Volume (1) No (7), 2025
Journal of Global Research in Multidisciplinary Studies (JGRMS)
Review Paper/Research Paper
Available online at https://saanvipublications.com/journals/index.php/jgrms/index

A Review on Large Language Models (LLMs) with RAG: Research Challenges and Opportunities

Dr. Parth Gautam
Associate Professor
Department of Computer Sciences and Applications
Mandsaur University
Mandsaur
parth.gautam@meu.edu.in

***Abstract*—Large Language Models (LLMs) are now essential to contemporary applications in natural language processing, and retrieval-augmented generation (RAG) approaches have emerged in response to the need for more factual, context-aware, and trustworthy outputs. In order to overcome constraints like hallucinations and static knowledge bounds, RAG improves generating quality by including external information retrieval into the LLM process. This paper systematically surveys the evolving landscape of RAG systems, highlighting foundational components such as sparse, dense, hybrid, and graph-based retrieval methods, and their impact on semantic richness, scalability, and reasoning capability. Challenges such as hallucination, data privacy, security vulnerabilities, and ethical risks are examined in depth, along with faithfulness issues in output generation. Evaluation frameworks, including both quantitative (e.g., Precision, BLEU, BERT Score) and qualitative metrics, are reviewed alongside emerging benchmarks like RAG-Bench, MIRAGE, and MTRAG. Further, the study identifies key research gaps in adaptive retrieval, sufficient context modeling, and privacy-preserving methods. As RAG continues to evolve, the integration of hybrid neuro-symbolic systems and reinforcement learning-based adaptation presents promising avenues for future research. The survey provides a very detailed insight into the technical progress of RAG, its assessment systems, and deployment issues, and it can be viewed as a guide towards researchers and developers who seek to develop reliable, interpretable, and robust retrieval-augmented generation systems.**

***Keywords*—Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Knowledge Retrieval, Hallucination Mitigation, Evaluation Metrics, Privacy and Security, Hybrid Neuro-Symbolic Systems.**

I. INTRODUCTION

In the development of large language models (LLMs), retrieval-augmented generation (RAG) has emerged as a novel paradigm that gets around the main drawbacks of traditional generative models. As opposed to conventional LLMs, which employ a single set of parameters and, hence, knowledge [1], RAG adds external knowledge retrieval sources to improve the quality of natural language generation. This hybrid model significantly improves factual accuracy, contextual relevance, and domain-specific flexibility.

In RAG, the basic idea is to add dynamically identified content in external sources (e.g. document repositories, knowledge bases, or indexed corpora) [2] to the generative process. Such an integration allows the real-time availability of updated, context-aware and verifiable information, at the time of inference. Consequently, RAG diminishes hallucinations greatly, alleviates using stale training data, and improves performance in highly specific areas such as medicine, law, and finance, in which precision is critical.

Natural language processing (NLP) applications the GPT and other LLMs, have also been used to advance NLP by supporting tasks like content generation, translation, summarization, and question answering. Even with these accomplishments, conventional models tend to have difficulties when handling domain-specific or difficult queries because they lack the capability of grounding in context and maintaining the consistency of facts [3]. RAG conceptualization is used to handle such issues by pairing retrieval modules, usually of a dense or sparse search-based nature, with strong generative transformers to provide more knowledgeable and consistent answers.

The scalability and potential of retrieval-based models have also been further illustrated by newer advances such as the Retrieval-Enhanced Transformer (RETRO) [4]. As an example, RETRO has demonstrated capacity to access stores of knowledge involving trillions of tokens, along with competitively fast performance [5]. Nonetheless, problems like retrieval latency, integration difficulty, and sensitivity to the quality of the retrievers remain in the way of overall efficacy and resiliency of the system.

An in-depth study of RAG techniques in large language models [6]. Different components such as retrievers, document stores, generation strategies are discussed and their contribution to enhancing performance in different NLP tasks [7]. Issues that appear to be unresolved and present possible avenues of improvement of RAG systems in scalable, domain-specific and real-time tasks are also brought out in the discussion.

A. *Structure of the Paper*

The structure of this paper is as follows: Section II presents the fundamentals of Retrieval-Augmented Generation (RAG) in LLMs. Section III discusses core RAG techniques. Section IV outlines recent advancements. Section V highlights evaluation methods and future directions. Section VI reviews related literature. Section VII concludes with key insights and future work.

II. Fundamentals of Retrieval-Augmented Generation in LLMS

In the realm of large language models (LLMs), retrieval-augmented generation (RAG) is a groundbreaking paradigm that enhances accuracy, relevance, and factuality by fusing generating elements with external retrieval systems. RAG leverages modules of retrieval, augmentation of prompts, and advanced generation abilities to eliminate some of the main shortcomings of conventional LLMs, including outdated information and hallucinations [8]. It has been used in a variety of areas such as healthcare, legal, enterprise and multimodal. In contrast to the static models that depend on previously devoted information, RAG can extract the real-time data, which enhances reliability and functionality to a great extent. This questionnaire examines the structure, uses and relative merits of RAG, making special reference to the fact that it is central in development of smart language systems.

A. RAG Architecture and Workflow

One recently created hybrid architecture that attempts to address the drawbacks of pure generative models is called RAG. The foundation of RAG consists of two main parts: (i) a retrieval mechanism that finds pertinent documents or information stored in an external knowledge source, and (ii) a generation module that digests this data and produces a text that is a human-like manner [9]. RAG, which enhances job generating and is a significant advancement in the art of LLMs (see Figure 1). RAG incorporates an additional stage whereby the LLMs search a third-party data source for relevant information before producing text or answering enquiries [10]. RAG's usage of external data sources to create data is one of its drawbacks, but it gets over this by integrating external data retrieval into the appropriate generative process, which increases the accuracy and relevance of the output that is produced.

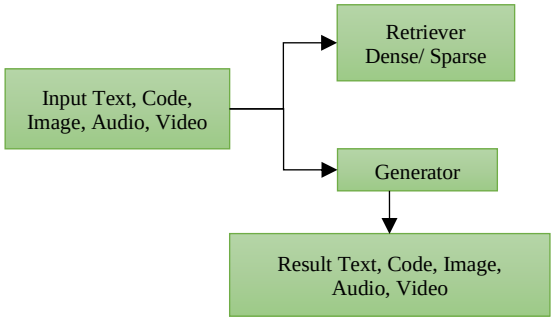

Fig. 1. A generic RAG architecture

A RAG workflow where input (text, code, image, audio, or video) is processed by a retriever to fetch relevant data, which is then used by a generator to produce a final result in the same or different modality. Here are the Key Components of Architecture Workflow of RAG are as follows:

- **Retrieval (Retriever Module):** Takes the input query, and encodes it (dense/sparse embeddings or keywords). Queries the index (e.g. vector DB) to retrieve K document fragments.
- **Augmentation (Prompt Construction):** Some of these passages are prefixed or suffixed to the original query [11]. This forms a prompt situation and is basically delivering live type knowledge to the LLM.
- **Generation (LLM Decoder):** The LLM creates the current output based not just on the user's question but the evidence discovered, making the result less psychedelic and more factual.

B. Key Applications Across Domains

Retrieval-Augmented Generation (RAG) has also come along as a flexible model on which to enhance LLMs with more knowledge relevant to a specific field. Its usages can be found in a variety of spheres, all of which can be enhanced by a higher level of factual accuracy and the decreased number of hallucinations:

- **Healthcare & Biomedicine:** RAG can strengthen the ability of the diagnostic systems and clinical-support systems. That causes more knowledgeable, competent proposals of the way of treatment, interaction, or reports about cases.
- **Law & Compliance:** Hallucinations risks are mitigated by legal assistants created with RAG who can retrieve relevant statutes. According to experts, RAG systems provide greater trust worthiness because they anchor the response to those documents.
- **Enterprise Knowledge Management:** Corporations also incorporate internal record management system documents, such as policies, manuals, knowledge base, into vector stores. A good example is the use of RAG which Workday adopted in order to present policy answers to staff inquiries.
- **Multimodal Applications:** In addition to text, RAG is extended to these in addition to appearance and sound: image-grounded RAG models refer to and merge images or pieces of transcript [12]. Multiple publications are being devoted to such outcomes as Multimodal RAG applied to tasks such as VQA, video-captioning, and medical imagery Q&A.

C. Comparison with Traditional LLM Approaches

Retrieval-Augmented Generation (RAG) has a different source centralization as the traditional LLM pipelines-pure generation, prompt-engineering, and fine-tuning-because it integrates the retrieval elements into language models to generate grounded and up-to-date outputs [13]. The following are the core difference of RAG and Traditional LLM approaches in Table I:

TABLE I. Compare of Traditional vs LLM approach

Aspect	Traditional LLMs	RAG (Retrieval-Augmented Generation)
Real-time Knowledge vs. Static Training	Based only on pre-training data; may produce outdated or incorrect information.	Incorporates real-time, external knowledge sources at inference time for up-to-date information.
Factual Accuracy and Hallucination Reduction	Prone to hallucinations due to reliance on internal memory and lack of citation.	Significantly reduces hallucinations by grounding responses in retrieved documents with source traces.
Retrieval vs. Fine-Tuning	Fine-tuning requires weight updates and is costly; suitable for narrow domains.	Retrieval avoids retraining, supports dynamic knowledge access, and performs well without model updates.
Versus Prompt Engineering	Uses internal knowledge only; prompt tuning is low-cost but limited in scope.	Augments prompts with retrieved content, enabling deeper and more contextually rich responses.

III. Techniques of RAG in Large Language Model

The LLM are improved by RAG approaches, which use external information sources to provide more pertinent and accurate replies [14]. More dependable and contextually aware AI systems are produced by optimizing the retrieval and production processes using advanced RAG methodologies, which go beyond simple approaches.

A. Advanced Retrieval Techniques

Retrieval is the backbone of RAG, and recent techniques focus on enhancing efficiency and relevance:

- **Query Optimization:** Techniques like query expansion (e.g., Multi-Query, Sub-Query, Chain-of-Verification) and transformation (e.g., HyDE Hypothetical Document Embeddings, Step-back Prompting) refine user queries for better retrieval. Query routing, using semantic routers (Semantic Router), directs queries to appropriate data sources based on metadata or semantics.
- **Embedding Models:** Advanced embedding models, such as BGE (BGE Embeddings), Voyage (Voyage Docs), and AngIE, improve dense retrieval. These are evaluated on benchmarks like MTEB (8 tasks, 58 datasets) and C-MTEB (6 tasks, 35 datasets) (MTEB Leaderboard), ensuring robust performance across diverse scenarios.
- **Retrieval Sources and Granularity:** RAG now supports unstructured data (e.g., Wikipedia Dump: Hotpot QA 1st October 2017, DPR 20 December 2018), semi-structured data (PDFs), structured data (Knowledge Graphs, KGs), and LLM-generated content (e.g., SKR, Gen Read, Self mem). Retrieval granularity ranges from tokens to documents, with Knowledge Graphs using entities and sub-graphs, adapted for tasks like item IDs and sentence pairs (RAG Survey).
- **Indexing Optimization:** Strategies include chunking text into fixed sizes (100, 256, 512 tokens), attaching metadata (e.g., page number, timestamp), and hierarchical indexing. Knowledge Graph indexing enhances structured data retrieval, crucial for complex queries (Indexing in RAG).

B. Advances in Generation Techniques

Post-retrieval generation has seen significant enhancements to ensure relevance and coherence:

- **Context Curation:** Reranking methods, both rule-based (e.g., Diversity, Relevance, MRR) and model-based (e.g., BERT series, cohere rerank Cohere Rerank, bge-raranker-large, GPT) [15], prioritize relevant contexts. Context selection and compression, like Long LLM Lingua using SLMs (e.g., GPT-2 Small, LLaMA-7B), and RECOMP with contrastive learning (1 positive, 5 negative samples), reduce noise and enhance input quality (Context Curation).
- **LLM Fine-tuning:** Fine-tuning enhances domain-specific performance, adjusting input/output formats. Techniques like SANTA's tripartite training use contrastive learning for embeddings, while RA-DIT aligns scoring functions using KL divergence. Reinforcement learning aligns models with human or retriever preferences, and distillation from models like GPT-4 improves efficiency (Fine-tuning RAG).

C. Multimodal and Specialized RAG Techniques

RAG is expanding to handle diverse data types and specialized domains:

- **Multimodal RAG:** Retrieval-augmented multimodal language modeling (Multimodal RAG Paper) integrates text, images, and videos, enhancing cross-modal retrieval. RULE (RULE Paper) is a multimodal RAG for medical Vision-Language Models (Med-LVLM), enhancing medical RAG factual correctness.
- **Domain-specific RAG:** RAFT (RAFT Paper) trains to disregard distractor documents, citing relevant sources for PubMed, Hotpot, and Gorilla datasets. HyPA-RAG (HyPA-RAG Paper) adapts for legal/policy contexts (e.g., NYC Local Law 144), enhancing correctness. LA-RAG (LA-RAG Paper) enhances automatic speech recognition (ASR) in LLMs with token-level speech data stores, managing accents in Mandarin and Chinese dialects.

IV. Advances and Enhancements in RAG

The RAG systems in LLM have seen rapid technical advances, leveraging a range of retrieval strategies from sparse and dense textual methods to sophisticated graph-based approaches to provide richer, more factual, and context-aware outputs [16]. These innovations aim to improve reasoning, scalability, and semantic depth, but also introduce new challenges. Persistent issues include hallucinations, where outputs may deviate from source truth, and critical concerns around data privacy, security vulnerabilities, and ethical risks, especially in sensitive domains.

A. Retrieval Techniques: From Textual to Graph-Based Methods

RAG integrates domain-specific information to guarantee factuality and trustworthiness while combining external knowledge with LLMs to enhance task performance. Most RAG systems today employ a variety of retrieval strategies, including the following: more conventional text similarity to the more graph-based techniques, in order to successfully provide relevant context to LLMs. Such approaches differ in terms of scalability, semantic richness and capability to reason. Here are the Textual Retrieval Methods are as follows:

- **Sparse Retrieval:** These are based on keyword search and inverted index. Neural augmentations such as SPLADE v2 employ word level expansions to enhance recall on retrievals.
- **Dense Retrieval with ANN:** The queries and documents is embedded in dense vector spaces. This is possible through nearest neighbor searching which allows semantically rich scaling.
- **Hybrid Retrieval:** A combination of semantic matching (dense) and lexical precision (sparse) in which reranking.

Graph-Based Retrieval

Textual retrieval techniques are able to handle textual reasoning well but lack structural reasoning capabilities, and that is where the Graph RAG comes. Graph-Based Retrieval Techniques includes:

- **Non-parametric:** Searching k-hop path (e.g. k-hop path search, Prize-Collecting Steiner Tree).
- **LM-based retrieval:** The abovementioned LLMs output nodes or relations as a pool of candidate nodes or relations (e.g., Subgraph Retriever, KG GPT).
- **GNN-based retrieval:** Graph Neural Networks rank edges or paths according to the query matching (e.g. GNN-RAG, EtD).

B. Hallucination and Faithfulness in RAG Outputs

The degree to which the output produced by LLM stays accurate, dependable, and consistent with the source material or ground truth is referred to as fidelity [17]. LLM-generated outputs frequently contain "hallucinations," which are situations in which the LLM introduces false claims, either by misattributing a fact or, more frequently, by fabricating one entirely.

The term "faithfulness hallucination" describes inconsistencies with context, user instructions, or internal consistency. Examples of these include logical inconsistencies in responses, context inconsistencies (responses that don't match the context given), and instruction inconsistencies (responses that don't follow the original instructions). Hallucinations were further divided into six categories by them: time warp, geographic erratum, created golem, acronym ambiguity, and numerical annoyance.

In retrieval-augmented LLMs, retrieving external information reduces the hallucinations brought on by the LLMs' intrinsic knowledge barrier, although hallucinations might still happen [18]. Factual hallucinations can happen when the information that is retrieved is inaccurate, out-of-date, or deviates from the model's internal parameters.

C. Data Privacy, Security, and Ethical Considerations

RAG systems may face significant privacy issues if the retrieval procedure involves sensitive data. Humanized Text: While RAG systems are utilized to scale, and applied more and more to sensitive areas, such as healthcare, law and enterprise fields, they create considerable data privacy and security concerns, as well as ethical issues, that are worthy of close consideration.

- **Privacy Risks:** RAG systems suffer an unintentional loss of sensitive or proprietary information on system membership or retrieval-based generation. In a case where the corpus also has some stored information, personal or confidential, they can be retrieved through queries.
- **Security Vulnerabilities:** RAG outputs can be corrupted by prompt injection, adversarial manipulation of retrieval and insecure storage [19]. Such assaults may result in the provision of false information, inappropriate use of systems, or IP infringements.
- **Ethical Challenges:** Potentially, biases in the retrieval corpus can directly affect the generation outputs. Trust can also be diminished because of the absence of source attribution transparency and the danger of hallucination. In addition, the legal frameworks like GDPR demand such aspects as consent and data protection.
- **Mitigation Approaches:** These are the techniques of user trust-protection: differential privacy [20], secure embedding storage, timely filtering, synthetic retrieval conditions and audit logging.

V. Evaluation, Benchmarks, and Future Directions

The assessment of RAG systems in LLM needs particular quantitative and qualitative measures to evaluate retrieval precision and generation quality. The recent findings emphasize that benchmarks and various datasets, like RAG Bench and MIRAGE, are required to measure the accuracy, faithfulness, and user satisfaction in the context of complex real-life settings. Although impressive improvement can be noticed, certain open issues still remain, such as the adequacy of context, scalability, and privacy risks [21]. This survey examines the latest evaluation frameworks, emerging benchmarks, and unresolved research opportunities, providing a comprehensive overview of how RAG techniques are assessed and pointing toward future directions for reliable and effective RAG deployment in LLMs.

A. Quantitative and Qualitative Evaluation Metrics

The accurateness of the retrieval and the quality of generation is evaluated through the assessment of RAG systems. In contrast to regular LLMs, RAG pipelines involve two layers of evaluation, the first one covering the assessment of the ability of system to access and the second one covering the assessment of applying that knowledge successfully during the generation. Figure 2 shows the LLM-based output validation to complete their framework.

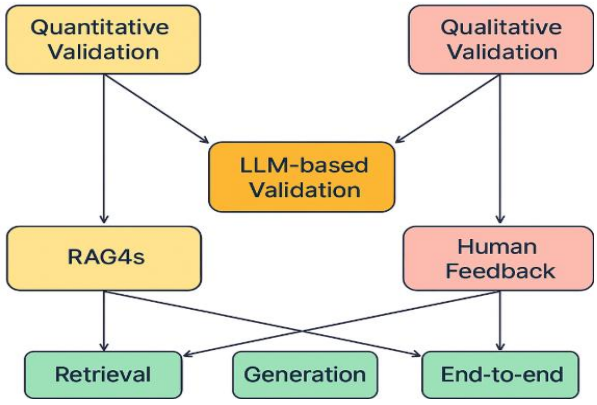

Fig. 2. Output validation framework for RAG

There are some evaluation metrices are as follows:

- Quantitative evaluation has measures that measure precision and recall including Precision, Recall, MRR, and NDCG to gauge the capability of the retriever to retrieve useful context. To measure fluency and semantic match to reference output, common approximations of generation quality such as BLEU, ROUGE, METEOR and BERT-Score are applied to generation.
- Qualitative evaluation concentrates on the parameters like contextual relevance, truthful consistency, semantic congruity as well as user satisfaction. These tend to be evaluated either by human grading or large language models as the graders.

Furthermore, composite and reference-free scores, e.g., RAGAs and RAG-Bench, are on the rise, which provides an end-to-end score on the relevance of the answer and attribution and detection of hallucinations.

B. Emerging Datasets and Benchmarking Efforts

To judge RAG systems correctly a number of new datasets and benchmarks have been created. These evaluate retrieval precision, creation quality and systems robustness.

Key Benchmarks

There are some key benchmarks are as follows:

- **RAG Bench:** Provides 100K examples in domains which can be explained by Trace.
- **MIRAGE:** Concentrates on QA, noise robustness and interpretability.
- **Benchmarked:** Automates big benchmarking of text and graph-based RAG.
- **MTRAG:** Focuses on multi-turn dialogue generation, such as in the field of finance and IT.
- **WixQA:** Real-world snapshot knowledge base enterprise-oriented QA.
- **Emerging Themes:** Multilinguality (Indic RAG Suite), multi-hop reasoning (Multi Hop-RAG) [22], long-context handling (LONG2RAG) and robustness (Safe RAG) are becoming significant.

Such activities allude to thorough and practical assessment of RAG systems.

C. Open Problems and Research Opportunities

In spite of its wonderful advances, RAG systems have not reached the point of perfection, and this leaves the domain potentially ripe to innovate.

1) Key Challenges

There are key challenges are as follows:

- **Context Sufficiency & Retrieval Quality:** The challenge of models is that they tend to get the right context with a little more; this results in hallucinations whereas more context incurs too much noise and expenses.
- **Scalability & Indexing Efficiency:** Such applications require efficient indexing techniques (e.g. HNSW) [23], the ability to update the index incrementally and fault tolerance on the high-dimensional retrieval over large corpora that are increasing rapidly.
- **Privacy and Security Vulnerabilities:** Membership inference, data poisoning and timely injection attacks are underdeveloped.

2) Research Opportunities

There are some are opportunities as follows:

- **Sufficient-Context Modeling:** Design mechanisms to know when the information that had been retrieved is sufficient, or rather, not sufficient [24], to generate with confidence.
- **Adaptive & Self-Improving RAG:** To effectively integrate different retrieval strategies, apply reinforcement learning and meta-learning that allows dynamically adapt the strategies to the feedback loops and goal-specific requirements of a task.
- **Hybrid Neuro-Symbolic and Graph Methods:** Combine symbolic reasoning, knowledge graphs and GNNs to improve multi-hop inference and provenance tracking.

VI. LITERATURE OF REVIEW

A survey of the literature on retrieval-augmented generation (RAG) methods in large language models (LLMs) is presented in this part, with a focus on key architectural elements, retrieval-generation tactics, domain-specific implementations, assessment frameworks, and technological developments.

Cheng et al. (2025) intends to give a thorough introduction to RAG by looking at its essential elements, such as retrieval methods, generation processes, and how the two are integrated. elucidate the key characteristics of RAG, such as its ability to augment generative models with dynamically obtained external information and the challenge of matching returned data to generative objectives. provide a taxonomy of RAG techniques as well, ranging from basic retrieval-augmented models to more complex ones that include multi-modal data and reasoning [25].

Li et al. (2025) recent studies on RAG's integration into learning environments. describe RAG and its operation. Then, depending on RAG's indexing pattern, outline the several types of retrievers and optimization techniques that may be used for generation. As the primary focus of their study, examine real-world applications of RAG in education, such as interactive learning platforms, assessment and creation of educational material, and extensive deployment in learning settings. The challenges and future research areas covered in this work include reducing hardware computational needs, improving multimodal support for RAG-based educational applications, minimizing hallucinations, and making retrieved knowledge timely and comprehensive [26].

Abbas and Rasool (2025) explores the ways in which intelligent data modelling and retrieval-augmented generation (RAG) might be used in tandem to address these issues and provide further financial insights into the Salesforce ecosystem. RAG is a powerful paradigm that combines real-time data retrieval with the power of LLM and offers a fresh viewpoint on improving the decision-making process. RAG allows creating satisfactory responses to sophisticated requests. This, when further combined with strong data model design specific to the financial sector, can create a vast improvement in the quality and detail of the provided insights to the financial analysts, advisors and executives. The paper proposes a conceptual model of RAG implementation in Salesforce focusing on the compatibility between the current Financial Cloud tools and data structure and RAG [27].

Gan et al. (2025) provides a summary of RAG assessment structures and methods, methodically reviewing both new and traditional evaluation techniques in terms of computing efficiency, safety, factual correctness, and system performance in the LLM era. RAG, which uses LLM and external information retrieval to generate text on a variety of applications in a verifiable, current, and accurate manner, has revolutionized NLP. A meta-analysis of assessment procedures in the high-impact RAG research is also carried out, along with the collection and classification of RAG-specific datasets and evaluation frameworks [28].

Gupta, Ranjan and Singh (2024) explain RAG's basic structure, which is focused on combining generation and retrieval in knowledge-intensive jobs. Retrieval mechanisms and generative language models are used in RAG to improve output accuracy and overcome major LLM drawbacks. The substantial technological developments in RAG are

thoroughly reviewed, with particular attention paid to important breakthroughs in retrieval-augmented language models and applications in a range of fields, including knowledge-based tasks, question-answering, and summarization [9].

Fan et al. (2024) this assessment, thoroughly examine previous Retrieval-Augmented Large Language Models (RA-LLMs) research works from three main technical angles. Although LLMs have shown ground-breaking capabilities in language creation and comprehension, they nevertheless have intrinsic drawbacks such as hallucinations and outdated internal information. RA-LLMs have emerged to leverage authoritative and external knowledge bases instead of depending only on the model's internal knowledge to improve the quality of the content generated by LLMs. This is due to RAG's strong capabilities in providing the most recent and useful auxiliary information [29].

Table II summarizes key studies on RAG techniques in language models, highlighting focus areas, retrieval-generation approaches, main findings, existing challenges, and future research directions.

TABLE II. COMPARATIVE ANALYSIS OF RECENT STUDIES ON RAG TECHNIQUES IN LANGUAGE MODELS

Reference	Study On	Approach	Key Findings	Challenges	Future Direction
Cheng et al. (2025)	Core components and taxonomy of RAG	Categorization of RAG models by retrieval and generation strategies	Provides a detailed taxonomy and highlights RAG's ability to integrate external knowledge	Alignment of retrieved content with generation remains complex	Exploration of multi-modal RAG and improved reasoning capabilities
Li et al. (2025)	RAG in educational applications	Indexed retrieval and optimization techniques for educational content	RAG enhances personalized learning and large-scale content generation	Hallucinations, knowledge completeness, and computational overhead	Strengthen multimodal retrieval and improve knowledge freshness
Abbas et al. (2025)	RAG in financial analytics	RAG integrated with intelligent data modelling in Salesforce	Improves insight quality and decision-making in financial systems	Adapting RAG to domain-specific financial data is complex	Develop domain-adaptive RAG for financial cloud environments
Gan et al. (2025)	Evaluation frameworks for RAG	Survey of RAG evaluation metrics and benchmark datasets	Reviews factuality, efficiency, and safety metrics in RAG systems	Lack of unified evaluation benchmarks	Establish standard evaluation pipelines and benchmark datasets
Gupta et al. (2024)	RAG architecture and technical advancements	Analysis of retrieval-generation fusion in LLMs	Demonstrates RAG's utility in QA, summarization, and knowledge tasks	Addressing LLM limitations like outdated internal knowledge	Enhance adaptive retrieval mechanisms in dynamic contexts
Fan et al. (2024)	RA-LLMs and knowledge augmentation	Technical review from 3 perspectives on LLM-RAG integration	Shows how RAG addresses outdated and hallucinated outputs	Maintaining source accuracy and authority	Advance retrieval precision and real-time knowledge injection

VII. CONCLUSION AND FUTURE WORK

Recent breakthroughs in language model design have spotlighted the need for more reliable, fact-aware, and context-sensitive generation methods. The integration of external information retrieval has improved the ability of existing systems to provide dynamic and grounded responses. The potential for real-time and domain-specific applications is enormous with such developments. The revolutionary framework known as RAG has emerged to address the drawbacks of conventional large language models by providing dynamic access to external knowledge sources. This survey reviewed core components, including retrieval strategies (sparse, dense, graph-based), generation architectures, and evaluation metrics. Despite its strengths, such as improved factual accuracy and reduction of hallucinations, RAG systems still face challenges in scalability, context sufficiency, and privacy risks. Benchmarking tools like RAG-Bench and MIRAGE have become essential to assess their performance across domains. Future directions point to adaptive learning, hybrid neuro-symbolic systems, and privacy-preserving mechanisms. As RAG continues to evolve, it paves the way for more robust and intelligent NLP systems.

Future research in RAG should focus on developing adaptive retrieval mechanisms that dynamically adjust to task-specific requirements and user feedback. Exploring hybrid neuro-symbolic and graph-based techniques can enhance multi-hop reasoning and factual consistency. Additionally, efforts are needed to improve context sufficiency modeling, scalable indexing, and privacy-preserving methods to ensure secure, accurate, and efficient RAG deployment across real-world, sensitive, and multilingual applications.

REFERENCES

[1] B. Peng *et al.*, "Graph Retrieval-Augmented Generation: A Survey," *J. ACM*, vol. 37, no. 4, 2024.

[2] S. Kresevic, M. Giuffrè, M. Ajcevic, A. Accardo, L. S. Crocè, and D. L. Shung, "Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework," *npj Digit. Med.*, vol. 7, no. 1, 2024, doi: 10.1038/s41746-024-01091-y.

[3] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," *Medicina (B. Aires).*, vol. 60, no. 3, 2024, doi: 10.3390/medicina60030445.

[4] J. Gu, "A Research of Challenges and Solutions in Retrieval Augmented Generation (RAG) Systems," *Highlights Sci. Eng. Technol.*, vol. 124, pp. 132–138, 2025, doi: 10.54097/364hex16.

[5] S. R. Thota, S. Arora, and S. Gupta, "Al-Driven Automated Software Documentation Generation for Enhanced Development Productivity," in *2024 International Conference on Data Science and Network Security (ICDSNS)*, IEEE, Jul. 2024, pp. 1–7. doi: 10.1109/ICDSNS62112.2024.10691221.

[6] V. Prajapati, "Advances in Software Development Life Cycle Models : Trends and Innovations for Modern Applications," *J. Glob. Res. Electron. Commun.*, vol. 1, no. 4, pp. 1–6, 2025.

[7] C. Yao and S. Fujita, "Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags," *Electronics*, vol. 13, no. 23, 2024, doi: 10.3390/electronics13234643.

[8] N. Malali, "The Role of Retrieval-Augmented Generation (RAG) in Financial Document Processing: Automating Compliance and Reporting," *Int. J. Manag. Technol.*, vol. 12, no. 3, pp. 26–46, 2025, doi: 10.37745/ijmt.2013/vol12n32646.

[9] S. Gupta, R. Ranjan, and S. N. Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions," 2024, doi: 10.48550/arXiv.2410.12837.

[10] S. R. P. Madugula and N. Malali, "AI-Powered Life Insurance Claims Adjudication Using LLMs and RAG Architectures," *Int. J. Sci. Res. Arch.*, vol. 15, no. 1, pp. 460–470, Apr. 2025, doi: 10.30574/ijsra.2025.15.1.0867.

[11] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, 2020.

[12] J. He *et al.*, "Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications." 2025. doi: 10.48550/arXiv.2505.01146.

[13] Q. Li *et al.*, "Using fine-tuned conditional probabilities for data transformation of nominal attributes," *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2019.08.024.

[14] S. Pahune and M. Chandrasekharan, "Several Categories of Large Language Models (LLMs): A Short Survey," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 7, pp. 615–633, 2023, doi: 10.22214/ijraset.2023.54677.

[15] P. Choudhary, R. Choudhary, and S. Garaga, "Enhancing Training by Incorporating ChatGPT in Learning Modules: An Exploration of Benefits, Challenges, and Best Practices," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 11, 2024.

[16] S. Pahune and Z. Akhtar, "Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models," *Information*, vol. 16, no. 2, Jan. 2025, doi: 10.3390/info16020087.

[17] B. Malin, T. Kalganova, and N. Boulgouris, "A review of faithfulness metrics for hallucination assessment in Large Language Models." 2024. doi: 10.48550/arXiv.2501.00269.

[18] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," *Mathematics*, vol. 13, no. 5, 2025, doi: 10.3390/math13050856.

[19] S. Zeng *et al.*, "The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)," 2024.

[20] Y. Zhou *et al.*, "Trustworthiness in Retrieval-Augmented Generation Systems: A Survey," pp. 1–22, 2024.

[21] S. Pahune, Z. Akhtar, V. Mandapati, and K. Siddique, "The Importance of AI Data Governance in Large Language Models," *Preprints*, 2025.

[22] D. Cohen *et al.*, "WixQA: A Multi-Dataset Benchmark for Enterprise Retrieval-Augmented Generation," 2025, doi: 10.48550/arXiv.2505.08643.

[23] P. Chatterjee, "Real-Time Payment Systems and their Scalability Challenges," *Iconic Res. Eng. Journals*, vol. 6, no. 12, pp. 1461–1470, 2023.

[24] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven Failure Points When Engineering a Retrieval Augmented Generation System," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, ACM, Apr. 2024, pp. 194–199. doi: 10.1145/3644815.3644945.

[25] M. Cheng *et al.*, "A Survey on Knowledge-Oriented Retrieval-Augmented Generation," pp. 1–50, 2025, doi: 10.48550/arXiv.2503.10677.

[26] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Comput. Educ. Artif. Intell.*, vol. 8, 2025, doi: 10.1016/j.caeai.2025.100417.

[27] A. Abbas and M. Rasool, "Empowering Financial Insights in the Cloud: Leveraging Retrieval- Augmented Generation and Intelligent Data Modeling on Salesforce." 2025. doi: 10.13140/RG.2.2.30679.38569.

[28] A. Gan *et al.*, "Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey." 2025. doi: 10.48550/arXiv.2504.14891.

[29] W. Fan *et al.*, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 6491–6501, 2024, doi: 10.1145/3637528.3671470.

