

Retrieval-Augmented Generation (RAG): A Two-Decade Survey (2005–2020)

Harshal Narayanrao Ranotkar ¹, Gaurav Rameshrao Savant ² Milind V. Lande ³

^{1,2} Lecturer in Computer Engg, Department of Computer Engineering, Government Polytechnic, Arvi, Maharashtra India, 442201.

³ Lecturer in Computer Engg, Department of Computer Engineering, Government Polytechnic, Gadchiroli, Maharashtra India, 442605.

Abstract- The rapid advancement of Artificial Intelligence (AI) and Large Language Models (LLMs) has transformed natural language processing applications across diverse domains. However, traditional language models often suffer from limitations such as hallucination, outdated knowledge, and lack of transparency in decision-making. Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm that combines information retrieval techniques with generative models to enhance factual accuracy, contextual relevance, and knowledge grounding. This survey presents a comprehensive review of the evolution of RAG-related technologies over the past two decades, from early information retrieval and question-answering systems to modern neural retrieval and large language model architectures. We examine the fundamental components of RAG frameworks, including retrieval mechanisms, embedding models, vector databases, reranking strategies, and generation modules. Furthermore, we compare prominent RAG architectures, discuss their strengths and limitations, and analyze their applications in healthcare, education, enterprise knowledge management, scientific research, and conversational AI[1-2]. Finally, we identify current challenges and future research directions, including multimodal retrieval, adaptive retrieval strategies, continual learning, and trustworthy AI systems. This survey aims to provide researchers and practitioners with a structured understanding of the development, state-of-the-art methods, and future prospects of Retrieval-Augmented Generation systems.

Keyword:- Retrieval-Augmented Generation, RAG, Information Retrieval, Natural Language Processing, Neural Retrieval, Dense Passage Retrieval.

1. INTRODUCTION

The field of Artificial Intelligence has witnessed significant progress in natural language understanding and generation during the past two decades. Early information retrieval systems focused primarily on locating relevant documents from large repositories, while subsequent advances in machine learning enabled automated extraction and synthesis of information. The emergence of transformer-based architectures and Large Language Models (LLMs), such as GPT, BERT, and T5, further revolutionized the ability of machines to generate coherent and contextually relevant text.

Natural language processing has traditionally followed two complementary directions: information retrieval (IR), which focuses on finding relevant information from large collections, and language generation, which focuses on producing coherent natural-language text. Conventional retrieval systems identify documents or passages relevant to a query, whereas generative neural models produce text based primarily on knowledge encoded in their parameters. The development of neural ranking methods provided an important bridge between these areas. Burges et al.

introduced RankNet in 2005, demonstrating how neural networks could learn ranking functions for information retrieval. Later, semantic retrieval approaches learned distributed representations of queries and documents, enabling systems to identify semantic relationships beyond exact keyword matching. Huang et al.'s Deep Structured Semantic Models (DSSM), for example, projected queries and documents into a common semantic space.

The emergence of large-scale question-answering datasets further encouraged systems that could retrieve evidence and reason over textual information. SQuAD, introduced in 2016, provided more than 100,000 questions for machine reading comprehension. Reading Wikipedia for open-domain question answering subsequently demonstrated the importance of combining document retrieval with machine comprehension.

Transformer-based pretrained models significantly changed NLP. BERT demonstrated that deep bidirectional representations could be pretrained on unlabeled text and subsequently adapted to tasks such as question answering. However, pretrained language models store much of their factual knowledge implicitly in model parameters. This creates challenges when knowledge must be updated, verified, or explicitly cited.

By 2020, several approaches began integrating retrieval directly into pretrained language models. REALM introduced a latent knowledge retriever that could retrieve documents from a large corpus during language-model pretraining and inference. Dense Passage Retrieval demonstrated that dense neural representations could retrieve relevant passages effectively for open-domain question answering. These developments culminated in Retrieval-Augmented Generation, proposed by Lewis et al., which combined a pretrained sequence-to-sequence generator with a dense vector index of external knowledge. Despite these achievements, standalone language models exhibit several limitations. Their knowledge is restricted to training data, making them susceptible to outdated

information. Additionally, they often generate plausible but incorrect responses, a phenomenon commonly referred to as hallucination. These challenges have motivated the development of Retrieval-Augmented Generation (RAG), which integrates external knowledge retrieval with text generation[1-5].

The concept underlying RAG originates from traditional information retrieval and open-domain question-answering systems developed during the 2000s. Modern RAG architectures retrieve relevant documents or passages from external knowledge sources and use them as contextual evidence for generative models. This approach enhances factual correctness, interpretability, and adaptability to dynamic knowledge environments.

Since the introduction of the formal RAG framework in 2020, research has expanded rapidly, incorporating dense retrieval methods, vector databases, knowledge graphs, multimodal retrieval systems, and agent-based architectures. These developments have established RAG as a key technology for enterprise AI applications, intelligent search systems, and domain-specific assistants[6-10].

This survey reviews the historical evolution of RAG-related technologies from 2005 to 2020, presents a taxonomy of major approaches, compares representative methods, and highlights open challenges and future research opportunities.

II. OVERVIEW OF RAG EVOLUTION (2005–2020)

Table 1. Overview of RAG Evolution (2005–2020)

PERIOD	TECHNOLOGY	CONTRIBUTION TO RAG
2005	RankNet	Learned document ranking
2013	DSSM	Semantic query-document representations
2014	Attention	Selective contextual information
2016	SQuAD	Large-scale machine reading
2017	Open-domain QA	Retrieval + reading

2017	TriviaQA	Evidence-based QA
2017	Large-scale vector search	Efficient similarity retrieval
2018	BERT	Strong contextual representations
2019	Latent retrieval	Learned evidence retrieval
2019	Natural Questions	Real-world open-domain QA
2020	REALM	Retrieval integrated with language-model pretraining
2020	DPR	Dense neural passage retrieval
2020	ColBERT	Efficient contextual retrieval
2020	RAG	Retrieval integrated with text generation

Table 2. Key Contribution of RAG Evolution (2005–2020)

Period	Major Development	Key Contribution
2005–2010	Traditional Information Retrieval	TF-IDF, BM25, search engines
2010–2015	Statistical & Semantic Retrieval	Latent Semantic Analysis, Word Embeddings
2015–2018	Neural Information Retrieval	Deep learning-based ranking models
2018–2020	Transformer Era	BERT, GPT, Dense Passage Retrieval

III- COMPARISON OF MAJOR RAG METHODS

Table 3. Comparison of Major RAG Methods

Method	Retrieval Type	Generator	Advantages	Limitations
TF-IDF + Seq2Seq[11]	Spars e	RNN/Seq2Seq	Simple, efficient	Poor semantic understanding
BM25 + Transformer	Spars e	Transformer	Strong baseline, interpretable	Keyword dependency
DPR + BART RAG	Dens e	BART	Better semantic retrieval	High computational cost
REALM[12]	Dens e	BERT-based	End-to-end training	Complex optimization
RETRO	Chunk	Transformer	Scalable knowledge	Large retrieval infrastructure

	Retrieval		access	
Atlas[13]	Dense Retrieval	T5	Strong QA performance	Expensive training
GraphRAG	Knowledge Graph	LLM	Better reasoning	Graph construction complexity
Hybrid RAG[14]	Spars e + Dense	LLM	Improved recall and precision	Higher latency
Agentic RAG	Dynamic Retrieval	LLM Agents	Adaptive and autonomous	Increased complexity
Multimodal RAG[15]	Text + Images + Documents	Multimodal LLM	Rich knowledge integration	Storage and processing overhead

IV- LIMITATIONS OF CURRENT RAG SYSTEMS

4.1 Retrieval Errors

- Relevant documents may not be retrieved.
- Sensitive to embedding quality.
- Domain-specific knowledge may be missing.

4.2 Context Window Constraints

- LLMs have limited context capacity.
- Long retrieved documents may be truncated.

4.3 Hallucination Persists

- Even with retrieval, models may misinterpret evidence.
- Contradictory documents can confuse generation.

4.4 Computational Cost

- Vector search and reranking increase latency.
- Large-scale deployment requires significant infrastructure.

4.5 Data Freshness and Maintenance

- Knowledge bases require continuous updates.
- Outdated retrieval sources affect answer quality.

4.6 Security and Privacy Concerns

- Data leakage risks in enterprise environments.

- Vulnerable to prompt injection and retrieval poisoning attacks.

4.7 Evaluation Challenges

- No universally accepted benchmark.
- Difficult to jointly measure retrieval quality and generation quality.

V-RESEARCH GAP

The literature from 2005–2020 demonstrates a clear progression from retrieval-oriented systems toward retrieval-generation integration.

However, several gaps remained at the end of 2020:

1. Retrieval and generation were not always jointly optimized.
2. Retrieval quality could strongly constrain generation quality.
3. Evidence selection remained challenging for ambiguous queries.
4. Hallucination was not completely eliminated by retrieval.
5. Knowledge freshness required continuous index maintenance.
6. Evaluation needed to consider both retrieval and generation.
7. Evidence attribution and provenance required further development.
8. Large-scale RAG systems required efficient vector indexing and retrieval infrastructure.

These gaps established the foundation for subsequent research into advanced RAG architectures, adaptive retrieval, hybrid retrieval, reranking, query expansion, knowledge graphs, multimodal retrieval, and agentic retrieval.

VI- FUTURE RESEARCH DIRECTIONS

1. Adaptive Retrieval Mechanisms
2. Multimodal RAG Systems
3. Knowledge Graph-Augmented RAG
4. Continual Learning RAG
5. Trustworthy and Explainable RAG
6. Federated and Privacy-Preserving RAG
7. Agentic and Autonomous RAG Architectures
8. Domain-Specific RAG for Healthcare, Education, and Scientific Research

CONCLUSION

Retrieval-Augmented Generation has evolved from traditional information retrieval foundations into a central paradigm for knowledge-grounded artificial intelligence. By combining external knowledge retrieval with powerful generative models, RAG significantly improves factual accuracy, transparency, and adaptability. Retrieval-Augmented Generation did not emerge suddenly in 2020. Rather, it represents the convergence of approximately two major research streams: information retrieval and neural language generation. From RankNet's learned ranking in 2005, through semantic retrieval, attention mechanisms, machine reading comprehension, open-domain question answering, BERT, latent retrieval, REALM, and Dense Passage Retrieval, each development contributed an important component of modern RAG. The 2020 work by Lewis et al. provided a particularly influential formulation by combining a pretrained sequence-to-sequence generator with a dense external knowledge index. Although challenges related to retrieval quality, computational cost, and robustness remain, ongoing advancements in hybrid retrieval, agentic systems, and multimodal AI indicate that RAG will continue to play a pivotal role in the next generation of intelligent systems. This survey provides a comprehensive understanding of the evolution, methodologies, limitations, and future directions of RAG research over the last two decades.

REFERENCES

- [1] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to Rank Using Gradient Descent," *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 89–96, 2005. DOI: 10.1145/1102351.1102363.
- [2] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data," *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2013. DOI: 10.1145/2505515.2505665.
- [3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *International Conference on Learning Representations (ICLR)*, 2015.
- [4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016. DOI: 10.18653/v1/D16-1264.
- [5] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1870–1879, 2017. DOI: 10.18653/v1/P17-1171.

- [6] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017. DOI: 10.18653/v1/P17-1147.
- [7] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *arXiv preprint arXiv:1702.08734*, 2017.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [9] J. K. Lee, M.-W. Chang, and K. Toutanova, "Latent Retrieval for Weakly Supervised Open Domain Question Answering," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, 2019. DOI: 10.18653/v1/P19-1612.
- [10] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 452–466, 2019. DOI: 10.1162/tacl_a_00276.
- [11] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, pp. 3929–3938, 2020.
- [12] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. DOI: 10.18653/v1/2020.emnlp-main.550.
- [13] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *arXiv preprint arXiv:2004.12832*, 2020.
- [14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. DOI: 10.18653/v1/2020.acl-main.703.
- [15] J. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.