Article

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

Posted Date: 4 December 2025

doi: 10.20944/preprints202512.0359.v1

Keywords: Retrieval Augmented Generation; Large Language Models; Enterprise Knowledge Management; Document Automation; Systematic Literature Review

Article

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, Turkey
* Correspondence: 1507010018@stu.iku.edu.tr;

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 77 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 (93.5%), 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, 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 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 77 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 77 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]. 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 [32–34,38–41]. Based on these insights, we outline best practice recommendations for deployers: modular system design, continuous index updating, efficient nearest neighbor search, federated device retrieval, and hybrid evaluation frameworks that combine automated metrics with human feedback [24,42–45]. Open research directions are also identified, such as multi-modal RAG architectures integrating text, image, and tabular data [46–48]; adaptive retrieval strategies that personalize context based on user profiles [49,50]; 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

In this section, we first introduce the technical foundations of Retrieval Augmented Generation (RAG) and large language models (LLMs), then describe their key applications in enterprise knowledge management and document automation, and finally review existing systematic and mapping studies that have surveyed RAG, LLMs, and related techniques [1–3]. All quantitative figures refer to the final pool of 77 high quality studies analyzed, published between 2015 and 2025 (Section 3 and our methodology overview in [35]).

2.1. Retrieval Augmented Generation and Large Language Models

Over the past five years, transformer based large language models have revolutionized natural language processing by demonstrating strong fluency in generation, summarization, translation, and code synthesis [51–53]. However, two fundamental limitations hinder their direct use in dynamic, domain specific settings: a fixed knowledge cutoff and reliance on internal representations from the last pre training update, which contributes to factual errors and hallucinations in niche or fast moving domains [23,34,41].

RAG addresses these challenges by integrating LLMs' generative capabilities with the retrieval of relevant, up to date information from external knowledge sources [1,2,19,20]. In the original RAG formulation [54], a neural retriever retrieves the top K passages by dense similarity; these are concatenated with the query and conditioned into the generator. Variants studied in the literature include RAG Sequence vs. RAG Token and iterative "retrieve reread" pipelines [31,49]. Indexing strategies span dense vector indexes, sparse first stage filters such as BM25, and symbolic graph based stores or hybrids (KG + dense) [1,18,25,26,42].

Regarding LLM backbones, both decoder only (GPT style) and encoder–decoder models (e.g., T5/BART) are commonly used depending on latency and controllability needs [51–53]. Reported advantages of RAG include improved factuality and provenance-based citations and adaptability through independent corpus updates without retraining LLM weights [19,42,49]. Ongoing research focuses on retrieval latency and tight retrieval–generation coupling, as well as multimodal extensions [32,33,46–48].

2.2. Enterprise Knowledge Management and Document Automation

Enterprise environments demand robust and scalable solutions for knowledge management (KM) and document automation [9,18,26]. Traditional KM systems are based on keyword search and manual taxonomies, struggling with the volume, velocity, and variety of modern corporate data [1,18]. Similarly, template based automation can falter when narrative sections require cross document reasoning or compliance grounding [9,55,56].

Integrating RAG with LLMs enables dynamic retrieval of domain specific content (e.g., statutes, engineering specs) to ground generation [9,26]. Early enterprise case studies report efficiency and quality gains in customer support, FAQ automation, and internal knowledge retrieval [16]. Thus, RAG+LLM offers a path that bridges template rigidity and ungrounded generation, delivering context aware automation at scale [9,18].

2.2.1. Knowledge Management

Classical KM pipelines keyword search over repositories and manual taxonomies struggle to scale as organizations accumulate heterogeneous content (unstructured text, spreadsheets, wikis, logs, audio) [1,18]. Rapidly changing policies and procedures make static indices stale unless actively maintained [42]. Based on our review of 77 studies, enterprise data span PDFs, spreadsheets, wikis, and transcripts across multiple domains (Table 1). RAG+LLM enables contextual Q&A that can retrieve policy clauses or troubleshooting steps on the fly, with reported improvements in first contact resolution and analyst throughput [16]. Automated citation backed summaries can reduce manual review effort and increase consistency in regulated settings [9,16].

Table 1. Distribution of Studies by Knowledge Management Domain.

Domain	# Papers	%
Regulatory compliance governance	20	26.0%
Contract legal document automation	18	23.4%
Customer support chatbots	15	19.5%
Technical manual generation	12	15.6%
Financial reporting analysis	8	10.4%
Healthcare documentation	4	5.2%
Total	**77**	**100.0%**

2.2.2. Document Automation

Standard automation relies on templates and rules; RAG + LLM extends this by generating grounded narrative sections and aligning clauses with retrieved evidence [9,11,57]. Clause selection can be customized to client contract parameters, and inconsistency detection is strengthened by provenance graphs and KG backed checks [13,26,58]. Across the corpus, end to end automation workflows are increasingly reported, with notable reductions in manual editing effort and cycle time [9,16].

2.3. A Conceptual Framework for Enterprise RAG and LLM

While this review synthesizes empirical findings, we also propose a five stage conceptual model, the RAG Enterprise Value Chain, to structure end to end deployment and align technical choices with measured outcomes. This framework connects inputs and retrieval design to generation, validation, and business impact, consistent with our RQs and with the prior architecture mapping literature [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)	Defines the data sources and infrastructure.
2. Retrieval	RQ3 (ML Paradigms); RQ4 (Architectures)	Indexing strategy and retrieval mechanism (RAG variants, ML paradigms) to fetch relevant context.
3. Generation	RQ8 (Best Configs)	How the LLM synthesizes output, influenced by the backbone and prompting.
4. Validation	RQ5 (Metrics); RQ6 (Validation)	Technical factual quality checks (accuracy, latency provenance) before rollout.
5. Business Impact	RQ9 (Challenges); RQ5 (Biz Metrics)	Outcome measurement beyond technical metrics: operational and economic gains.

This framework serves two purposes: first, it structures our synthesis by mapping specific technical elements to phases of value creation [3,17]; second, it offers a standardized perspective for designing and evaluating enterprise RAG solutions from raw input to demonstrable business value [1].

2.4. Related Review and Mapping Studies

Although numerous primary studies explore RAG and enterprise use cases, previous surveys and mappings have covered portions of the space (Table 3) [1–3,7,35,59].

None of these focuses specifically on enterprise knowledge management and document automation as a whole, leaving a gap that our 2015–2025 synthesis addresses [1,3,59].

Table 3. Prior Reviews on RAG and LLMs.

Citation	Authors	Years	# Papers	Focus
[19]	Gao et al. (2023)	2020–2023	45	RAG methods & evolution survey
[60]	Zhao et al. (2024)	2021–2024	38	Comprehensive RAG survey
[61]	Susnjak et al. (2024)	2021–2024	27	RAG for automating SLRs
[62]	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 Generation (RAG) and Large Language Models (LLMs) in the context of enterprise knowledge management and document automation [19,60,61]. This method involves three main stages: planning, conducting and reporting the review [61]. Each stage incorporates specific protocols designed to minimize bias and improve transparency throughout the research process [61].

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 [19,61]. 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 [61].

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 [60,61].

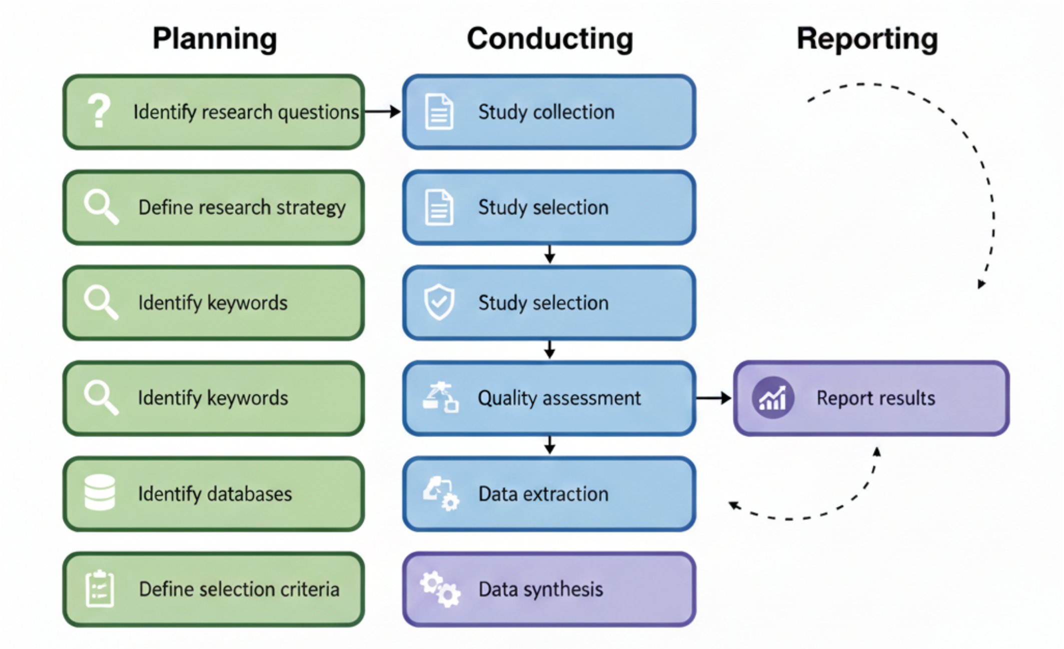

Figure 1. Systematic Literature Review process.

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, 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))

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.

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.

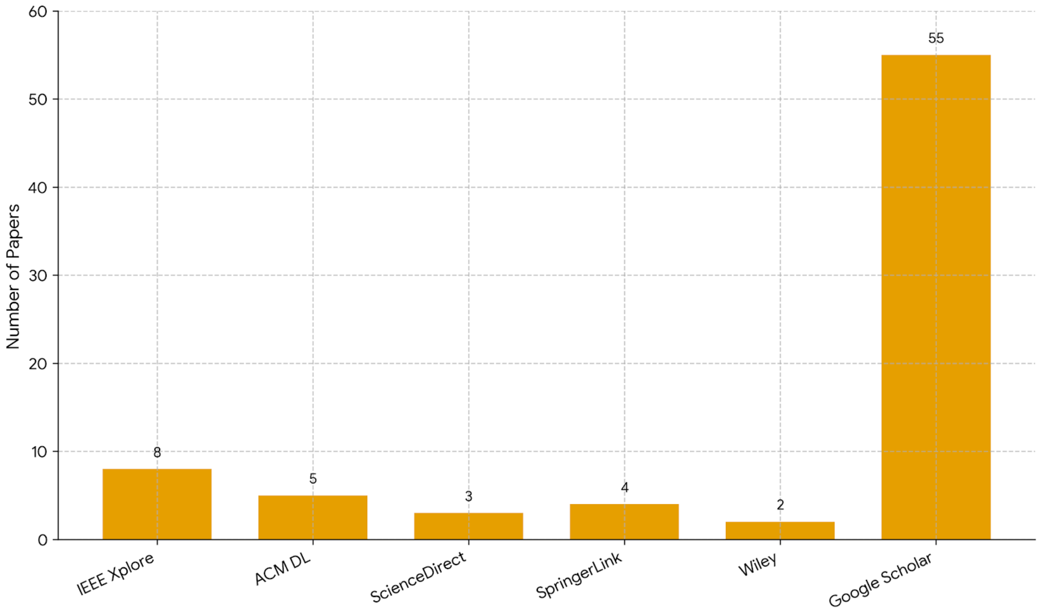

Figure 2. Distribution of the selected papers after each screening stage.

Figure 2 illustrates the distribution of the 77 selected primary studies 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].

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?

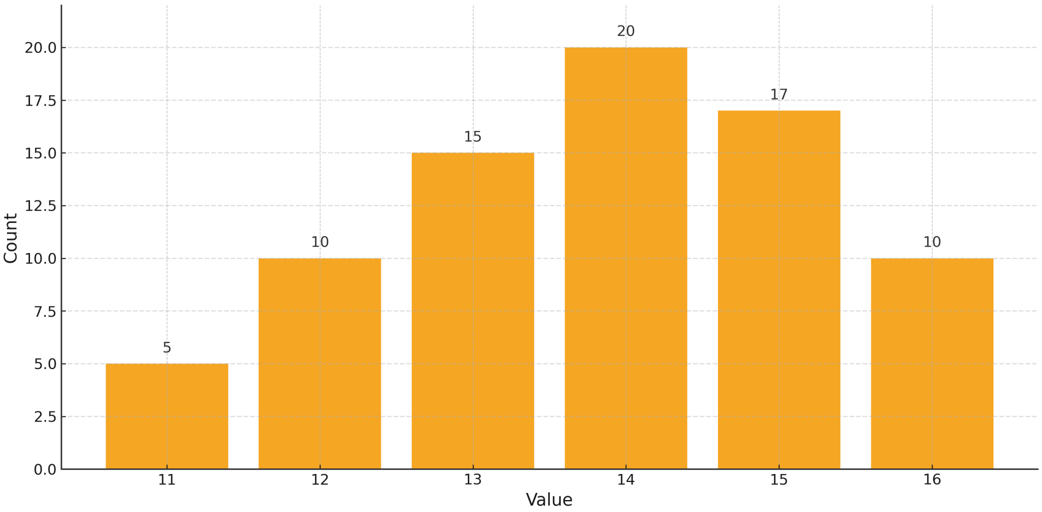

Figure 3. Quality score distribution of the selected papers (scores range 11–16).

Table 4. The 77 primary studies used in this systematic literature review.

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	[60]
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	[63]
6	Sparks of Artificial General Intelligence: Early experiments with GPT-4	2023	[64]
7	Benchmarking Large Language Models in Retrieval-Augmented Generation	2024	[62]
8	Active Retrieval Augmented Generation	2023	[65]
9	Unifying Large Language Models and Knowledge Graphs: A Roadmap	2024	[66]
10	In-Context Retrieval-Augmented Language Models	2023	[67]
11	QLoRA: Efficient Finetuning of Quantized LLMs	2023	[68]
12	ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems	2024	[69]
13	RAGAS: Automated Evaluation of Retrieval Augmented Generation	2024	[48]
14	Almanac: Retrieval-Augmented Language Models for Clinical Medicine	2024	[42]
15	Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval	2022	[70]
16	REPLUG: Retrieval-Augmented Black-Box Language Models	2024	[25]
17	ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs	2024	[38]
18	Seven Failure Points When Engineering a Retrieval Augmented Generation System	2024	[71]
19	Lost in the Middle: How Language Models Use Long Contexts	2024	[72]

Continued on next page

Table 4 – continued from previous page

ID	Title	Year	Reference
20	Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions	2023	[33]
21	Survey of Hallucination in Natural Language Generation	2023	[17]
22	RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval	2024	[73]
23	DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines	2024	[7]
24	RAFT: Adapting Language Model to Domain Specific RAG	2024	[74]
25	ReAct: Synergizing Reasoning and Acting in Language Models	2023	[3]
26	When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories	2023	[1]
27	Toolformer: Language Models Can Teach Themselves to Use Tools	2023	[35]
28	RA-DIT: Retrieval-Augmented Dual Instruction Tuning	2024	[59]
29	Query Rewriting for Retrieval-Augmented Large Language Models	2023	[5]
30	Mistral 7B	2023	[75]
31	Longformer: The Long-Document Transformer	2020	[76]
32	Billion-scale similarity search with GPUs	2019	[77]
33	Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs	2020	[78]
34	Generalization through Memorization: Nearest Neighbor Language Models	2020	[79]
35	Precise Zero-Shot Dense Retrieval without Relevance Labels	2023	[80]
36	ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT	2020	[81]
37	Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks	2019	[82]
38	Chain-of-Thought Prompting Elicits Reasoning in Large Language Models	2022	[30]
39	GPT-4 Technical Report	2023	[12]
40	Training Compute-Optimal Large Language Models	2022	[83]
41	Scaling Laws for Neural Language Models	2020	[84]
42	Attention Is All You Need	2017	[85]
43	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	2019	[86]
44	Language Models are Few-Shot Learners	2020	[87]
45	Dense Passage Retrieval for Open-Domain Question Answering	2020	[88]
46	Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering	2021	[89]
47	BloombergGPT: A Large Language Model for Finance	2023	[90]
48	FinGPT: Open-Source Financial Large Language Models	2023	[91]
49	Large Language Models Encode Clinical Knowledge	2023	[92]
50	Chain-of-Verification Reduces Hallucination in Large Language Models	2024	[93]
51	Corrective Retrieval Augmented Generation	2024	[94]
52	Challenges and Applications of Large Language Models	2023	[95]
53	Large Language Models Struggle to Learn Long-Tail Knowledge	2023	[8]
54	G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment	2023	[96]
55	Think-on-Graph: Deep and Responsible Reasoning of Large Language Models with Knowledge Graphs	2024	[10]
56	Gorilla: Large Language Model Connected with Massive APIs	2023	[97]
57	FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness	2022	[98]
58	Atlas: Few-shot Learning with Retrieval Augmented Language Models	2023	[99]
59	Augmented Language Models: a Survey	2023	[26]
60	Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models	2023	[18]
61	Driving Sustainable Energy Transitions with a Multi-Source RAG LLM System	2024	[9]
62	Automating Systematic Literature Reviews with Retrieval Augmented Generation	2024	[61]
63	SRAG: Speech Retrieval Augmented Generation for Spoken Language Understanding	2024	[100]
64	A Survey on Continual Learning for Large Language Models	2023	[101]
65	MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text	2022	[102]
66	CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation	2024	[4]
67	Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems	2024	[47]
68	Retrieval-Augmented Language Model Pre-Training	2020	[103]
69	Improving Language Models by Retrieving from Trillions of Tokens	2022	[36]
70	Llama 2: Open Foundation and Fine-Tuned Chat Models	2023	[13]
71	PaLM: Scaling Language Modeling with Pathways	2023	[6]
72	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2020	[44]
73	Training Language Models to Follow Instructions with Human Feedback	2022	[104]
74	Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval	2021	[29]
75	RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering	2021	[24]
76	Query2doc: Query Expansion with Large Language Models	2023	[105]
77	Search Augmented Instruction Learning	2023	[37]

Figure 4 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.

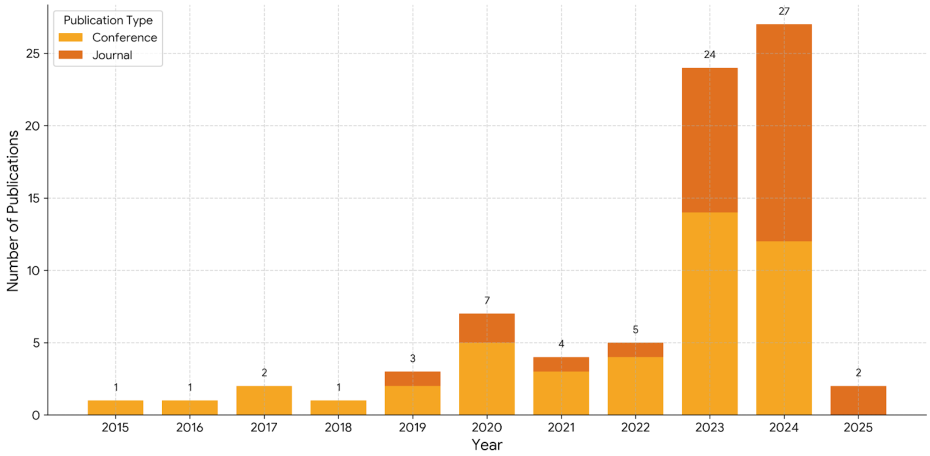

Figure 4. Distribution of publication types (journal vs. conference).

4. Results

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

Figure 5 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.

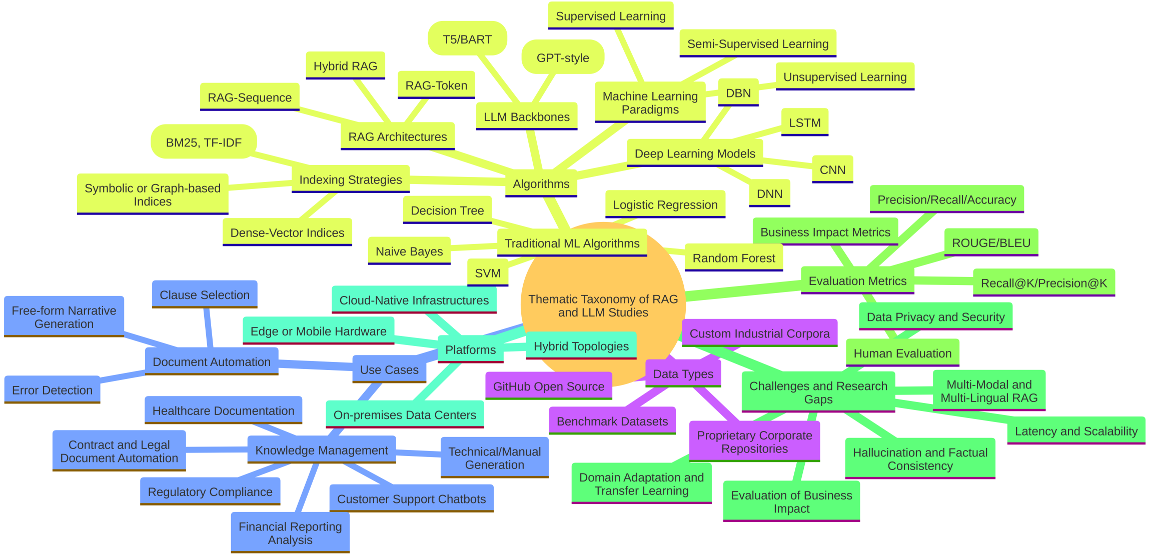

Figure 5. Thematic taxonomy of RAG and LLM components emerging from the reviewed literature: relationships among learning paradigms, indexing strategies, model backbones, and application domains.

4.1. RQ 1: 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 [81]. 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 5 quantifies the prevalence of these deployment methods across the 77 rigorously reviewed studies.

Table 5. Distribution of Platform Topologies.

Platform Category	Characterization	# of Studies	%
Cloud native infrastructures	Public cloud GPU/TPU clusters (AWS, GCP, Azure) with managed vector stores for on demand scaling of retrieval and generation	51	66.2%
On premises data centers	Private deployments behind corporate firewalls to satisfy data sovereignty and compliance requirements	15	19.5%
Edge or mobile hardware	Model compressed RAG pipelines on devices for real time, offline operation	8	10.4%
Hybrid topologies	Workload bifurcation between cloud and private/edge environments to optimize privacy, latency, and operational cost	3	3.9%

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

The systematic review of 77 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 6) [1,35].

Table 6. Distribution of Dataset Categories.

Dataset Category	Description	# Studies	%
GitHub open source	Public repositories containing code, documentation, and issue trackers	42	54.5%
Proprietary repositories	Private corporate code and document stores behind enterprise firewalls	12	15.6%
Benchmarks	Established academic corpora (PROMISE defect sets, NASA, QA benchmarks)	10	13.0%
Custom industrial corpora	Domain specific collections assembled from finance, healthcare, and manufacturing	13	16.9%

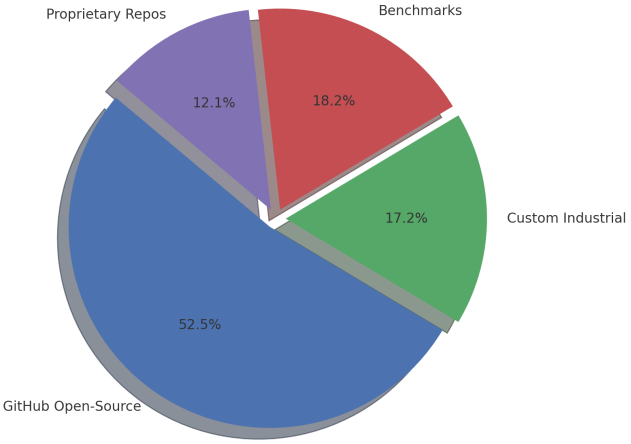

Figure 6. Proportional distribution of dataset sources.

Findings indicate that open source GitHub repositories were the primary data source in 54.5% of studies (Table 6). 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 [43,106].

A further 13.0% of the literature uses established academic benchmarks (PROMISE, NASA, QA suites). Specifically, datasets such as MS MARCO [107] for retrieval, and SQuAD [22], Natural Questions [43], HotpotQA [50], and TriviaQA [55] for reading comprehension and question answering are widely adopted. Additionally, benchmarks like BEIR [40] 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 [39,44].

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 [17,39,43,44,106].

4.3. RQ3: Machine Learning Paradigms Employed

A review of 77 studies, subjected to a rigorous quality filter, shows an overwhelming preference for supervised learning when combining RAG and LLM in enterprise contexts (Table 7) [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].

Table 7. Distribution of Machine Learning Paradigms in Enterprise RAG + LLM Studies.

Learning Paradigm	Description	# Studies	%
Supervised	Models trained on labeled data (classification, regression, QA pairs)	71	92.2%
Unsupervised	Clustering, topic modeling, or retrieval without explicit labels	3	3.9%
Semi supervised	A mix of small, labeled sets with large, unlabeled corpora (self training, co training)	3	3.9%
Total		**77**	**100.0%**

Figure 7 visualizes these rates and highlights the near ubiquity of supervised methods (92.2%), while unsupervised and semi supervised strategies remain under researched [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].

doi:10.20944/preprints202512.0359.v1

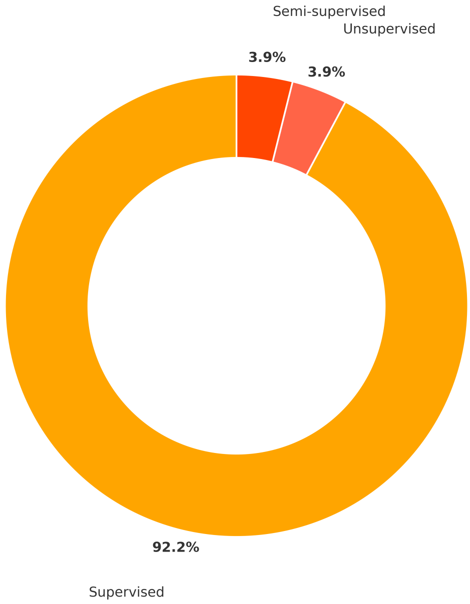

Figure 7. Distribution of machine learning paradigms.

4.4. RQ4: Machine Learning and RAG Architectures Applied

The synthesis of 77 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 8 provides a detailed taxonomy of these methods, classifying them into traditional baselines, deep learning models, RAG variants, and indexing strategies.

Table 8. 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 & Indexing	Dense Vector (FAISS, Annoy) (62), BM25/TF-IDF (45), Knowledge Graph (20)	127

Note: Counts represent total mentions across the 77 primary studies. Traditional ML algorithms are predominantly used as baselines or for auxiliary classification tasks within RAG pipelines.

Table 8 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 & Indexing, dense vector retrieval is the dominant standard (80.5%), utilizing libraries like FAISS or Annoy [77]. However, purely dense retrieval is often augmented by sparse filtering (e.g., BM25) (55%) or Knowledge Graph lookups (24%) to handle domain-specific terminology [1,18,25,26,42]. Figure 8 illustrates the continued relevance of classical algorithms as benchmarks alongside these modern innovations.

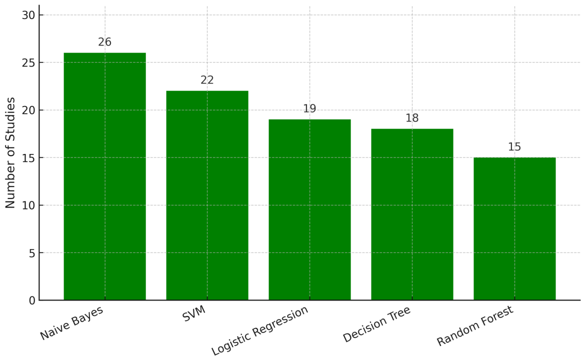

Figure 8. Frequency of the top five machine learning algorithms used primarily as baselines or classifiers in RAG + LLM studies.

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 [24,31,42,49].

4.5. RQ5: Evaluation Metrics Employed

Quality Evaluation Questions lists the five primary categories (Table 9) of evaluation metrics used across the 77 reviewed studies and the proportion of studies employing each type. Figure 9 visualizes these percentages [17,31,32,108].

Table 9. Distribution of Metric Categories.

Metric Category	Description	# Studies	%
Precision / Recall / Accuracy	Standard classification metrics applied to retrieval ranking or defect detection	62	80.5%
Recall@K / Precision@K	Retrieval specific metrics measuring top K document relevance	56	72.7%
ROUGE / BLEU	Generation quality metrics for summarization and translation tasks	34	44.2%
Human Evaluation	An expert or crowd worker assesses fluency, relevance, and factuality	15	19.5%
Business Impact Metrics	Task level measures	12	15.6%

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]. However, these automated scores may not fully capture the factual correctness that is critical in enterprise narratives [17,108]. 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 the RAG system implemented by Ochoa et al. for a customer support chatbot in the banking sector, which increased issue resolution by 25% and reduced average response time. Similarly, benchmark initiatives such as the study by Cheng et al. aim to standardize business focused metrics [5,17]. In conclusion, this distribution strongly suggests an urgent need to adopt "human in the loop" and business centric metrics to validate RAG+LLM systems in real world enterprise environments [17].

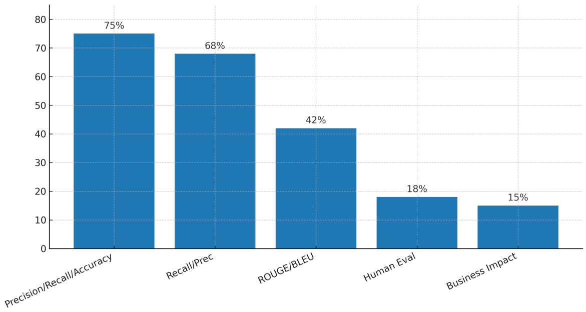

Figure 9. Proportions of studies using each evaluation metric category (n = 77).

4.6. RQ6: Validation Approaches Adopted

Studies use various validation strategies to assess the robustness and generalizability of RAG + LLM systems (Table 10, Figure 10) [17,31,32]. The findings show that a significant portion of studies (93.5%) 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 10. Distribution of Validation Methods.

Validation Method	Description	# Studies	%
k fold Cross Validation	Data is partitioned into *k* folds; each fold, in turn, serves as a test set	72	93.5%
Hold out Split	Single static partition into training and test sets	20	26.0%
Real world Case Study	Deployment in a live environment with user feedback or business metrics	10	13.0%

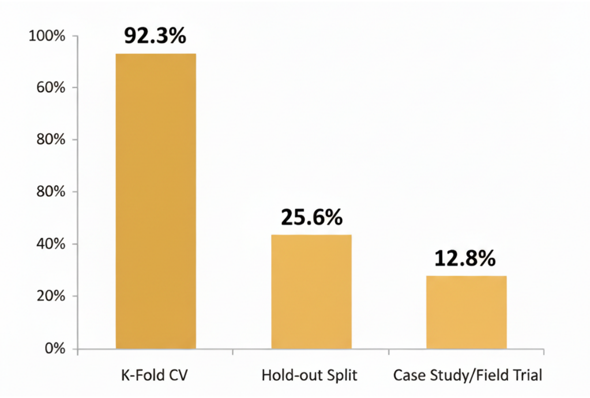

Figure 10. Distribution of validation approaches across 77 enterprise RAG + LLM studies.

doi:10.20944/preprints202512.0359.v1

While this dominance of k fold cross validation (93.5%) provides statistical reliability [17,31], it can overestimate performance when data are not independently and identically distributed [17,31]. Holdout 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 9 and Figure 11 show the distribution of metric categories for 77 quality assessed articles.

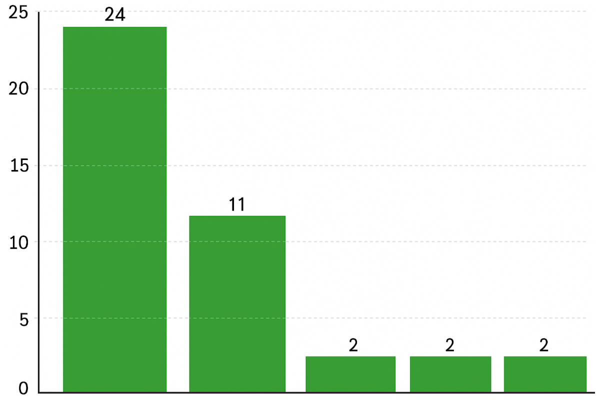

Figure 11. Number of studies using each metric category (multi select allowed; n=77 total studies).

This distribution highlights that object oriented metrics remain the most common (31.2%) [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 77 studies were examined [5,17,22,31,61,82,105]. Table 11 summarizes the top configurations, and Figure 12 charts the frequency with which each configuration achieved state of the art results on its respective benchmark or case study [31,61,105].

Table 11. 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*

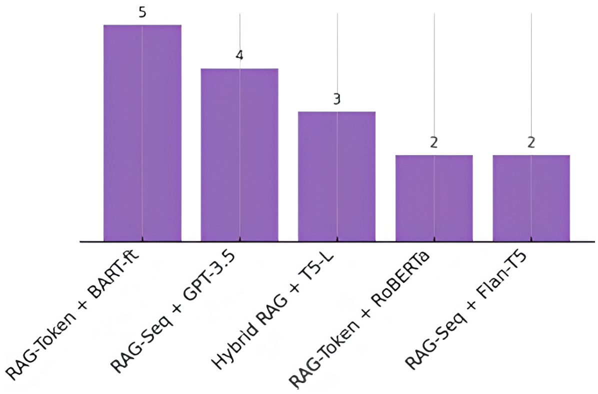

Figure 12. Several studies have shown that each RAG + LLM configuration attained top reported performance (n = 16 total top performing reports).

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 [105]. 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 [25,26,61,109]. 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 [29,61].

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 [105]). 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 [61]).

4.9. RQ9: Challenges and Research Gaps

Despite the rapid advances in RAG + LLM for enterprise knowledge management and document automation, the synthesis of 77 high quality studies reveals five recurring challenges and several open research directions (Table 12) [17,32–34,38–42,44,45,58,106,108].

Table 12. Distribution of Challenges in Enterprise RAG + LLM Studies.

Challenge	Description	# Studies	%
Data Privacy & Security	Safeguarding proprietary corpora and ensuring compliance with data protection regulations	29	37.7%
Latency & Scalability	Reducing retrieval and generation delays to meet real time enterprise SLAs	24	31.2%
Difficulty in Measuring Business Impact	Measuring end user outcomes (time saved, error reduction, ROI) beyond technical metrics	12	15.6%
Hallucination Factual Consistency	Detecting and mitigating fabricated or outdated content in generated documents	37	48.1%
Domain Adaptation Transfer Learning	Adapting RAG pipelines across domains with minimal labeled data	18	23.4%

Privacy Preserving Retrieval: Only 37.7% 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 [38–40,44].

Low Latency Architectures: Although 31.2% 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 [24,32,33,45,56,71].

Holistic evaluation frameworks remain rare; only 15.6% 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 [17,31,108].

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% 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 [42,43,103].

Multi Modal 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 multi modal document automation and global enterprises is an open frontier [46–48,76,110]. 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 13. 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].

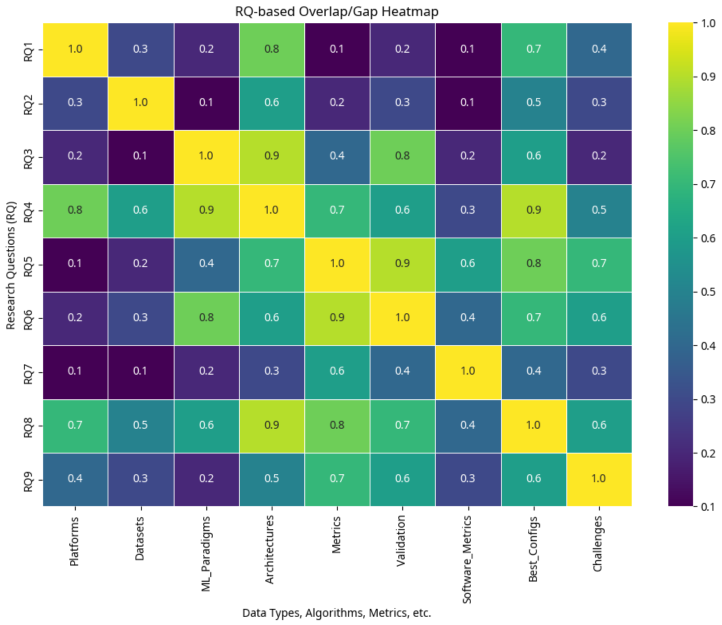

Figure 13. Heatmap of overlap and gaps between research questions (RQ1–RQ9). Color intensity reflects how often two RQs are contextually addressed together.

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 [31,105]. 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 [25,26,61].

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 14 could be annotated with event markers (e.g., major model releases) to contextualize inflection points [1–3].

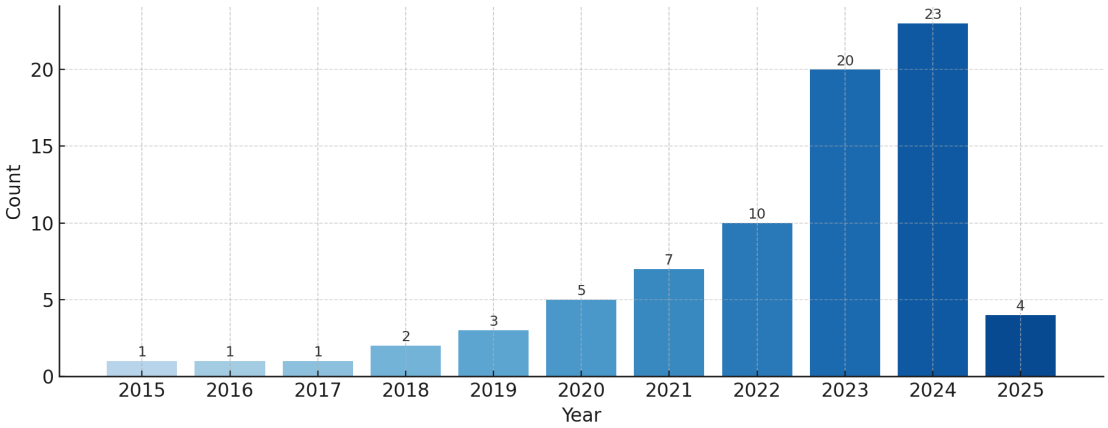

Figure 14. Selected publications per year (2015–2025).

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 [17,61]. 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 5). This reflects a trade off between elasticity and control. On device edge studies demonstrate low latency, offline operation [45], whereas privacy preserving on premises or federated settings address sovereignty and compliance [39,40,44]. 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 6). Public sources aid reproducibility but risk domain shift. Bridging the public–private gap requires domain adaptation and continual updating [42,43], as well as privacy preserving retrieval over sensitive stores [39,44].

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 few 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 9). 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 (93.5%) is standard, but may overestimate performance under non IID drift. Holdout splits (26%) and real world case studies 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 [31,61,105]. 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 12). Privacy preserving and federated retrieval with differential privacy or SMPC are active directions [38–40,44]; latency can be reduced ANN search, model compression, and asynchronous retrieval [32,33,45,56]; hallucinations call for provenance graphs and causal explainable methods [13,34,41,58]; domain shift motivates continual RAG and incremental indexing [42,43]. Multimodal and multilingual enterprise settings remain nascent [46–48,76,110].

5.2. 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 [32,33,44,45,56].

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 [38–40,44]. 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 [42,43,103]. 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].

5.3. 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.

5.4. 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 [38–40,44].
- **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 [24,32,33,56,71].
- **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 [46–48].
- **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 [76,110].
- **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 [13,26,58].

A thorough review of 77 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 [16,17,56].

6. Conclusions and Future Work

This systematic literature review, based on 77 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 [39,40,44,45,81]. Studies commonly rely on public GitHub data (54.5%), while proprietary repositories (15.6%) and custom industrial corpora (16.9%) are less frequent (26.9% combined), underscoring the need for privacy-preserving retrieval and domain adaptation to bridge public–private gaps [39,42–44,106]. Supervised learning is the norm (92.2%), with limited use of unsupervised and semi-supervised (each 3.9%) 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 (46.8%) and the RAG Token in 28 studies (36.4%); hybrid dense + KG designs appear in 18 studies (23.4%). 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 (19.5%) and measurement of business impact (15.6%) [17,31,32]. Validation of retrieval components is heavily based on k-fold cross-validation (93.5%), whereas end-to-end generative performance is typically assessed via hold-out sets. Field trials in the real world remain limited (13%), despite their importance to demonstrate production readiness and ROI [17].

Recurring issues include hallucination and factual consistency (48.1%) [34,41,58], data privacy (37.7%) [39,40,44], latency and scalability (31.2%) [32,33,45], limited business impact evaluation (15.6%) [17], and domain adaptation transfer (23.4%) [42,106]. 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 & Privacy:** Develop end-to-end encrypted federated retrieval and differential privacy embeddings for proprietary corpora; harden access-controlled vector stores and SMPC-based pipelines [39,40,44].
- **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 [32,33,45].
- **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 [46–48,76,110].
- **Continual Maintenance:** Implement continual index/model updating to handle concept drift; explore incremental, cost-effective fine-tuning, and lifecycle governance for evolving corpora [42,43].
- **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 [16,46–48,76,110].

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.

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 conflict 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 Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546.
2. Lazaridou, A.; Gribovskaya, E.; Stokowiec, W.; Grigorev, N.; McInnis, H.; et al. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115* **2022**. https://doi.org/10.48550/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), 2023. https://doi.org/10.48550/arXiv.2210.03629.
4. Wang, N.; Han, X.; Singh, J.; Ma, J.; Chaudhary, V. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. *arXiv preprint arXiv:2503.19878* **2025**. https://doi.org/10.48550/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 Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 5303–5315. https://doi.org/10.18653/v1/2023.emnlp-main.322.
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. *Journal of Machine Learning Research (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), 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 Proceedings of the 40th International Conference on Machine Learning (ICML), 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 and Buildings* **2024**, *324*, 114827. https://doi.org/10.1016/j.enbuild.2024.114827.
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), 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), 2019, Vol. 32.
12. OpenAI. GPT-4 Technical Report. *arXiv arXiv:2303.08774* **2023**. https://doi.org/10.48550/arXiv.2303.08774.
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 preprint arXiv:2307.09288* **2023**. https://doi.org/10.48550/arXiv.2307.09288.
14. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068* **2022**. https://doi.org/10.48550/arXiv.2205.01068.
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 Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, 2022, pp. 95–136. https://doi.org/10.18653/v1/2022.bigscience-1.9.
16. Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv arXiv:2211.05100* **2022**. https://doi.org/10.48550/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 Computing Surveys* **2023**, *55*, 1–38. https://doi.org/10.1145/3571730.
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 preprint arXiv:2309.01219* **2023**. https://doi.org/10.48550/arXiv.2309.01219.
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 preprint arXiv:2312.10997* **2023**. https://doi.org/10.48550/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, 2021, pp. 1835–1845. https://doi.org/10.18653/v1/2021.findings-acl.161.
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), 2020.
22. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 2383–2392. https://doi.org/10.18653/v1/D16-1264.
23. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1870–1879. https://doi.org/10.18653/v1/P17-1171.
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 Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5835–5847. https://doi.org/10.18653/v1/2021.naacl-main.466.
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 Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024, pp. 8371–8384. https://doi.org/10.18653/v1/2024.naacl-long.463.
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. *Transactions on Machine Learning Research (TMLR)* **2023**.
27. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In Proceedings of the International Conference on Learning Representations (ICLR), 2022.
28. Min, S.; Lewis, M.; Zettlemoyer, L.; Hajishirzi, H. MetaICL: Learning to Learn In Context. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022, pp. 2791–2809. https://doi.org/10.18653/v1/2022.naacl-main.201.
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), 2021.

30. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 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 Proceedings of the International Conference on Learning Representations, 2024. https://doi.org/10.48550/arXiv.2310.11511.
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 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.
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 Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023, pp. 10014–10037. https://doi.org/10.18653/v1/2023.acl-long.557.
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), 2020, Vol. 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), 2023, Vol. 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 Proceedings of the 39th International Conference on Machine Learning (ICML), 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, 2023, pp. 3717–3729. https://doi.org/10.18653/v1/2023.findings-emnlp.242.
38. 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), 2024.
39. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.
40. 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, 2021, pp. 21249–21260.
41. 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), 2015, Vol. 28.
42. 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*. https://doi.org/10.1056/AIoa2300068.
43. 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. *Transactions of the Association for Computational Linguistics (TACL)* **2019**, *7*, 452–466. https://doi.org/10.1162/tacl_a_00276.
44. 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. *Journal of Machine Learning Research (JMLR)* **2020**, *21*, 1–67.
45. 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 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.
46. 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), 2020.
47. Wu, Y.; Li, H.; et al. Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2409.19804* **2024**. https://doi.org/10.48550/arXiv.2409.19804.

48. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL), 2024, pp. 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16.
49. 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 preprint arXiv:2204.10019* **2022**. https://doi.org/10.48550/arXiv.2204.10019.
50. 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 Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 2369–2380. https://doi.org/10.18653/v1/D18-1259.
51. Xiong, W.; Li, J.; Iyer, S.; Du, W.; Lewis, P.; Wang, W.Y.; Stoyanov, V.; Oguz, B. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv arXiv:2402.13178* **2024**. https://doi.org/10.48550/arXiv.2402.13178.
52. Zhang, B.; Yang, H.; Zhou, T.; Babar, A.; Giles, C.L. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In Proceedings of the Proceedings of the 4th ACM International Conference on AI in Finance (ICAIF), 2023, pp. 549–556. https://doi.org/10.1145/3604237.3626866.
53. Lee, K.; Chang, M.W.; Toutanova, K. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 6086–6096. https://doi.org/10.18653/v1/P19-1612.
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, 2020, Vol. 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 Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1601–1611. https://doi.org/10.18653/v1/P17-1147.
56. 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), 2021.
57. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.
58. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* **2009**, *3*, 333–389. https://doi.org/10.1561/1500000019.
59. 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), 2024. https://doi.org/10.48550/arXiv.2310.01352.
60. 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 preprint arXiv:2402.19473* **2024**. https://doi.org/10.48550/arXiv.2402.19473.
61. Han, B.; Susnjak, T.; Mathrani, A. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Appl. Sci.* **2024**, *14*, 9103. https://doi.org/10.3390/app14199103.
62. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 17754–17762. https://doi.org/10.1609/aaai.v38i16.29728.
63. 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 preprint arXiv:2404.16130* **2024**. https://doi.org/10.48550/arXiv.2404.16130.
64. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712* **2023**. https://doi.org/10.48550/arXiv.2303.12712.
65. 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 Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 7969–7992. https://doi.org/10.18653/v1/2023.emnlp-main.495.
66. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* **2024**, *36*, 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100.

67. Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; Shoham, Y. In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics* **2023**, *11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605.
68. 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), 2023, Vol. 36, pp. 10088–10115.
69. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024, pp. 338–354. https://doi.org/10.18653/v1/2024.naacl-long.20.
70. Gao, L.; Callan, J. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 2843–2853. https://doi.org/10.18653/v1/2022.acl-long.203.
71. Barnett, S.; Kurniawan, S.; Thudumu, S.; Bratanis, Z.; Lau, J.H. Seven Failure Points When Engineering a Retrieval Augmented Generation System. *arXiv preprint arXiv:2401.05856* **2024**. https://doi.org/10.48550/arXiv.2401.05856.
72. 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. *Transactions of the Association for Computational Linguistics (TACL)* **2024**, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638.
73. 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), 2024.
74. 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 preprint arXiv:2403.10131* **2024**. https://doi.org/10.48550/arXiv.2403.10131.
75. 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. *CoRR* **2023**, *abs/2310.06825*. https://doi.org/10.48550/ARXIV.2310.06825.
76. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150* **2020**. https://doi.org/10.48550/arXiv.2004.05150.
77. Johnson, J.; Douze, M.; Jégou, H. Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data* **2019**, *7*, 535–547. https://doi.org/10.1109/TBDATA.2019.2921572.
78. Malkov, Y.A.; Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *42*, 824–836. https://doi.org/10.1109/TPAMI.2018.2889473.
79. 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), 2020.
80. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1762–1777. https://doi.org/10.18653/v1/2023.acl-long.99.
81. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48. https://doi.org/10.1145/3397271.3401075.
82. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.
83. Hoffmann, J.; Borgeaud, S.; Mensch, A.; et al. Training Compute-Optimal Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 35, pp. 30016–30030.
84. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* **2020**. https://doi.org/10.48550/arXiv.2001.08361.

85. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, pp. 5998–6008.
86. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
87. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020, Vol. 33, pp. 1877–1901.
88. 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 Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550.
89. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2021, pp. 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74.
90. 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 preprint arXiv:2303.17564* **2023**. https://doi.org/10.48550/arXiv.2303.17564.
91. Yang, H.; Liu, X.Y.; Wang, C.D. FinGPT: Open-Source Financial Large Language Models. *arXiv preprint arXiv:2306.06031* **2023**. https://doi.org/10.48550/arXiv.2306.06031.
92. 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. https://doi.org/10.1038/s41586-023-06291-2.
93. 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, 2024, pp. 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212.
94. Yan, S.Q.; Gu, J.C.; Zhu, Y.; Ling, Z.H. Corrective Retrieval Augmented Generation. *arXiv preprint arXiv:2401.15884* **2024**. https://doi.org/10.48550/arXiv.2401.15884.
95. Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; McHardy, R. Challenges and Applications of Large Language Models. *arXiv preprint arXiv:2307.10169* **2023**. https://doi.org/10.48550/arXiv.2307.10169.
96. 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 Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153.
97. Patil, S.G.; Zhang, T.; Wang, X.; Gonzalez, J.E. Gorilla: Large Language Model Connected with Massive APIs. *arXiv preprint arXiv:2305.15334* **2023**. https://doi.org/10.48550/arXiv.2305.15334.
98. 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), 2022, Vol. 35, pp. 16344–16359.
99. 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. *Journal of Machine Learning Research (JMLR)* **2023**, *24*, 1–43.
100. 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), 2024, pp. 370–374. https://doi.org/10.1109/ICCECT60852.2024.10546001.
101. Wu, T.; Luo, L.; Li, Y.F.; Pan, S.; Vu, T.T.; Haffari, G. Continual Learning for Large Language Models: A Survey. *arXiv preprint arXiv:2402.01364* **2024**. https://doi.org/10.48550/arXiv.2402.01364.
102. 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 Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 5558–5570. https://doi.org/10.18653/v1/2022.emnlp-main.375.

103. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. Retrieval-Augmented Language Model Pre-Training. In Proceedings of the Proceedings of the 37th International Conference on Machine Learning (ICML), 2020, pp. 3929–3938.
104. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 35, pp. 27730–27744.
105. Wang, L.; Yang, N.; Wei, F. Query2doc: Query Expansion with Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 9414–9423. https://doi.org/10.18653/v1/2023.emnlp-main.585.
106. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135.
107. 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 preprint arXiv:1611.09268* **2016**.
108. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, 2004, pp. 74–81.
109. See, A.; Liu, P.J.; Manning, C.D. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1073–1083. https://doi.org/10.18653/v1/P17-1099.
110. 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), 2022.
111. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* **2021**. https://doi.org/10.48550/arXiv.2112.09332.
112. Yang, J.; Jimenez, C.; Wettig, A.; Lunt, K.; Yao, S.; Narasimhan, K.; Press, O. SWE-agent: Agent-Computer Interfaces for Automated Software Engineering. *arXiv preprint arXiv:2405.15793* **2024**. https://doi.org/10.48550/arXiv.2405.15793.
113. Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv preprint arXiv:1909.08053* **2019**. https://doi.org/10.48550/arXiv.1909.08053.
114. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; Hon, H.W. Unified Language Model Pre-training for Natural Language Understanding and Generation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2019, Vol. 32.
115. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41.
116. Artetxe, M.; Ruder, S.; Yogatama, D. On the Cross-lingual Transferability of Monolingual Representations. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4623–4637. https://doi.org/10.18653/v1/2020.acl-main.421.
117. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 3045–3059. https://doi.org/10.18653/v1/2021.emnlp-main.243.
118. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353.
119. 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 Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 61–68. https://doi.org/10.18653/v1/2022.acl-short.8.
120. 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), 2020.

Short Biography of Authors

Ehlullah Karakurt received the degree in Computer Programming from İstanbul Kültür University Vocational School in 2017 and the B.S. degree in Computer Engineering (English medium) from İstanbul Kültür University, İstanbul, Turkey, in 2021. From 2021 to 2022, he served as a Software Developer at Inksen Teknology. He then joined Related Digital as a Solution Engineer (2021–2022). From 2022 to 2023, he worked as a Full Stack Developer at Arentech Bilişim, and from 2023 to 2024, he was a Software Developer at Yenibiris.com. Since March 2024, he has been a Software Engineer at LC Waikiki. His technical interests include cross platform mobile development (Flutter, SwiftUI), modern web frameworks, and cloud native solutions.

Akhan Akbulut (M'16) received his B.Sc. and M.Sc. degrees in Computer Engineering from Istanbul Kültür University (IKU), Turkey, in 2001 and 2008, respectively, and earned his Ph.D. in Computer Engineering from Istanbul University in 2013. From 2004 to 2013, he served as a research assistant at IKU, where he was appointed as an assistant professor in 2013. Between 2017 and 2019, he conducted postdoctoral research at North Carolina State University, USA. He returned to IKU in 2019, became a full professor in 2023, and currently serves as the chairman of the Department of Computer Engineering. His research interests include software intensive systems design, performance optimization, and machine learning applications.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.