LLMs in the Enterprise: A Systematic Review of Security Architectures for RAG-Augmented Chatbots

Fatimah Alali[1], Haya Aldawsari[1], Saad Ezzini[1,2*], Sajjad Mahmood[1,3]
[1]Information and Computer Science Department, KFUPM
[2]IRC for Intelligent Manufacturing and Robotics, KFUPM
[3]IRC for Intelligent and Secure Software, KFUPM
Dhahran, Saudi Arabia
[*]saad.ezzini@kfupm.edu.sa

Abstract

Retrieval-Augmented Generation (RAG) grounds Large Language Models (LLMs) in domain-specific data, but its security architectures in enterprise environments, such as SAP-based planning suites, remain underexplored. We present a Systematic Literature Review (SLR) of 50 empirical RAG-augmented chatbot studies (2020–2025) conducting quality assessment and mapping findings against established threat catalogues. While RAG consistently outperforms base LLMs, security coverage in the reviewed literature is limited and inconsistent. No study presents an end-to-end secure RAG architecture covering retrieval, generation, and access-control simultaneously. We map observed treatments, highlight critical security gaps (e.g., indirect prompt injection and corpus poisoning), and conclude that detailed architectural guidance and benchmarking against cloud-managed toolkits are essential for safe enterprise adoption.

1 Introduction

AI-powered assistant tools have become central to operational efficiency in modern industry, particularly for tasks such as analyzing natural-language requirements (Ezzini et al., 2023). Organizations increasingly rely on them for tasks ranging from routine document navigation to the resolution of domain-specific technical issues. A consistently reported challenge, particularly in IT-intensive environments, is that employees spend significant time searching large, unstructured documentation repositories. In environments centered on enterprise resource-planning suites such as SAP, a family of integrated business software products covering finance, logistics, human resources, and procurement, recurring support requests consume valuable IT department capacity that could be redirected toward higher-impact work. A purpose-built chatbot that lets users query their own documents directly would eliminate much of this overhead, delivering immediate, precise responses and reducing reliance on human escalation.

Large language models have established a clear presence in this space, demonstrating strong capabilities in conversational applications across many sectors. However, LLMs have well-documented limitations when deployed in knowledge-intensive organizational contexts. Their internal knowledge is fixed at the time of training, making it impractical to keep them up to date with rapidly changing enterprise data. They are also prone to generating plausible but factually incorrect responses, a phenomenon known as hallucination, which is particularly problematic when accurate technical guidance is required.

The RAG framework addresses both of these shortcomings. Rather than relying on what the model learned during training, RAG retrieves relevant content from an external knowledge base when a query is submitted and incorporates it into the response generation process. This means responses are grounded in actual organizational documents, the knowledge base can be updated independently of the model, and the system can be adapted to any domain without retraining (Feridooni et al., 2024). RAG works well with document-heavy knowledge bases such as PDFs, Word files, and structured technical manuals, making it a natural fit for enterprise use cases.

Despite the significant growth in RAG-related research, the field is not yet well consolidated. Studies vary widely in methodology, evaluation approach, and domain of application. (Ko et al., 2024a) report a relevance improvement of 13-25% over base LLM, while (Zhou et al., 2024) demonstrate 95% context recall and 93.73% faithfulness in a medical chatbot setting. (Ko et al., 2024b) propose ReRAG, a modified RAG design aimed at reducing hallucination. At the same time, challenges related to computational cost, retrieval latency, and

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

the availability of sufficient training data remain recurring concerns in the literature (Feridooni et al., 2024; Chen and Tran, 2024; Leung et al., 2024).

To the best of our knowledge, this is among the first systematic literature reviews to explicitly frame RAG-based chatbot development around enterprise security requirements in the context of SAP-type industrial deployments. We do not claim that no prior survey discusses RAG: comprehensive technical surveys exist (Gao et al., 2023), and chatbot-security SLRs have been published in adjacent areas (Jing et al., 2023). Our contribution lies in positioning the two literatures against each other and identifying where enterprise-oriented secure-RAG guidance is missing. The findings of this SLR informed a companion empirical study, in which a RAG-LLM chatbot was implemented and evaluated against Microsoft Azure in a real SAP industrial setting (Alali et al., 2025).

The paper is organized as follows. Section 2 describes the three-phase SLR methodology. Section 3 presents the results and findings mapped to each research question. Section 4 provides a synthesized overview of the RAG pipeline drawn from the reviewed studies. Section 5 discusses implications for research and practice. Section 6 presents the limitations of this review. Section 7 concludes with directions for future work.

2 Method

The review follows the Kitchenham and Charters guidelines for systematic literature reviews in software engineering (Kitchenham and Charters, 2007) and reports its identification, screening, eligibility, and inclusion steps consistent with the PRISMA 2020 flow (Page et al., 2021). The pipeline has three main phases: plan, conduct, and document, each comprising three steps: (1) specifying research questions and developing and validating the protocol; (2) identifying candidate studies, selecting primary studies, and applying quality assessment; (3) extracting, documenting, and synthesizing data. The remainder of this section details each step.

2.1 Phase 1: Plan Review

1.1 Specify Research Questions. The research questions were derived from an initial investigation of industry needs, including informal interviews with practitioners from three companies operating SAP-based environments. The following are the RQs and motivation for each one. **RQ1:** *To what extent do the reviewed studies compare RAG-augmented chatbots with cloud-based enterprise toolkits, and how does RAG perform against standalone LLM baselines when such comparisons exist?* (RQ1.1: Which metrics are used to assess performance? RQ1.2: Which LLM backbones and retrieval toolkits are used?). **RQ2:** *How does the RAG chatbot facilitate industry work?* (RQ2.1: Why does the industry not rely on LLM alone? RQ2.2: What are the views of industry professionals on developing secure, custom chatbots? RQ2.3: Which practices do practitioners follow to implement an effective RAG chatbot?). **RQ3:** *Which security threats, controls, and architectural components are reported in the primary studies, and how do they map onto established threat catalogues for LLM-integrated applications (Greshake et al., 2023; Vassilev et al., 2024)?* (RQ3.1: Which security threats are explicitly addressed? RQ3.2: Which mitigation strategies and architectural components are reported?).

1.2 Protocol development. The protocol documents the background, objectives, inclusion/exclusion criteria, search strategy, data-extraction template, and synthesis method (narrative synthesis).

1.3 Protocol validation. The protocol was iteratively reviewed by two authors over multiple sessions. Disagreements on criteria wording and on the search-string scope were resolved by majority decision and recorded in a change log. A pilot extraction on five randomly selected papers was conducted before the main extraction to calibrate the criteria.

2.2 Phase 2: Conduct a review

2.1 Search Strategy and Identification. The search query was designed with an enterprise focus using RAG, LLM, and security keywords. We searched five primary databases: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerNature Link, and Google Scholar, yielding 1,791 records initially. After deduplication and title/abstract screening, we retrieved 71 full-text candidates. The detailed search string and database-specific counts are available in our online repository[1].

2.2 Study Selection and Quality Assessment. To select eligible primary studies, we applied rig-

[1] https://github.com/ezzini/LLMs-Enterprise

orous inclusion and exclusion criteria (detailed in our online repository). We conducted quality assessment using a 7-item instrument adapted from Kitchenham and Charters (Kitchenham and Charters, 2007), retaining only studies scoring ≥ 5 out of 7. Rating was performed independently by two authors, achieving a substantial inter-reviewer agreement (average Cohen's $\kappa = 0.78$ (McHugh, 2012)). Disagreements were resolved through a third-author adjudication round. Ultimately, 50 studies satisfied the eligibility criteria and formed the final corpus. The complete quality criteria checklist and scores are available in our online repository.

This section illustrates the details of how the data was extracted, synthesized and documented.

3.1 Data Extraction & Documentation: The purpose of data extraction from this systematic literature review is to ensure data consistency, accuracy and relevancy following a systematic approach. The extracted data from 50 studies were split between two authors, each taking 25. After that, each author revised the data correctness that wasn't extracted by the author. So each study was extracted once by the author and revised once by the second author. The extracted data was documented in an Excel sheet, tracking publication details (year, type), system configurations (model, platform, comparison model, retrieval method), methodology details (testing method, data source, evaluation metrics), security treatment, and key findings/limitations. The data extracted focused especially on answering RQ1.1, RQ1.2, RQ2.2. During this process, it was noticed that the retrieval method has a different reporting approach among the studies where some of them only stated database instead of referring to the technicality. This SLR clarifies that the retrieval method will focus on details of retrieval such as embedding and vector database type. Some extracted data varies on testing and evaluation methods, where some relied on user experience evaluation of the system and others used numerical metrics such as accuracy and scores. In addition, not all the research questions were answered in the collected studies, for that reason, each study marked with the question that has been answered mapped in our online repository.

2.3 Data Synthesis. We adopted a narrative synthesis approach, which is appropriate given the heterogeneity of metrics, domains, and protocols across the corpus. Findings were synthesized along three main axes: (i) comparative performance gains of RAG over standalone LLM configurations, (ii) recurrent engineering and operational challenges (e.g., latency, dataset constraints, query complexity), and (iii) security and privacy treatments. These synthesized outcomes are discussed in detail in Section 3.

Threats to validity. To comply with strict page limitations, the detailed analysis of the four categories of threats to validity (internal, external, construct, and conclusion validity) and their corresponding mitigations has been moved to our online repository.

3 Results and discussion

This section presents the findings of the collected studies. We first summarize the corpus and then discuss the results for each research question.

3.1 Overview of selected studies

The selected studies comprise conferences, theses, and journals. The inclusion criterion on publication year was 2020-2025. In practice, every study that satisfied all remaining criteria such as empirical, applies RAG with an LLM, scored ≥ 5 on the quality instrument was published in 2024 or 2025. The earlier years of the window returned candidate records but none survived the full eligibility pipeline. A substantive finding in itself, indicating that the RAG-with-LLM line of empirical work is very recent. The studies are empirical and develop chatbots using RAG with an LLM in various domains. Of the 71 full-text candidates, 50 scored $\geq 5/7$ on the quality instrument and form the final corpus. Figure 1 shows the distribution by publication type, and Figure 2 shows the distribution by publication year.

The reviewed studies span a broad range of application domains, consistently demonstrating RAG's practical value. In healthcare, (Sree et al., 2024; Zhou et al., 2024; Steybe et al., 2025; Bhattacharyya et al., 2024; Azam et al., 2024; Gu et al., 2025; Xu et al., 2024; Fink et al., 2025; Masanneck et al., 2025) collectively show improvements in user satisfaction, response precision, and contextual relevance following RAG integration. Zhou et al. (Zhou et al., 2024) achieve a 95% context recall rate. In (Azam et al., 2024), RAG reduces hallucinations in cancer-related queries compared to a GPT-4 baseline. Clinical decision support

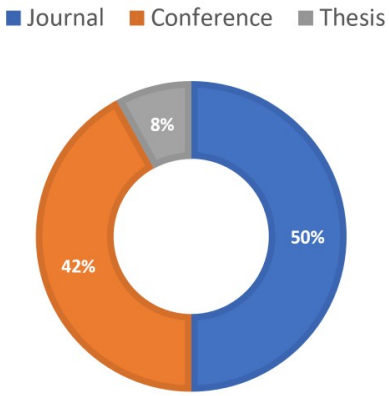

Figure 1: Distribution of the 50 included studies by publication type.

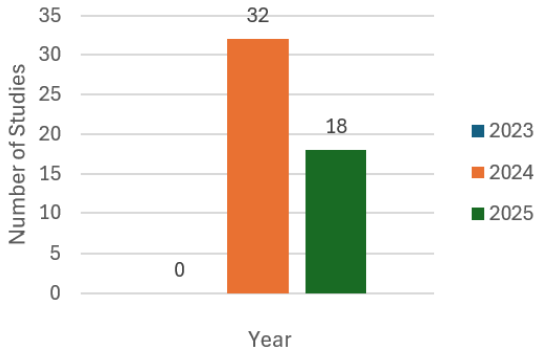

Figure 2: Distribution of the 50 included studies by publication year.

for stroke assessment, radiology classification, and breast cancer nursing care is addressed in (Gu et al., 2025; Jiang et al., 2025), underscoring RAG's capacity to strengthen diagnostic accuracy in high-stakes medical settings. Education is another area of strong RAG adoption. Studies (Dakshit, 2024; Matar and Mohammad, 2024; Alario-Hoyos et al., 2024) all report accuracy improvements and high student satisfaction; the rate in (Alario-Hoyos et al., 2024) specifically reached 93.3%. RAG has also been applied to textbook question-answering systems requiring reasoning over extended educational texts (Alawwad et al., 2025). Table 1 summarizes key metrics findings by domain. Industry-oriented applications are prominently represented. Manufacturing benefits from RAG in quality control and real-time production monitoring (Heredia Álvaro and Barreda, 2025). Construction-sector work targets contract risk identification and safety compliance (Lee et al., 2024; Wong et al., 2024). Supply chain operations report gains in agility and efficiency (Zhu and Vuppalapati, 2024), and the automotive sector uses locally deployed RAG models to automate large-scale technical document processing (Liu et al., 2024). Broader organizational contexts include ESG reporting extraction (Zou et al., 2025) and sustainable energy transition planning (Arslan et al., 2024a), demonstrating RAG's handling of regulation-driven, multi-source queries. For e-commerce and customer support, (Analytics and Sukhwal, 2024; Benita et al., 2024) focus on retrieval quality and response latency, with (Analytics and Sukhwal, 2024) showing that simpler query formulations directly improve retrieval performance. General-purpose RAG systems in (Ko et al., 2024a; Xu and Liu, 2024; Chang et al., 2024; Chen and Tran, 2024; Ko et al., 2024b; Putri et al., 2024; Lu et al., 2024; Leung et al., 2024; Muthyala et al., 2024; Andersson, 2024; Kim et al., 2024; Vakayil et al., 2024; Zhang et al., 2024) range from culturally-inclusive chatbots to secure document interaction tools. All share common findings: improved response quality, reduced hallucination, and stronger domain-specific grounding. Ko et al. (Ko et al., 2024b) propose ReRAG as an architecture specifically targeting chunking optimization and hallucination control, while Andersson (Andersson, 2024) conducts a comparative evaluation of RAG with GPT-3.5 Turbo and LLaMA 3 70B.

3.2 RAG Chatbot vs. Cloud-Based Toolkits-RQ1

RQ1 asks to what extent reviewed studies compare RAG-augmented chatbots with cloud-based enterprise toolkits, and how RAG compares with standalone-LLM baselines where such evidence exists. The first part of the answer is an explicit negative finding: none of the fifty reviewed studies benchmarks RAG against a managed cloud service such as Microsoft Azure AI Search, Amazon Bedrock, or Google Vertex AI. We therefore report RQ1 as an open evidence gap and frame the comparison RQ1 invites as future work; our companion empirical study (Alali et al., 2025) addresses this gap on a single SAP deployment but does not generalise across vendors. For the standalone-LLM comparison, the reviewed studies are consistent: RAG outperforms base-LLM on response precision, relevance, and domain-specific accuracy. Ko et al. (2024a) report a 13-25% relevance improvement over a base-LLM baseline. Ko et al. (2024b) report a 12.2% improvement using their ReRAG architecture. Zhou et al. (2024) report 95% context recall and 93.73% faithfulness in a clinical chatbot. The sub-question was *What are the metrics that as-*

sess the performance?, from the studies, has been found the metrics for performance assessment are discussed below.Table 1 shows all the metrics used in the studies The primary evaluation metrics surfaced across the corpus include: **Mean Reciprocal Rank (MRR)** for retrieval performance (Analytics and Sukhwal, 2024); **BLEU and ROUGE** scores for generation quality (Leung et al., 2024; Alario-Hoyos et al., 2024); the **System Usability Scale (SUS)** for user satisfaction (Zhou et al., 2024); alongside other common metrics such as accuracy, relevance score, context recall, and faithfulness.

Furthermore, various fields apply different evaluation measures based on their domain requirements. Medical applications typically measure trustworthiness, clinical validity, and diagnostic accuracy using domain-specific benchmark models (Fink et al., 2025). Legal and construction applications assess the effectiveness of chatbots through contract risk identification accuracy and regulatory compliance benchmarks (Wong et al., 2024). Industrial applications prioritize retrieval precision, process optimization effectiveness, and real-time decision accuracy (Jeon et al., 2025). Sustainability and ESG reporting systems are quantified based on accuracy in data extraction, completeness of disclosures, and regulatory compliance rates (Zou et al., 2025). In education, RAG models are evaluated for consistency in facts, context-dependent reasoning, and correctness in multiple-choice questions (Alawwad et al., 2025).

The answer to the second sub-question *What toolkit/LLMs will be used in this study* is that there are different LLMs and tools used, such as Llama 3.1 and Phi-3 (Ko et al., 2024a), GLM-4 (Xu and Liu, 2024), Coze Platform (Chang et al., 2024), Azure AI Search (Andersson, 2024).
Retrieval frameworks such as LangChain, knowledge graphs, and vector-based retrieval mechanisms are widely used to optimize retrieval pipelines (Sun et al., 2024). Several studies implement graph-based retrieval systems to complement LLM-generated responses, ensuring context-sensitive retrieval and regulatory compliance (Wong et al., 2024). In the legal, healthcare, and ESG domains, the deployment of domain-specific knowledge bases guaranties that retrieval-enhanced chatbots remain factually current and domain-specific (Zou et al., 2025; Arslan et al., 2024a). The discussion summarizing these questions is that the RAG-chatbot performs better than base-LLM, especially in domain-specific tasks. There are many tools and studies to develop and evaluate RAG chatbots effectively. However, some challenges remain, such as computational efficiency and latency, that need to be addressed in future work (Ko et al., 2024a; Chen and Tran, 2024; Leung et al., 2024). Table 2 summarizes the LLM/tools used in the studies.

3.3 RAG-Chatbot&Industry- RQ2

This section is to discuss the findings related to the research questions: *How does the RAG chatbot facilitate industry work?* RQ2.1: Why does the industry not rely on LLM only to load their document? RQ2.2: What are the opinions of industry professionals on developing secure, custom chatbots? RQ2.3:What are the practices of the practitioners' practices to implement an effective RAG chatbot? The studies found that RAG has a significant impact on facilitating industry work and that RAG shows high performance in terms of response quality and retrieval of real-time data. To illustrate, (Zhu and Vuppalapati, 2024) reduces the distribution that occurs and improves overall operation efficiency, while (Liu et al., 2024), benefits the automotive industry in their high-performance technical report automation. In (Benita et al., 2024), highlight the advantage of having RAG-chatbot in the customer support area. Industries avoid excessive reliance on LLMs to process documents due to their risks of hallucination, outdated information, and susceptibility to attacks (Shusterman et al., 2025). In healthcare, single LLMs do not learn real-time information well, producing erroneous or misleading diagnostic recommendations (Gu et al., 2025). The construction industry identifies that contractual legal contracts require precise interpretation, which single LLMs cannot do without external retrieval systems (Wong et al., 2024). Sustainability and business applications identify that timely policy updates and ESG compliance require the convergence of several sources of retrieval (Arslan et al., 2024a). Experts in the industry are strongly committed to the development of trustworthy, application-specific chatbots that apply proprietary knowledge bases to ensure trustworthiness and secrecy (Arslan et al., 2024b). Healthcare practitioners observe that clinical chatbots should align with professional ethics and moral values, avoiding hallucinated clinical advice (Shusterman et al., 2025). Experts in law approve of AI-backed con-

Domain	Key Findings	Metrics Used	Toolkits/Platforms	References
Healthcare	Accuracy and relevance improvement (95% context recall)	SUS, BLEU, ROUGE	GLM-4, DIAG-GPT	(Sree et al., 2024)
Education	Student Satisfaction Rate: 93.3%	SUS, BLEU	GPT-3.5, GPT-4	(Zhou et al., 2024)
Supply Chain	Operational agility and efficiency	MRR	Azure AI Search	(Steybe et al., 2025)
E-Commerce	User experience satisfaction and reduced response times	MRR, Recall@n	Falcon-RW-1B, GPT-3.5	(Benita et al., 2024)

Table 1: Summary of Key Metrics Findings Across Different Domains

tract review but prioritize retrieval-based risk assessment functionalities to ensure observance of statutory provisions (Wong et al., 2024). Corporate sustainability leaders claim that methods to pull out ESG information must be transparent, regulation-compliant, and able to handle unstructured firm reports (Zou et al., 2025). The deep understanding of industry LLM adoption and requirements to develop a chatbot remains unclear, which requires deep analysis in the industry settings, but the general need is a secure, customized chatbot that provides data in real-time. To make RAG-based chatbots effective, practitioners adhere to the following best practices: fine-tuning retrieval mechanisms for better search precision (Feng et al., 2025), tweaking chunking methods to ensure retrieval relevance in long-text processing applications (Alawwad et al., 2025), applying knowledge graphs to hierarchically organize retrieved information (Sun et al., 2024), rendering chatbots more versatile by using multi-agent architectures (Arslan et al., 2024b), and maintaining knowledge bases up-to-date in real-time to prevent outdated responses in dynamic industries (Arslan et al., 2024a).

Tool/Large Language Model
Llama 3.1, Phi-3
LLaMA-2-7B-chat-hf
DIAG-GPT
LLaMA-2-7B, Coze platform (coze.com)
Gemini LLM
RAG+Llama-2-13b
GLM-4
GPT-3.5-Turbo
OpenAI GPT-3.5 Turbo
Falcon-RW-1B, GPT-3.5
Chatglm3-6B-Chat
Baichuan2-13B-Chat
Owen-14B-Chat

Table 2: Toolkits/Platforms Integrated with RAG

4 Security Taxonomy for Enterprise RAG - RQ3

RQ3 asks which security threats, controls, and architectural components are reported in the primary studies, and how these map onto established threat catalogues for LLM-integrated applications. To answer this question while avoiding unsupported generalizations beyond the reviewed corpus, we proceed in two steps. We first compile an external threat catalogue from outside the primary corpus, drawing on the NIST taxonomy of adversarial machine learning (Vassilev et al., 2024), recent secure-RAG literature (Greshake et al., 2023; Liu et al., 2023; Zhang et al., 2025; Chaudhari et al., 2024; Naseh et al., 2025), the privacy-preserving generative-AI review of Feretzakis et al. (2024), the safety/trust survey of Huang et al. (2024), the attack–defence survey of Liao et al. (2026), and the chatbot-security SLR of Jing et al. (2023). We then check, threat by threat, which of the fifty primary studies (if any) discuss it. The result of this mapping is the conservative claim of Section 5: only two studies engage with security as a primary object of analysis, and no study presents an end-to-end secure-RAG architecture.

4.1 Threats

We group threats by where they enter the RAG pipeline. *Retrieval-time threats* target the indexed knowledge base. Knowledge-base or corpus poisoning inserts adversarial passages so that the retriever surfaces attacker-controlled content for chosen queries; practical attacks of this form are demonstrated by Zhang et al. (2025) for textual RAG and generalised by Chaudhari et al. (2024) to backdoor settings. *Generation-time threats* target the prompt assembled from the user query and the retrieved passages. Direct prompt injection rewrites the system's instructions through the user-

supplied text (Liu et al., 2023); *indirect* prompt injection embeds those instructions inside the documents themselves, so that the injection fires once retrieved into context (Greshake et al., 2023). *Privacy and confidentiality threats* target the knowledge base and the model's behaviour. Membership-inference attacks reveal whether a given document is present in the RAG index (Naseh et al., 2025); sensitive-data leakage and exfiltration arise when the model regurgitates personal data, credentials, or proprietary text contained in retrieved chunks (Feretzakis et al., 2024; Vassilev et al., 2024). *Behavioural threats* target the LLM's alignment: jailbreaking elicits responses that violate provider or organisational policies (Huang et al., 2024; Liao et al., 2026).

4.2 Architectural controls

The controls that recur across the secure-RAG literature can be organised along the same pipeline. At ingestion: provenance tracking, signed sources, and content validation reduce exposure to corpus-poisoning. At retrieval: input sanitisation, ranking with provenance-aware re-rankers, and per-query access control enforce a need-to-know boundary; Gummadi et al. (2024) discusses transport-layer controls for the embedding/retrieval channel. At generation: system-prompt hardening, output filtering, and grounding checks reduce the success rate of direct and indirect prompt injection (Greshake et al., 2023). At deployment: locally hosted models (the design pattern adopted by Chen and Tran (2024) and Liu et al. (2024)) reduce data residency risk relative to cloud APIs at the cost of operational overhead. At governance: secure logging, audit trails, and incident response are baseline expectations for enterprise systems and are explicitly discussed by Andersson (2024) for the Azure deployment case, but are absent from forty-eight of the fifty included studies.

4.3 Coverage in the primary corpus

Mapping the external taxonomy back onto the primary corpus produces a sparse picture. Direct prompt injection is mentioned only implicitly, as a motivation for reliability checks (Benita et al., 2024); indirect prompt injection is not addressed in any reviewed study. Corpus poisoning is not addressed. Membership inference and sensitive-data leakage are mentioned as concerns by Chen and Tran (2024) and Andersson (2024) but no technical mitigations are evaluated. Local-deployment is

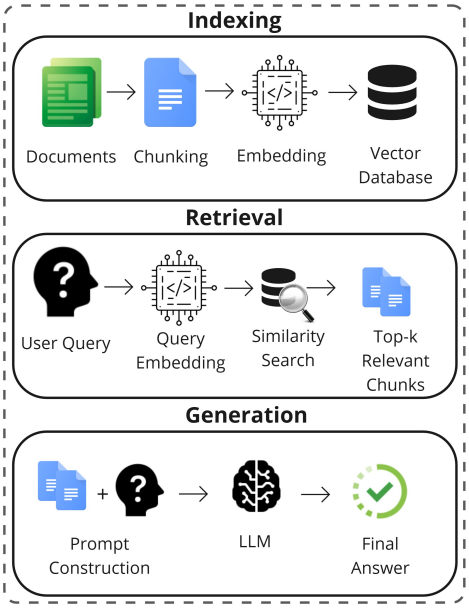

Figure 3: Retrieval-Augmented Generation pipeline.

the most frequently discussed control (Chen and Tran, 2024; Liu et al., 2024). Access control, audit logging, and output validation are mentioned in passing in (Matar and Mohammad, 2024; Zhang et al., 2024; Benita et al., 2024; Muthyala et al., 2024) but never instantiated. The evidence base for security architectures in enterprise RAG remains limited. Accordingly, this review frames secure RAG as an emerging research gap in enterprise settings.

5 RAG Pipeline

Based on a reading of the collected studies, this section synthesizes an overview of the standard RAG architecture. The pipeline is relevant both for researchers seeking to replicate prior work and for practitioners planning a chatbot deployment. Three core stages are consistently reported across the reviewed literature: *indexing*, *retrieval*, and *generation*. Figure 3 illustrates the full flow.

- *Indexing*: Before any query can be answered, all documents in the knowledge base must be prepared in a form that supports semantic search. This begins by dividing each document into smaller passages, or chunks, each of which is independently searchable. The appropriate chunk size is a key design decision: chunks that are too large may dilute relevance signals, while overly small chunks can lack the context needed for a meaningful answer.

Each chunk is then passed through an embedding model that maps its textual content to a dense numerical vector, encoding the semantic meaning of the passage so that content covering similar topics is represented nearby in the vector space. All resulting vectors are stored in a vector database such as Chroma, which supports fast similarity-based lookups. This stage is foundational because the ceiling on retrieval quality is determined here (Ko et al., 2024b; Lu et al., 2024).

- *Retrieval*: When a user submits a query, it is transformed using the same embedding model applied during indexing, producing a query vector in the same high-dimensional space (Analytics and Sukhwal, 2024). The system then computes similarity between this query vector and every stored document vector using cosine similarity, which reflects semantic relatedness rather than literal word overlap. Documents whose vectors align most closely with the query are ranked as the most relevant, and the top-k results are selected for the generation stage. The quality of the embedding model and the value of k are both engineering decisions that directly determine how useful the retrieved context will be for generating a correct answer (Ko et al., 2024b; Andersson, 2024).

- *Generation*: The retrieved document chunks are combined with the original user query and provided to the LLM through a structured prompt template. The model generates a response that draws on both the retrieved evidence and its pre-trained language knowledge. This grounding in retrieved content is what fundamentally distinguishes RAG from a direct LLM call: answers are anchored to material that exists in the knowledge base, reducing hallucination and improving factual reliability. The LLM can be any capable model, including OpenAI's models, LLaMA variants, or locally deployed options such as Ollama, depending on the privacy and infrastructure requirements of the deployment.

6 Implications for Research and Practice

This SLR establishes a baseline for designing RAG-augmented chatbots, outlining distinct directions for research and practice. For **researchers**, the lack of standardized evaluation metrics across studies points to the need for a unified evaluation suite. Additionally, unresolved optimization challenges—such as hybrid sparse-dense retrieval and hierarchical semantic segmentation—require deeper investigation (Vakayil et al., 2024; Lu et al., 2024). Critically, future research must address enterprise security concerns, including defences against indirect prompt injection, corpus poisoning, and membership inference (Greshake et al., 2023; Zhang et al., 2025; Vassilev et al., 2024). For **practitioners**, RAG provides substantial workflow gains in healthcare, supply chain, and manufacturing (Zhu and Vuppalapati, 2024). Successful industrial deployment requires fine-tuning LLMs, optimizing retrieval mechanisms to balance cost and latency (Liu et al., 2024), and integrating user-friendly interfaces with robust access controls and prompt engineering to ensure reliability (Alario-Hoyos et al., 2024).

7 Conclusion and Future Work

This Systematic Literature Review maps the current landscape of RAG-LLM chatbot research against the requirements of industrial deployment, particularly in enterprise planning contexts like SAP. By analyzing 50 empirical studies, we confirm that RAG-augmented chatbots consistently outperform standalone LLM configurations in precision and relevance across healthcare, education, and manufacturing. However, two critical gaps remain: the absence of direct comparisons between customized RAG chatbots and cloud-managed services (e.g., Azure), and the lack of technical depth regarding enterprise security controls and access-management architectures.

To address these gaps, future work must focus on: (1) expanding evaluation datasets to improve model generalization; (2) optimizing retrieval using hybrid sparse-dense methods; (3) treating security as a first-class design requirement, specifically evaluating defences against indirect prompt injection, corpus poisoning, and data leakage (Vassilev et al., 2024; Greshake et al., 2023); (4) benchmarking RAG chatbots against production-ready cloud services in realistic industrial settings; and (5) establishing unified evaluation standards to enable cross-study meta-analyses.

Limitations

This SLR has several key limitations. First, the primary corpus is relatively small (50 studies) and exclusively covers 2024–2025, which limits longitudinal claims. Second, the high heterogeneity of evaluation metrics (e.g., MRR, BLEU, SUS, accuracy) prevents quantitative meta-analysis. Third, most studies report controlled laboratory experiments rather than long-term production deployments. Fourth, none of the included studies evaluates a direct, native SAP-ERP integration. Finally, as noted in Section 4, security and data privacy remain the thinnest areas of coverage across the reviewed literature.

Ethical Statements

This systematic literature review does not involve human subjects or primary data collection that would require ethical approval. All analyzed papers are publicly available and properly cited.

Acknowledgments

We acknowledge the support of the Information and Computer Science Department at KFUPM.

References

Fatimah Alali, Abul Bashar, Haya Aldawsari, and Sajjad Mahmood. 2025. Comparative analysis of industrial sap chatbots: Rag-llm and cloud-based approaches. In *Proceedings of the International Conference on Social Sciences and Business (ICSSB 2025)*, volume 1, pages 1–8.

Carlos Alario-Hoyos, Rebiha Kemcha, Carlos Delgado Kloos, Patricia Callejo, Iria Estévez-Ayres, David Santín-Cristóbal, Francisco Cruz-Argudo, and José Luis López-Sánchez. 2024. Tailoring your code companion: Leveraging llms and rag to develop a chatbot to support students in a programming course. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 1–8. IEEE.

H. A. Alawwad, A. Alhothali, U. Naseem, et al. 2025. Enhancing textual textbook question answering with large language models and retrieval augmented generation. In *Proceedings of Pattern Recognition*, volume 162, page 111332. Elsevier.

Data Analytics and Dishant Sukhwal. 2024. Retrieval augmented generation: An evaluation of rag-based chatbot for customer support. *Retrieval Augmented Generation: An Evaluation of RAG-based Chatbot for Customer Support*.

Henrik Andersson. 2024. Retrieval-augmentedgeneration with azure open ai.

M. Arslan, L. Mahdjoubi, and S. Munawar. 2024a. Driving sustainable energy transitions with a multi-source rag-llm system. In *Proceedings of Energy and Buildings*, volume 324, page 114827. Elsevier.

M. Arslan, S. Munawar, and C. Cruz. 2024b. Sustainable digitalization of business with multi-agent rag and llm. In *Proceedings of Procedia Computer Science*, volume 246, pages 4722–4731. Elsevier.

Ayesha Azam, Zubaira Naz, and Muhammad Usman Ghani Khan. 2024. Cancerbot: A retrieval-augmented generation based cancer chatbot using large language models. In *2024 18th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE.

J Benita, Kosireddy Vivek Charan Tej, E Vinay Kumar, G Venkata Subbarao, and CH Venkatesh. 2024. Implementation of retrieval-augmented generation (rag) in chatbot systems for enhanced real-time customer support in e-commerce. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 1381–1388. IEEE.

M. Bhattacharyya, R. Joy, and S. Sarkar. 2024. Comparative analysis of retrieval-augmented generation (rag) based large language models (llm) for medical chatbot applications. ResearchGate technical report. https://www.researchgate.net/publication/385167532. Accessed: 2025.

Chen-Chi Chang, Han-Pi Chang, and Hung-Shin Lee. 2024. Leveraging retrieval-augmented generation for culturally inclusive hakka chatbots: Design insights and user perceptions. In *2024 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, pages 1–6. IEEE.

Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Anca Cretu, and Alina Oprea. 2024. Phantom: General backdoor attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*.

Andre Chen and Sieu Tran. 2024. Supercharging document composition with generative ai: A secure, custom retrieval-augmented generation approach. In *2024 11th IEEE Swiss Conference on Data Science (SDS)*, pages 123–130. IEEE.

Sagnik Dakshit. 2024. Faculty perspectives on the potential of rag in computer science higher education. In *Proceedings of the 25th Annual Conference on Information Technology Education*, pages 19–24.

Saad Ezzini, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. 2023. Ai-based question answering assistance for analyzing natural-language requirements. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1277–1289. IEEE.

J. Feng, Q. Wang, H. Qiu, and L. Liu. 2025. Retrieval in decoder benefits generative models for explainable complex question answering. In *Proceedings of Neural Networks*, volume 181, page 106833. Elsevier.

Georgios Feretzakis, Konstantinos Papaspyridis, Aris Gkoulalas-Divanis, and Vassilios S. Verykios. 2024. Privacy-preserving techniques in generative AI and large language models: A narrative review. *Information*, 15(11):697.

Tiam Feridooni, Arshia P Javidan, Daniyal N Mahmood, Zoya Gomes, Andrew Dueck, Mark Wheatcroft, and David Szalay. 2024. Development of a vascular surgery–specific artificial intelligence chat interface using retrieval-augmented generation: Vasc. ai, a specialized vascular surgery chatbot. *JVS-Vascular Insights*, page 100137.

A. Fink, J. Nattenmüller, S. Rau, et al. 2025. Retrieval-augmented generation improves precision and trust of a gpt-4 model for emergency radiology diagnosis and classification: a proof-of-concept study. In *Proceedings of European Radiology*. Springer.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, pages 79–90.

Z. Gu, W. Jia, M. Piccardi, and P. Yu. 2025. Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. In *Proceedings of Artificial Intelligence in Medicine*, volume 162, page 103078. Elsevier.

Venkata Gummadi, Pamula Udayaraju, Venkata Rahul Sarabu, Chaitanya Ravulu, Dhanunjaya Reddy Seelam Ratnam, and S. Chandragandhi. 2024. Enhancing communication and data transmission security in RAG using large language models. In *2024 International Conference on Sustainable Energy Solutions (ICSES)*, pages 1–6.

J. A. Heredia Álvaro and J. G. Barreda. 2025. An advanced retrieval-augmented generation system for manufacturing quality control. In *Proceedings of Advances in Engineering Informatics*, volume 64, page 103007. Elsevier.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175.

J. Jeon, Y. Sim, H. Lee, et al. 2025. Chatcnc: Conversational machine monitoring via large language model and real-time data retrieval augmented generation. In *Proceedings of Journal of Manufacturing Systems*, volume 79, pages 504–514. Elsevier.

J. Jiang, L. Yan, H. Liu, et al. 2025. Knowledge assimilation: Implementing knowledge-guided agricultural large language model. In *Proceedings of Knowledge-Based Systems*. Elsevier.

Yang Jing, Yen-Lin Chen, Lip Yee Por, and Chin-Soon Ku. 2023. A systematic literature review of information security in chatbots. *Applied Sciences*, 13(11):6355.

Minsoo Kim, Doyoun Kim, Yunji Park, and Doowon Jeong. 2024. Development of an expert chatbot for digital forensics using rag model implementation. In *2024 International Conference on Platform Technology and Service (PlatCon)*, pages 182–187. IEEE.

B.A. Kitchenham and S. Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE-2007-01, Keele University and University of Durham.

Khant Ko, Thwet Yin Nyein, Khine Khine Oo, Thant Zin Oo, and Thet Thet Zin. 2024a. Retrieval augmented generation for document query automation using open source llms. In *2024 5th International Conference on Advanced Information Technologies (ICAIT)*, pages 1–6. IEEE.

Robin Ko, Mustafa Kağan Gürkan, and Fatoş T Yarman Vural. 2024b. Rerag: A new architecture for reducing the hallucination by retrieval-augmented generation. In *2024 9th International Conference on Computer Science and Engineering (UBMK)*, pages 961–965. IEEE.

J. Lee, S. Ahn, D. Kim, and D. Kim. 2024. Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. In *Proceedings of Automation in Construction*, volume 168, page 105846. Elsevier.

Chun-hung Jonas Leung, Yicheng Yi, Le Kuai, Zongxi Li, Siu-kei Au Yeung, Kwok-wah John Lee, Kahim Kelvin Ho, and Kevin Hung. 2024. Rag for question-answering for vocal training based on domain knowledge base. In *2024 11th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6. IEEE.

Zhiyu Liao, Kang Chen, Yuanguo Lin, Zhaoyu Yang, Wei Kang, Fan Lin, and Jianbing Sun. 2026. Attack and defense techniques in large language models: A survey and new perspectives. *Neural Networks*, 196:108388.

Fei Liu, Zejun Kang, and Xing Han. 2024. Optimizing rag techniques for automotive industry pdf chatbots: A case study with locally deployed ollama modelsoptimizing rag techniques based on locally deployed ollama modelsa case study with locally deployed ollama models. In *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Intelligent Information Processing*, pages 152–159.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against LLM-integrated applications. *arXiv preprint arXiv:2306.05499*.

Yanyan Lu, Jiao Peng, Xing Xu, Yue He, Tao Li, Jie Wei, Hongyu Jing, Heqing Wang, Bo Xu, and Hui Song. 2024. A retrieval-augmented generation framework for electric power industry question answering. In *Proceedings of the 2024 2nd International Conference on Electronics, Computers and Communication Technology*, pages 95–100.

L. Masanneck, S. G. Meuth, and M. Pawlitzki. 2025. Evaluating base and retrieval augmented llms with document or online support for evidence based neurology. In *Proceedings of NPJ Digital Medicine*, volume 8, page 137. Springer.

Khaled Matar and Yousef Mohammad. 2024. Improving the reliability ofeducational ai chatbots usingretrieval-augmented generation.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Mayank P Muthyala, Claire Lauer, and Stephen Carradini. 2024. So you want to build a chatbot?: A systematic case study comparing the design and development of two water chatbots. In *Proceedings of the 42nd ACM International Conference on Design of Communication*, pages 128–137.

Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. 2025. Riddle me this! stealthy membership inference for retrieval-augmented generation. *arXiv preprint arXiv:2502.00306*.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Mindi Richia Putri, Ario Yudo Husodo, and Budi Irmawati. 2024. Simplification of embedding process in retrieval augmented generation for optimizing question answering chatbot model. In *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 665–670. IEEE.

R. Shusterman, A. C. Waters, S. O'Neill, et al. 2025. An active inference strategy for prompting reliable responses from large language models in medical practice. In *Proceedings of NPJ Digital Medicine*, volume 8, page 119. Springer.

Y Bhanu Sree, Addicharla Sathvik, Damarla Sai Hema Akshit, Omrender Kumar, and Bandaru Sai Pranav Rao. 2024. Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health. In *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, pages 1–8. IEEE.

David Steybe, Philipp Poxleitner, Suad Aljohani, Bente Brokstad Herlofson, Ourania Nicolatou-Galitis, Vinod Patel, Stefano Fedele, Tae-Geon Kwon, Vittorio Fusco, Sarina EC Pichardo, et al. 2025. Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. *Journal of Cranio-Maxillofacial Surgery*.

Y. Sun, X. Li, C. Liu, et al. 2024. Development of an intelligent design and simulation aid system for heat treatment processes based on llm. In *Proceedings of Materials and Design*, volume 248, page 113506. Elsevier.

Sonia Vakayil, D Sujitha Juliet, Sunil Vakayil, et al. 2024. Rag-based llm chatbot using llama-2. In *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–5. IEEE.

Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 2024. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical Report NIST AI 100-2e2023, National Institute of Standards and Technology (NIST).

S. Wong, C. Zheng, X. Su, and Y. Tang. 2024. Construction contract risk identification based on knowledge-augmented language models. In *Proceedings of Computers and Industry*, volume 157–158, page 104082. Elsevier.

Liwei Xu and Jiarui Liu. 2024. A chat bot for enrollment of xi'an jiaotong-liverpool university based on rag. In *2024 8th International Workshop on Control Engineering and Advanced Algorithms (IWCEAA)*, pages 125–129. IEEE.

R. Xu, Y. Hong, F. Zhang, and H. Xu. 2024. Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. In *Proceedings of Scientific Reports*, volume 14. Springer.

Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Wang, Xiaoyu Cao, Neil Zhenqiang Gong, Jiliang Liu, and Yuqing Yang. 2025. Practical poisoning attacks against retrieval-augmented generation. *arXiv preprint arXiv:2504.03957*.

Huaxiao Zhang, Zhigang Li, Fang Liu, Yilan He, Zhichen Cao, and Yunbo Zheng. 2024. Design and implementation of langchain-based chatbot. In *2024 International Seminar on Artificial Intelligence, Computer Technology and Control Engineering (ACTCE)*, pages 226–229. IEEE.

Qingqing Zhou, Can Liu, Yuchen Duan, Kaijie Sun, Yu Li, Hongxing Kan, Zongyun Gu, Jianhua Shu, and Jili Hu. 2024. Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. *Frontiers in Medicine*, 11:1392555.

Beilei Zhu and Chandrasekar Vuppalapati. 2024. Enhancing supply chain efficiency through retrieve-augmented generation approach in large language models. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pages 117–121. IEEE.

Y. Zou, M. Shi, Z. Chen, et al. 2025. Esgreveal: An llm-based approach for extracting structured data from esg reports. In *Proceedings of Journal of Cleaner Production*, volume 489, page 144572. Elsevier.