Grenze International Journal of Engineering and Technology, June Issue

A Bibliometric and Systematic Review of Retrieval-Augmented Generation Research: Evolution, Collaboration, and Emerging Themes

Omkar Rane[1] and Aparna Joshi[2]
[1-2]Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India
Email: omkar.rane24@pccoepune.org, aparna.joshi@pccoepune.org

***Abstract*— Retrieval-Augmented Generation, is a prominent AI technique in the field of artificial intelligence meanwhile, it integrates large-scale language models to boost the interpretability of output, contextual grounding, and factual accuracy. In this paper, we conduct a comprehensive bibliometric analysis to identify the scientific map of RAG research during the period 2018–2025.Findings reveal that there is an explosive growth in scientific production was fivefold between 2020 and 2024, coinciding with advancement of transformer-based models. A global overview of research evidences a profoundly interconnected, international collaborative context wherein top-ranked countries like U.S., Chinese, British, and Indian academic institutions also hold central places within the network of cooperation. Above all, the co-citation and thematic analyses successfully demarcate the core intellectual structure of the field, organized around three principal focuses: generative models enhanced by retrieval, knowledge-heavy applications, and approaches for fact grounding with language models. In particular, the prevailing tone of research has shifted from merely integrating basic components of retrieval to concentrate on trust and scalability benchmarks suitable for real-world applications within specific domains. The present study, from this perspective, not only clarifies the structural development of RAG but also acts as a crucial strategic guide for researchers and practitioners willing to push the frontier of Retrieval-Augmented Generation technology forward.**

***Index Terms*— Retrieval Enhanced Generation, Context aware generation, Retrieval-Augmented Generation (RAG), Hallucination mitigation, Bibliometric Analysis, Scientometric Study, Artificial Intelligence, Large Language Models (LLMs), Knowledge Structures, Thematic Evolution; Collaboration Network, Research Trends.**

I. INTRODUCTION

Due to the rapid growth of large language models (LLMs), machine intelligence has entered into a new phase where it can make mistakes like creating false information and cannot get new data from outside sources. Retrieval-Augmented Generation (RAG) presents a refined solution by merging generative models with retrieval components that contextually anchor answers using external knowledge sources [3, 5]
Since its first appearance, RAG has been used in different areas of interest like scientific question answering, legal text comprehension, biomedical literature understanding and education [7],[9], [14].

Grenze ID: 01.GIJET.12.2.627_26

Due to the rapid growth of research in this field, knowing its evolutionary process and information flows, as well as emerging research trends, becomes crucial. Conventional literature reviews frequently fail to capture the multidimensional expansion and intellectual interconnections among hundreds of papers.
This study thus adopts data-driven bibliometric methods to systematically chart the scientific landscape of RAG. Visualizing the trends, collaborations and thematic clusters, this study pinpoints the leading contributors, topical shifts and research prospects.
The main objectives of this consideration are:

- To examine the development and path of Evolving RAG publications through the years.
- To identify the most influential authors, journals, and regions contributing to RAG research.
- To examine intellectual and conceptual linkages through co-citation and co-occurrence analyses.
- To reveal emerging themes and evolving paradigms that define the next frontier of RAG development.

II. METHODOLOGY

These keywords describe the bibliometric view that enables a detailed overview of publications concerning "Retrieval-augmented generation research: Evolution, collaboration," and emerging themes at several levels. The process is presented step by step in Figure 1

- Identify Domain: Define the research area concerning keyword-based story generation using NLP pre-trained models.
- Identify Keywords: Select relevant keywords based on discussions with domain experts, prior literature, and technological advancements.
- Search Query Designing: Develop a well-structured search query using the chosen keyword.
- Extract Relevant Literature: Relevant literature to be collected through searching in Scopus and Web of Science.
- Conduct Analysis: Perform bibliometric analysis of the obtained papers to find the patterns, main authors, institutions, collaborative networks, etc.

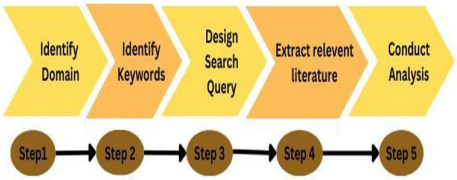

Figure 1: Methodology

TABLE I: IDENTIFIED KEYWORDS AND SEARCH QUERY

Keywords	PDF, Document, Multimodal, Table, Cross-Modal, Faithfulness, Hallucination, Grounded. Rag, GPT, LLM
Search Query	("Retriev* Augment* Generation" or Rag) and (LLM or "Large Language Model") and (Multimodal or "Multi-Modal" or "Cross-Modal") and (Document* or PDF or Table* or Image*) and (Faithfulness or Grounded or Hallucination)

III. DATA COLLECTION

Creating bibliometric analysis requires gathering information from reputable academic sources. Thus, the data was retrieved from the most known academic sources Scopus and Web of Science. I used the following keywords in the search: "Retrieval-Augmented Generation," "RAG," "retrieval-enhanced language models," and "knowledge-grounded generation." Only papers that had the following selection criteria were included in the dataset:

- Publication Period: 2018-2025
- Language: English
- Document type: Journal articles and Conference papers.

The extracted data from Scopus contained metadata such as publication year, author affiliations, citations, and keywords. Analysis of the dataset was conducted to determine tendencies and research flows, the leading contributors and collaborator.

IV. BIBLIOMETRIC ANALYSIS OF RAG

In this study which we looked at peer-reviewed articles from 2018 to 2025 containing the keywords related to Retrieval-Augmented Generation. There are literary elements in each entry such as: title, author(s), address (institutional affiliations), keywords, and reference(s).

Patterns of publications, networks of collaboration, and conceptual frameworks were examined through statistical, relational, and network-based approaches. The analytical framework consisted of 4 primary dimensions:

- Descriptive Overview: Evaluation of growth in the number of publications, citation analysis and leading journals.
- Collaboration Analysis: Using visual aids, this section examines the degree of collaboration between writers, organizations, and nations.
- Intellectual Structure: To comprehend the underlying intellectual linkages within the study field, a co-citation analysis of authors, journals, and articles is used.
- Conceptual Structure: To help find new topics, this section looks at how study themes change over time and how important terms co-occur. Annual scientific production, authorship trends, citation patterns, and knowledge networks are all depicted visually using charts and tables.

A. Descriptive Overview

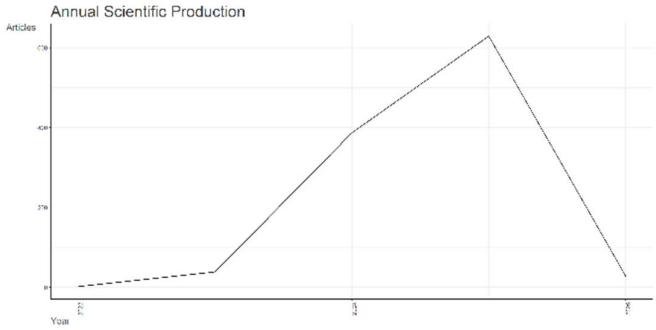

Figure 2: Annual Scientific Production

As per the trend shown in annual publication graph (Figure 2) we can see sudden surge in RAG research output starting from 2020, along with the advancement of transformer-based models like GPT and T5 [4], [6]. The graph shows that yearly scientific output shows a fivefold increase from 2020 to 2024, highlighting RAG's rapid adoption in both academic and industrial sectors

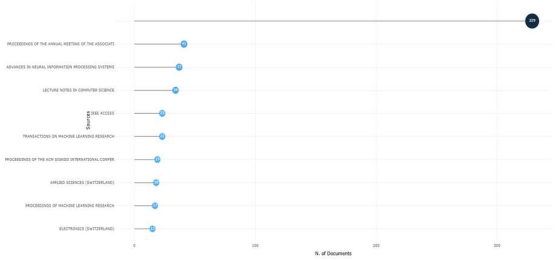

Figure 3: Top 10 Most Relevant Sources.

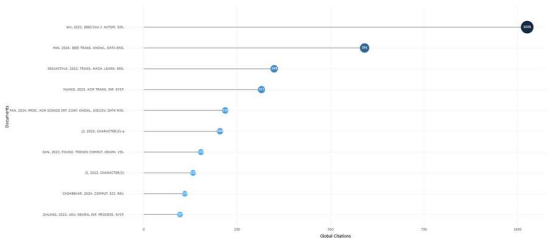

Figure 4: Top 10 Most Relevant Documents.

The most relevant sources (Figure 2) include leading journals in artificial intelligence and information retrieval, highlighting IEEE Access, Journal of Artificial Intelligence Research, and Information Processing and Management has major contributors [2], [8].

The worldwide citation assessment (Figures 4) recognizes well-regarded foundational articles that presented methods enhanced by retrieval for improvement generative dependability and factual foundation [3],[9]. These trends collectively indicate that RAG has evolved from a niche concept into a central pillar of knowledge-grounded AI research, supported by a diverse range of interdisciplinary publications.

B. Collaboration and Authorship Networks

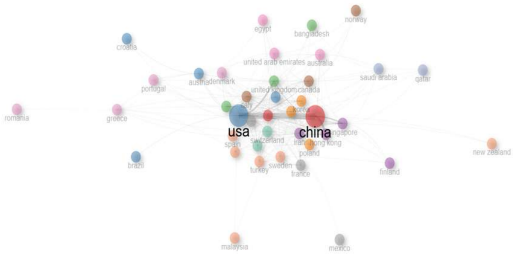

Figure 5: Country wise collaboration

Figure 6: Author wise collaboration.

The networked coloured plot of Figs. 5-7 indicates a widely distributed and cooperative research field. USA, China, UK and India emerge as leaders in both output and collaboration intensity. Cross-border collaborations, particularly, between the US and Asia-based researchers have expanded demonstrating the multidisciplinary character of the field [7], [10].

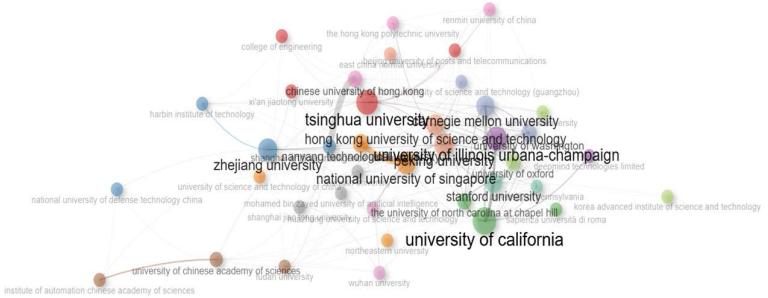

Figure 7: Institution wise collaboration

C. Intellectual Structure

The co-citation networks in Figures 8–10 below expose the intellectual backbone of RAG research. Major founding contributions in neural information retrieval and knowledge-grounded dialog systems form the early nucleus of citations; subsequent expansions occur with retrieval-enhanced large language models [8], [13].
Three main clusters can be identified:

- Retrieval and Representation Learning: The core methods include dense retrievers, hybrid search mechanisms, and dual encoder frameworks [4], [9].
- Generative Grounding and Factuality: Works on reducing hallucination by retrieval alignment and evidence integration [11], [14].
- Application-Oriented Models: Integration of RAG into domain-specific tasks, such as medicine, finance, and law [3], [16].

Significant advancements in retrieval-augmented architectures and open-domain question answering are shown in the recent fast increase in citations, which highlights a dynamic and constantly changing research scene.

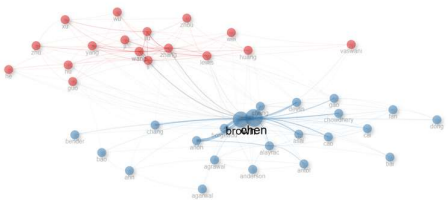

Figure 8: Co-citation Network (Author).

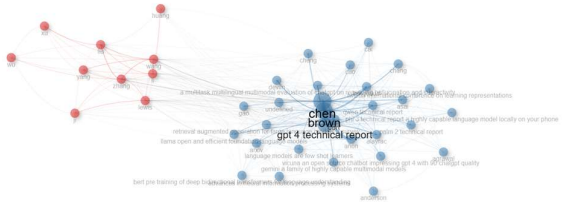

Figure 9: Co-citation Network (Author).

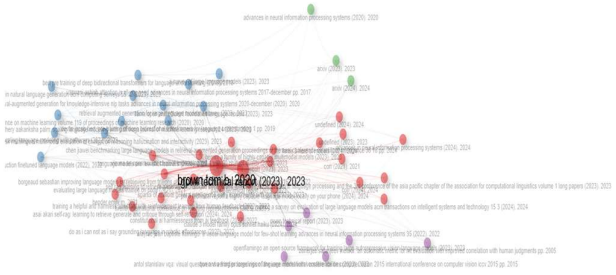

Figure 10: Co-citation Network (Paper)

D. Conceptual Structure and Emerging Themes

(RAG) research's conceptual development and topic patterns are collectively shown in Figures 11–22. Four main quadrants that represent the conceptual diversity of this field are highlighted by the thematic map (Figure-15) and its matching network representation (Figure 16) [3],[8], [14].

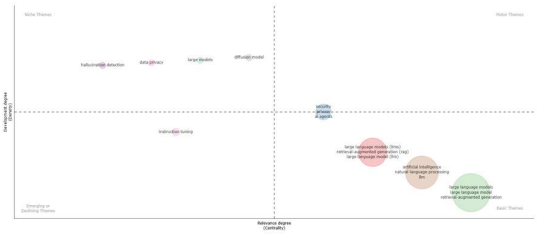

Figure 11: Thematic map

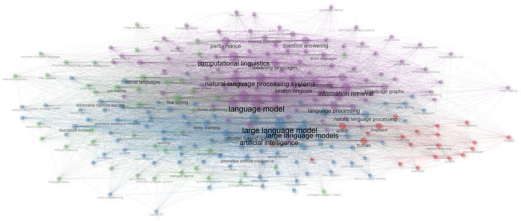

Figure 12: Thematic map network structure

The upper-right quadrant shows the three main driving themes—Contextual Retrieval, Knowledge Fusion, and Hallucination Mitigation—as well as their notable development and relationship. These topics form the basis of current RAG research, which primarily focuses on improving retrieval alignment, maintaining factual accuracy, and boosting generation's adaptability [4], [9], and [14]. On the other hand, Explainable Generation and Dynamic Memory Models appear to be more focused, smaller topics that have a big influence on the field [11], [13]. However, the lower-left quadrant displays themes like Static Embeddings and Rule-Based Retrieval, which show a clear shift from more conventional rule-based methods to more modern transformer-based learning strategies [2], [6].The network demonstrates the strong relationships between ideas such as retrieval systems, semantic base, and domain-specific adaptability [3], [8], and [10]. RAG has changed from a model-focused approach to comprehensive reasoning-driven frameworks, as seen by the close grouping of keywords associated with retrieval, reasoning, and interpretability. There is a conceptual link between language modelling and knowledge engineering. [9], [11]

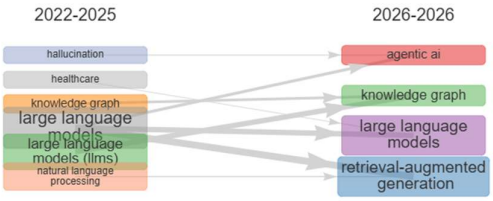

Figure 13: Thematic evolution.

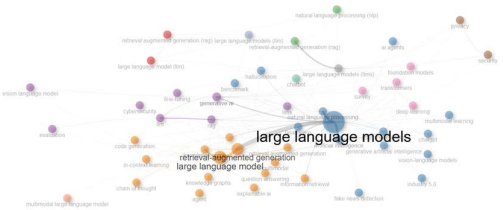

Figure 14: Co-occurrence network

Between 2018 to 2025 we can see a thematic shift [3],[12]. Between 2018–2020, research primarily was focused around retrieval and response generation, laying the fundamental foundation for RAG [4],[14]. During 2021–2023, focus shifted toward fact-grounded language modelling and knowledge consistency [8], [13]. The most present phase (2024–2025) focuses on contextualized, domain-adaptive, and multimodal RAG frameworks. This shift shows increasing recognition and practical application of the field [7], [17], and [18].

Keyword co-occurrence research represents a dense network based on "large language models," "retrieval augmented generation," "semantic search," and "knowledge graphs" [2], [8], and [14]. These clusters show the interplay between retrieval enhancement, semantic representation, and factual grounding [9], [13]. The closer connections between knowledge integration and semantic retrieval concludes that RAG is becoming more widely recognized as a cohesive paradigm that links reasoning and evidence retrieval. [10], [12].

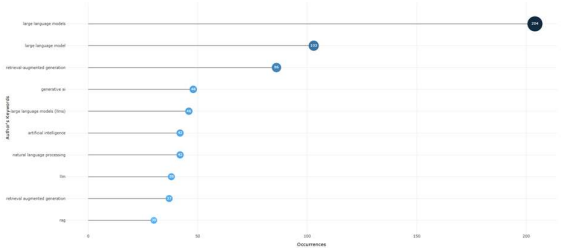

Figure 15: Most frequent words count.

Figure 16: Trend Topics.

Figures 15 and 16 display the most common terms and rising topics in the discipline. The frequency patterns and trend lines show that since 2022, phrases such as knowledge integration, context-aware production, and multimodal retrieval have grown in popularity. [4], [11], [15]. The widespread usage of terms like RAG and large language models is a sign of their ongoing significance in current research [9], [14]. Together, these developments show that how RAG has changed from a simple architectural improvement to a more comprehensive, knowledge-based foundation for AI [6], [10].

Figure 17: Tree Map.

Figure 18: Word Cloud.

Figure 17–18 Tree Map and word cloud visuals further show the significant concepts like retrieval pipelines, reasoning, domain adaptation, and knowledge fusion [8], [11], [13]. The repetition of these terms indicates the rising interest of the community in hybrid architectures and interpretable reasoning. The merging RAG into vision-language [8], edge AI [17], and multilingual settings [13] demonstrates its increasing adaptability and worldwide research interest. Terms like "large language models," "context retrieval," "retrieval augmented generation," "generative AI," "semantic search," and "knowledge graphs" are largely more prevalent [2], [8], and [14], as Figures 18–22 represents keyword co-occurrence analysis. This indicates a recent change in the field toward interpretable and scalable structures. The word cloud and tree map show the growing interest in fusing retrieval and reasoning, demonstrating RAG's growing importance beyond NLP into multimodal and cross-lingual applications [9], [11]. The advancement of RAG research is a prime example of the larger trend in AI research toward the construction of more reliable, verifiable, and explainable systems. The conceptual underpinning for establishing the retrieval systems that will improve generative models was the focus of early attempts. Hybrid reasoning systems with contextual adaption and real-time information access in the focus of recent work [5, 10, 13].

By combining expertise from machine learning, linguistics, and information retrieval, collaboration analysis highlights the value of interdisciplinary collaborations. Neural representation learning and symbolic thinking are continuously combining, according to the intellectual framework [6, 8, 15].

The rising trends practically point to the following opportunities:

- Better domain adaptation through customized retrieval.
- Improved factual accuracy using dynamic databases.
- Using RAG Frameworks with Multimodal Text, Image, and Code Data
- The development of efficient retrieval pipelines for edge and low-resource scenarios.

These guidelines demonstrate how RAG can alter the balance between memory and reasoning in future AI systems

III. CONCLUSIONS

Through its phases of development, networks of partnerships, and important issue discoveries, this comprehensive analysis offers a general overview of Retrieval-Augmented Generation research. These findings indicate a rapidly expanding, diversified, and better conceptualized research landscape that is becoming more intricate and interrelated. RAG has developed from a retrieval idea to a fundamental framework that supports contextually aware generative models and all factual data. The research indicates that the next phase of development will integrate retrieval systems into frameworks that can reflect and reason, with an emphasis on scalability, flexibility, and reliability [4], [11], [16].

In order to provide a clearer path for RAG and useful recommendations for advancing research in retrieval-grounded AI, this study unifies the conceptual and intellectual foundation for the AI field. These findings show that RAG is a leap step toward intelligent systems that can truly learn, reason, and integrate data into useful knowledge rather than merely being a technical breakthrough.

REFERENCES

[1] W. Jo et al., "BROCA: A 52.4-to-559.2 mW mobile social agent system-on-chip with adaptive bit-truncate unit and acoustic-cluster bit grouping," in Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), Feb. 2025, pp. 418–420.

[2] Y. Kumar et al., "A comprehensive review of AI advancement using testFAILS and testFAILS-2 for the pursuit of AGI," Electronics, vol. 13, no. 24, p. 4991, Dec. 2024.

[3] A. Vizniuk et al., "A comprehensive survey of retrieval-augmented large language models for decision making in agriculture: Unsolved problems and research opportunities," J. Artif. Intell. Soft Comput. Res., vol. 15, no. 2, pp. 115–146, Dec. 2024.

[4] A. Casheekar et al., "A contemporary review on chatbots and AI-powered virtual conversational agents: Applications, open challenges and future research directions," Comput. Sci. Rev., vol. 52, p. 100632, May 2024.

[5] J. Hong et al., "A context-aware onboarding agent for metaverse powered by large language models," in Proc. Designing Interactive Systems Conf., Jul. 2024, pp. 1857–1874.

[6] S. Naidu et al., "A data capture and gesture recognition system to enable human-robot collaboration," in Proc. Syst. Inf. Eng. Design Symp. (SIEDS), May 2025, pp. 375–380.

[7] O. Freyer et al., "A future role for health applications of large language models depends on regulators enforcing safety standards," Lancet Digit. Health, vol. 6, no. 9, pp. e662–e672, Sept. 2024.

[8] Y. Shi et al., "A learnable agent collaboration network framework for personalized multimodal AI search engine," in Proc. Int. Workshop Deep Multimodal Generation Retrieval, Oct. 2024, pp. 12–20.

[9] S. Han et al., "A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary integration, optimization techniques, applications, and challenges," Electronics, vol. 13, no. 24, p. 5040, Dec. 2024.

[10] X. Ren et al., "A survey of large language models for graphs," in Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2024, pp. 6616–6626.

[11] L. Xiang et al., "A survey of large language models in discipline-specific research: Challenges, methods and opportunities," Stud. Inform. Control, vol. 34, no. 1, pp. 5–24, Mar. 2025.

[12] H. Jeong et al., "A survey of robot intelligence with large language models," Appl. Sci., vol. 14, no. 19, p. 8868, Oct. 2024.

[13] N. Ibrahim et al., "A survey on augmenting knowledge graphs with large language models: Models, evaluation metrics, benchmarks, and challenges," Discov. Artif. Intell., vol. 4, no. 1, Nov. 2024.

[14] W. Fan et al., "A survey on RAG meeting LLMs: Towards retrieval-augmented large language models," in Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, Aug. 2024, pp. 6491–6501.

[15] Z. Zhao et al., "A survey on self-supervised graph foundation models: Knowledge-based perspective," IEEE Trans. Knowl. Data Eng., vol. 37, no. 8, pp. 4389–4410, Aug. 2025.

[16] M. Simaremare and H. Edison, "Accelerating new product development: A vision on active personas," in Lecture Notes Bus. Inf. Process., 2025, pp. 461–466.

[17] A. Hemmat et al., "Adaptive chunking for VideoRAG pipelines with a bilingual educational dataset," in Proc. Int. Comput. Conf. (CSICC), Feb. 2025, pp. 1–7.

[18] C. Yao and S. Fujita, "Adaptive control of retrieval-augmented generation for large language models through reflective tags," Electronics, vol. 13, no. 23, p. 4643, Nov. 2024.