Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026482489

Retrieval-Augmented Generation for Large Language Model-Based Intelligent Assistants: A Review

Zilang Shao*

School of Artificial Intelligence, Shenyang University of Technology, Shenyang, China

**Corresponding author: Zilang Shao.*

Abstract

Large language models have accelerated the development of intelligent assistants by providing flexible natural-language understanding and generation. However, hallucination, knowledge staleness, and limited coverage of domain-specific information continue to restrict their reliability in knowledge-intensive tasks. This review examines how Retrieval-Augmented Generation (RAG) can strengthen LLM-based intelligent assistants by connecting generative capability with external, maintainable knowledge. It synthesizes research on the technical foundations of RAG, key components and optimization strategies, and applications and challenges in intelligent-assistant settings. The review finds that RAG can improve knowledge accuracy and timeliness by grounding responses in retrieved evidence and allowing knowledge resources to be updated independently of the base model. These benefits are conditional: unreliable retrieval, poorly maintained sources, ineffective use of context, and fragmented evaluation can still produce unsupported or unsafe answers. Reliable deployment, therefore, requires coordinated retrieval quality, knowledge management, generation control, and trustworthy evaluation. Future RAG-based assistants should combine these capabilities to become scalable, secure, evidence-aware, and verifiable systems.

Keywords

retrieval-augmented generation, large language models, intelligent assistants, knowledge augmentation, hallucination mitigation, information retrieval, trustworthy AI

1. Introduction

LLM-based intelligent assistants are increasingly used to support knowledge-intensive interactions because they can interpret varied requests and produce flexible responses [1, 2]. Their practical value, however, depends less on general fluency than on whether they can deliver accurate, up-to-date information relevant to the user's task and appropriate to a specific domain. This requirement is especially important in office, enterprise, and professional settings, where users may rely on an assistant's response to make decisions. Yet a stand-alone LLM draws primarily on knowledge encoded in its parameters, which cannot always reflect changing or private information. The resulting gap between broad language capability and dependable knowledge access motivates the use of RAG as a grounding mechanism for intelligent assistants.

Reliance on parametric knowledge limits dependable assistance in three connected ways. LLMs can produce fluent but unsupported claims, making hallucinations difficult for users to detect [3, 4]. Their knowledge also reflects a fixed training boundary, so changes in policies, products, and professional information are not incorporated automatically. An assistant may therefore present obsolete information with the same confidence as current information. General models also cannot fully cover private organizational resources or the terminology and evidential standards of every domain [5]. This mismatch is especially consequential when responses must follow organization-specific policies or professional evidence standards. These weaknesses reduce trust in knowledge-intensive settings, where apparently authoritative answers may guide decisions. RAG provides external grounding by retrieving query-relevant knowledge before generation, improving access to current and specialized information [6]. Its benefit, nevertheless, remains contingent on the relevance and quality of the retrieved evidence.

Existing RAG reviews organize architectures, retrieval and generation methods, evaluation approaches, and general applications [7, 8, 9]. Rather than reproducing a component catalog, this review examines how these technical choices affect the reliability of LLM-based intelligent assistants, particularly their knowledge accuracy and timeliness. The review, therefore, considers how RAG enhances the accuracy and reliability of knowledge for LLM-based intelligent assistants, how different components and optimization strategies influence system performance, and how external knowledge access improves timeliness in intelligent-assistant scenarios. It also examines the challenges and future directions that remain for RAG-based assistants. Together, these concerns connect technical design with the practical conditions under which retrieval improves assistant behavior.

2. Literature Review

2.1 Technical Foundations of RAG

2.1.1 Evolution from Large Language Models to Retrieval-Augmented Generation

Although LLMs offer strong generation capabilities [1, 2], their parametric knowledge is difficult to update and cannot reliably cover private, dynamic, or specialized information. Knowledge-intensive tasks, therefore, require access to external evidence. RAG emerged by combining a generative model with retrieved non-parametric knowledge, allowing relevant evidence to guide responses and expanding knowledge coverage without retraining the model [6].

2.1.2 Basic Architecture and Workflow of RAG

A basic RAG workflow connects four elements: an external knowledge source, a retrieval process, retrieved context, and a generator. The knowledge source contains information beyond the LLM's parameters and is structured so that relevant units can be searched. When a user submits a query, the retriever identifies passages judged relevant to the request. These passages are then supplied with the query as context for generation, allowing the LLM to interpret the evidence and formulate a natural-language answer. This architecture differs from stand-alone search because retrieval results directly condition what the generator produces. It also differs from purely parametric generation because the evidence available for a query can change without retraining the base model. The resulting answer may therefore be more attributable and responsive to current information, although its quality depends on whether the retrieval and generation stages work coherently [6, 10].

2.1.3 Role of RAG in Reducing Hallucination and Knowledge Staleness

RAG can reduce unsupported generation by providing external evidence against which an answer can be conditioned. When the retrieved context is relevant, and the generator follows it, the model relies less on uncertain parametric memory, thereby addressing one important source of hallucination [3, 11]. Separating the knowledge source from the LLM also supports timeliness: documents can be added, corrected, or removed without repeatedly retraining the model. However, RAG does not eliminate hallucination or guarantee current answers. A retriever may miss the required evidence, return irrelevant passages, or draw from an outdated or low-quality source; the generator may also misinterpret valid context. Even when retrieval is correct and relevant, the generator may misunderstand the context, incorrectly combine separate pieces of evidence, ignore supporting passages, or draw conclusions that go beyond what the sources establish. Reliability, therefore, depends not only on retrieval quality but also on the utilization of evidence and faithfulness in generation.

RAG should therefore be viewed as a reliability-enhancement mechanism whose effectiveness remains conditional on retrieval quality, source maintenance, and evidence-aware generation [6].

2.2 Key Components and Optimization Strategies

2.2.1 Retrieval Strategies

Retrieval determines which evidence reaches the generator, thereby placing an upper bound on answer quality. Sparse retrieval uses lexical overlap; it is efficient, interpretable, and well-suited to queries containing distinctive names, codes, or policy terms. Sparse indexes are also relatively straightforward to maintain and diagnose. Their weakness is lexical mismatch, because relevant passages expressed with different wording may be missed. Dense retrieval represents queries and passages as vectors, capturing semantic relevance more effectively for complex natural-language questions. However, its performance depends on embedding quality and generally incurs greater computational cost. DPR and contrastive retrieval illustrate these strengths and dependencies [12, 13]. Hybrid retrieval combines keyword precision with semantic coverage and may add reranking, but this balance increases pipeline complexity and tuning cost. No approach is uniformly superior: selection should reflect the assistant domain, query form, latency budget, and need for interpretability. Because irrelevant or missing evidence can produce unsupported answers even when generation is competent, retrieval strategy is a reliability decision rather than a preliminary search choice [10]. Stable, terminology-heavy repositories may favor sparse retrieval, whereas exploratory questions often benefit more from dense or hybrid matching.

2.2.2 Knowledge Storage and Indexing Methods

External knowledge must be organized into retrievable units before it can support an assistant. Document processing includes cleaning, chunking, metadata assignment, and embedding-based indexing. Chunk size creates a direct trade-off: larger chunks preserve surrounding context but may dilute relevant evidence with unrelated material, whereas smaller chunks can improve retrieval precision while separating claims from qualifiers or supporting explanations. Overlap can maintain continuity across adjacent chunks, but excessive overlap duplicates evidence, increases the index size, and raises retrieval costs. Metadata enables filtering by source, date, document type, authority, or user permission, improving control in enterprise knowledge bases. Knowledge-oriented RAG, therefore, treats source integration, segmentation, and maintenance as determinants of evidence quality rather than as preprocessing details [14]. Vector indexes provide scalable semantic access, while graph-based organization can expose entities, links, and global themes that isolated chunks may miss [15]. Storage technology alone cannot ensure retrieval quality: chunk boundaries, overlap, metadata accuracy, and update policies jointly determine whether the retrieved context is precise, complete, current, and authorized. These parameters should be validated against representative queries rather than fixed globally for every document collection.

2.2.3 Generation and Prompt Optimization

Generation optimization can be organized into four complementary strategies. Prompt constraints instruct the model to rely on retrieved evidence, avoid unsupported claims, and decline to answer when evidence is insufficient; wording and output constraints materially affect task interpretation [16]. Context compression removes redundant or weakly relevant material to increase the proportion of useful evidence, although aggressive compression may discard qualifiers or details that change a conclusion [17]. Evidence ordering determines which passages receive attention: models may underuse information placed in the middle of long contexts, so ranking and placement affect factual consistency [18]. Citation generation adds source identifiers or passage references to the answer, making claims easier for users to verify, but a citation is useful only when it actually supports the associated statement. These strategies address different failure points and should be combined deliberately. The objective is not greater fluency but faithful use of sufficient evidence; prompt instructions or citations cannot make missing, distorted, or unreliable context factual. The appropriate combination depends on context length, evidence density, and the level of verification expected from the assistant.

2.2.4 Retrieval-Generation Collaboration

Traditional RAG passes retrieved passages to a generator in a largely linear sequence, allowing retrieval errors to propagate without correction. Collaborative designs introduce feedback, but they intervene at

different levels. Corrective RAG focuses on retrieval quality: it evaluates the available evidence and can filter results, revise the query, or retrieve again when the support is inadequate [19]. Self-RAG focuses on model self-reflection by determining when retrieval is needed and assessing whether the evidence supports the generated answer [20]. AssistRAG addresses the broader intelligent-assistant workflow, coordinating question understanding, information acquisition, evidence processing, and response production [21]. These approaches are therefore complementary rather than simple substitutes: Corrective RAG repairs evidence acquisition, Self-RAG regulates retrieval and generation decisions, and AssistRAG coordinates end-to-end assistance. Their feedback can reduce unnecessary context and unsupported answers, but it also adds latency and potential evaluator errors. Reliable deployment requires choosing the collaboration mechanism that matches the dominant failure point while optimizing retrieval quality and generation reasoning as connected processes. A system dominated by retrieval noise requires a different intervention from one that retrieves well but reasons unreliably.

2.3 Applications and Challenges in Intelligent Assistants

2.3.1 Applications in Office Assistants and Knowledge-Based Question Answering

RAG is well-suited to intelligent assistants because many user requests depend on knowledge external to the model, and that changes over time. In knowledge-based question answering, retrieval grounds responses in selected documents rather than relying entirely on parametric memory, improving access to task-relevant evidence [6]. For enterprise and office assistants, this means searching internal policy documents, project records, product manuals, and customer support histories so that answers reflect organizational knowledge rather than relying solely on a general-purpose LLM. Knowledge-oriented RAG treats the integration, maintenance, and authorized use of these private sources as part of the assistant's knowledge capability [14]. The same principle extends to education, where assistants can retrieve updated course materials and learning resources for curriculum-aligned, personalized question answering. In healthcare, retrieval can facilitate access to medical guidelines and domain-specific evidence, but such assistance should enhance information access rather than replace clinical judgment. Conversational systems can further adapt retrieval to the current query, user role, and permitted resources. AssistRAG illustrates coordination among question understanding, information acquisition, evidence processing, and response generation [21]. Across these settings, the practical benefit is improved grounding and traceability: users can inspect the source basis of an answer, while the assistant can work with updated or domain-specific knowledge without assuming that every fact is encoded in model parameters.

2.3.2 Current Challenges and Limitations

Real-world deployment presents three connected classes of challenge. Technical failures can occur at any point in the retrieval-generation chain: ambiguous queries, incomplete indexes, or irrelevant evidence may mislead the generator, while hallucination can propagate even after retrieval. Corrective RAG can filter results or retrieve again when support is inadequate, but correction introduces latency and may itself fail [19]. Evaluation is equally difficult because final-answer accuracy does not reveal whether an error originated in retrieval, context use, or generation; evidence relevance, factuality, faithfulness, and task usefulness must therefore be assessed together [22]. Governance and security challenges concern the trustworthiness and controllability of external knowledge. Poisoned or unreliable documents can influence answers, while enterprise stores require privacy protection, permission-aware retrieval, and access control [23]. Cost and efficiency challenges arise from retrieval, reranking, knowledge-base updates, and longer contexts, all of which increase computation and response time as the system scales [10]. More elaborate safeguards and verification may improve reliability but add operational cost. RAG deployment, therefore, requires an explicit balance among evidence quality, reliability, security, latency, and cost, rather than optimizing answer quality alone.

3. Discussion

3.1 Improving Knowledge Accuracy

Rather than treating knowledge accuracy as a single outcome, RAG-based assistants should be evaluated across three linked levels. Evidence availability asks whether an authoritative and sufficient source exists for the query; without such evidence, neither retrieval nor generation can produce a well-grounded answer [6].

Retrieval effectiveness asks whether the system identifies that evidence despite query ambiguity, indexing choices, and competing passages. Evidence utilization and generation faithfulness then ask whether the model interprets retrieved material correctly, preserves qualifications, and grounds each important claim in the evidence. Corrective and self-reflective methods address these latter levels by detecting weak retrieval, initiating additional search, and assessing whether evidence supports the answer [19, 20]. This layered view localizes failure: relevant evidence may exist but remain unretrieved, or retrieved evidence may be ignored or misused. Reliability, therefore, emerges from the complete retrieval-generation chain rather than from the performance of any single module.

3.2 Improving Knowledge Timeliness

RAG addresses knowledge staleness by separating the assistant's knowledge resource from the base LLM. Documents can be added, corrected, or removed without retraining the model, allowing retrieval to reflect information available at query time. This is particularly useful when organizational policies or domain knowledge change faster than model development cycles. Knowledge-oriented RAG emphasizes that timeliness depends not only on retrieval but also on source integration, representation, and maintenance [14]. An updateable store can nevertheless remain stale in practice if ingestion is delayed, obsolete passages are retained, or metadata does not distinguish current from superseded information. Frequent updates also create indexing, validation, and governance costs. RAG surveys, therefore, treat external knowledge access as an update mechanism rather than as proof that an answer is current [7, 10]. For intelligent assistants, improved timeliness requires clear source ownership, update policies, and retrieval rules that prioritize authoritative, up-to-date evidence.

3.3 Existing Research Gaps

Three gaps continue to limit reliable deployment. First, evaluation remains fragmented. Final-answer accuracy cannot reveal whether failure originated in retrieval, context use, or generation, while retrieval relevance alone does not establish factuality. Existing surveys call for joint measures of evidence quality, answer correctness, faithfulness, and task utility, but datasets and automatic metrics remain difficult to compare across domains [9, 22]. Second, reliability and security are insufficiently integrated. Incorrect retrieval is often studied as a quality problem, whereas poisoned sources, unauthorized access, and information leakage require threat-aware data governance and system controls. Because attacks can target the knowledge base, retrieval, prompt assembly, or output, isolated safeguards leave important vulnerabilities [23]. Third, system optimization lacks a stable account of trade-offs among accuracy, latency, and cost. Additional retrieval, reranking, verification, or reflection may improve evidence support but also slow responses and increase computational cost. These gaps obstruct intelligent-assistant deployment because reliability must be demonstrated under realistic queries, evolving knowledge, security constraints, and resource limits, not only through benchmark improvements in a single component. Without such evidence, deployment decisions remain difficult to justify or reproduce.

3.4 Future Development Trends

Future development can be separated into short-term practical improvements and longer-term research. In the short term, existing RAG systems can become more reliable and deployable through adaptive retrieval that adjusts effort to task complexity, refines queries, and uses feedback; through better indexing, source maintenance, metadata, and governance; and through evaluation that distinguishes retrieval failure from generation failure. Self-RAG, AssistRAG, and agentic RAG illustrate mechanisms for coordinating retrieval decisions, evidence processing, critique, and workflow control [20, 21, 24]. Near-term work should prioritize measurable reliability, latency, and operational control rather than autonomy for its own sake. In the longer term, structured and knowledge-aware RAG can use graph-based representations and richer knowledge organization to expose relationships that isolated chunks miss [14, 15]. More autonomous agentic systems may combine planning, self-reflection, and multi-step reasoning, but they will also require trustworthy deployment through stronger security, uncertainty reporting, and auditability. The short-term objective is to improve current system reliability and deployability; the long-term objective is to build more autonomous, structured, and trustworthy intelligent assistants.

4. Conclusion

Large language models have accelerated the development of intelligent assistants, but hallucinations, knowledge staleness, and limited coverage of domain-specific information prevent parametric knowledge alone from consistently supporting reliable answers. This review has shown that Retrieval-Augmented Generation provides an important connection between general language capability and external, maintainable knowledge. By examining its technical foundations, key components, optimization strategies, applications, and deployment challenges, the review finds that RAG can improve knowledge accuracy and timeliness when relevant evidence is retrieved and used faithfully. However, retrieval does not guarantee reliability. Its benefits depend on the quality and currency of knowledge sources, effective knowledge organization, appropriate context construction, controlled generation, and evaluation that distinguishes retrieval failure from generation failure. Future RAG-based assistants must therefore combine technical performance with trustworthy evaluation, efficient operation, secure knowledge management, and clear traceability of evidence. Progress toward these goals will determine whether RAG systems can move beyond promising demonstrations to become reliable, scalable, and verifiable intelligent assistants for real knowledge-intensive tasks.

References

[1] T. Brown *et al.*, 'Language Models are Few-Shot Learners', *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[2] W. X. Zhao *et al.*, 'A Survey of Large Language Models', *Front. Comput. Sci.*, vol. 20, no. 12, p. 2012627, May 2026, doi: 10.1007/s11704-026-60308-3.

[3] L. Huang *et al.*, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', *ACM Trans. Inf. Syst.*, vol. 43, no. 2, p. 42:1-42:55, Jan. 2025, doi: 10.1145/3703155.

[4] Y. Zhang *et al.*, 'Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models', *Computational Linguistics*, vol. 51, no. 4, pp. 1373–1418, Dec. 2025, doi: 10.1162/COLI.a.16.

[5] A. Kostikova, Z. Wang, D. Bajri, O. Pütz, B. Paaßen, and S. Eger, 'LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models', *ACM Comput. Surv.*, vol. 58, no. 11, p. 282:1-282:33, Apr. 2026, doi: 10.1145/3801096.

[6] P. Lewis *et al.*, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9459–9474. Accessed: Jul. 14, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[7] Y. Huang and J. X. Huang, 'A Survey on Retrieval-Augmented Text Generation for Large Language Models', *ACM Comput. Surv.*, vol. 58, no. 12, p. 300:1-300:38, May 2026, doi: 10.1145/3805774.

[8] C. Sharma, 'Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers', May 28, 2025, *arXiv*: arXiv:2506.00054. doi: 10.48550/arXiv.2506.00054.

[9] A. Brown, M. Roman, and B. Devereux, 'A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges', Sep. 09, 2025, *arXiv*: arXiv:2508.06401. doi: 10.48550/arXiv.2508.06401.

[10] Y. Gao *et al.*, 'Retrieval-Augmented Generation for Large Language Models: A Survey', Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[11] Y. Li, X. Fu, G. Verma, P. Buitelaar, and M. Liu, 'Mitigating Hallucination in Large Language Models (LLMs): An Application-Oriented Survey on RAG, Reasoning, and Agentic Systems', Oct. 28, 2025, *arXiv*: arXiv:2510.24476. doi: 10.48550/arXiv.2510.24476.

[12] G. Izacard *et al.*, 'Unsupervised Dense Information Retrieval with Contrastive Learning', Aug. 29, 2022, *arXiv*: arXiv:2112.09118. doi: 10.48550/arXiv.2112.09118.

[13] V. Karpukhin *et al.*, 'Dense Passage Retrieval for Open-Domain Question Answering', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds, Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[14] M. Cheng *et al.*, 'A Survey on Knowledge-Oriented Retrieval-Augmented Generation', Mar. 17, 2025, *arXiv*: arXiv:2503.10677. doi: 10.48550/arXiv.2503.10677.

[15] D. Edge *et al.*, 'From Local to Global: A Graph RAG Approach to Query-Focused Summarization', Feb. 19, 2025, *arXiv*: arXiv:2404.16130. doi: 10.48550/arXiv.2404.16130.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, 'Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing', *ACM Comput. Surv.*, vol. 55, no. 9, p. 195:1-195:35, Jan. 2023, doi: 10.1145/3560815.

[17] S. Verma, 'Contextual Compression in Retrieval-Augmented Generation for Large Language Models: A Survey', Oct. 02, 2024, *arXiv*: arXiv:2409.13385. doi: 10.48550/arXiv.2409.13385.

[18] N. F. Liu *et al.*, 'Lost in the Middle: How Language Models Use Long Contexts', *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.

[19] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, 'Corrective Retrieval Augmented Generation', Oct. 2024, Accessed: Jul. 15, 2026. [Online]. Available: https://openreview.net/forum?id=JnWJbrnaUE

[20] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, 'Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection', *International Conference on Learning Representations*, vol. 2024, pp. 9112–9141, May 2024.

[21] Y. Zhou, Z. Liu, and Z. Dou, 'AssistRAG: Boosting the Potential of Large Language Models with an Intelligent Information Assistant', Nov. 11, 2024, *arXiv*: arXiv:2411.06805. doi: 10.48550/arXiv.2411.06805.

[22] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, 'Evaluation of Retrieval-Augmented Generation: A Survey', in *Big Data*, W. Zhu, H. Xiong, X. Cheng, L. Cui, Z. Dou, J. Dong, S. Pang, L. Wang, L. Kong, and Z. Chen, Eds, Singapore: Springer Nature, 2025, pp. 102–120. doi: 10.1007/978-981-96-1024-2_8.

[23] B. Palanisamy, G. S. S. Chalapathi, V. Hassija, and R. Buyya, 'Security and Privacy in Retrieval-Augmented Generation: Architectures, Threats, Defenses, and Future Directions for Building Trustworthy Systems', Jun. 24, 2026, *arXiv*: arXiv:2606.25533. doi: 10.48550/arXiv.2606.25533.

[24] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, and A. V. Vasilakos, 'Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG', Apr. 01, 2026, *arXiv*: arXiv:2501.09136. doi: 10.48550/arXiv.2501.09136.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access