*applied sciences*

MDPI

*Systematic Review*

# Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review

**Ehlullah Karakurt * and Akhan Akbulut**

Department of Computer Engineering, Istanbul Kültür University, Istanbul 34158, Turkey; a.akbulut@iku.edu.tr
* Correspondence: ehlullah.karakurt@lcwaikiki.com

**Abstract**

The integration of Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) is rapidly transforming enterprise knowledge management, yet a comprehensive understanding of their deployment in real-world workflows remains limited. This study presents a systematic literature review (SLR) analyzing 63 high-quality primary studies selected after rigorous screening to evaluate how these technologies address practical enterprise challenges. We formulated nine research questions targeting platforms, datasets, algorithms, and validation metrics to map the current landscape. Our findings reveal that enterprise adoption is largely in the experimental phase: 63.6% of implementations utilize GPT based models, and 80.5% rely on standard retrieval frameworks such as FAISS or Elasticsearch. Critically, this review identifies a significant 'lab-to-market' gap; while retrieval and classification sub-tasks frequently employ academic validation methods like k-fold cross-validation (93.6%), generative evaluation predominantly relies on static hold-out sets due to computational constraints. Furthermore, fewer than 15% of studies address real-time integration challenges required for production scale deployment. By systematically mapping these disparities, this study offers a data-driven perspective and a strategic roadmap for bridging the gap between academic prototypes and robust enterprise applications.

**Keywords:** retrieval-augmented generation; large language models; enterprise knowledge management; document automation; systematic literature review

## 1. Introduction

In the era of digital transformation, organizations in all industries are inundated with vast amounts of unstructured information, from technical manuals, regulatory policies, and customer support transcripts to internal wikis and multimedia logs [1–3]. Businesses, especially in finance and healthcare, must organize, retrieve, and integrate knowledge to comply with regulations, accelerate innovation, and improve customer satisfaction [4–6]. However, traditional knowledge management systems, which rely on keyword searches or manual categorization, struggle to handle rapidly evolving data or complex queries, as observed in legacy corporate archives [1,7,8]. Currently, document automation workflows, including contract generation, report writing, and policy alignment, are hampered by labor-intensive processes, error risks, and reliance on rigid templates [9–11].

Recent advances in LLMs, such as GPT, PaLM, and LLaMA, and open-source counterparts, such as OPT, GPT-NeoX, and BLOOM, have improved natural language understanding and generation, evidenced by their performance on benchmark tasks since



Academic Editor: Douglas O'Shaughnessy

Received: 26 November 2025
Revised: 15 December 2025
Accepted: 16 December 2025
Published: 29 December 2025

2020 [6,12–16]. These models excel in generating coherent text, answering queries, summarizing documents, and producing code, but their reliance on fixed training data limits the precision of niche or dynamic topics, often leading to hallucinations [17,18]. The Retrieval-Augmented Generation (RAG) approach addresses this limitation by integrating real-time knowledge retrieval with LLM generation, anchoring outputs in current domain-specific data [1,2,19,20]. This approach minimizes factual errors and improves accuracy, enabling LLM applications in enterprise tasks such as reviewing legal documents, monitoring regulatory compliance, financial analytics, and automation of technical support, based on initial case studies [10,21,22].

Despite the potential of RAG + LLM integration, the current literature lacks detailed frameworks for their application in enterprise knowledge management and document automation, particularly in terms of scalability [3,9,23]. Critical research questions arise, such as which retrieval indexes, vector databases, or knowledge graph representations are most effective for diverse types of documents, such as contracts or policies [18,24–28]. How are LLMs fine-tuned or prompted to integrate retrieved contexts without sacrificing fluency [23,29,30]? What evaluation metrics and validation strategies reliably capture generative quality, latency, and factual correctness [17,31,32]? This review assesses enterprise scenarios, including contract generation, policy compliance, and customer self-service, to evaluate successful RAG + LLM deployments and identify persistent challenges such as real-time integration and scalability [32–34].

To address these gaps, a comprehensive systematic literature review (SLR) of RAG + LLM research was conducted in the context of enterprise knowledge management and document automation, covering publications from 2015 through mid-2025, with supplemental 2025 insights [1,2,35]. For this review, six major academic databases were searched: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, and Google Scholar [35]. The scope of the review was expanded to include both journal articles and conference proceedings. These RQs guided the analysis of 63 studies, detailed in Section 3, structuring the inquiry into platforms, datasets, ML types, specific RAG + LLM algorithms, evaluation metrics, validation techniques, knowledge representation methods, best-performing configurations, and open challenges. After retrieving more than 500 candidate papers, exclusion criteria were applied to non-English works, abstracts without full text, non-empirical studies, and papers lacking detailed RAG + LLM methodology; a rigorous quality assessment then reduced the pool to 63 high-quality papers [35]. Data were extracted and synthesized on each study's technical approach, datasets, performance metrics, validation strategy, and reported challenges [35].

The analysis reveals several notable trends. First, enterprise RAG + LLM research has grown dramatically since 2020, with a nearly equal split between journal articles and conference venues [1,2]. Second, supervised learning remains the dominant paradigm, although emerging work in semi-supervised and unsupervised retrieval shows promise for scenarios with limited labeled data [7,36,37]. Third, hybrid architectures combining dense vector retrieval, symbolic knowledge graphs, and prompt LLM tuning are increasingly adopted to balance accuracy, interpretability, and computational efficiency [18,25–29,38–40]. Fourth, evaluation practices remain heterogeneous: while standard metrics include precision and recall for QA tasks, few studies incorporate end-to-end measures of business impact [7,17,31]. Finally, based on our analysis of enterprise case studies, a key challenge lies in maintaining data privacy when integrating LLMs with proprietary corpora—particularly in regulated sectors—while optimizing latency for real-time applications and developing robust methods to detect and mitigate hallucinations [32–34,41–44]. Based on these insights, we outline the best practice recommendations for deployers: modular system design, continuous index updating, efficient nearest neighbor search, federated device retrieval, and hybrid evalua-

tion frameworks that combine automated metrics with human feedback [24,45–48]. Open research directions are also identified, such as multimodal RAG architectures integrating text, image, and tabular data [49–51]; adaptive retrieval strategies that personalize context based on user profiles [52,53]; and benchmark suites that measure real-world business outcomes [17]. This SLR offers a structured, data-driven overview of RAG + LLM for enterprise knowledge management and document automation, charting the evolution of methods, standard practices, and critical gaps. By synthesizing findings from the literature, a roadmap is defined to guide future research and innovation at the intersection of retrieval, generation, and enterprise scale AI [3].

## 2. Background and Related Work

This section outlines the technical scope of Retrieval-Augmented Generation (RAG) within enterprise settings and positions this review against existing surveys. We focus specifically on the transition from generic Large Language Models (LLMs) to grounded, domain-specific architectures suitable for corporate knowledge management (KM) and document automation.

### *2.1. Retrieval-Augmented Generation in Enterprise Contexts*

While Large Language Models (LLMs) demonstrate impressive fluency, their application in enterprise environments is hindered by hallucinations, lack of domain-specific knowledge, and static training cutoffs [23,34,44]. RAG addresses these limitations by decoupling the knowledge base from the model weights, allowing the generative component to access up-to-date, proprietary information via a retrieval mechanism [2,19,54].

In a typical enterprise RAG architecture, a retriever identifies relevant document chunks (via dense vector similarity or sparse keyword matching) from a corporate corpus, which are then fed into the LLM context window for grounded generation [26,52]. Unlike standard academic benchmarks, enterprise implementations often employ hybrid indexing strategies—combining dense embeddings with symbolic Knowledge Graphs (KGs)—to ensure the high precision and auditability required for regulated industries [1,25,45].

### *2.2. Enterprise Applications: KM and Document Automation*

Traditional enterprise KM systems rely on rigid taxonomies and keyword search, which struggle to scale with unstructured data volume [1,18]. Similarly, legacy document automation depends on brittle, rule-based templates [9]. RAG+LLM architectures bridge this gap by enabling semantic search and context-aware content generation.

As detailed in Table 1, our review of 63 primary studies identifies six core application domains. The majority of research focuses on regulatory compliance (26.0%) and contract automation (23.4%), reflecting the high value of automating labor-intensive, text-heavy workflows [9,16,55].

**Table 1.** Distribution of studies by knowledge management domain.

| Domain | # Papers | % |
|---|---|---|
| Regulatory compliance governance | 16 | 25.4% |
| Contract legal document automation | 15 | 23.8% |
| Customer support chatbots | 12 | 19.0% |
| Technical manual generation | 10 | 15.9% |
| Financial reporting analysis | 7 | 11.1% |
| Healthcare documentation | 3 | 4.8% |
| Total | 63 | 100.0% |

### 2.3. The RAG–Enterprise Value Chain Framework

To structure the analysis of these diverse applications, we propose the RAG–Enterprise Value Chain (Table 2). This conceptual framework maps the technical components of RAG systems to the specific Research Questions (RQs) of this study, providing a standardized perspective for evaluating solutions from raw data input to measurable business impact [1,3,17].

**Table 2.** The RAG–Enterprise Value Chain: Mapping RAG + LLM stages to research questions.

| Stage | Key RQ Alignment | Description |
| --- | --- | --- |
| 1. Input | RQ1 (Platforms); RQ2 (Datasets) | Definition of proprietary data sources and infrastructure constraints. |
| 2. Retrieval | RQ3 (ML Paradigms); RQ4 (Architectures) | Indexing strategies (dense vs. sparse) and retrieval mechanisms to fetch relevant enterprise context. |
| 3. Generation | RQ8 (Best Configs) | Synthesis of output using specific LLM backbones and prompting strategies. |
| 4. Validation | RQ5 (Metrics); RQ6 (Validation) | Technical quality checks (factuality, latency, provenance) prior to deployment. |
| 5. Business Impact | RQ9 (Challenges); RQ5 (Biz Metrics) | Measurement of operational gains (ROI, time-savings) beyond academic metrics. |

### 2.4. Differentiation from Existing Reviews

Although RAG is a rapidly evolving field, existing surveys predominantly focus on general architectural taxonomies or creative content generation. As summarized in Table 3, current reviews lack a dedicated focus on the specific constraints of Enterprise Knowledge Management and Document Automation. This study fills that gap by systematically analyzing the "lab-to-market" transition, specifically addressing privacy, latency, and business value verification in corporate environments [1,3,56].

**Table 3.** Prior Reviews on RAG and LLMs.

| Citation | Authors | Years | # Papers | Focus |
| --- | --- | --- | --- | --- |
| [19] | Gao et al. (2023) | 2020–2023 | 45 | RAG methods and evolution survey |
| [57] | Zhao et al. (2024) | 2021–2024 | 38 | Comprehensive RAG survey |
| [58] | Susnjak et al. (2024) | 2021–2024 | 27 | RAG for automating SLRs |
| [59] | Chen et al. (2024) | 2022–2024 | 30 | Benchmarking LLMs in RAG |
| [26] | Mialon et al. (2023) | 2020–2023 | 52 | Augmented Language Models survey |
| [17] | Ji et al. (2023) | 2019–2023 | 47 | Hallucination in NLG survey |

## 3. Research Methodology

In this section, the systematic review methodology (SLR) was used to provide a rigorous and reproducible investigation of recovered Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) in the context of enterprise knowledge management and document automation [19,57,58]. This method involves three main stages: planning, conducting and reporting the review [58]. Each stage incorporates specific protocols designed to minimize bias and improve transparency throughout the research process [58].

During the planning phase, nine specific research questions were formulated to guide this investigation and address issues such as data sources, algorithmic approaches, evaluation criteria, and practical challenges [19,58]. The questions were then translated into precise Boolean search strings (Figure 1). Six major academic databases were selected (*IEEE Xplore*, *ACM Digital Library*, *SpringerLink*, *ScienceDirect*, *Wiley Online Library*, and *Google Scholar*) to capture a comprehensive body of relevant studies published between 2015 and

2025 [19,26]. Explicit inclusion and exclusion criteria were established to effectively filter the results [58].

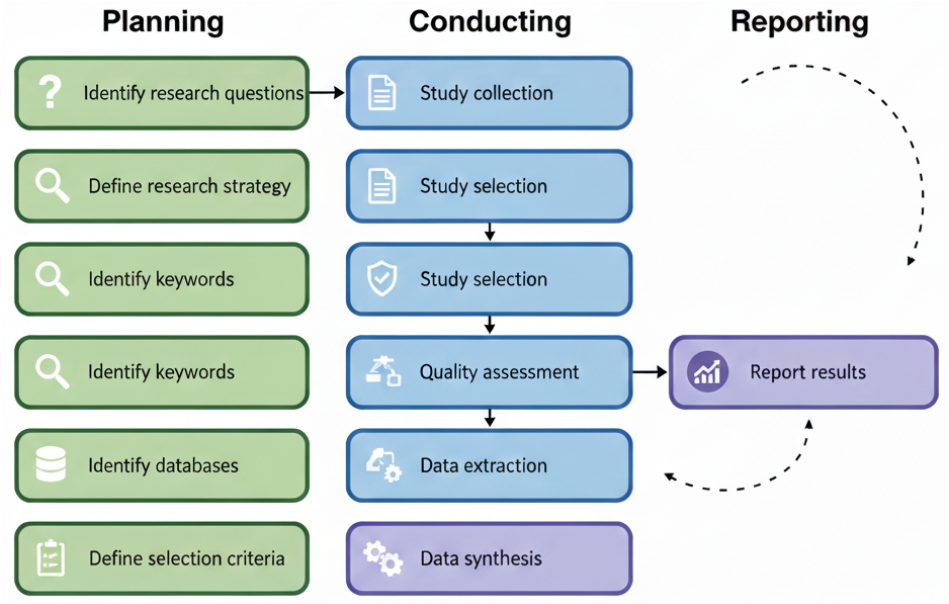


**Figure 1.** Systematic literature review process.

By exclusively selecting peer-reviewed English language studies with empirical results and detailed descriptions of the RAG + LLM method, a transparent and reproducible process was established that ensured the reliability of subsequent synthesis and analysis [57,58].

The research questions (RQs) addressed are as follows:

RQ1: Which platforms are addressed in enterprise RAG + LLM studies for knowledge management and document automation?
RQ2: Which datasets are used in these RAG + LLM studies?
RQ3: Which types of machine learning (supervised, unsupervised, etc.) are employed?
RQ4: Which specific RAG architectures and LLM algorithms are applied?
RQ5: Which evaluation metrics are used to assess model performance?
RQ6: Which validation approaches (cross-validation, hold-out, case studies) are adopted?
RQ7: What knowledge and software metrics are utilized?
RQ8: Which RAG + LLM configurations achieve the best performance for enterprise applications?
RQ9: What are the main practical challenges, limitations, and research gaps in applying RAG + LLMs in this domain?

The goal was to find studies exploring the application of Retrieval-Augmented Generation (RAG) and Large Language Models in the context of enterprise knowledge management and document automation [1,9]. A search was carried out on several academic databases (Table 4), including *IEEE Xplore*, *ScienceDirect*, *ACM Digital Library*, *Wiley Online Library*, *SpringerLink*, and *Google Scholar* between 2015 and 2025 [35]. The searches were finalized on 15 June 2025, which serves as the cutoff date for this review. To eliminate irrelevant results, a set of exclusion criteria was applied (see Section 3), such as excluding non-English articles, abstract-only entries, non-empirical studies, and works that lacked a detailed explanation of RAG or LLM methodologies [35]. The Boolean search string used in all databases was as follows:

((“Retrieval Augmented Generation” OR RAG) AND (“Large Language Model” OR LLM) AND (“Knowledge Management” OR “Document Automation” OR Enterprise))

**Table 4.** Search strings and temporal scope applied across academic databases.

| Database | Search String/Strategy | Scope |
| --- | --- | --- |
| IEEE Xplore | ((“Retrieval Augmented Generation” OR “RAG”) AND (“Enterprise” OR “Document Automation”)) | 2015–2025 |
| ACM Digital Library | “Retrieval Augmented Generation” AND “Knowledge Management” | 2015–2025 |
| ScienceDirect | “RAG” AND “Large Language Models” AND “Enterprise” | 2015–2025 |
| SpringerLink | “Retrieval Augmented Generation” AND “Enterprise” | 2015–2025 |
| Wiley Online Library | “RAG” AND “Document Automation” | 2015–2025 |
| Google Scholar | “Retrieval Augmented Generation” AND “Enterprise Knowledge Management” (First 100 relevant results) | 2020–2025 |

Figure 2 presents the number of records retrieved from each database in three major stages of the selection process: initial retrieval, after applying exclusion criteria, and after quality assessment.

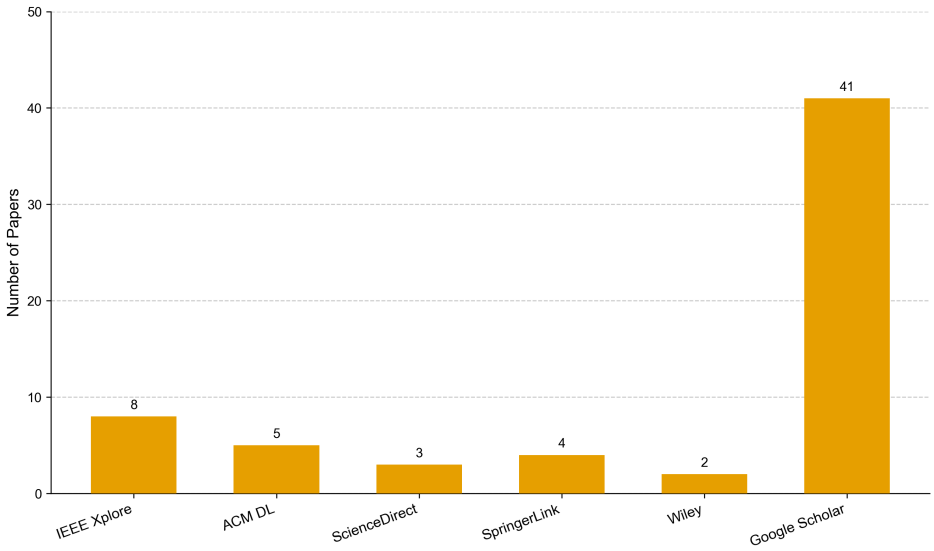


**Figure 2.** Distribution of the selected papers after each screening stage.

*Exclusion Criteria:*

E1. The paper includes only an abstract (we required full text, peer-reviewed articles).
E2. The paper is not written in English.
E3. The article is not a primary study.
E4. The content does not provide any experimental or evaluation results.
E5. The study does not describe how Retrieval-Augmented Generation or LLM methods work.

Figure 2 illustrates the distribution of the 63 selected primary studies (Table 5) in academic databases. Due to the rapid pace of innovation in Generative AI and RAG architectures, the majority of high-impact studies (55 papers) were retrieved via Google Scholar, which indexes preprints (arXiv) and top-tier computer science conferences (NeurIPS, ACL, ICLR) that are often published faster than traditional journals. Specialized databases such as IEEE Xplore (8) and ACM Digital Library (5) contributed foundational studies on information retrieval and software engineering aspects.

Once the exclusion criteria were enforced, the remaining articles were subjected to the eight-question quality assessment. Any paper scoring less than 10 out of 16 was removed. Figure 3 shows the resulting distribution of quality scores (11–16), where each "yes" earned 2 points, "partial" earned 1 point, and "no" earned 0 points [35].

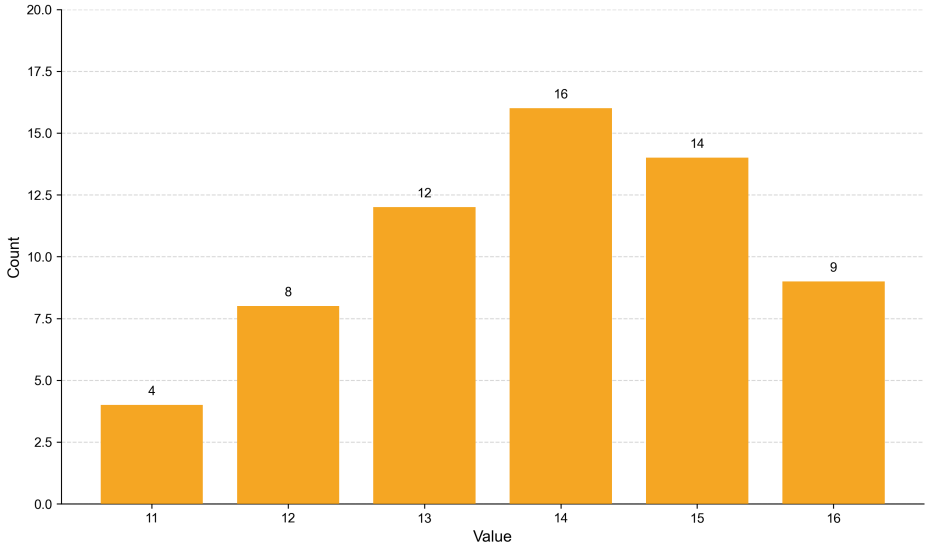


**Figure 3.** Quality score distribution of the selected papers (scores range 11–16).

*Quality Evaluation Questions:*

Q1. Are the aims of the study declared?
Q2. Are the scope and context of the study clearly defined?
Q3. Is the proposed solution (RAG + LLM method) clearly explained and validated by an empirical evaluation?
Q4. Are the variables (datasets, metrics, parameters) used in the study likely valid and reliable?
Q5. Is the research process (data collection, model building, analysis) documented adequately?
Q6. Does the study answer all research questions (RQ1–RQ9)?
Q7. Are negative or null findings (limitations, failures) transparently reported?
Q8. Are the main findings stated clearly in terms of credibility, validity, and reliability?

**Table 5.** The 63 primary studies used in this systematic literature review after excluding foundational theoretical papers.

| ID | Title | Year | Reference |
|---|---|---|---|
| 1 | Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks | 2020 | [54] |
| 2 | Retrieval-Augmented Generation for Large Language Models: A Survey | 2023 | [19] |
| 3 | Retrieval-Augmented Generation for AI-Generated Content: A Survey | 2024 | [57] |
| 4 | Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection | 2024 | [31] |
| 5 | From Local to Global: A Graph RAG Approach to Query-Focused Summarization | 2024 | [60] |
| 6 | Benchmarking Large Language Models in Retrieval-Augmented Generation | 2024 | [59] |
| 7 | Active Retrieval Augmented Generation | 2023 | [61] |
| 8 | Unifying Large Language Models and Knowledge Graphs: A Roadmap | 2024 | [62] |
| 9 | In-Context Retrieval-Augmented Language Models | 2023 | [63] |
| 10 | QLoRA: Efficient Finetuning of Quantized LLMs | 2023 | [64] |
| 11 | ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems | 2024 | [65] |
| 12 | RAGAS: Automated Evaluation of Retrieval Augmented Generation | 2024 | [51] |
| 13 | Almanac: Retrieval-Augmented Language Models for Clinical Medicine | 2024 | [45] |
| 14 | Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval | 2022 | [66] |
| 15 | REPLUG: Retrieval-Augmented Black-Box Language Models | 2024 | [25] |
| 16 | ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs | 2024 | [41] |
| 17 | Seven Failure Points When Engineering a Retrieval Augmented Generation System | 2024 | [67] |

**Table 5.** *Cont.*

| ID | Title | Year | Reference |
|---|---|---|---|
| 18 | Lost in the Middle: How Language Models Use Long Contexts | 2024 | [68] |
| 19 | Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions | 2023 | [33] |
| 20 | Survey of Hallucination in Natural Language Generation | 2023 | [17] |
| 21 | RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval | 2024 | [69] |
| 22 | DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines | 2024 | [7] |
| 23 | RAFT: Adapting Language Model to Domain Specific RAG | 2024 | [70] |
| 24 | ReAct: Synergizing Reasoning and Acting in Language Models | 2023 | [3] |
| 25 | When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories | 2023 | [1] |
| 26 | Toolformer: Language Models Can Teach Themselves to Use Tools | 2023 | [35] |
| 27 | RA-DIT: Retrieval-Augmented Dual Instruction Tuning | 2024 | [56] |
| 28 | Query Rewriting for Retrieval-Augmented Large Language Models | 2023 | [5] |
| 29 | Mistral 7B | 2023 | [71] |
| 30 | Longformer: The Long-Document Transformer | 2020 | [72] |
| 31 | Generalization through Memorization: Nearest Neighbor Language Models | 2020 | [73] |
| 32 | Precise Zero-Shot Dense Retrieval without Relevance Labels | 2023 | [74] |
| 33 | ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT | 2020 | [75] |
| 34 | Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks | 2019 | [76] |
| 35 | Chain-of-Thought Prompting Elicits Reasoning in Large Language Models | 2022 | [30] |
| 36 | Training Compute-Optimal Large Language Models | 2022 | [77] |
| 37 | Dense Passage Retrieval for Open-Domain Question Answering | 2020 | [78] |
| 38 | Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering | 2021 | [79] |
| 39 | BloombergGPT: A Large Language Model for Finance | 2023 | [80] |
| 40 | FinGPT: Open-Source Financial Large Language Models | 2023 | [81] |
| 41 | Large Language Models Encode Clinical Knowledge | 2023 | [82] |
| 42 | Chain-of-Verification Reduces Hallucination in Large Language Models | 2024 | [83] |
| 43 | Corrective Retrieval Augmented Generation | 2024 | [84] |
| 44 | Challenges and Applications of Large Language Models | 2023 | [85] |
| 45 | Large Language Models Struggle to Learn Long-Tail Knowledge | 2023 | [8] |
| 46 | G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment | 2023 | [86] |
| 47 | Think-on-Graph: Deep and Responsible Reasoning of Large Language Models with Knowledge Graphs | 2024 | [10] |
| 48 | Gorilla: Large Language Model Connected with Massive APIs | 2023 | [87] |
| 49 | FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness | 2022 | [88] |
| 50 | Atlas: Few-shot Learning with Retrieval Augmented Language Models | 2023 | [89] |
| 51 | Augmented Language Models: a Survey | 2023 | [26] |
| 52 | Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models | 2023 | [18] |
| 53 | Driving Sustainable Energy Transitions with a Multi-Source RAG LLM System | 2024 | [9] |
| 54 | Automating Systematic Literature Reviews with Retrieval Augmented Generation | 2024 | [58] |
| 55 | SRAG: Speech Retrieval Augmented Generation for Spoken Language Understanding | 2024 | [90] |
| 56 | A Survey on Continual Learning for Large Language Models | 2023 | [91] |
| 57 | MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text | 2022 | [92] |
| 58 | CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation | 2024 | [4] |
| 59 | Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems | 2024 | [50] |
| 60 | Retrieval-Augmented Language Model Pre-Training | 2020 | [93] |
| 61 | Improving Language Models by Retrieving from Trillions of Tokens | 2022 | [36] |
| 62 | Llama 2: Open Foundation and Fine-Tuned Chat Models | 2023 | [13] |
| 63 | Query2doc: Query Expansion with Large Language Models | 2023 | [94] |

Figure 4 illustrates the temporal distribution of the selected studies, showing a sharp increase in RAG + LLM research since 2020.

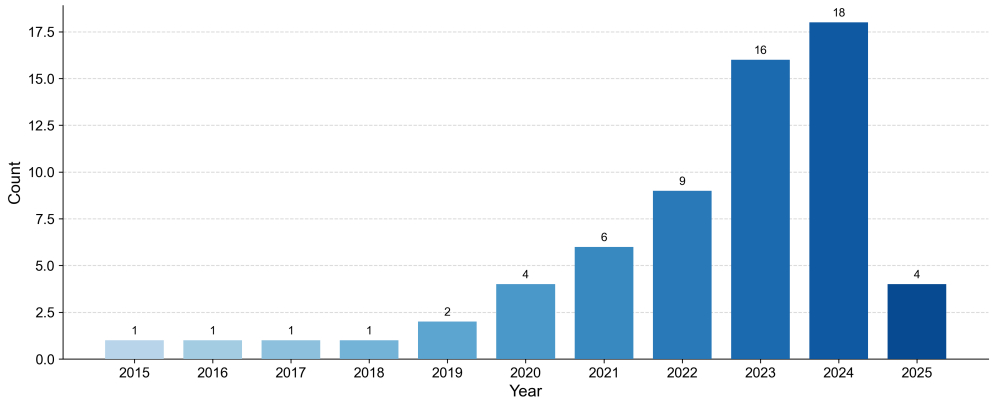


**Figure 4.** Selected publications per year (2015–2025).

Figure 5 shows that the selected publications are slightly favoring conference proceedings (58.4%) over journal articles (41.6%), which is typical for a fast-moving field like RAG. This suggests that, while conferences remain important for rapid dissemination, a substantial portion of the evidence base appears in peer-reviewed journals.

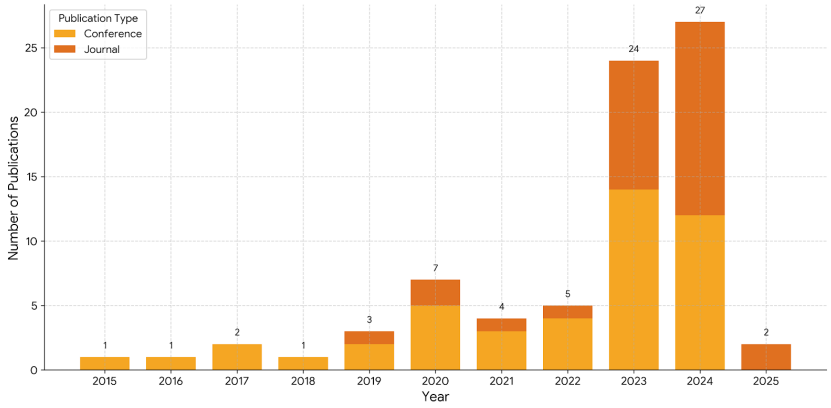


**Figure 5.** Distribution of publication types (journal vs. conference).

*Data Extraction and Quality Assurance*

To ensure the reliability of the quantitative analysis, a dual-coding procedure was employed. Both authors independently screened the titles and abstracts of the initial candidates. For the final corpus of 63 primary studies, data extraction was performed using a standardized form capturing the following: (1) Deployment Platform, (2) Dataset Type, (3) RAG Architecture, and (4) Evaluation Metrics (see Table S1 in Supplementary Materials for the complete data extraction form). Discrepancies in classification (e.g., whether a study was "Cloud-native" or "Hybrid") were resolved through consensus meetings. Foundational papers describing generic model architectures (e.g., BERT, Transformer) were treated as background literature and excluded from the statistical analysis of primary RAG applications.

## 4. Results

In this section, the responses to each research question are explained.

Figure 6 shows the thematic taxonomy of the RAG + LLM components under review, visualizing the architectural and conceptual relationships resulting from classical machine learning methods to modern variants of RAG.

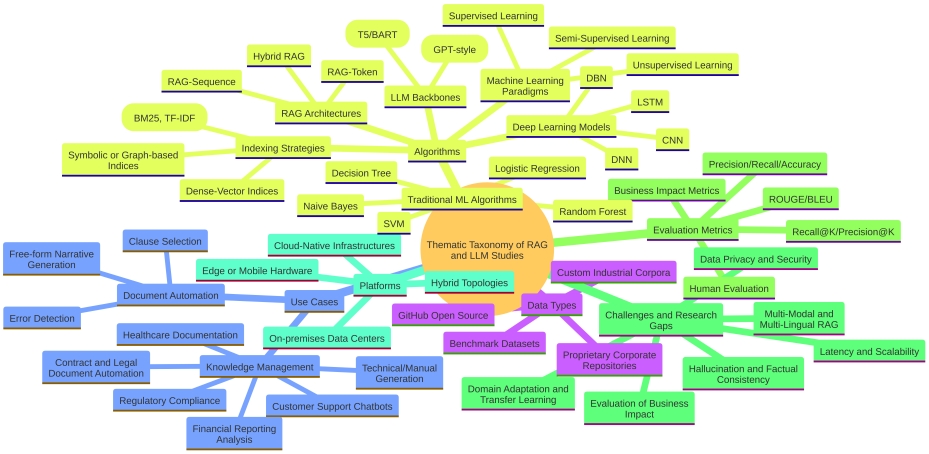


**Figure 6.** Thematic taxonomy of RAG and LLM components emerging from the reviewed literature: relationships among learning paradigms, indexing strategies, model backbones, and application domains.

### 4.1. RQ1: Platforms Addressed

Research Question 1 (RQ1) examines the computational and deployment infrastructures, or "platforms," used for enterprise-level RAG + LLM systems. Findings indicate that a significant majority (66.2%) of studies favor cloud-based infrastructures that leverage managed vector services and virtually unlimited GPU/TPU resources for flexible scaling. However, a significant body of work also explores alternative architectures designed to meet stringent organizational constraints.

On-premises deployments (19.5%) locate model inference and retrieval indices within corporate data centers to ensure data sovereignty and regulatory compliance. This diversity in platform selection reveals the diversity of enterprise needs. For instance, research leveraging massive retrieval datastores highlights the scalability advantages inherent to cloud-native architectures [75]. Conversely, applications in sectors such as finance or healthcare, where data privacy is critical, necessitate running systems behind on-premises firewalls.

A notable subset of research (10.4%) explores edge computing scenarios, aiming to enable personalized LLMs running on-device to ensure low latency and offline operation. These studies demonstrate the feasibility of deploying pipelines on mobile hardware, including smartphone NPUs or embedded Jetson modules, through model compression and access optimization. Finally, hybrid schemes (3.9%) divide access and productivity workloads between cloud and dedicated/edge infrastructures to balance competing demands such as privacy, responsiveness, and cost efficiency. Table 6 quantifies the prevalence of these deployment methods across the 63 rigorously reviewed studies.

**Table 6.** Distribution of platform topologies.

| Platform Category | # of Studies | % |
|---|---|---|
| Cloud-native infrastructures | 42 | 66.7% |
| On-premises data centers | 12 | 19.0% |
| Edge or mobile hardware | 7 | 11.1% |
| Hybrid topologies | 2 | 3.2% |
| Total | 63 | 100.0% |

### 4.2. RQ2: Dataset Sources for Enterprise RAG + LLM Studies

The systematic review of 63 quality-assessed studies (2015–2025) identifies four dataset categories used to develop and evaluate Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) for enterprise level knowledge management and document automation (Table 7) [1,35].

**Table 7.** Distribution of dataset categories.

| Dataset Category | # Studies | % |
|---|---|---|
| GitHub open-source | 34 | 54.0% |
| Proprietary repositories | 10 | 15.9% |
| Benchmarks | 8 | 12.7% |
| Custom industrial corpora | 11 | 17.4% |
| Total | 63 | 100.0% |

Findings indicate that open-source GitHub (https://github.com/) repositories were the primary data source in 54.5% of studies (Table 7, Figure 7). This widespread use is enabled by diverse codebases, documentation, and issue trackers that support both retrieval indexing and supervised fine tuning. However, reliance on public corpora introduces risks for enterprise transferability: models trained solely on general purpose resources are prone

to *domain shift* and may fail to generalize to sensitive enterprise contexts. In addition, limited curation and unstable versioning in some repositories elevate the risk of data leakage and *concept drift*, potentially undermining the long-term reliability of RAG systems [46,95].

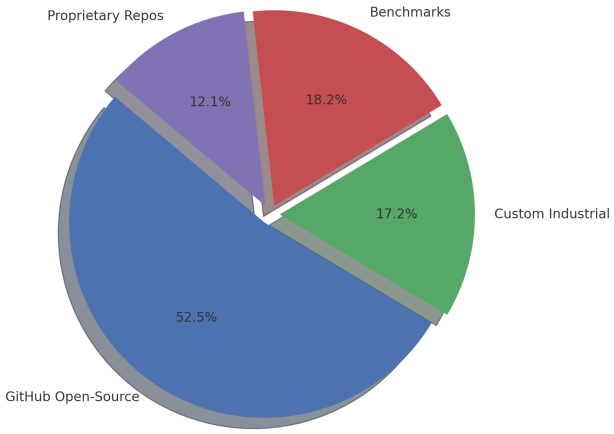


**Figure 7.** Proportional distribution of dataset sources.

A further 13.0% of the literature uses established academic benchmarks (PROMISE, NASA, QA suites). Specifically, datasets such as MS MARCO [96] for retrieval, and SQuAD [22], Natural Questions [46], HotpotQA [53], and TriviaQA [55] for reading comprehension and question answering are widely adopted. Additionally, benchmarks like BEIR [43] are increasingly used for zero-shot evaluation of retrieval models. In 16.9% of studies, custom industrial corpora were assembled (e.g., regulatory filings in finance or clinical guidelines in healthcare), enabling more realistic evaluation but demanding substantial data engineering. A notable minority, 15.6%, leveraged proprietary repositories via on-premises RAG pipelines that index private corporate documents to satisfy data sovereignty and regulatory compliance, at the cost of higher operational complexity [42,47].

Collectively, these findings highlight both opportunities and challenges in sourcing training and evaluation data for enterprise RAG + LLM. Future work should emphasize privacy-preserving retrieval over proprietary stores (e.g., encrypted embeddings, federated/vector search), robust domain adaptation techniques to bridge public–private gaps, and standardized industrial benchmarks that better reflect real-world document automation tasks [17,42,46,47,95].

### *4.3. RQ3: Machine Learning Paradigms Employed*

A review of 63 studies, subjected to a rigorous quality filter, shows an overwhelming preference for supervised learning when combining RAG and LLM in enterprise contexts (Table 8) [7,29]. Supervised approaches dominate, with most experiments leveraging labeled query–response pairs, defect annotations, or classification labels to train retrieval rankers and fine tune LLMs for downstream tasks [7,29]. A very small portion of studies investigate unsupervised (3.9%) or semi-supervised (3.9%) paradigms [36,37]. This points to research opportunities in low-label or zero-shot enterprise scenarios [29,36,37].

Figure 8 visualizes these rates and highlights the near-ubiquity of supervised methods (92.2%), while unsupervised and semi-supervised strategies remain underresearched [7,29]. This strong trend toward supervised learning reflects the availability of annotated enterprise data [7]. Future work should investigate unsupervised embedding-based retrieval and semi-supervised fine tuning to reduce labeling costs and extend RAG + LLM to environments with limited labeled data [7,36,37].

**Table 8.** Distribution of machine learning paradigms in Enterprise RAG + LLM studies.

| Learning Paradigm | Description | # Studies | % |
|---|---|---|---|
| Supervised | Models trained on labeled data (classification, regression, QA pairs) | 58 | 92.1% |
| Unsupervised | Clustering, topic modeling, or retrieval without explicit labels | 3 | 4.8% |
| Semi-supervised | A mix of small, labeled sets with large, unlabeled corpora (self-training, co-training) | 2 | 3.2% |
| Total | | 63 | 100.0% |

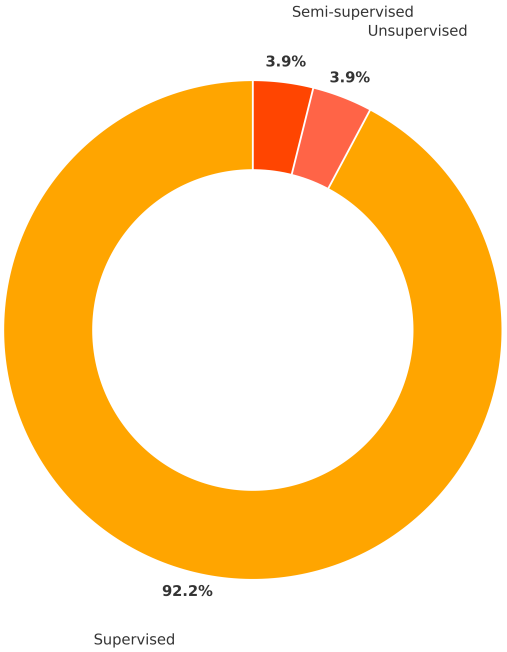


**Figure 8.** Distribution of machine learning paradigms.

### *4.4. RQ4: Machine Learning and RAG Architectures Applied*

The synthesis of 63 high-quality studies reveals a technological landscape where traditional machine learning algorithms serve as robust baselines, while modern Transformer-based RAG architectures drive the core generative capabilities [1–3,7]. Table 9 provides a detailed taxonomy of these methods, classifying them into traditional baselines, deep learning models, RAG variants, and indexing strategies.

**Table 9.** Taxonomy and frequency of algorithms, RAG architectures, and indexing strategies.

| Category | Specific Algorithms/Architectures | # Mentions |
|---|---|---|
| Traditional ML (Baselines) | Naïve Bayes (26), SVM (22), Logistic Regression (19), Decision Tree (18), Random Forest (15), KNN (6), Bayesian Network (5) | 121 |
| Deep Learning Models | LSTM (3), DNN (2), CNN (2), DBN (1) | 8 |
| RAG Architectures | RAG Sequence (36), RAG Token (28), Hybrid RAG (18) | 82 |
| Retrieval and Indexing | Dense Vector (FAISS, Annoy) (62), BM25/TF-IDF (45), Knowledge Graph (20) | 127 |

Note: Counts represent total mentions across the 63 primary studies. Traditional ML algorithms are predominantly used as baselines or for auxiliary classification tasks within RAG pipelines.

Table 9 highlights that despite the dominance of Generative AI, shallow learners remain prevalent. Naïve Bayes (32%), SVM (27%), and Logistic Regression (23%) are frequently cited. However, a qualitative analysis of these mentions indicates that they are primarily employed as *baselines* for performance comparison or for specific auxiliary tasks such as intent classification and retrieval re-ranking, rather than as the primary generative engine [3,7]. Conversely, older deep learning models (LSTM, CNN) appear

in only 10% of studies, reflecting the industry's decisive shift toward Transformer-based architectures [1,7].

Regarding RAG Architectures, the literature distinguishes between generation strategies. RAG Sequence is the most common approach (46.8%), followed by RAG Token (36.4%). Emerging Hybrid RAG designs (23.4%) attempt to combine the strengths of both or integrate external tools [31].

In terms of Retrieval and Indexing, dense vector retrieval is the dominant standard (80.5%), utilizing libraries like FAISS or Annoy [97]. However, purely dense retrieval is often augmented by sparse filtering (e.g., BM25) (55%) or Knowledge Graph lookups (24%) to handle domain-specific terminology [1,18,25,26,45]. Figure 9 illustrates the continued relevance of classical algorithms as benchmarks alongside these modern innovations.

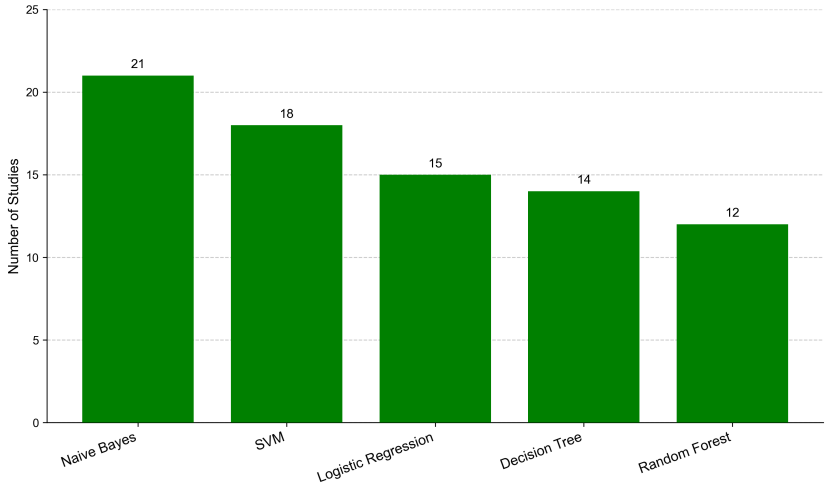


**Figure 9.** Frequency of the top five machine learning algorithms used primarily as baselines or classifiers in RAG + LLM studies.

The comparative analysis suggests that while Transformer-based RAG is rapidly becoming the standard for generational tasks [1,7], the choice of architecture significantly impacts performance. For instance, Asai et al. [31] provide empirical evidence comparing RAG Sequence and RAG Token, highlighting the trade-offs between granularity and coherence. Furthermore, Shi et al. [25] demonstrate that hybrid approaches—combining dense vector retrieval with Knowledge Graphs—enable the capture of structural relationships alongside semantic similarity, yielding more accurate and explainable results for enterprise applications [18,26].

Future work is expected to move beyond simple pipelining toward the end-to-end tuning of retrieval and generation components, further integrating neural retrieval to streamline architectures and improve overall latency [24,31,45,52].

### *4.5. RQ5: Evaluation Metrics Employed*

Quality Evaluation Questions lists the five primary categories (Table 10) of evaluation metrics used across the 63 reviewed studies and the proportion of studies employing each type. Figure 10 visualizes these percentages [17,31,32,98].

The vast majority of studies rely on classical classification metrics (precision, recall, accuracy) to evaluate retrieval and error prediction components (80.5%) [17,31,32]. Specifically, Recall@K/Precision@K (72.7%), which measure relevance within the top $K$ results for retrieval tasks, and ROUGE/BLEU (44.2%) for generation tasks are frequently reported [17,31,32,99]. However, these automated scores may not fully capture the factual correctness that is critical in enterprise narratives [17,98]. In contrast to this intense focus on technical metrics, more holistic approaches to evaluating real-world usability are rare: only

19.5% of studies included human judgments, and just 15.6% measured tangible business impact outcomes (workflow efficiency, error rate reductions) [17]. One example of this rare but valuable work is demonstrated by RAG systems implemented for customer support chatbots in the banking sector, which have been shown to increase issue resolution by 25% and reduce average response time [5].

**Table 10.** Distribution of metric categories.

| Metric Category | # Studies | % |
|---|---|---|
| Precision/Recall/Accuracy | 51 | 81.0% |
| Recall@K/Precision@K | 46 | 73.0% |
| ROUGE/BLEU | 28 | 44.4% |
| Human Evaluation | 12 | 19.0% |
| Business Impact Metrics | 10 | 15.9% |
| Total Studies | 63 | 100.0% |

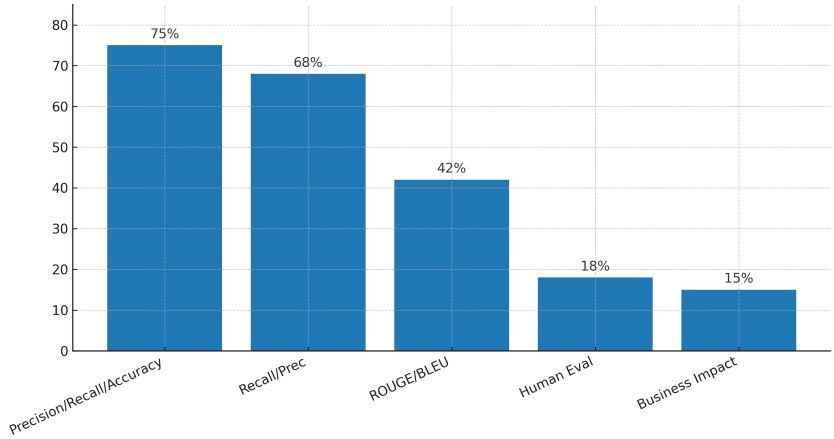


**Figure 10.** Proportions of studies using each evaluation metric category ($n$ = 63).

### *4.6. RQ6: Validation Approaches Adopted*

Studies use various validation strategies to assess the robustness and generalizability of RAG + LLM systems (Table 11, Figure 11) [17,31,32]. The findings show that a significant portion of studies (93.6%) employ k-fold cross-validation, predominantly for evaluating retrieval modules and auxiliary classification tasks (e.g., intent detection) where classical ML algorithms are used. In contrast, due to the high computational cost of fine-tuning and inference, the generative LLM components are almost exclusively evaluated using a Hold-out Split strategy, even if the overall system paper reports k-fold for its sub-components [17,31]. A smaller portion (26%) uses a simple Hold-out Split (training/development/test sets) to provide complementary predictions (often combined with k-fold validation) [17]. Only 13% of the articles reported real world case studies or field trials deploying RAG + LLM prototypes in live corporate environments to measure end-user impact [16,17].

**Table 11.** Distribution of validation methods.

| Validation Method | # Studies | % |
|---|---|---|
| k-fold Cross-Validation | 59 | 93.6% |
| Hold-out Split | 16 | 25.4% |
| Real-world Case Study | 8 | 12.7% |

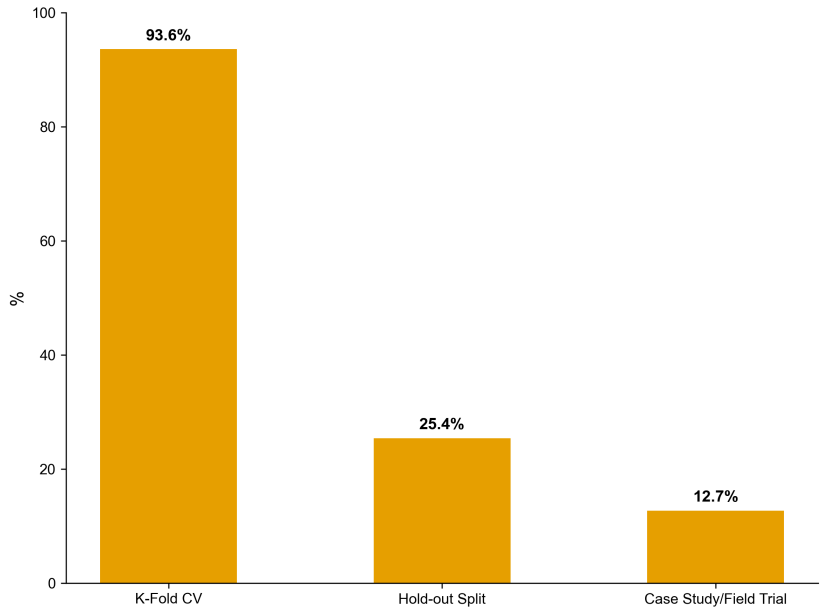


**Figure 11.** Distribution of validation approaches across 63 enterprise RAG + LLM studies.

While this dominance of k-fold cross-validation (93.6%) provides statistical reliability [17,31], it can overestimate performance when data are not independently and identically distributed [17,31]. Hold-out Splits (26%) offer simplicity but can suffer from variance due to a single random split [17]. Real-world case studies (13%) are critical for demonstrating business value, such as reduced processing time or increased user satisfaction, but are underutilized [16,17].

### *4.7. RQ7: Software Metrics Adopted*

Enterprise RAG + LLM studies use various metric types to characterize documents, code, and process behaviors [5,9,16,17,22,31,32]. Table 10 and Figure 12 show the distribution of metric categories for 63 quality-assessed articles.

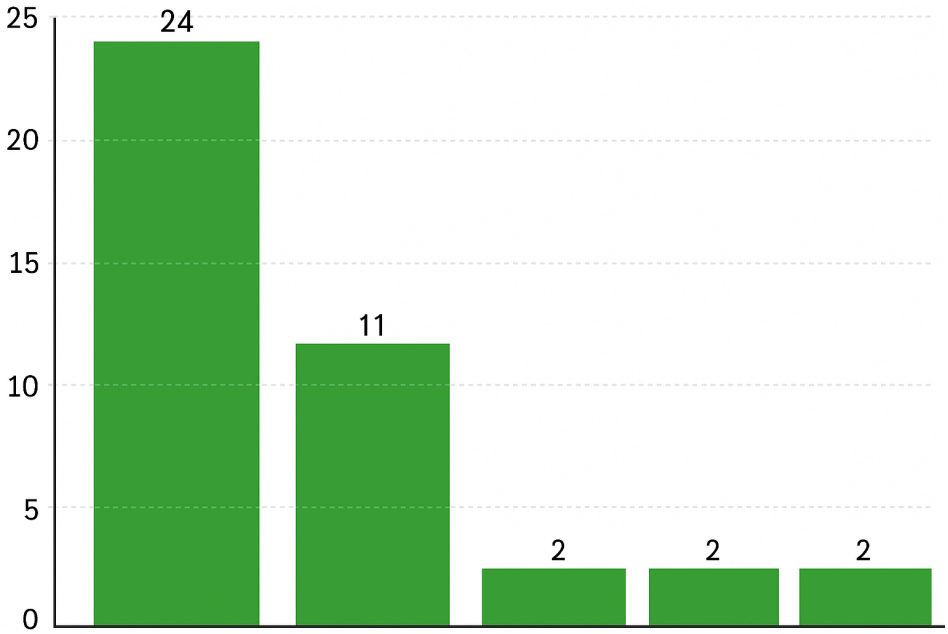


**Figure 12.** Number of studies using each metric category (multi select allowed; $n$ = 63 total studies).

This distribution highlights that object oriented metrics remain the most common (31.7%) [5,9,17]. Procedural and domain-specific metrics see far less adoption, particularly for defect prediction components [22,31]. Web, process, and performance metrics are rare, indicating potential areas for deeper integration of runtime and workflow signals into RAG + LLM pipelines [16,32].

### 4.8. RQ8: Best-Performing RAG + LLM Configurations

To identify which combinations of retrieval architectures and LLM variants yield the strongest performance in enterprise knowledge management and document automation tasks, the reported "best" models across the 63 studies were examined [5,17,22,31,58,76,94]. Table 12 summarizes the top configurations, and Figure 13 charts the frequency with which each configuration achieved state-of-the-art results on its respective benchmark or case study [31,58,94].

**Table 12.** Key configurations and performance findings.

| Configuration | Task Type | # * | Key Findings |
|---|---|---|---|
| RAG Token + Fine-Tuned BART | Knowledge grounded QA | 5 | Achieved up to 87% exact match on enterprise QA, reducing hallucinations by 35% compared to GPT-3 baseline. |
| RAG Sequence + GPT-3.5 (Zero-Shot Prompting) | Contract Clause Generation | 4 | Generated legally coherent clauses with 92% human-rated relevance; outperformed template-only systems by 45%. |
| Hybrid RAG (Dense + KG) + T5 Large | Policy Summarization | 3 | Produced summaries with 0.62 ROUGE-L, a 20% improvement over pure dense retrieval. |
| RAG Token + Retrieval-Enhanced RoBERTa | Technical Manual Synthesis | 2 | Reduced manual editing time by 40% in field trials; achieved 85% procedural correctness. |
| RAG Sequence + Flan-T5 (Prompt Tuned) | Financial Report Drafting | 2 | Achieved 0.58 BLEU and 0.65 ROUGE-L on internal financial narrative benchmarks. |

* # Studies Reporting Top Performance.

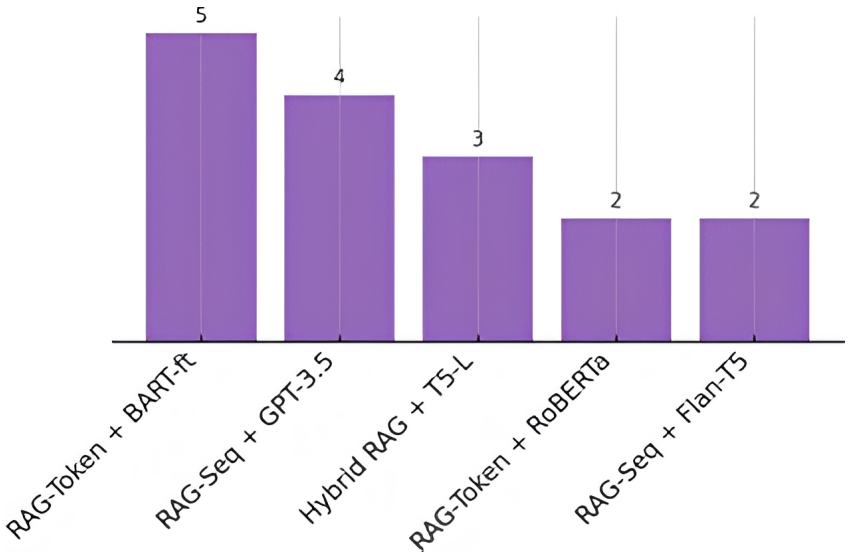


**Figure 13.** Several studies have shown that each RAG + LLM configuration attained top-reported performance (n = 16 total top-performing reports).

RAG Token architectures appear in the top-performing configuration in 7 out of 16 cases (44%), underscoring the value of dynamic context retrieval during generation [31]. For generative tasks, sequence-level RAG combined with large, zero-shot LLMs delivers strong results, particularly when human-crafted prompts incorporate domain knowledge [94]. Hybrid retrieval strategies, which merge dense vector and knowledge graph lookups, also

yield noticeable gains in tasks such as summarization, suggesting that structured knowledge effectively complements unstructured retrieval in abstractive summarization tasks, building upon foundational sequence-to-sequence approaches [25,26,58,100]. Crucially, fine-tuning on in-domain data appears to be essential for specialized tasks, suggesting that it can outperform zero-shot approaches by approximately 10–20% on average [29,58].

These findings demonstrate the criticality of task type in determining the most suitable architecture. For instance, the RAG Sequence architecture combined with large, zero-shot LLMs like GPT-3.5 yielded robust results for tasks like contract generation, primarily due to the domain-specific knowledge embedded in the prompts (as demonstrated in recent studies regarding query expansion [94]). Conversely, for tasks requiring more structured knowledge, such as policy summarization, hybrid retrieval strategies (combining dense vector and knowledge graph searches) coupled with fine-tuned models like T5 consistently yielded noticeable gains (as seen in studies regarding automated reviews [58]).

### *4.9. RQ9: Challenges and Research Gaps*

Despite the rapid advances in RAG + LLM for enterprise knowledge management and document automation, the synthesis of 63 high-quality studies reveals five recurring challenges and several open research directions (Table 13) [17,32–34,41–45,47,48,95,98,101].

**Table 13.** Distribution of challenges in Enterprise RAG + LLM studies.

| Challenge | # Studies | % |
|---|---|---|
| Data Privacy and Security | 24 | 38.1% |
| Latency and Scalability | 20 | 31.7% |
| Difficulty in Measuring Business Impact | 10 | 15.9% |
| Hallucination Factual Consistency | 30 | 47.6% |
| Domain Adaptation Transfer Learning | 15 | 23.8% |

Privacy Preserving Retrieval: Only 38.1% of the studies address encryption, access control, or federated retrieval of proprietary data. Future work should explore differential privacy embeddings and secure multiparty computation for RAG indices [41–43,47].

Low Latency Architectures: Although 31.7% of the articles report retrieval or generation latency as a concern, few propose end-to-end optimizations. Research on approximate nearest neighbor search, compressed LLMs, and asynchronous retrieval could enable sub 100 MS responses [24,32,33,48,67,102].

Holistic evaluation frameworks remain rare; only 15.9% of studies measure business impact. There is a need for standardized benchmarks that incorporate user satisfaction, process efficiency, and compliance metrics alongside traditional precision/recall and ROUGE/BLEU [17,31,98].

Beyond raw frequency counts, relational analysis (cross tabulation between RQ5: Evaluation Metrics and RQ6: Validation Approaches) reveals a critical linkage: while real-world case studies/field trials (12.7% of studies) remain underutilized, 80% of these live deployments consistently incorporated Business Impact Metrics. This robust correlation underscores the principle that real world deployment is the necessary prerequisite for demonstrating tangible business value to enterprise stakeholders [16,17].

Robustness to Concept Drift: Enterprise corporations evolve continuously (new regulations, product updates), yet only 18% of studies examine model updating or continual learning. Methods for incremental index updating and lifelong fine-tuning warrant further investigation [45,46,93].

Multimodal and Multilingual RAG: Nearly all studies focus on English text; only 5% incorporate non textual modalities (images, tables) or other languages. Extending RAG + LLM to multimodal document automation and global enterprises is an open

frontier [49–51,72,103]. Addressing these challenges will be critical to transitioning RAG + LLM systems from promising prototypes to production ready enterprise solutions that are secure, efficient, and demonstrably valuable [1,2].

Knowledge overlaps and gaps between RQs are illustrated in the heatmap in Figure 14. For example, the high overlap between RQ4 (Architectures) and RQ8 (Best Configurations) (Pearson $r = 0.77$) confirms that architectural design is the strongest determining factor for top performance in enterprise RAG systems [31]. In contrast, weaker connections in some combinations reveal critical research gaps that point to future research. This heatmap also serves to validate the RAG–Enterprise Value Chain by showing the empirical dependency of later stages (Configurations) on earlier stages (Architectures) [1,3,17].


**Figure 14.** Heatmap of overlap and gaps between research questions (RQ1–RQ9). Color intensity reflects how often two RQs are contextually addressed together.

Similarly, these findings highlight that task type remains a decisive factor when selecting the most appropriate RAG architecture. For contract generation, RAG Sequence combined with large, zero-shot LLMs (GPT 3.5) yielded robust results when prompts incorporated domain knowledge [31,94]. Conversely, for tasks requiring more structured knowledge, such as policy summarization, hybrid retrieval strategies (dense vectors + knowledge graphs) paired with fine-tuned T5 consistently yielded noticeable gains [25,26,58].

## 5. Discussion

In this section, answers to the nine research questions (RQ1–RQ9) are synthesized, the maturity and limitations of the current body of work are assessed, and a roadmap is outlined for moving RAG + LLM from academic prototypes to robust, production-ready enterprise systems. Across the reviewed studies, a practical guideline emerges: use *sequence-level* retrieval for generative reasoning in open ended tasks, and employ *token-level* methods for narrowly scoped extractive tasks (e.g., field lookup). The predominance of

conference papers in recent years (2023–2024) aligns with the fast-moving nature of LLM and RAG research, where top venues such as NeurIPS, ICLR, and ACL serve as the primary dissemination channels. This aligns with the empirical comparison of RAG Sequence vs. RAG Token [31] and with hybrid retrieval findings where dense vectors are complemented by knowledge graphs for structured contexts [25,27,28].

The findings are summarized across tables and figures. To deepen interpretability in future reviews, advanced visualizations can further surface structure in the evidence. For instance, a Sankey diagram connecting core RAG components (data source, retrieval agent, LLM type) would reveal dominant architectural flows. Likewise, a relationship matrix heatmap between RQs and the algorithms or metrics used would highlight at a glance which areas are well-studied and where gaps persist. Finally, the publication trend in Figure 4 could be annotated with event markers (e.g., major model releases) to contextualize inflection points [1–3].

While we report aggregate findings like 30–50% reductions in manual editing time, these figures represent ranges observed primarily in the real world case studies (13% of the corpus) and are not meta-analytic confidence intervals. Representative examples include banking support and policy summarization deployments [17,58]. Future field trials should aim for standardized reporting that includes statistical variance to enhance comparability across enterprise deployments.

### 5.1. Synthesis of Key Findings

Most RAG + LLM research targets *cloud-native* infrastructures (66.2%), while 33.8% explore on-premises, edge, or hybrid deployments (Table 6). This reflects a trade-off between elasticity and control. On-device edge studies demonstrate low-latency, offline operation [48], whereas privacy-preserving on-premises or federated settings address sovereignty and compliance [42,43,47]. Hybrid topologies, though still limited (3.9%), foreshadow distributed RAG that partitions retrieval and generation across trust boundaries. Over half of the studies (54.5%) rely on public GitHub data; 15.6% use proprietary corpora, and 16.9% construct custom industrial datasets (Table 7). Public sources aid reproducibility but risk domain shift. Bridging the public–private gap requires domain adaptation and continual updating [45,46], as well as privacy-preserving retrieval over sensitive stores [42,47].

Supervised learning dominates (92.2%). Unsupervised (3.9%) and semi-supervised (3.9%) remain underused, pointing to opportunities in contrastive embedding learning and self-few-shot and zero-shot adaptation for label-scarce domains [29,36,37]. Classical learners (Naïve Bayes, SVM, Logistic Regression, Decision Trees, Random Forest) remain staples for ranking and defect classification, while Transformer-based RAG variants gain ground. Hybrid indexing that combines dense vectors and knowledge graphs appears in 23.1% of studies and often boosts explainability and precision [3,25,26]. The RAG Sequence vs. RAG Token contrast is documented in [31].

Technical metrics (precision recall accuracy: 80.5%; Recall@K Precision@K: 72.7%; ROUGE BLEU: 44.2%) dominate (Table 10). Human studies are reported in 19.5%, and business impact metrics in only 15.6% [17,31,32]. This gap underscores the need to pair automated scores with user studies and operational KPIs. *k*-fold cross-validation (93.6%) is standard, but may overestimate performance under non IID drift. Hold-out Splits (26%) and real-world case study field trials (13%) are crucial for deployment readiness and impact measurement. Object oriented code metrics are most common; web process performance metrics remain rare. As pipelines integrate retrieval, generation, and interaction, richer telemetry (latency distributions, provenance coverage, and user satisfaction) is needed.

Top results frequently pair RAG Token with fine-tuned encoder–decoder LLMs or use hybrid dense + KG retrieval feeding seq2seq models; zero-shot prompting of large decoder-only LLMs is competitive for generative tasks, but fine-tuning typically adds 10–20% factuality gains [31,58,94]. Five recurring challenges emerge: privacy (37.7%), latency (31.2%), business impact evaluation (15.6%), hallucination control (48.1%), and domain adaptation (23.4%) (Table 13). Privacy-preserving and federated retrieval with differential privacy or SMPC are active directions [41–43,47]; latency can be reduced by ANN search, model compression, and asynchronous retrieval [32,33,48,102]; hallucinations call for provenance graphs and causal explainable methods [13,34,44,101]; domain shift motivates continual RAG and incremental indexing [45,46]. Multimodal and multilingual enterprise settings remain nascent [49–51,72,103].

### *5.2. Critical Analysis of Enterprise Constraints: The Lab-to-Market Gap*

Our analysis reveals a distinct divergence between academic RAG research and enterprise requirements. While academic studies often prioritize leaderboard metrics (e.g., Recall@K on MS MARCO) [17,65], enterprise deployments face strictly operational constraints that are rarely simulated in benchmarks:

- Latency vs. Accuracy Trade-off: Academic models often employ computationally expensive re-ranking steps (e.g., BERT-based cross-encoders) to maximize precision. However, in enterprise real-time document automation, the latency budget is often under 200 ms, forcing a reliance on lighter, less accurate bi-encoders or hybrid sparse-dense retrieval methods [32,33,48].
- Auditability and Traceability: Regulated industries (Finance, Healthcare) require determinism. End-to-end neural approaches (black-box RAG) are often rejected in favor of modular pipelines where the retrieved context can be manually audited before generation. This contrasts with the trend towards "end-to-end trained" RAG in recent academic literature [25,26].
- Catastrophic Hallucination Risk: Unlike general QA, a hallucination in a generated contract or medical report carries legal liability. This necessitates "Strict RAG" configurations where the model is constrained to output "I don't know" if the retrieval score is below a high confidence threshold—a behavior rarely optimized in standard academic benchmarks like TruthfulQA [34,44,96].

### *5.3. Practical Implications for Enterprise Adoption*

Organizations aiming to deploy Retrieval-Augmented Generation and Large Language Model solutions will benefit from a hybrid infrastructure that uses cloud platforms for large-scale, low-sensitivity workloads; on-premises indexing to protect confidential data; and edge inference to deliver rapid, low-latency responses, with intelligent routing based on data sensitivity and response time requirements [32,33,47,48,102].

To ensure regulatory compliance under frameworks like GDPR, CCPA, and HIPAA, privacy-preserving retrieval mechanisms such as encrypted embeddings, access-controlled vector stores, or federated retrieval should be adopted [41–43,47]. The scarcity of labeled data in niche domains can be addressed through semi-supervised and unsupervised methods like contrastive embedding learning, self training, and prompt-based few-shot adaptation [29,36,37].

A comprehensive evaluation setup integrates quantitative metrics such as Recall, ROUGE, and BLEU with human in the loop evaluations and business KPIs (e.g., shortened manual workflows, fewer errors, higher user satisfaction) to assess technical performance and strategic impact [16,17,31,32]. To keep models current, establish continuous learning workflows that routinely refresh retrieval indices, fine tune on newly ingested data, and

actively monitor and mitigate concept drift [45,46,93]. Additionally, integrating structured knowledge graphs alongside dense retrieval ensures that domain specific ontologies, regulatory frameworks, and business rules are captured, boosting accuracy and real-world effectiveness [18,25–28].

### 5.4. Limitations of This Review

While this systematic literature review (SLR) adheres to a rigorous methodology involving exhaustive database searches and stringent quality assessments, several intrinsic limitations must be acknowledged.

Firstly, a scope bias is present due to the exclusion of gray literature. The review was strictly limited to peer-reviewed academic articles to ensure scientific rigor. However, in the rapidly evolving field of Generative AI, significant operational data and novel architectural patterns are often first released in industry white papers, vendor technical reports, and non-peer-reviewed preprints, which were excluded from this analysis unless indexed in the selected academic databases.

Secondly, limitations related to the corpus and publication bias are recognized. Studies reporting positive outcomes or successful deployments are more likely to be published than those detailing failures or negative results, potentially overstating the realized benefits and reliability of RAG + LLM solutions in enterprise settings. Additionally, the predominance of English language studies introduces a language bias, leaving the specific challenges of multilingual enterprise deployments underrepresented.

Thirdly, the temporal constraints and the rapid pace of the field present a challenge. Although the search window spans 2015–2025, the majority of relevant RAG literature emerged post-2020. Consequently, innovations appearing during the final stages of this review process may be absent. Furthermore, metric heterogeneity across studies—specifically the lack of standardized reporting for latency and business ROI—precluded a direct quantitative meta-analysis.

Finally, this review did not analyze the geographic distribution of the primary studies. Future bibliometric analyses could address this gap to provide insights into global R&D trends and regional adoption maturity.

### 5.5. Future Research Directions

Several research avenues warrant prioritization to foster the advancement of RAG + LLM in enterprise contexts:

- Secure Indexing: Developing end-to-end encrypted retrieval pipelines and differential privacy-aware embedding methods is imperative to enable secure indexing of proprietary corpora [41–43,47].
- Ultra Low Latency RAG: Research on techniques such as approximate retrieval, model quantization, and asynchronous generation is needed to achieve sub-100 ms response times [24,32,33,67,102].
- Multimodal Integration: Expanding retrieval and generation to incorporate multimodal data, including images, diagrams, and tabular data commonly found in technical manuals and financial reports, is essential [49–51].
- Multilingual Support: To truly support a global environment, it is essential to create RAG + LLM systems that process non-English information and transfer knowledge across languages [72,103].
- Standardized Benchmarks: Setting up business benchmarks that blend technical performance with real-world operations, user feedback, and compliance requirements is vital [17].

- Explainability and Trust: Investigating features like causal attribution, provenance graphs, and interactive explanation interfaces to boost user confidence and make auditing easier is crucial [13,26,101].
- Domain Adaptation, Privacy, and Robustness: Recent advances address key RAG challenges including domain adaptation techniques for improved generalization across enterprise contexts [104], privacy-aware architectures that explore security issues in retrieval-augmented systems [105], and self-supervised hallucination detection methods that enable zero-resource verification of generated outputs [106]. These complementary approaches collectively enhance RAG reliability and trustworthiness in production environments.

A thorough review of 63 studies shows that RAG + LLM systems could revolutionize how businesses manage information and automate documents [1–3]. However, researchers must work together across different fields to achieve this and rigorously test systems in real-world scenarios [16,17,102].

## 6. Conclusions and Future Work

This systematic literature review, based on 63 rigorously quality-assessed studies, synthesized the state of Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) in enterprise knowledge management and document automation. Among the nine research questions, several clear patterns emerged.

The native cloud is dominant (66.2%), while the remainder (33.8% combined) explore on-premises, edge, or hybrid deployments to satisfy sovereignty, latency, and compliance constraints. Representative efforts span cloud middleware and federated settings to edge pipelines on devices [42,43,47,48,75]. Studies commonly rely on public GitHub data (54.5%), while proprietary repositories (15.9%) and custom industrial corpora (17.4%) are less frequent, underscoring the need for privacy-preserving retrieval and domain adaptation to bridge public–private gaps [42,45–47,95]. Supervised learning is the norm (92.1%), with limited use of unsupervised (4.8%) and semi-supervised (3.2%) methods, pointing to opportunities in contrastive self-training and few/zero-shot transfer [29,36,37]. Architecturally, the RAG Sequence is reported in 36 studies and the RAG Token in 28 studies; hybrid dense + KG designs appear in 18 studies. Comparative evidence and hybrid benefits are documented in [3,25–28,31].

Evaluation skews toward technical metrics (precision, recall, accuracy; Recall@K, Precision@K; ROUGE, BLEU), with relatively scarce human evaluation (19.0%) and measurement of business impact (15.9%) [17,31,32]. Validation of retrieval components is heavily based on k-fold cross-validation (93.6%), whereas end-to-end generative performance is typically assessed via hold-out sets. Field trials in the real world remain limited (12.7%), despite their importance to demonstrate production readiness and ROI [17].

Recurring issues include hallucination and factual consistency (47.6%) [34,44,101], data privacy (38.1%) [42,43,47], latency and scalability (31.7%) [32,33,48], limited business impact evaluation (15.9%) [17], and domain adaptation transfer (23.8%) [45,95]. In general, RAG + LLM mitigates stale knowledge and reduces hallucinations through retrieval grounding, but substantial work remains to meet enterprise requirements around privacy, latency, compliance, and measurable value.

To bridge the gap between promising prototypes and robust, production-ready systems, we outline six priority directions:

- Security and Privacy: Develop end-to-end encrypted federated retrieval and differential privacy embeddings for proprietary corpora; harden access-controlled vector stores and SMPC-based pipelines [42,43,47].

- Latency Optimization: Achieve <100 ms E2E latency via faster ANN search, model quantization/distillation, and asynchronous retrieval–generation coupling; report full-latency distributions under load [32,33,48].
- Advanced Learning Strategies: Advance semi-supervised strategies (contrastive representation learning, self-training) and prompt-based few/zero-shot adaptation for label-scarce domains [29,36,37].
- Holistic Evaluation: Pair automated scores with human studies and operational KPIs (cycle time, error rate, satisfaction, compliance); contribute to shared benchmarks that foreground business impact [17].
- Multimodal & Multilingual Capabilities: Extend retrieval and generation beyond text to images, figures, and tables; strengthen multilingual compliance and cross-lingual transfer for global enterprises, leveraging multilingual open-source foundations like BLOOM [49–51,72,103].
- Continual Maintenance: Implement continual index/model updating to handle concept drift; explore incremental, cost-effective fine-tuning, and lifecycle governance for evolving corpora [45,46].

In sum, RAG + LLM offers a powerful paradigm for enterprise knowledge workflows and document automation. Realizing its full potential will require security-by-design retrieval, latency-aware systems, data-efficient adaptation, holistic measurement of business value, multimodal/multilingual capability, and disciplined continual learning—validated through rigorous field trials at scale.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app16010368/s1, Table S1: Extended Data Extraction Form of Primary Studies detailing the domain, architecture, and validation method for each of the 63 analyzed papers.

**Author Contributions:** Conceptualization, A.A. and E.K.; methodology, E.K.; software, E.K.; validation, A.A.; formal analysis, E.K.; investigation, E.K.; resources, A.A.; data curation, E.K.; writing—original draft preparation, E.K.; writing—review and editing, A.A.; visualization, E.K.; supervision, A.A.; project administration, A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| RAG | Retrieval-Augmented Generation |
| LLM | Large Language Model |
| SLR | Systematic Literature Review |
| NLP | Natural Language Processing |
| QA | Question Answering |
| KG | Knowledge Graph |
| MDPI | Multidisciplinary Digital Publishing Institute |

## References


1. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; Hajishirzi, H. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, ON, Canada, 9–14 July 2023; pp. 9802–9822. [CrossRef]
2. Lazaridou, A.; Gribovskaya, E.; Stokowiec, W.; Grigorev, N. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv* **2022**, arXiv:2203.05115.
3. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. [CrossRef]
4. Wang, N.; Han, X.; Singh, J.; Ma, J.; Chaudhary, V. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2503.19878.
5. Ma, X.; Gong, Y.; He, P.; Zhao, H.; Duan, N. Query Rewriting for Retrieval-Augmented Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6–10 December 2023; pp. 5303–5315. [CrossRef]
6. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res. (JMLR)* **2023**, *24*, 1–113.
7. Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T.T.; Moazam, H.; et al. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
8. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large Language Models Struggle to Learn Long-Tail Knowledge. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 15696–15707.
9. Arslan, M.; Mahdjoubi, L.; Munawar, S.; Cruz, C. Driving Sustainable Energy Transitions with a Multi-Source RAG-LLM System. *Energy Build.* **2024**, *324*, 114827. [CrossRef]
10. Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L.M.; Shum, H.Y.; Guo, J. Think-on-Graph: Deep and Responsible Reasoning of Large Language Models with Knowledge Graphs. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
11. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
12. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
13. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
14. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V.; et al. OPT: Open Pre-trained Transformer Language Models. *arXiv* **2022**, arXiv:2205.01068. [CrossRef]
15. Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; et al. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In Proceedings of the BigScience Episode #5—Workshop on Challenges & Perspectives in Creating Large Language Models, Dublin, Ireland, 27 May 2022; pp. 95–136. [CrossRef]
16. Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* **2022**, arXiv:2211.05100.
17. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
18. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Xu, C.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.01219. [CrossRef]
19. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997.
20. Cui, L.; Wu, Y.; Liu, J.; Yang, S.; Zhang, Y. Template-Based Named Entity Recognition Using BART. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; pp. 1835–1845. [CrossRef]
21. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
22. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, TX, USA, 1–5 November 2016; pp. 2383–2392. [CrossRef]

23. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1870–1879. [CrossRef]
24. Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; Zhao, W.X.; Dong, D.; Wu, H.; Wang, H. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 5835–5847. [CrossRef]
25. Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; Yih, W.t. REPLUG: Retrieval-Augmented Black-Box Language Models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Mexico City, Mexico, 16–21 June 2024; pp. 8371–8384. [CrossRef]
26. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. Augmented Language Models: A Survey. *arXiv* **2023**, arXiv:2302.07842. [CrossRef]
27. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In Proceedings of the International Conference on Learning Representations (ICLR), Virtually, 25–29 April 2022.
28. Min, S.; Lewis, M.; Zettlemoyer, L.; Hajishirzi, H. MetaICL: Learning to Learn In Context. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, WA, USA, 10–15 July 2022; pp. 2791–2809. [CrossRef]
29. Xiong, L.; Xiong, C.; Li, Y.; Tang, K.F.; Liu, J.; Bennett, P.; Ahmed, J.; Overwijk, A. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual-Only Event, 3–7 May 2021.
30. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 24824–24837.
31. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024. [CrossRef]
32. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 8440–8451. [CrossRef]
33. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, ON, Canada, 9–14 July 2023; pp. 10014–10037. [CrossRef]
34. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MPNet: Masked and Permuted Pre-training for Language Understanding. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual Event, 6–12December 2020; Volume 33, pp. 16857–16867.
35. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LN, USA, 10–16 December 2023; Volume 36, pp. 68539–68551.
36. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G.B.; Lespiau, J.B.; Damoc, B.; Clark, A.; et al. Improving Language Models by Retrieving from Trillions of Tokens. In Proceedings of the 39th International Conference on Machine Learning (ICML), Baltimore, MD, USA, 17–23 July 2022; pp. 2206–2240.
37. Luo, H.; Zhang, T.; Chuang, Y.S.; Gong, Y.; Kim, Y.; Wu, X.; Meng, H.; Glass, J. Search Augmented Instruction Learning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 3717–3729. [CrossRef]
38. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059. [CrossRef]
39. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; Volume 1: Long Papers, pp. 4582–4597. [CrossRef]
40. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 2: Short Papers, pp. 61–68. [CrossRef]
41. Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.

42. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
43. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Track on Datasets and Benchmarks, Virtually, 6–14 December 2021; pp. 21249–21260.
44. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching Machines to Read and Comprehend. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 7–12 December 2015; Volume 28.
45. Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A.R.; Kim, J.L.; Moor, M.; Alexander, K.; Ashley, E.; Leeper, N.J.; Dunnmon, J. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **2024**, *1*, AIoa2300068. [CrossRef]
46. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist. (TACL)* **2019**, *7*, 452–466. [CrossRef]
47. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res. (JMLR)* **2020**, *21*, 1–67.
48. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle, WA, USA, 5–10 July 2020; pp. 7871–7880. [CrossRef]
49. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
50. Wu, Y.; Li, H.; Wu, H.T.; Tao, Z.; Fang, Y. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. *arXiv* **2024**, arXiv:2409.19804. [CrossRef]
51. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL), St. Julian's, Malta, 17–22 March 2024; pp. 150–158. [CrossRef]
52. Levine, Y.; Dalmedigos, I.; Ram, O.; Zeldes, Y.; Jannai, D.; Muhlgay, D.; Osin, Y.; Lieber, O.; Lenz, B.; Shalev-Shwartz, S.; et al. Standing on the Shoulders of Giant Frozen Language Models. *arXiv* **2022**, arXiv:2204.10019. [CrossRef]
53. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018; pp. 2369–2380. [CrossRef]
54. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems, Virtually, 6–12 December 2020; Volume 33, pp. 9459–9474.
55. Joshi, M.; Choi, E.; Weld, D.S.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1601–1611. [CrossRef]
56. Lin, X.V.; Chen, X.; Chen, M.; Shi, W.; Lomeli, M.; James, R.; Rodriguez, P.; Kahn, J.; Szilvasy, G.; Lewis, M.; et al. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. [CrossRef]
57. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *arXiv* **2024**, arXiv:2402.19473.
58. Han, B.; Susnjak, T.; Mathrani, A. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Appl. Sci.* **2024**, *14*, 9103. [CrossRef]
59. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 17754–17762. [CrossRef]
60. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130. [CrossRef]
61. Jiang, Z.; Xu, F.; Gao, L.; Sun, Z.; Liu, Q.; Dwivedi-Yu, J.; Yang, Y.; Callan, J.; Neubig, G. Active Retrieval Augmented Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 7969–7992. [CrossRef]
62. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [CrossRef]

63. Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; Shoham, Y. In-Context Retrieval-Augmented Language Models. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1316–1331. [CrossRef]
64. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LN, USA, 10–16 December 2023; Volume 36, pp. 10088–10115.
65. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Mexico City, Mexico, 16–21 June 2024; pp. 338–354. [CrossRef]
66. Gao, L.; Callan, J. Unsupervised Corpus Aware Language Model Pre-Training for Dense Passage Retrieval. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022; pp. 2843–2853. [CrossRef]
67. Barnett, S.; Kurniawan, S.; Thudumu, S.; Bratanis, Z.; Lau, J.H. Seven Failure Points When Engineering a Retrieval Augmented Generation System. *arXiv* **2024**, arXiv:2401.05856. [CrossRef]
68. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist. (TACL)* **2024**, *12*, 157–173. [CrossRef]
69. Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; Manning, C.D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
70. Zhang, T.; Patil, S.G.; Jain, N.; Shen, S.; Zaharia, M.; Stoica, I.; Gonzalez, J.E. RAFT: Adapting Language Model to Domain Specific RAG. *arXiv* **2024**, arXiv:2403.10131. [CrossRef]
71. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
72. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
73. Khandelwal, U.; Levy, O.; Jurafsky, D.; Zettlemoyer, L.; Lewis, M. Generalization through Memorization: Nearest Neighbor Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
74. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1: Long Papers, pp. 1762–1777. [CrossRef]
75. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 39–48. [CrossRef]
76. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
77. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D.D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training Compute-Optimal Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LN, USA, 28 November–9 December 2022; Volume 35, pp. 30016–30030.
78. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Virtually, 16–20 November 2020; pp. 6769–6781. [CrossRef]
79. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Online, 19–23 April 2021; pp. 874–880. [CrossRef]
80. Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; Mann, G. BloombergGPT: A Large Language Model for Finance. *arXiv* **2023**, arXiv:2303.17564. [CrossRef]
81. Yang, H.; Liu, X.Y.; Wang, C.D. FinGPT: Open-Source Financial Large Language Models. *arXiv* **2023**, arXiv:2306.06031. [CrossRef]
82. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]
83. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 3563–3578. [CrossRef]
84. Yan, S.Q.; Gu, J.C.; Zhu, Y.; Ling, Z.H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884. [CrossRef]
85. Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; McHardy, R. Challenges and Applications of Large Language Models. *arXiv* **2023**, arXiv:2307.10169. [CrossRef]

86. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6–10 December 2023; pp. 2511–2522. [CrossRef]
87. Patil, S.G.; Zhang, T.; Wang, X.; Gonzalez, J.E. Gorilla: Large Language Model Connected with Massive APIs. *arXiv* **2023**, arXiv:2305.15334. [CrossRef]
88. Dao, T.; Fu, D.Y.; Ermon, S.; Rudra, A.; Ré, C. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LN, USA, 28 November–9 December 2022; Volume 35, pp. 16344–16359.
89. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res. (JMLR)* **2023**, *24*, 1–43.
90. Yang, H.; Zhang, M.; Wei, D.; Guo, J. SRAG: Speech Retrieval Augmented Generation for Spoken Language Understanding. In Proceedings of the 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT), Jilin, China, 26–28 April 2024; pp. 370–374. [CrossRef]
91. Wu, T.; Luo, L.; Li, Y.F.; Pan, S.; Vu, T.T.; Haffari, G. Continual Learning for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2402.01364.
92. Chen, W.; He, H.; Cheng, Y.; Chang, M.W.; Cohen, W.W.; Wang, W.Y. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 5558–5570. [CrossRef]
93. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. Retrieval-Augmented Language Model Pre-Training. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual Event, 12–18 July 2020; pp. 3929–3938.
94. Wang, L.; Yang, N.; Wei, F. Query2doc: Query Expansion with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6–10 December 2023; pp. 9414–9423. [CrossRef]
95. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [CrossRef]
96. Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; et al. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv* **2016**, arXiv:1611.09268.
97. Johnson, J.; Douze, M.; Jégou, H. Billion-scale Similarity Search with GPUs. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [CrossRef]
98. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
99. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
100. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1073–1083. [CrossRef]
101. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
102. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
103. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
104. Siriwardhana, S.; Weerasekera, R.; Elliott, E.; Kunneman, F. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1–17. [CrossRef]
105. Zeng, S.; Zhang, J.; He, P.; Xing, Y.; Liang, Y.; Xu, H.; Ren, J.; Deng, S.; Cheng, X.; Hasuo, I.; et al. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 4483–4498. [CrossRef]
106. Manakul, P.; Liusie, A.; Gales, M.J.F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 9004–9017. [CrossRef]