International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

From Retrieval-Augmented Generation to Agentic AI : A Longitudinal Bibliometric and Thematic Mapping of Autonomous Knowledge-Driven AI Systems (2020-2026)

Dr. Sabyasachi Saha
Techno Exponent, India
Data extraction cutoff: 14 August 2026

Abstract - Retrieval-Augmented Generation (RAG) has moved rapidly from a retrieval–generation architecture for grounding language models toward adaptive, tool-using and increasingly autonomous systems. This study quantifies that transition through a longitudinal bibliometric and thematic mapping of scholarly records published between 1 January 2020 and 14 August 2026. A reproducible OpenAlex search tracked an RAG family ("retrieval augmented generation", "advanced RAG", and "agentic RAG") and an agent family ("AI agent", "LLM agent", "language model agent", "tool-using language model", "autonomous AI agent", "agentic AI", "multi-agent LLM", and "multi-agent AI"), while "agentic RAG" was separately monitored as an explicit bridge concept. The combined corpus contains 73,862 OpenAlex records across all document types. Annual volume rose from 261 records in 2020 to 19,313 in 2025 and 47,588 in the partial 2026 window. RAG-family output accelerated first, reaching 2,971 records in 2024 and temporarily exceeding the agent-family count of 2,184; the balance then shifted toward agent terminology in 2025–2026, with 34,698 agent-family records already indexed by the 2026 cutoff. The agentic-RAG bridge expanded from 16 records in 2024 to 192 in 2025 and 522 by mid-August 2026. Preprints constitute 39.6% of the corpus, while journal articles and conference papers together account for 41.9%. Thematic mapping shows a transition from retrieval/search-centered work toward multi-agent systems, ethics, robustness, explainability, security, trust, and applied AI. The results support a three-stage interpretation: retrieval grounding, agentic orchestration, and a current shift toward reliable and governed autonomous systems. The paper concludes with a research agenda centered on trajectory-level evaluation, trust and security, memory governance, cost-aware model routing, multi-agent coordination, human oversight, and enterprise-grade benchmarks.

Index Terms - Agentic AI, Agentic RAG, bibliometric analysis, large language models, multi-agent systems, OpenAlex, retrieval-augmented generation, scientometrics, thematic mapping.

I. INTRODUCTION

Large language models (LLMs) have shifted the design space of artificial intelligence from task-specific prediction toward general-purpose language interfaces capable of reasoning over instructions, generating plans, invoking tools, and interacting with external information systems. Yet the knowledge encoded in model parameters is static, difficult to update, and not inherently attributable to an external source. Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [1], addressed this limitation by combining parametric generation with non-parametric retrieval. The resulting architecture made it possible to ground responses in an external corpus, update knowledge without retraining the full model, and provide a path toward provenance for knowledge-intensive generation.

The subsequent development of RAG has not been linear. Research moved from fixed retrieve-then-generate pipelines toward modular and adaptive architectures. Surveys of RAG distinguish naive, advanced, and modular paradigms [2], while methods such as Self-RAG [3] and Corrective RAG [4] introduced reflection, selective retrieval, retrieval-quality assessment, and corrective behavior. In parallel, the broader LLM literature developed mechanisms for reasoning and action. ReAct interleaved reasoning traces with environment actions [5], Toolformer investigated learned API use [6], and Reflexion showed that agents can improve behavior through linguistic feedback and memory without updating model weights [7]. These lines of work progressively weakened the boundary between information retrieval and autonomous action.

The term agentic RAG has emerged to describe architectures in which an agent or collection of agents dynamically decides what to retrieve, how to reformulate queries, whether evidence is sufficient, which tools to invoke, and how to iterate toward a goal. Recent surveys explicitly frame Agentic RAG as an evolution beyond static RAG and emphasize reflection, planning, tool

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

use, and multi-agent collaboration [8]. Meanwhile, the LLM-agent literature has developed its own taxonomies of agent construction, memory, planning, action, multi-agent coordination, and evaluation [9]–[11]. The result is not simply two parallel literatures. Retrieval has increasingly become a capability embedded inside a broader autonomous control loop, while agentic systems increasingly depend on retrieval to ground planning and decision-making.

This convergence is visible qualitatively in system architectures, but its longitudinal bibliometric signature has received less attention. A 2026 bibliometric study mapped the intellectual landscape of RAG as a standalone field [12], and domain-specific work has combined bibliometric and systematic analysis of LLM, RAG, and agentic workflows in clinical decision support [13]. These studies are valuable, but they do not directly quantify the field-level transition from retrieval-centric language systems to agent-oriented autonomous AI across disciplines. The present study addresses that gap by treating RAG and agentic AI as interacting concept families and by explicitly tracking "agentic RAG" as a bridge term.

The analysis is based on OpenAlex, a fully open scholarly knowledge graph that links works, authors, venues, institutions, topics, and citations [14]. The dataset covers the fixed interval from 1 January 2020 through 14 August 2026. Because 2026 is incomplete, all 2026 values are treated as partial-year observations and are never annualized in the primary results. The analysis emphasizes reproducibility: exact query strings, term families, document-type distributions, topic groupings, venues, countries, and sampled highly cited works are contained in the accompanying workbook.

This study contributes four elements. First, it provides a longitudinal empirical mapping of the RAG-to-agentic-AI transition using 73,862 OpenAlex records. Second, it distinguishes concept-family volume from unique combined-corpus volume, which is important because records can match both families. Third, it interprets thematic and dissemination patterns rather than treating publication counts as sufficient evidence of intellectual change. Fourth, it develops a future research agenda that connects the bibliometric findings with recent technical surveys on Agentic RAG, autonomous agents, multi-agent systems, reliability, and security.

A. Research Questions

RQ1: How did publication volume associated with RAG and agent-oriented AI change from 2020 to the fixed 2026 cutoff?

RQ2: How did the relative balance between retrieval-centric and agent-centric terminology change, and how did the explicit "agentic RAG" bridge concept evolve?

RQ3: Which document types, research topics, dissemination venues, countries, and influential works characterize the corpus?

RQ4: What research trajectory can be inferred from the combined quantitative and thematic evidence, and which open problems define the next phase of autonomous knowledge-driven AI?

II. CONCEPTUAL BACKGROUND AND RELATED WORK

A. From Foundational RAG to Adaptive Retrieval

RAG combines a retriever with a generator so that the generative model conditions its answer on externally retrieved passages. The original formulation demonstrated advantages on knowledge-intensive natural-language tasks and established a general architecture in which non-parametric memory can complement parametric model knowledge [1]. As LLMs became more capable, RAG was increasingly adopted as a practical approach for factual grounding, domain adaptation, and the integration of proprietary or frequently changing knowledge.

The field subsequently diversified. Gao et al. organized RAG research into naive, advanced, and modular paradigms [2]. Advanced and modular systems optimize retrieval, reranking, context construction, query transformation, generation, and feedback. Self-RAG introduced learned reflection behavior that can trigger retrieval on demand and critique both retrieved evidence and generated output [3]. Corrective RAG explicitly evaluates retrieval quality and can initiate corrective knowledge-acquisition actions when retrieval is unreliable [4]. Active Retrieval Augmented Generation similarly investigates retrieval that is triggered dynamically during generation rather than performed once at the beginning of the pipeline [15]. RAGAS and related evaluation work shifted attention from end-task accuracy toward retrieval relevance, answer faithfulness, and other component-level quality dimensions [16]. Benchmarking studies further established that RAG performance depends strongly on model, retriever, corpus, task, and evaluation protocol [17], [18].

B. From Tool Use to LLM Agents

Agentic AI extends language generation with an action loop. ReAct demonstrated that reasoning and acting can be interleaved so that observations from tools or environments update subsequent reasoning [5]. Toolformer explored how a language model can learn when to call external APIs and how to incorporate the resulting observations [6]. Reflexion added a memory of verbal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

feedback, enabling an agent to adapt future behavior from prior failures [7]. AgentBench later showed that interactive agent performance depends on long-horizon reasoning, decision-making, and instruction following, and that conventional LLM benchmarks do not adequately capture these properties [19].

Survey literature now treats memory, planning, tool use, perception, action, feedback, and multi-agent communication as recurring architectural components of LLM agents [9], [10]. Multi-agent research adds role specialization, communication protocols, collaboration, debate, task decomposition, and emergent social behavior [11]. These mechanisms create capabilities that static RAG alone does not provide, but they also introduce new failure modes, including error propagation across steps, unsafe tool use, context drift, memory contamination, coordination failure, and non-deterministic trajectories.

C. Agentic RAG as a Bridge Paradigm

Agentic RAG combines these two trajectories. Singh et al. characterize Agentic RAG as the embedding of autonomous agent behaviors inside the retrieval pipeline, including reflection, planning, tool use, and multi-agent collaboration [8]. More recent systematization work models Agentic RAG as a sequential decision process and emphasizes planning, memory, retrieval orchestration, tool invocation, trajectory evaluation, reliability, and oversight [20]. Reasoning-centric reviews similarly describe a shift from static retrieval toward systems that interleave reasoning and retrieval dynamically [21].

The significance of Agentic RAG is therefore architectural rather than terminological. Retrieval ceases to be a single pre-generation step and becomes an action that can be invoked repeatedly, conditionally, and strategically by an agent. Conversely, the agent becomes knowledge-grounded through the retrieval layer. The combined paradigm is particularly relevant in enterprise settings, where agents must work with proprietary documents, operational databases, policies, customer records, and tools while maintaining traceability and control.

D. Prior Bibliometric and Review Studies

Bibliometric analysis is used to map the structure, growth, influence, and thematic organization of research fields. Donthu et al. provide widely used methodological guidance covering performance analysis and science mapping [22], while VOSviewer [23] and bibliometrix [24] are established tools for network visualization and quantitative science mapping. OpenAlex has expanded the feasibility of reproducible open bibliometrics by providing a freely accessible scholarly graph and API [14].

The rapid growth of RAG has already motivated dedicated bibliometric research. Afarin et al. analyzed the intellectual landscape of RAG and documented the emergence of a recognizable scientometric field [12]. Domain-specific work has also studied the intersection of LLMs, RAG, and agentic workflows in clinical decision support [13]. At the same time, technical surveys now cover autonomous LLM agents [9], Agentic RAG [8], multi-agent systems [11], agent security [25], and increasingly specialized issues such as reasoning-driven retrieval and trustworthy orchestration [20], [21]. The present study complements those efforts by using bibliometric evidence to characterize the longitudinal transition linking these literatures rather than reviewing only one architecture or one application domain.

III. RESEARCH DESIGN

A. Data Source and Fixed Observation Window

All numerical results in this study are derived from OpenAlex. The extraction date was 14 August 2026, with a fixed publication window of 1 January 2020 through 14 August 2026. OpenAlex was selected because it was the only scholarly index that could be queried reliably in the working environment and because it provides structured aggregation by publication year, source, topic, work type, and institutional country. Scopus and Web of Science were not available through institutional subscription in the extraction environment, while direct bulk access to IEEE Xplore and ACM Digital Library was not available. Attempts to use Semantic Scholar, DBLP, and the arXiv export interface were rate-limited or blocked in that environment. The single-database design is therefore a material limitation rather than an intentional claim that OpenAlex is interchangeable with proprietary indexes.

OpenAlex is nevertheless a defensible source for exploratory bibliometric mapping. Priem et al. describe it as a fully open scholarly knowledge graph with works linked to authors, venues, institutions, and concepts [14]. Its openness makes the search strategy reproducible, but its automated entity resolution and topic assignment can generate metadata artifacts. The analysis therefore distinguishes aggregate trends, which are relatively robust to occasional record errors, from individual citation or venue claims, which require spot verification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

B. Search Vocabulary and Operational Query

The starting vocabulary was designed to capture both retrieval-centric and agent-centric terminology: “retrieval augmented generation”, “retrieval-augmented generation”, RAG, “advanced RAG”, “agentic RAG”, “AI agent”, “LLM agent”, “language model agent”, “tool-using language model”, “autonomous AI agent”, “agentic AI”, “multi-agent LLM”, and “multi-agent AI”. The bare acronym RAG was excluded from the operational OpenAlex phrase search because it is highly ambiguous outside the generative-AI context and would introduce unrelated red-amber-green, biological, and name-fragment matches. The remaining terms were grouped into two concept families, with “agentic RAG” separately tracked as a bridge term.

Concept group	Operational phrases
RAG-family	“retrieval augmented generation”; “advanced RAG”; “agentic RAG”
Agent-family	“AI agent”; “LLM agent”; “language model agent”; “tool-using language model”; “autonomous AI agent”; “agentic AI”; “multi-agent LLM”; “multi-agent AI”
Bridge	“agentic RAG” (also contained in RAG-family)

TABLE I. Concept families used in the operational search.

The RAG-family filter was executed against OpenAlex title_and_abstract.search using the three RAG phrases; the agent-family filter used the eight agent phrases; and the bridge query used “agentic RAG” alone. Every query included the same fixed date constraints. OpenAlex group_by operations were used for publication year, primary source, primary topic, institutional country code, and work type. Individual term totals were obtained through term-specific searches. Highly cited samples were restricted to articles and conference papers and sorted by cited_by_count.

C. Corpus Accounting and Overlap

The combined query returned 73,862 unique OpenAlex records for the full observation period. The RAG-family and agent-family totals are not additive because a work may mention terms from both families. This distinction matters for interpretation. Let R_t and A_t denote yearly RAG-family and agent-family counts, and C_t the unique combined count. The cross-family overlap is computed descriptively as $X_t = R_t + A_t - C_t$. X_t does not measure semantic integration directly, but it provides a transparent count of records that match both families under the operational search.

For relative framing after the LLM-agent transition, a family-normalized agent orientation can be written as $O_t = A_t/(R_t + A_t)$. This measure is used only descriptively and only with explicit caution: early “AI agent” literature includes robotics, marketing, negotiation, and other pre-LLM traditions, so high agent-family shares in 2020–2022 must not be interpreted as evidence of modern Agentic AI.

D. Bibliometric Dimensions

The analysis uses five descriptive dimensions: (1) annual publication volume; (2) concept-family and individual-term frequency; (3) document type; (4) OpenAlex primary-topic assignment as a scalable proxy for thematic structure; and (5) dissemination and participation, represented by hosting venue and institution-linked country counts. A sample of highly cited article and conference-paper records is used to identify intellectual anchors and diagnose search noise. Citation counts are treated as time-dependent OpenAlex values at the fixed extraction date, not as immutable properties of the papers.

No manual title/abstract screening or cross-database deduplication was performed because the corpus was generated from one index using server-side aggregate grouping. Accordingly, this work is a bibliometric and thematic mapping study rather than a PRISMA-compliant systematic review. This distinction is deliberate: the goal is to quantify a broad terminological and intellectual transition, not to make exhaustive claims about the effectiveness of individual systems.

E. Validity Safeguards

Four safeguards were applied. First, 2026 is labeled as partial in every temporal interpretation; no full-year annualization is used. Second, the ambiguous bare acronym RAG was excluded from the operational OpenAlex search. Third, document-type composition is reported explicitly so that the unusually large preprint share can be evaluated rather than conflated with peer-reviewed output. Fourth, anomalous highly cited records are not used as evidence of field influence without independent verification. One conspicuous RAG-family record in the supplied top-cited sample was flagged as a likely OpenAlex merge artifact and is excluded from the intellectual-anchor discussion below.

IV. RESULTS

A. Overall Publication Growth (RQ1)

The combined corpus grows from 261 records in 2020 to 343 in 2021, 398 in 2022, 905 in 2023, 5,054 in 2024, and 19,313 in 2025. By 14 August 2026, OpenAlex had already indexed 47,588 matching records for the partial year. The complete-year 2020–

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

2025 series therefore increased by approximately 74-fold, corresponding to a descriptive compound annual growth rate of roughly 136.5% over those five intervals. This rate should be read as a field-emergence indicator rather than a forecast because the underlying terminology and indexing environment change over time.

The corpus is strongly preprint-oriented, but formal publication is also substantial. Across the full 2020–2026 cutoff window, OpenAlex classifies 18,689 records as journal articles and 12,271 as conference papers, for 30,960 records (41.9% of the 73,862-record corpus). Because the supplied aggregate workbook does not contain a verified year-by-year cross-tabulation of document type, this study does not claim a separate annual growth curve for peer-reviewed output; the document-type analysis is used only to characterize corpus composition.

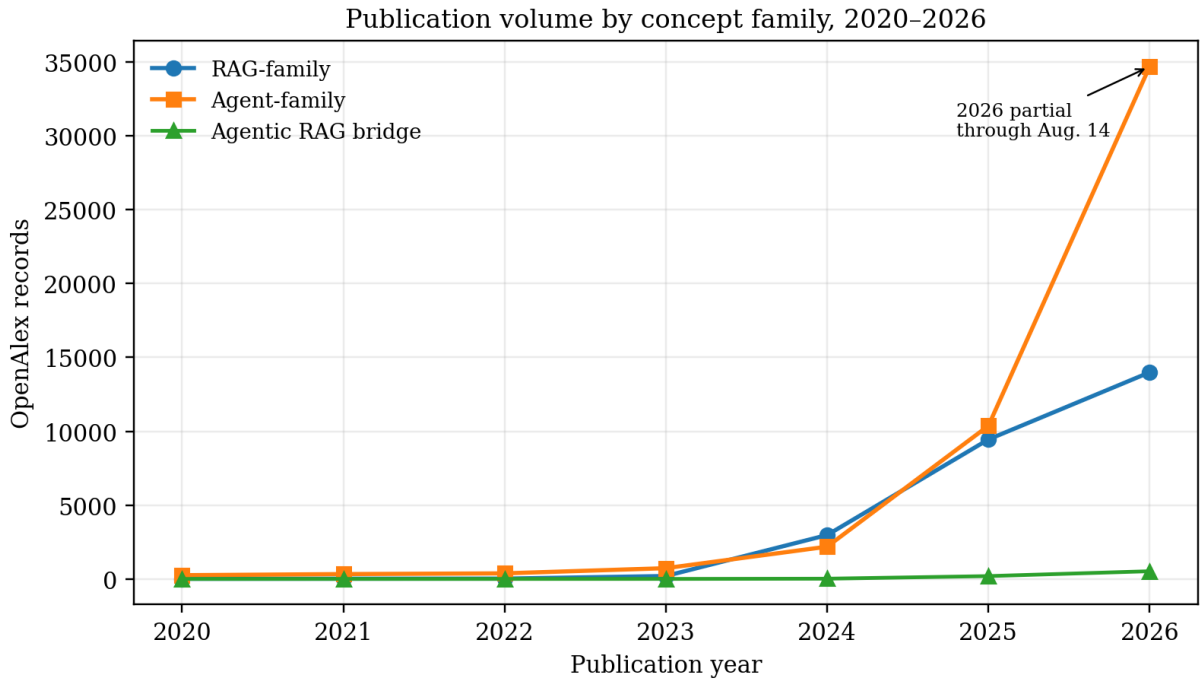

Fig. 1. Publication volume by concept family, 2020–2026. The 2026 observation is partial through 14 August. Source: OpenAlex extraction supplied with this study.

Year	RAG-family	Agent-family	Bridge	Unique combined
2020	9	252	0	261
2021	15	328	1	343
2022	21	377	0	398
2023	198	728	2	905
2024	2,971	2,184	16	5,054
2025	9,444	10,368	192	19,313
2026*	13,991	34,698	522	47,588

*TABLE II. Yearly all-document-type record counts. *2026 is partial through 14 August.*

B. The RAG-to-Agentic Shift and Cross-Family Convergence (RQ2)

The concept families exhibit different inflection points. RAG-family output remains very small through 2022, reaches 198 records in 2023, and then jumps to 2,971 in 2024. At that point it exceeds the agent-family count of 2,184. This is the clearest bibliometric signature of the RAG expansion phase. In 2025, the agent family again becomes larger (10,368 versus 9,444), and by the partial 2026 window the gap widens substantially (34,698 versus 13,991). The family-normalized agent orientation is therefore 42.4% in 2024, 52.3% in 2025, and 71.3% in partial 2026. These post-2023 values are consistent with a shift from retrieval-centered framing toward agent-centered framing.

The early period requires a different interpretation. Agent-family terminology accounts for more than 94% of family-normalized matches in 2020–2022, but the highly cited titles and topic composition show that much of this literature refers to older uses of “AI agent” in marketing, robotics, negotiation, and information systems rather than LLM-based autonomous agents. The corpus therefore contains a legacy-agent baseline that predates the modern Agentic AI meaning. The 2024–2026 transition is more informative because it coincides with the rapid emergence of “agentic AI”, “LLM agent”, and “agentic RAG” terminology.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Absolute cross-family overlap also rises sharply. X_t is zero in 2020–2022, 21 records in 2023, 101 in 2024, 499 in 2025, and 1,101 in the partial 2026 window. As a fraction of the unique combined corpus, the overlap remains near 2–2.6% after 2023; thus the evidence supports increasing absolute convergence but not a claim that the proportion of cross-family papers is accelerating monotonically. This distinction avoids overstating the bridge phenomenon.

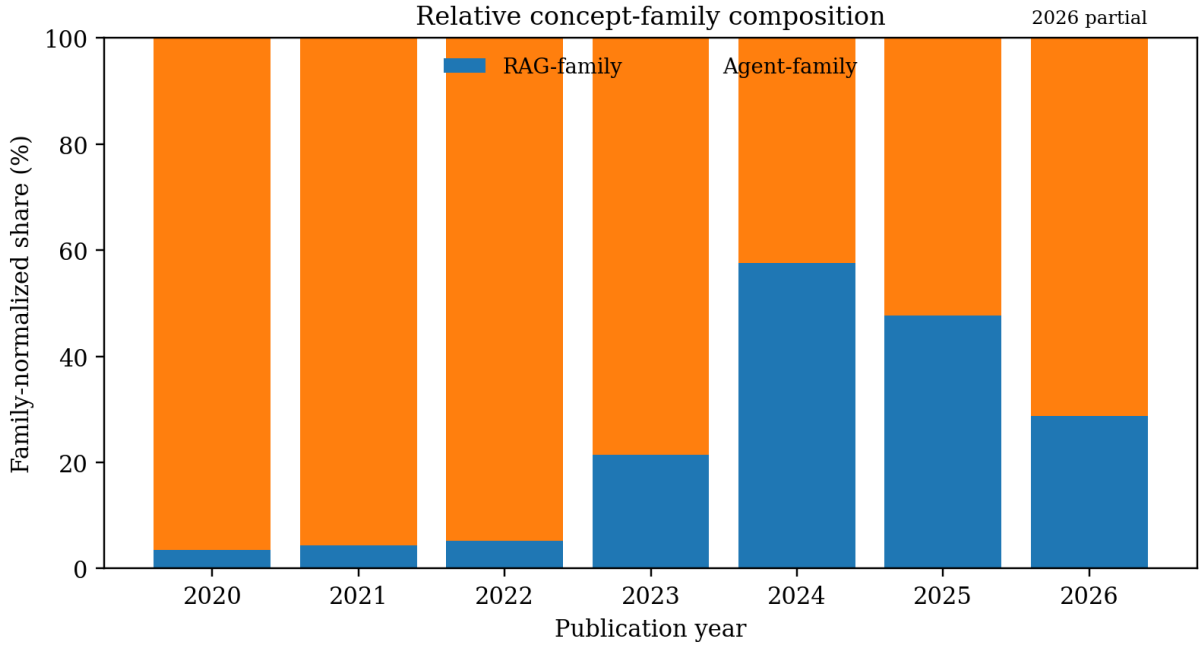

Fig. 2. Family-normalized composition. Percentages use RAG-family + agent-family as the denominator and therefore show relative framing rather than unique-corpus shares. Early agent values include legacy pre-LLM agent literature.

C. Term-Level Structure

The most frequent individual phrase is “retrieval augmented generation” with 26,376 records, followed by “AI agent” with 23,883, “agentic AI” with 14,366, and “LLM agent” with 11,368. The newer agent-specific terms are therefore already large at the fixed 2026 cutoff. By contrast, explicit compound phrases remain much smaller: “agentic RAG” yields 733 records, “advanced RAG” 150, and “tool-using language model” 51. These counts must not be summed because a single paper can match multiple phrases.

The difference between conceptual importance and literal phrase frequency is notable. “Agentic RAG” is a bridge label with fewer than 1,000 records, yet many systems that perform planning, iterative retrieval, reflection, and tool use may be described with other terminology. Similarly, “tool-using language model” is rare as an exact phrase even though tool use is central to the agent literature. Bibliometric phrase counts therefore trace naming conventions as well as technical content.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

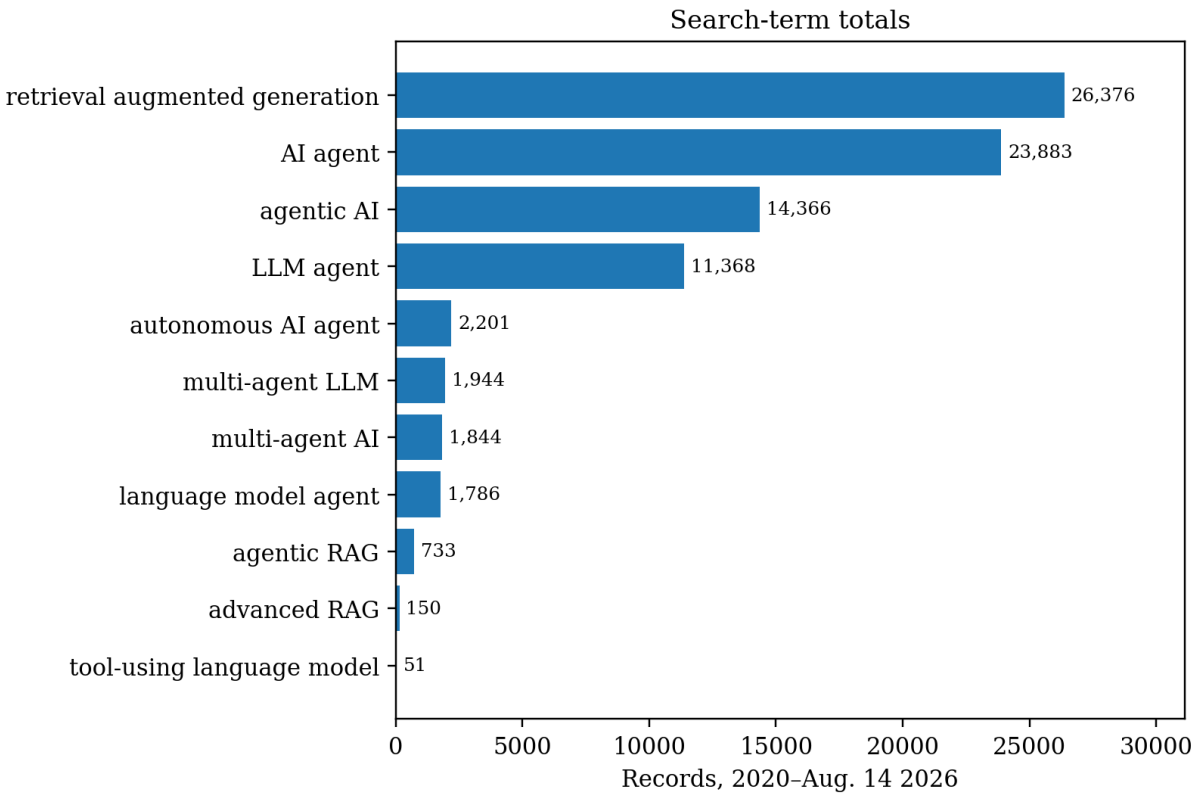

Fig. 3. Individual search-term totals for 2020–14 August 2026. A record may match more than one term.

D. Document-Type Composition and Publication Culture

Preprints are the largest document category, with 29,273 records (39.6% of the unique combined corpus). Journal articles account for 18,689 records (25.3%) and conference papers for 12,271 (16.6%). Articles and conference papers together therefore represent 30,960 records, or 41.9% of the corpus. The remaining records include software, datasets, dissertations, book chapters, reports, abstracts, reviews, and other types.

This distribution reflects the rapid dissemination cycle of contemporary AI research. arXiv and Zenodo dominate the source-level hosting results, while peer-reviewed outlets such as Lecture Notes in Computer Science, AAAI proceedings, IEEE Access, and Applied Sciences appear further down the ranking. The prevalence of preprints is methodologically important because it can inflate the apparent speed of field expansion relative to slower peer-review pipelines. At the same time, journal articles and conference papers together constitute 41.9% of the corpus, showing that the mapped literature includes a substantial formally published component even though the present aggregate data do not support a year-by-year peer-reviewed growth claim.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

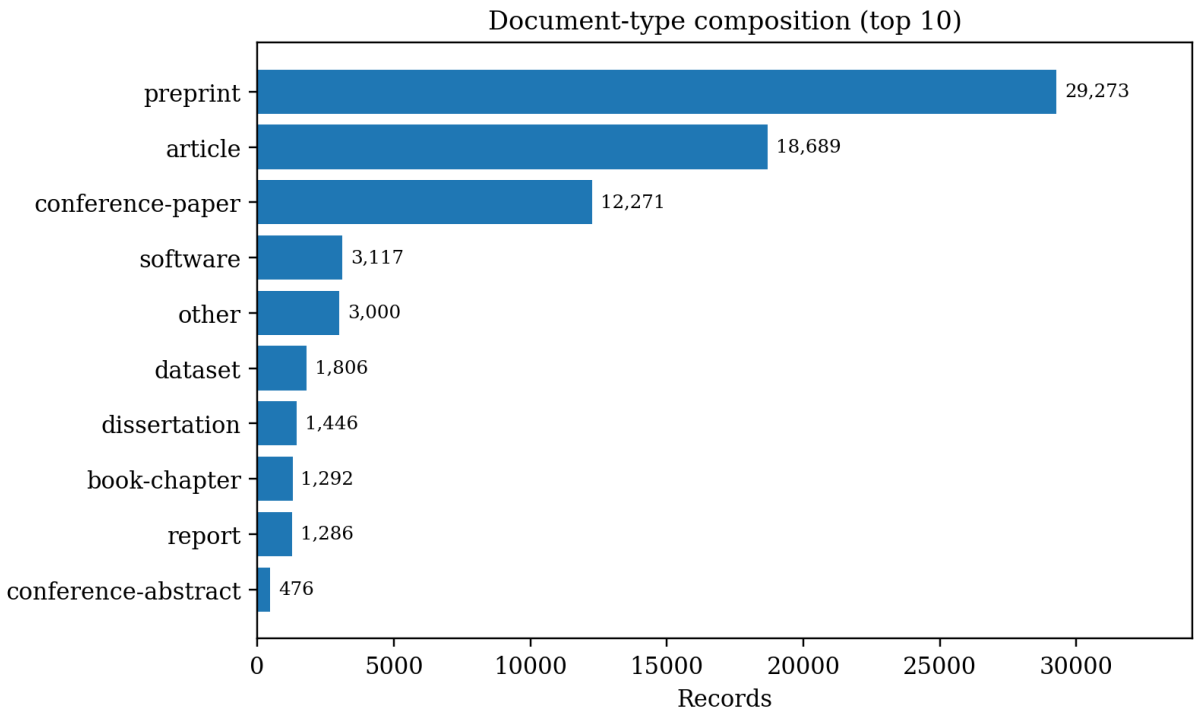

Fig. 4. Document-type composition of the combined corpus (top 10 categories).

E. Thematic Mapping

OpenAlex primary topics provide a scalable, though imperfect, thematic view. Across the combined corpus, the leading topics are Topic Modeling (6,268 records), Ethics and Social Impacts of AI (3,034), Multi-Agent Systems and Negotiation (2,965), Artificial Intelligence in Healthcare and Education (2,014), AI in Service Interactions (1,977), Multimodal Machine Learning Applications (1,925), Adversarial Robustness in Machine Learning (1,527), Scientific Computing and Data Management (1,320), Explainable Artificial Intelligence (1,311), and Advanced Graph Neural Networks (1,011).

The two concept families differ meaningfully. The RAG subset is dominated by Topic Modeling (4,699), AI in Service Interactions (1,031), Information Retrieval and Search Behavior (877), AI in Healthcare and Education (867), Multimodal Machine Learning Applications (848), and Advanced Graph Neural Networks (811). The agent subset is led by Ethics and Social Impacts of AI (2,897), Multi-Agent Systems and Negotiation (2,867), Topic Modeling (1,707), Adversarial Robustness (1,323), Healthcare and Education (1,245), multimodal applications (1,128), Explainable AI (1,077), and service interactions (978). Agent-family topics additionally include Security and Verification in Computing (581), Access Control and Trust (561), and Software System Performance and Reliability (533).

These contrasts support an interpretation in which RAG research remains closely tied to retrieval, knowledge organization, graph methods, and domain grounding, while agent research allocates more attention to coordination, ethics, robustness, trust, and reliability. The Agentic RAG bridge contains smaller but strategically relevant topic counts for AI-based problem solving and planning, multi-agent systems, access control and trust, multimodal systems, and information retrieval. Because OpenAlex assigns topics algorithmically, the analysis should be treated as thematic indication rather than human-coded content analysis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

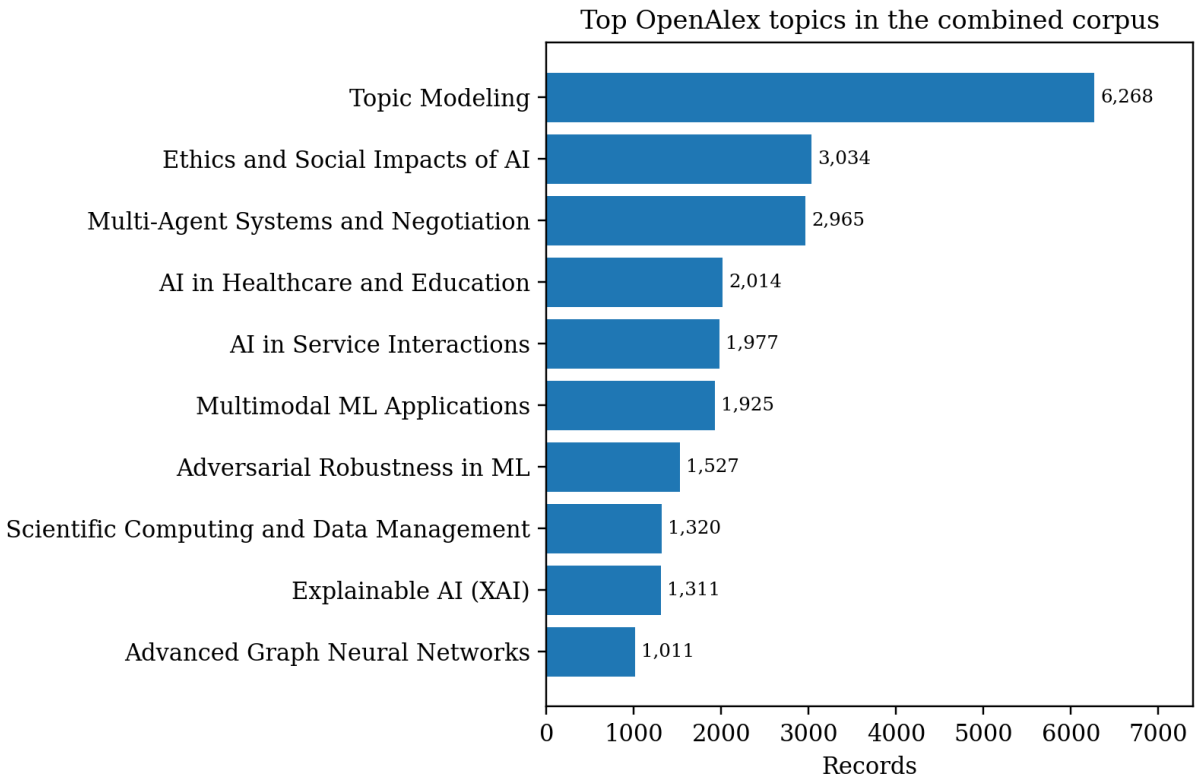

Fig. 5. Top OpenAlex primary topics in the combined corpus.

F. Venues and Dissemination

The combined venue/repository ranking is led by arXiv with 18,344 records and Zenodo with 16,325, followed by Open MIND (4,182) and SSRN Electronic Journal (2,706). Lecture Notes in Computer Science contributes 877 records, Figshare 682, Underline Science 608, Research Square 459, Preprints.org 392, Communications in Computer and Information Science 359, Lecture Notes in Networks and Systems 299, AAAI proceedings 262, and IEEE Access 210. arXiv and Zenodo together account for 34,669 primary-location records, equivalent to 46.9% of the unique combined corpus, illustrating the importance of repository-first dissemination.

The source ranking should not be read as a journal-quality ranking. OpenAlex primary locations mix repositories, journals, proceedings, and other hosts. Instead, the pattern indicates that the RAG/agentic-AI literature is disseminated through a hybrid ecosystem in which preprints, repositories, conferences, and journals coexist. For an emerging AI field, this has consequences for citation timing, versioning, duplicate manifestations, and reproducibility.

G. Geographic Distribution

Affiliation-linked counts are led by the United States (12,684), China (5,694), the United Kingdom (3,118), India (2,966), Canada (2,078), Germany (1,702), Mexico (1,142), South Korea (942), Italy (910), Japan (856), Australia (851), Russia (786), Hong Kong (691), Singapore (672), and France (671). Country counts are not exclusive because a multinational paper can contribute to multiple institutional countries. They therefore indicate participation rather than mutually exclusive national shares.

The concept-family split is also informative. The RAG family has particularly strong counts from the United States (3,889), China (3,093), India (1,529), and the United Kingdom (978), whereas the agent family is more heavily led by the United States (9,152) followed by China (2,743), the United Kingdom (2,278), India (1,568), and Canada (1,490). The pattern suggests broad international participation with different regional emphases across retrieval-centric and agent-centric work.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

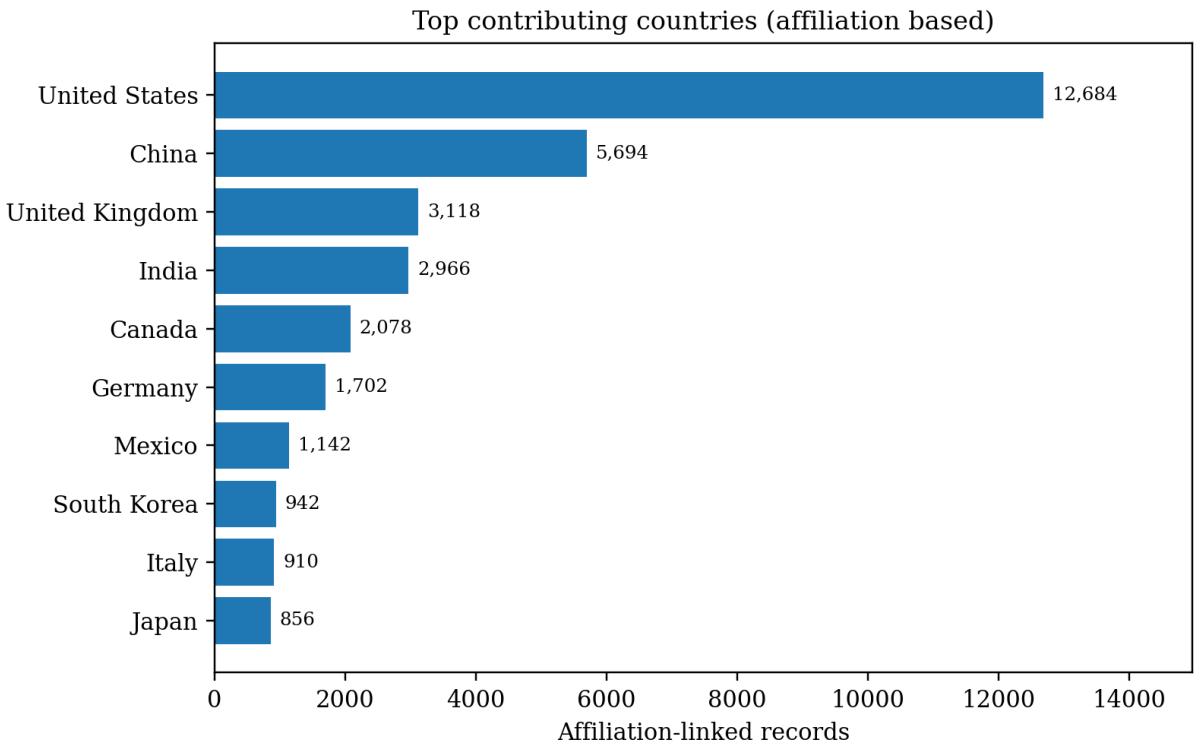

Fig. 6. Top contributing countries by institutional affiliation. Counts are non-exclusive across multinational papers.

H. Intellectual Anchors and Search-Noise Diagnosis

The top-cited sample contains recognizable RAG intellectual anchors. Fan et al.'s survey of RAG and LLMs [26] had 683 OpenAlex citations at the cutoff, Active Retrieval Augmented Generation [15] had 413, RAGAS [16] had 376, the AAAI RAG benchmarking paper [17] had 360, domain-adaptation work by Siriwardhana et al. [27] had 301, and medical RAG benchmarking [18] had 247. These works indicate a field moving beyond the introduction of RAG toward evaluation, benchmarking, adaptation, and domain-specific validation.

The agent-family top-cited sample illustrates both intellectual transition and search ambiguity. A 2025 Agentic AI survey in IEEE Access [28] had 537 OpenAlex citations at the cutoff, while the survey on the rise and potential of LLM-based agents [29] had 516. At the same time, high-citation records from 2020–2022 include marketing, customer-service, robotics, and information-system studies that use "AI agent" in older senses. Rather than deleting these records post hoc, the study uses them as evidence that the semantic history of "agent" predates LLM-based agency and must be controlled in interpretation.

Strand	Representative work	Year	OpenAlex cites*
RAG	A Survey on RAG Meeting LLMs [26]	2024	683
RAG	Active Retrieval Augmented Generation [15]	2023	413
RAG	RAGAS [16]	2024	376
RAG	Benchmarking LLMs in RAG [17]	2024	360
RAG	Domain Adaptation of RAG [27]	2023	301
RAG	RAG for Medicine [18]	2024	247
Agent	Agentic AI: Autonomous Intelligence… [28]	2025	537
Agent	Rise and Potential of LLM-Based Agents [29]	2025	516
Bridge	Agentic RAG survey [8]	2025	Corpus identifies bridge

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Strand	Representative work	Year	OpenAlex cites*
			literature

*TABLE III. Representative intellectual anchors. *Citation counts are OpenAlex values at the 14 August 2026 cutoff where supplied by the dataset.*

V. DISCUSSION

A. A Three-Stage Evolution: Grounding, Orchestration, Governance

Taken together, the temporal and thematic evidence supports a three-stage interpretation. The first stage is retrieval grounding. RAG begins as a mechanism for combining model generation with external non-parametric knowledge [1] and then expands into a broad design space of retrieval optimization, modularity, adaptive retrieval, reflection, and evaluation [2]–[4]. The bibliometric inflection is visible in 2023–2024, when RAG-family publication volume rises from 198 to 2,971 records and briefly exceeds the agent-family count.

The second stage is agentic orchestration. Tool use, reasoning-action loops, memory, reflection, and multi-agent coordination provide control mechanisms that can call retrieval as one action among many [5]–[11]. The agent-family resurgence in 2025 and especially 2026 is consistent with this broader framing. Crucially, the paper does not infer this transition from the raw "AI agent" count alone; the whole-period prominence of the more specific terms "agentic AI" and "LLM agent", the rise of the explicit Agentic RAG bridge, and recent survey literature collectively support the interpretation.

The third stage is reliability and governance. The topic distribution within the agent family includes ethics, adversarial robustness, explainability, security and verification, access control and trust, and software reliability. Recent Agentic RAG and agent-security surveys likewise emphasize cascading errors, memory poisoning, tool-execution vulnerabilities, evaluation inconsistency, and oversight [20], [25]. As autonomous systems gain the ability to change external state, reliability becomes a systems property rather than only a factuality property.

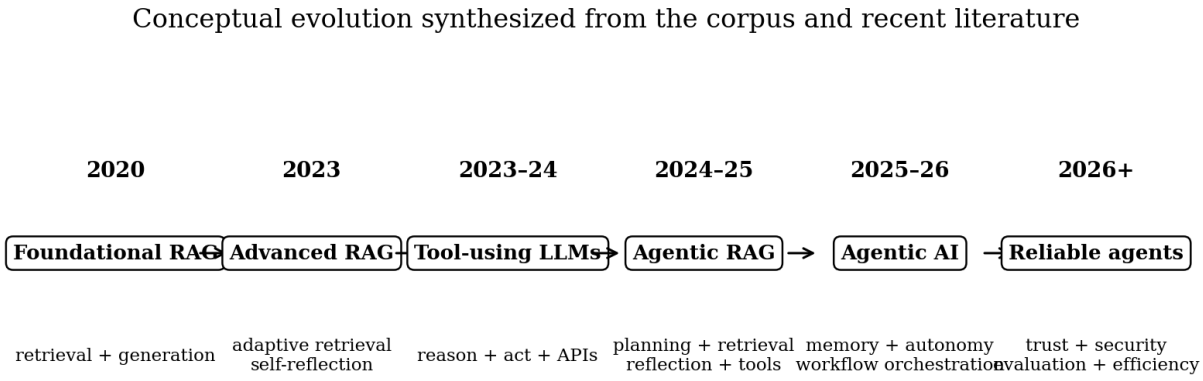

Fig. 7. Conceptual evolution synthesized from the bibliometric evidence and recent technical literature. Dates indicate dominant emergence, not exclusive boundaries.

B. Agentic RAG Is a Bridge, Not Necessarily the Dominant Label

The phrase "agentic RAG" accounts for only 733 records across the full period, much less than "retrieval augmented generation", "agentic AI", or "LLM agent". This does not imply that Agentic RAG is unimportant. Bridge concepts can be intellectually influential before a stable label becomes dominant. Many systems that dynamically retrieve, plan, call tools, and reflect may be published under RAG, LLM-agent, reasoning, multi-agent, or application-specific terminology. Accordingly, the 733-record bridge count is best interpreted as a lower-bound indicator of explicit naming rather than a census of all agentic retrieval architectures.

The cross-family overlap statistic reinforces this view. The absolute number of records matching both concept families grows from 21 in 2023 to 1,101 in partial 2026, yet the overlap remains only about 2–2.6% of the unique combined corpus. The two literatures are therefore interacting but not collapsing into a single vocabulary. A useful future bibliometric study could apply record-level embedding or citation-network clustering to identify implicit bridges that phrase matching misses.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

C. A Preprint-First Field With Rapid Formalization

Nearly 40% of the corpus consists of preprints, and arXiv plus Zenodo account for almost 47% of primary-location records. This is characteristic of fast-moving AI fields, where technical claims often circulate before journal publication. Formal venues such as AAAI, IEEE Access, LNCS, and domain journals are also visible in the venue and highly cited samples, while journal articles and conference papers together make up 41.9% of the corpus. The field is therefore better described as preprint-first rather than preprint-only, although the current aggregate dataset does not permit a verified year-by-year comparison of preprint and peer-reviewed growth.

This dissemination structure has practical implications for bibliometric interpretation. Citation counts can accumulate before final publication; multiple versions may coexist; venue metadata may be incomplete; and automated indexes can merge records incorrectly. For reviewers and researchers, individual landmark claims should therefore be verified against DOI-level metadata even when aggregate OpenAlex counts are used for field mapping.

D. Thematic Broadening Beyond Information Retrieval

The topic distribution demonstrates diffusion beyond core NLP and information retrieval. Healthcare, education, customer/service interaction, multimodal learning, graph methods, scientific computing, ethics, robustness, and explainability all appear among the leading topics. This breadth is compatible with the architectural role of RAG and agents as horizontal technologies: retrieval provides domain grounding, while agentic control provides a mechanism for orchestrating domain tools and workflows.

The agent-family concentration in ethics, robustness, security, trust, and reliability is particularly important. These themes emerge when the model is no longer merely answering but acting. An incorrect retrieval may yield a poor answer; an autonomous agent may transform the same error into an external tool call, a persisted memory, a downstream message, or a workflow transition. The future of Agentic RAG is therefore likely to be shaped as much by control, verification, evaluation, and governance as by retrieval quality.

E. Global Research Participation

The United States, China, the United Kingdom, India, Canada, and Germany form the leading affiliation-linked group. The RAG family shows relatively high contributions from China and India, while the agent family is more strongly led by the United States and has higher counts from the United Kingdom and Canada. These patterns likely reflect differences in institutional specialization, publication venues, language, and indexing coverage as well as actual research activity. Because OpenAlex coverage and affiliation resolution vary across regions, the country results should not be interpreted as a definitive national ranking.

VI. RESEARCH AGENDA FOR 2027–2030

The bibliometric evidence identifies where research activity is concentrated; recent technical literature clarifies which problems remain unresolved. Combining the two produces a forward research agenda. The agenda is not presented as a forecast of exact calendar-year breakthroughs. Instead, 2027–2030 is used as a planning horizon for research programs that follow naturally from the transition observed through August 2026.

A. Trajectory-Level Reliability and Evaluation

RAG evaluation historically focused on retrieval relevance, answer correctness, and faithfulness. Agentic systems require a broader unit of analysis: the trajectory. A system can retrieve correctly yet plan incorrectly, select the wrong tool, propagate an early hallucination, or fail after a valid intermediate step. AgentBench already demonstrates the need for interactive evaluation [19], and recent Agentic RAG systematization highlights trajectory-level risks [20]. Future benchmarks should therefore record planning decisions, retrieval actions, tool calls, state transitions, retries, and human interventions, not only final outputs.

A central research problem is selective autonomy: when should an agent continue, retry, use a stronger model, request clarification, or escalate to a human? This requires calibrated uncertainty and risk-aware evaluation. The corpus already shows meaningful attention to robustness, explainability, trust, and reliability, but those themes remain fragmented across topic categories. A mature Agentic RAG benchmark should unify them at the workflow level.

B. Security, Trust, and Access Control

Agentic systems expand the attack surface because retrieved content can influence planning and tools can modify external state. Recent agent-security surveys describe threats involving prompt injection, compromised tools, poisoned memory, unauthorized actions, and coordination vulnerabilities [25]. The agent-family topic map already contains substantial activity in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

adversarial robustness, security and verification, and access control and trust. Future work should formalize trust boundaries between model, retriever, memory, tool, user, and agent peers.

Particularly important are deterministic controls outside the LLM: least-privilege tool permissions, schema validation, action allowlists, cryptographic or identity-backed authorization, audit logs, and human approval for high-impact actions. Research should evaluate how these controls interact with probabilistic model confidence rather than treating safety as a prompt-engineering problem.

C. Memory and Knowledge Lifecycle Governance

As systems become long-lived agents, memory becomes both a capability and a liability. Persistent memory can improve personalization and long-horizon reasoning, but it can also preserve stale, incorrect, or malicious information. RAG already provides a mechanism for accessing dynamic external knowledge; future agent architectures need explicit policies for what enters memory, when it expires, how provenance is retained, and how conflicts are resolved between memory, retrieved evidence, and model priors.

A promising direction is provenance-aware memory in which every persistent fact carries source, timestamp, confidence, and revision lineage. Research should examine memory poisoning, forgetting, contradiction management, temporal knowledge, and organization-specific retention policies. This agenda links RAG's knowledge-management roots with the autonomy requirements of future agents.

D. Cost-, Latency-, and Energy-Aware Orchestration

Agentic workflows can require many model calls, retrieval operations, rerankers, verifiers, and tool invocations. Repeated reflection and multi-agent debate improve capability in some settings but can multiply latency and inference cost. Future research should treat model selection as a routing problem: small language models can handle deterministic or narrow tasks, while larger models are reserved for complex planning, ambiguity, or high-risk reasoning. The optimization objective should include task success, reliability, latency, monetary cost, and energy rather than accuracy alone.

This creates an opportunity for adaptive SLM–LLM systems, cache-aware retrieval, early-exit policies, uncertainty-triggered escalation, and budget-constrained planning. Agentic RAG is especially suitable for such work because the controller can decide not only what to retrieve but also which model or verifier to invoke at each stage.

E. Multi-Agent Coordination and Agent-to-Agent Trust

Multi-agent systems are already among the largest topics in the agent-family corpus. Future research must move beyond demonstrations of role-playing or debate toward measurable coordination protocols. Open questions include task decomposition, communication bandwidth, shared versus private memory, conflict resolution, reputation, consensus, specialization, and failure containment. A multi-agent system can amplify both capability and error; correlated hallucinations can create false consensus, while excessive communication can erase the cost advantage of decomposition.

Agent-to-agent trust is therefore likely to become a distinct research area. Systems need mechanisms to determine which agent is authoritative for which task, how claims are verified across agents, and when a coordinator should override or isolate a faulty participant. Retrieval provenance can become a shared evidence layer for such trust decisions.

F. Human Oversight and Organizational Governance

Enterprise autonomy is not binary. Systems can draft, recommend, simulate, execute reversible actions, or execute high-impact actions under different levels of supervision. Research should therefore define autonomy levels and evaluate the relationship between automation coverage, error rate, human workload, and consequence. Human oversight should be designed around information asymmetry: the reviewer needs concise evidence, uncertainty, provenance, and predicted impact rather than an opaque request to "approve" an action.

This work also intersects with auditability and policy. Organizations need to know which model, prompt, retrieval sources, tool versions, and intermediate decisions produced an outcome. Bibliometric growth in ethics, explainability, security, and trust suggests that these concerns are moving from peripheral discussion into the core architecture of agentic systems.

G. Enterprise-Grade Benchmarks and Reproducible Evaluation

The application breadth visible in the corpus implies that generic question-answering benchmarks are insufficient. Enterprise Agentic RAG requires realistic workflows involving CRM, finance, HR, IT operations, software engineering, customer service, and regulated domains. Benchmarks should include incomplete data, contradictory evidence, changing policies, permission constraints, API failures, malicious documents, and tasks where the correct action is to abstain or request clarification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Reproducibility requires frozen model identifiers, prompt versions, tool schemas, retrieval configuration, random seeds, and evaluation code. Because commercial models change over time, benchmark reports should record exact evaluation dates. Public benchmark suites should also separate model capability from scaffolding quality so that progress in orchestration is not confused with improvements in the underlying foundation model.

Research priority	Near-term objective	Key evaluation question
Trajectory reliability	Step-level logging, calibrated abstention, repair policies	Can errors be detected before external execution?
Security and trust	Prompt-injection resilience, least privilege, agent trust	Does autonomy remain bounded under adversarial inputs?
Memory governance	Provenance, expiry, contradiction and poisoning controls	Can long-term memory remain correct and auditable?
Efficient orchestration	SLM/LLM routing, cache-aware retrieval, budgeted planning	What is the cost/reliability frontier?
Multi-agent coordination	Specialization, communication, consensus, fault isolation	When does collaboration outperform one strong agent?
Human oversight	Risk-based approval and evidence presentation	How much autonomy is safe at a target error tolerance?
Enterprise benchmarks	Realistic tools, state, failures, permissions and abstention	Do systems transfer from demos to operational workflows?

TABLE IV. Research agenda derived from the bibliometric and technical synthesis.

VII. IMPLICATIONS FOR RESEARCHERS AND PRACTITIONERS

For researchers, the results suggest that “RAG” and “agentic AI” should no longer be treated as independent topic silos. Retrieval is becoming one decision within an agent trajectory, while agent research increasingly depends on grounding, provenance, and knowledge access. Literature searches, benchmarks, and taxonomies should therefore include both retrieval and agent terminology to avoid missing cross-disciplinary work.

For practitioners, the transition changes the engineering problem. A basic RAG assistant can often be evaluated as an information system: does it retrieve relevant evidence and generate a faithful response? An agentic system must additionally be evaluated as a control system: does it select the right action, respect authorization, recover from faults, avoid unsafe persistence, and know when not to act? This difference explains why security, trust, reliability, and governance rise in thematic prominence as agentic terminology expands.

For research managers and funding bodies, the corpus indicates a field that is still expanding rapidly enough that terminology and evaluation standards are unstable. Funding programs that focus only on larger models or more complex agent architectures risk underinvesting in evaluation infrastructure, governance, and reproducibility. The research agenda in Section VI therefore prioritizes the enabling systems needed to convert experimental autonomy into dependable autonomy.

VIII. LIMITATIONS AND THREATS TO VALIDITY

The study has several important limitations. First, every numerical result is derived from OpenAlex. Scopus, Web of Science, IEEE Xplore, and ACM Digital Library were not directly queried. Coverage, citation counts, source classification, and affiliation resolution can differ across indexes, so the exact values should not be assumed to reproduce in another database. A confirmatory Scopus or Web of Science replication is recommended for venues that explicitly require proprietary-index bibliometrics.

Second, OpenAlex title_and_abstract.search uses stemmed phrase matching rather than the full proximity and Boolean field operators available in some bibliographic databases. The operational vocabulary can therefore produce both false positives and false negatives. The phrase “AI agent” is the most important source of semantic noise because it predates LLM agents by decades and appears in marketing, robotics, economics, and information systems. The analysis explicitly treats 2020–2022 agent-family values as a legacy baseline, but some older meanings may remain in later years.

Third, the dataset is aggregate rather than a manually screened record-level systematic review. No cross-database deduplication, title/abstract eligibility screening, or full-text coding was performed. OpenAlex’s own deduplication is used. Consequently, the paper maps a broad research discourse; it does not estimate the prevalence of particular architectures with the precision of a manually curated systematic review.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Fourth, 2026 covers only 1 January through 14 August, approximately 62% of the calendar year. The partial-year count is unusually large, but indexing lag and within-year growth make linear annualization inappropriate. The study therefore reports the observed partial value and does not extrapolate a 2026 total.

Fifth, document types are heterogeneous. Preprints, journal articles, conference papers, software, datasets, dissertations, and other records are included in the main corpus. The aggregate document-type table is used to describe composition, but a verified year-by-year document-type cross-tabulation was not available in the supplied dataset and is therefore not inferred. Sixth, OpenAlex primary topics are automated classifications rather than manually coded themes. “Topic Modeling”, for example, may function as a broad machine-learning category rather than a literal statement that each work develops topic-modeling methods.

Seventh, affiliation-based country counts are non-exclusive and subject to missing or incorrect affiliation metadata. They indicate participation, not national market share or quality. Finally, citation counts are time-sensitive and susceptible to record merges. One implausible RAG-family citation outlier was flagged in the supplied dataset and excluded from interpretive claims. Researchers using the companion workbook should independently verify individual high-impact records against DOI-level sources before citing them.

IX. CONCLUSION

This study provides a longitudinal bibliometric and thematic mapping of the transition from Retrieval-Augmented Generation toward Agentic AI between January 2020 and 14 August 2026. Across 73,862 OpenAlex records, the quantitative trajectory is unmistakably nonlinear. RAG-family research accelerates first, rising from 198 records in 2023 to 2,971 in 2024 and temporarily exceeding the agent-family count. Agent-centered terminology then grows faster, reaching 10,368 records in 2025 and 34,698 in the partial 2026 window. The explicit “agentic RAG” bridge grows from 16 records in 2024 to 192 in 2025 and 522 by the 2026 cutoff, while cross-family matches increase from 21 in 2023 to 1,101 in partial 2026.

The thematic evidence complements the volume trend. RAG research remains strongly associated with retrieval, graph methods, knowledge-intensive applications, and domain grounding. Agent research is more heavily associated with multi-agent systems, ethics, robustness, explainability, security, trust, and reliability. The dissemination structure is preprint-first but increasingly formalized through journal and conference publication, and the geographic distribution shows broad global participation led by the United States, China, the United Kingdom, India, Canada, and Germany.

The central interpretation is that retrieval is being absorbed into a broader agentic architecture. The transition is not a replacement of RAG by agents; rather, retrieval increasingly becomes one adaptive capability inside systems that plan, use tools, maintain memory, coordinate multiple agents, and act in external environments. That architectural shift changes the dominant research question from “How can a model retrieve better evidence?” to “How can an autonomous system retrieve, reason, act, verify, and remain controllable over a full trajectory?”

The next phase of research is therefore likely to depend on trajectory-level evaluation, security and trust boundaries, memory governance, cost-aware model routing, multi-agent fault containment, human oversight, and realistic enterprise benchmarks. In this sense, the bibliometric transition from RAG to Agentic AI is also a transition from grounding-centric AI toward governed autonomy.

DATA AVAILABILITY

The structured bibliometric workbook supporting this study is supplied as supplementary material under the filename “RAG_to_AgenticAI_bibliometric_dataset.xlsx”. It contains the methodology notes, yearly trend, term breakdown, document types, top OpenAlex topics, top venues, top countries, sampled highly cited papers, and exact OpenAlex search log. Before external publication, the author is encouraged to deposit the workbook in a persistent repository and replace this statement with the final DOI or repository URL.

ACKNOWLEDGMENT

The author acknowledges the use of OpenAlex as the primary bibliographic data source. Any remaining errors in search design, interpretation, or manuscript preparation are the responsibility of the author.

REFERENCES

[1] 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,” in Advances in Neural Information Processing Systems, vol. 33, 2020.

[2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv:2312.10997, 2023.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

[3] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proc. International Conference on Learning Representations (ICLR), 2024.
[4] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective Retrieval Augmented Generation," arXiv:2401.15884, 2024.
[5] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," in Proc. International Conference on Learning Representations (ICLR), 2023.
[6] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language Models Can Teach Themselves to Use Tools," in Advances in Neural Information Processing Systems, 2023.
[7] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," in Advances in Neural Information Processing Systems, 2023.
[8] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, and A. V. Vasilakos, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," arXiv:2501.09136, 2025.
[9] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A Survey on Large Language Model Based Autonomous Agents," Frontiers of Computer Science, vol. 18, no. 6, 2024, doi: 10.1007/s11704-024-40231-1.
[10] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, and X. He, "Exploring Large Language Model Based Intelligent Agents: Definitions, Methods, and Prospects," arXiv:2401.03428, 2024.
[11] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges," arXiv:2402.01680, 2024.
[12] M. A. B. Afarin, M. A. bin Isa, S. Z. M. Hashim, H. N. A. Hamed, and A. Matthew, "Mapping the Intellectual Landscape of Retrieval-Augmented Generation (RAG): A Bibliometric Analysis," KSII Transactions on Internet and Information Systems, vol. 20, no. 1, pp. 284–307, 2026, doi: 10.3837/tiis.2026.01.013.
[13] S. Kara and A. A. Karcioglu, "LLMs, RAG, and Agentic Workflow in Clinical Decision Support Systems: Bibliometric and Systematic Analysis," Neural Computing and Applications, vol. 38, art. 509, 2026, doi: 10.1007/s00521-026-12279-6.
[14] J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," arXiv:2205.01833, 2022.
[15] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active Retrieval Augmented Generation," in Proc. 2023 Conference on Empirical Methods in Natural Language Processing, 2023, doi: 10.18653/v1/2023.emnlp-main.495.
[16] S. Es, J. James, L. E. Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, doi: 10.18653/v1/2024.eacl-demo.16.
[17] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," in Proc. AAAI Conference on Artificial Intelligence, vol. 38, no. 16, 2024, doi: 10.1609/aaai.v38i16.29728.
[18] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking Retrieval-Augmented Generation for Medicine," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, doi: 10.18653/v1/2024.findings-acl.372.
[19] X. Liu et al., "AgentBench: Evaluating LLMs as Agents," arXiv:2308.03688, 2023.
[20] S. Mishra, S. Niroula, U. Yadav, D. Thakur, S. Gyawali, and S. Gaire, "SoK: Agentic Retrieval-Augmented Generation (RAG): Taxonomy, Architectures, Evaluation, and Research Directions," arXiv:2603.07379, 2026.
[21] J. Liang, G. Su, H. Lin, Y. Wu, R. Zhao, and Z. Li, "Reasoning RAG via System 1 or System 2: A Survey on Reasoning Agentic Retrieval-Augmented Generation for Industry Challenges," arXiv:2506.10408, 2025.
[22] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to Conduct a Bibliometric Analysis: An Overview and Guidelines," Journal of Business Research, vol. 133, pp. 285–296, 2021, doi: 10.1016/j.jbusres.2021.04.070.
[23] N. J. van Eck and L. Waltman, "Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, 2010, doi: 10.1007/s11192-009-0146-3.
[24] M. Aria and C. Cuccurullo, "bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis," Journal of Informetrics, vol. 11, no. 4, pp. 959–975, 2017, doi: 10.1016/j.joi.2017.08.007.
[25] J. Kim, X. Liu, Z. Wang, S. Qiu, B. Li, W. Guo, and D. Song, "The Attack and Defense Landscape of Agentic AI: A Comprehensive Survey," arXiv:2603.11088, 2026.
[26] W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, doi: 10.1145/3637528.3671470.
[27] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," Transactions of the Association for Computational Linguistics, 2023, doi: 10.1162/tacl_a_00530.
[28] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3532853.
[29] Z. Xi et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," Science China Information Sciences, 2025, doi: 10.1007/s11432-024-4222-0.
[30] A. K. Pati, "Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications," IEEE Access, vol. 13, pp. 151824–151837, 2025, doi: 10.1109/ACCESS.2025.3585609.