Naman., Küttler, Heinrich., Lewis, Mike., Yih, Wen-tau., Rocktäschel, Tim., Riedel, Sebastian., Kiela, Douwe. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *In:* Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

[17] Xu, Y., Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. *In:* Proceedings *of the Association for Computational Linguistics* (pp. 1970–1979).

[18] Shi, L., Teng, Z., Wang, L., Zhang, Y., Binder, A. (2019). DeepClue: Visual interpretation of text-based deep stock prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31(6), 1094–1108.

[19] Cheng, D., Yang, F., Xiang, S., Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition,* 121, 108218.

[20] Xu, W., Liu, W., Wang, L., Xia, Y., Bian, J., Yin, J., Liu, T. (2021). *HIST: A graph-based framework for stock trend forecasting via mining* concept-oriented shared information. arXiv preprint arXiv:2110.13716.