Federated Retrieval-Augmented Generation: A Systematic Mapping Study

Abhijit Chakraborty[*1] **Chahana Dahal**[2] **Vivek Gupta**[*1]
[1]Arizona State University [2]University of Nevada, Las Vegas
{achakr40,vgupt140}@asu.edu {chahana.dahal}@unlv.edu

Abstract

Federated Retrieval-Augmented Generation (Federated RAG) combines Federated Learning (FL),which enables distributed model training without exposing raw data, with Retrieval-Augmented Generation (RAG), which improves the factual accuracy of language models by grounding outputs in external knowledge. As large language models are increasingly deployed in privacy-sensitive domains such as healthcare, finance, and personalized assistance, Federated RAG offers a promising framework for secure, knowledge-intensive natural language processing (NLP). To the best of our knowledge, this paper presents the first systematic mapping study of Federated RAG, covering literature published between 2020 and 2025. Following Kitchenham's guidelines for evidence-based software engineering, we develop a structured classification of research focuses, contribution types, and application domains. We analyze architectural patterns, temporal trends, and key challenges, including privacy-preserving retrieval, cross-client heterogeneity, and evaluation limitations. Our findings synthesize a rapidly evolving body of research, identify recurring design patterns, and surface open questions, providing a foundation for future work at the intersection of RAG and federated systems.

1 Introduction

Large language models (LLMs) are increasingly deployed in domains such as healthcare, finance, and personalized assistance, raising concerns about data privacy, ownership, and regulatory compliance. In response, **Federated Learning** (FL) has emerged as a compelling paradigm for training models across distributed clients without exchanging raw data (Kairouz et al., 2021). In parallel, **Retrieval-Augmented Generation** (RAG) (Lewis et al., 2020a) enhances LLMs by grounding their outputs in dynamically retrieved external knowledge, reducing hallucinations and improving the accuracy of facts (Lewis et al., 2020b).

FL trains models across clients without sharing raw data, using methods like FedAvg (McMahan et al., 2017) and secure aggregation, widely applied in privacy-sensitive (Kairouz et al., 2021) domains.

RAG is a hybrid method that combines a text retriever with a generator. Before producing a response, the model first retrieves relevant text from an external knowledge base or document store. This process helps reduce hallucinations and improves factual consistency in the generated output. Popular implementations include the original RAG model (Lewis et al., 2020b) and Fusion-in-Decoder (FiD) (Izacard and Grave, 2020), which demonstrate how integrating retrieval can enhance open-domain question answering.

At the intersection of these paradigms lies **Federated RAG**, a hybrid approach that enables LLMs to access distributed knowledge sources in a privacy-preserving manner. It combines the strengths of data localization, personalized retrieval, and context-aware generation. Conceptually, Federated RAG builds on earlier federated search methods (Shokouhi and Si, 2011), which aggregated results from siloed sources without centralized indexing, and extends them to support complex generative tasks.

Federated RAG is distinctive because federated learning safeguards training-time privacy by restricting data to local silos, while retrieval-augmented generation grounds inference-time outputs in external evidence, reducing hallucinations. Their integration enables capabilities such as source attribution, local index maintenance, and dynamic document unlearning that neither paradigm achieves alone.

As LLMs have grown in capability, driven by innovations such as the Transformer architecture (Vaswani et al., 2017), pre-training methods (De-

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7362–7374
November 4-9, 2025

vlin et al., 2019; Radford et al., 2019; Brown et al., 2020), and prompting strategies (Schick and Schütze, 2021; Liu et al., 2021; Debnath et al., 2025), so has the demand for integrating them with heterogeneous private data contexts. Since 2019, the parallel evolution of LLMs, RAG, and FL has laid the groundwork for Federated RAG to emerge as both a viable and increasingly necessary framework.

In this paper, we present the first systematic mapping study of Federated RAG, following the methodology of evidence-based software engineering proposed by Kitchenham (Kitchenham et al., 2011). Our objective is to map how retrieval-augmented generation is being adapted and deployed across federated architectures, and to surface recurring trends, gaps, and design patterns in the literature. To this end, we address the following research questions:

- **RQ1:** What are the dominant architectural patterns used to integrate Retrieval-Augmented Generation (RAG) into federated systems?
- **RQ2:** What are the primary research focuses and contribution types in the literature on federated RAG systems?
- **RQ3:** In which application domains (e.g., healthcare, finance, education) has federated RAG been explored, and what problems are these systems designed to solve?
- **RQ4:** What key challenges, open issues, and underexplored areas emerge from current research at the intersection of RAG and federated systems?

These questions structure our analysis of the literature published between 2020 and 2025, allowing us to classify prior work, identify underexplored areas, and provide a foundation for future research at the intersection of privacy, retrieval, and generation.

2 Methodology

We searched the best NLP, ML and security venues (2020-2025) using terms such as *"federated learning"*, *"retrieval-augmented generation"* and *"federated search"*, including backward references. From 50 papers, 18 met criteria. Following Kitchenham's (Kitchenham et al., 2011) method, we coded each by (a) research focus, (b) contribution type, (c) application domain.

We also recorded the publication year of each study to observe temporal trends and emerging research fronts. Based on this categorization, we developed the unified classification scheme (Figure 1) to summarize the distribution of studies across focus areas and contribution types. Furthermore, this schema informed the design of our architectural taxonomy (Figure 2), which visually organizes the Federated RAG landscape by contribution, application, and system goals.

3 Analysis of Recent Studies

Retrieval ranges from naive aggregation to secure encrypted search (e.g., FRAG (Zhao, 2024)). Generation spans centralized, client-specific, or hybrid (e.g., GPT-FedRec (Zeng et al., 2024)). These choices balance scalability, personalization, and privacy (Figure 2).

These architectural choices require balancing scalability, performance, personalization, and privacy.

3.1 Research Focus

Foundational privacy work from 2020–2022 paved the way for Federated RAG. *PPDA* (Jeon et al., 2021) introduced lightweight decentralized aggregation without differential privacy (DP) or homomorphic encryption (HE). *DP-FedKGE* (Peng et al., 2021) added entity-level DP for federated knowledge-graph embedding. *DeTrustFL* (Xu et al., 2022) used cryptographic consensus to block disaggregation attacks, and *FedX* (Han et al., 2022) employed cross-client distillation for unsupervised representation learning.

From 2023 through 2025, privacy mechanisms became system-level. *C-FedRAG* (Addison et al., 2024) executes retrieval *and* generation entirely inside trusted enclaves, securing clinical QA on *MKP-QA*. *FRAG* (Zhao, 2024) realises IND-CPA-secure kNN search via single-key HE while keeping latency practical.

Parallel work targets efficiency. *RAGRoute* (Guerraoui et al., 2025) learns to route queries to high-utility silos, cutting 75 % redundant traffic on MIRAGE. *FedB4RAG* (Wang et al., 2024b) benchmarks routing policies over 16 BEIR datasets, exposing accuracy–cost trade-offs.

Integration efforts replace isolated modules with end-to-end federated stacks. *FedE4RAG* (Mao et al., 2025) trains dense retrievers collaboratively, combining HE-protected parameter exchange with

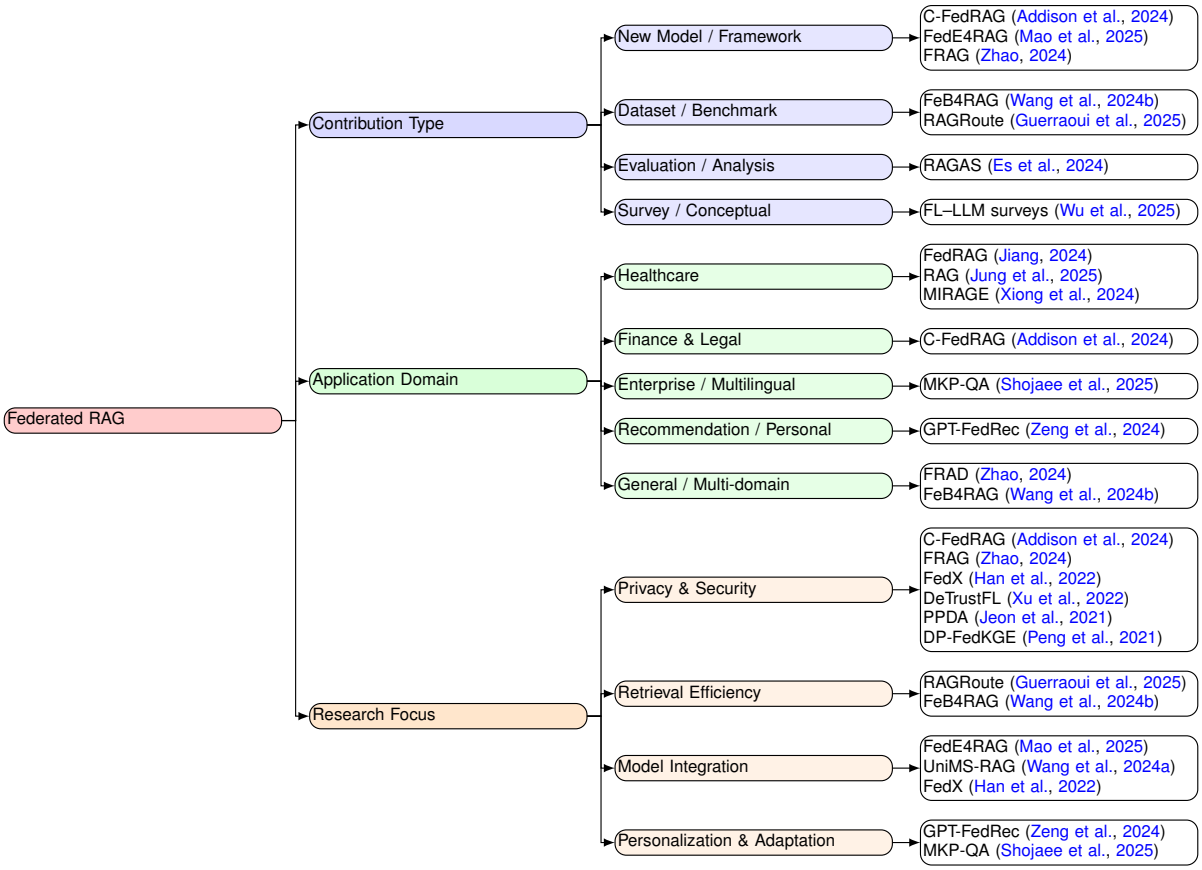

Figure 1: **Classification Scheme:** Summarizes the scheme with example categories and the distribution of studies (one study may fall into multiple categories).

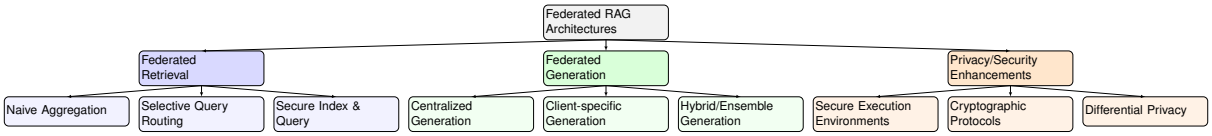

Figure 2: **Taxonomy** of Federated RAG Architectures. A conceptual map to classify the diverse system designs in the literature.

Research Focus	**# of Studies**
Privacy & Security	6
Personalization & Adaptation	2
Retrieval Efficiency	2
Model Integration	3

Table 1: Classification of primary research focus in federated RAG studies, aligned with the taxonomy shown in Figure 1.

distillation to align heterogeneous embeddings. *UniMS-RAG* unifies client-side retrieval with a shared generator, allowing adaptive control over privacy and latency. Table 1 summarizes these trajectories in the primary studies *18* on map. Half still prioritize *privacy and security*, reflecting the FL heritage of the field. *Personalisation* (e.g., *GPT-FedRec*, *MKP-QA*) and *retrieval efficiency* (e.g., *RAGRoute*, *FedB4RAG*) are rising, while *model-level integration* (*FedE4RAG*, *UniMS-RAG*) signals a shift toward deployable cross-silo systems.

Overall, research has progressed from algorithmic proofs of concept (2020–2022) to holistic architectures (2023–2025) that simultaneously safeguard data, route queries intelligently, and harmonise retrieval with generation, marking a decisive step toward scalable Federated RAG. To complement these qualitative descriptions, Table 5 in the Appendix synthesizes empirical metrics reported across representative systems, highlighting trade-offs between privacy, efficiency, and personalization.

3.2 Contribution Type

The literatures on Federated RAG reflect a balanced evolution across four primary contribution

Contribution Type	# of Studies
New Model / Framework	3
Dataset / Benchmark	2
Evaluation / Analysis	1
Survey / Conceptual	1

Table 2: Categorization of studies by primary contribution type. Majority propose new models; others contribute benchmarks, evaluation metrics, or conceptual overviews.

types: *new models or frameworks*, *datasets and benchmarks*, *evaluation and analysis tools*, and *conceptual overviews*. As shown in Table 2, a substantial portion of the work focuses on developing novel system architectures for privacy-preserving generation and retrieval. *FedE4RAG*(Mao et al., 2025) introduces a modular design that integrates homomorphic encryption and knowledge distillation to train federated dense retrievers, offering a decentralized alternative to traditional centralized systems. *C-FedRAG*(Addison et al., 2024) builds a confidential execution pipeline using Trusted Execution Environments (TEEs), enabling secure generation and retrieval particularly suited to sensitive domains like healthcare. *FRAG* (Zhao, 2024) contributes a cryptographic kNN retrieval system backed by IND-CPA-secure homomorphic encryption, facilitating privacy-preserving vector search.

Beyond system-level proposals, several contributions focus on datasets and benchmarking. *FeB4RAG*(Wang et al., 2024b) presents the first federated retrieval benchmark across 16 BEIR-derived domains, enabling rigorous evaluation of routing strategies and retrieval quality under resource constraints. *MKP-QA*(Shojaee et al., 2025) introduces a domain-specific benchmark for multilingual enterprise QA, designed to assess the selection of cross-silo documents and the preservation of context. Additionally, *RAGRoute* (Guerraoui et al., 2025) offers a specialized evaluation testbed for its adaptive query planning mechanism, allowing performance measurement across silos under constrained budgets.

Evaluation and analysis contributions aim to establish consistent standards for measuring retrieval and generation quality in federated setups. *RAGAS* (Es et al., 2024) proposes a suite of evaluation metrics tailored to federated RAG pipelines, accounting for inconsistencies in retrieved evidence, hallucinations, and robustness to noise injection. *FeB4RAG* extends this line of work by benchmarking retrieval degradation when shifting from naive to intelligent routing policies, illustrating the trade-offs between accuracy and resource consumption.

Finally, *survey and conceptual* literature remains sparse. The only entry in this category is *FL-LLM Surveys* (Wu et al., 2025), which provides a broader discussion on federated LLMs and briefly references RAG mechanisms. In contrast, our mapping study represents to the best of our knowledge, the first structured synthesis of architectural, empirical, and deployment-oriented work in Federated RAG, addressing a clear gap in the literature.

3.3 Application Domains and Use Cases

Federated RAG is gaining traction across a variety of domains that require both strong privacy guarantees and access to distributed knowledge sources. Among these, *healthcare* has emerged as the most prominently explored application vertical, driven by high data sensitivity and institutional silos. Jung et al. (2025) demonstrated that federated RAG pipelines outperform traditional FL setups on clinical question answering tasks by retrieving from a shared medical literature corpus while preserving patient privacy. Complementing this, Jiang (2024) introduced a hierarchical retrieval approach that selects relevant hospitals before accessing localized documents, facilitating fine-grained medical inference. *MIRAGE* (Xiong et al., 2024) further formalizes this direction through a clinical QA benchmark tailored for federated evaluation, underscoring the domain's methodological maturity.

Outside of healthcare, *finance and legal* sectors have been identified as strong candidates for Federated RAG due to similar compliance and confidentiality constraints. *C-FedRAG*(Addison et al., 2024) targets regulation-aware summarization and document generation, although large-scale evaluation in these domains remains limited. In *enterprise and multilingual QA*, *MKP-QA*(Shojaee et al., 2025) introduces a federated benchmark for retrieving internal product knowledge across silos and languages without centralizing sensitive documentation. Meanwhile, *GPT-FedRec* (Zeng et al., 2024) explores *personalized recommendation* using a hybrid framework that combines collaborative filtering with retrieval-augmented generation, showing reductions in hallucination and data sparsity.

In addition to these domain-specific efforts, several studies adopt *general or multi-domain* settings. *FeB4RAG*(Wang et al., 2024b) provides a comprehensive evaluation across 16 BEIR-derived tasks, while *FRAG*(Zhao, 2024) focuses on architectural generalization under cryptographic constraints. These studies emphasize benchmarking

Application Domain	# of Studies
Healthcare	3
Finance & Legal	1
Enterprise / Multilingual	1
Recommendation / Personal	1
General / Multi-domain	2

Table 3: Application domains addressed in federated RAG research. Studies are counted by explicit domain targeting, with several works remaining domain-agnostic or generalized.

and scalable deployment over specific industry contexts.

As summarized in Table 3, healthcare dominates the literature with three focused studies, followed by singular works in finance/legal, enterprise/multilingual QA, and recommendation systems. This distribution aligns with real-world concerns over privacy, compliance, and heterogeneous data silos. Earlier Figure 2 illustrated how such deployment settings are reflected in the system-level architecture choices across the literature. While many models are still domain-agnostic, healthcare and enterprise NLP tasks are emerging as leading targets for practical Federated RAG adoption, with future work likely to expand into education, public sector governance, and legal tech.

Beyond these domain categorizations, several privacy-critical scenarios further illustrate the practical relevance of Federated RAG. In healthcare, systems such as C-FedRAG and MIRAGE enable *hospital QA across silos*, allowing institutions to answer clinical questions from shared medical knowledge while ensuring that patient records never leave local storage. In financial services, approaches like RAGRoute support *cross-division portfolio analysis*, where sensitive compliance and investment documents remain within organizational boundaries yet can be queried collaboratively. Finally, in personal and consumer-facing settings, *on-device assistants* equipped with Federated RAG can retrieve from local files or personal notes without exposing them to cloud servers, offering a concrete pathway to privacy-preserving personal AI.

These scenarios underscore that Federated RAG extends beyond abstract architectural design to address urgent real-world concerns, where the dual imperatives of factual accuracy and data sovereignty are non-negotiable. By grounding retrieval and generation within privacy-preserving frameworks, such applications demonstrate the transformative potential of Federated RAG in bridging regulatory, ethical, and operational requirements.

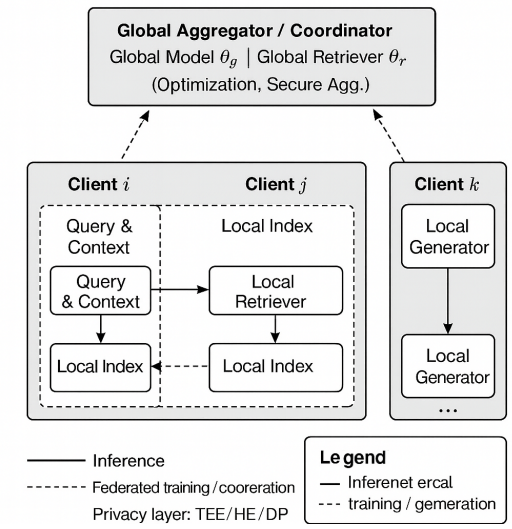

Figure 3: Conceptual architecture of Federated RAG, showing local query processing and index management at each client, coordinated with global model updates.

3.4 Unified Objective Function and Architecture

While the taxonomy in Figure 2 organizes existing designs across retrieval, generation, and privacy axes, it is also useful to formalize how these components interact within a single optimization framework. We introduce a conceptual objective for Federated RAG see Figure 3 that captures joint training across clients:

$$L_{\text{FedRAG}}(\theta) = \sum_{i=1}^{M} \frac{n_i}{N} \left(L^{(i)}_{\text{retrieve}}(\theta_r) + L^{(i)}_{\text{generate}}(\theta_r, \theta_g) \right)$$

where M is the number of clients, n_i is the number of samples on client i, N is the total number of samples, θ_r are retriever parameters, and θ_g are generator parameters. This loss highlights that retrieval and generation can be jointly optimized in a federated setting, balancing local personalization with global consistency.

This abstraction complements the taxonomy: retrieval modules may use selective routing or encrypted kNN, generation may be centralized, client-specific, or hybrid, and privacy layers such as secure enclaves or differential privacy wrap the pipeline. Together, the unified loss and architecture diagram illustrate how system design choices can be framed within a coherent optimization perspective.

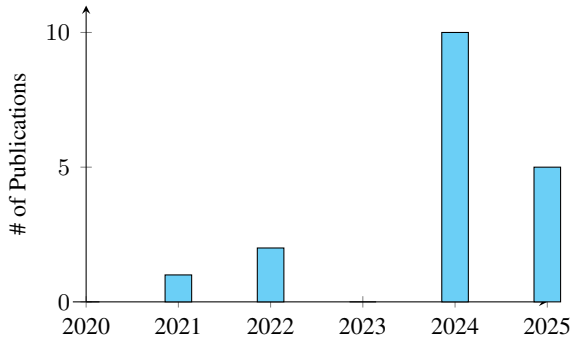

Figure 4: Publication trends on federated RAG from 2020 to 2025, showing a significant rise in research activity in recent years.

4 Research Activity and Publication Trends

Research in Federated Retrieval-Augmented Generation (RAG) has accelerated notably in recent years, driven by rising demand for privacy-preserving NLP and the maturation of retrieval-augmented systems. Before 2022, federated learning (FL) and RAG evolved independently, FL focused on decentralized optimization and privacy, while RAG addressed hallucinations in large language models (LLMs) (Lewis et al., 2020a). Their integration began in 2022–2023 with early architectural proposals such as *UniMS-RAG* (Wang et al., 2024a) and *FedE4RAG* (Mao et al., 2025), which outlined how to combine distributed training with retrieval-based augmentation.

Figure 4 shows a sharp rise in 2024, driven by high-stakes LLM use (healthcare/finance) and regulatory pressure. Industry (e.g., NVIDIA's C-FedRAG (Addison et al., 2024), Adobe's MKP-QA (Shojaee et al., 2025)) and open-source toolkits signal consolidation around benchmarks like MIRAGE (Xiong et al., 2024) and FeB4RAG.

Alongside this growth, the field is also beginning to consolidate around shared metrics and benchmarks, which provide clearer evaluation protocols and practical entry points for researchers and practitioners. Notable examples include the *MIRAGE* clinical QA benchmark (Xiong et al., 2024), the *FeB4RAG* federated BEIR benchmark (Wang et al., 2024b), the *RAGAS* evaluation toolkit for hallucination and consistency (Es et al., 2024), and the open-source *FedRAG toolkit* (Vector Institute). Together, these resources mark an inflection point: Federated RAG research is evolving from isolated system proposals to a field with standardized evaluation pipelines. A comprehensive summary of these toolkits and datasets is provided in Appendix 5.

4.1 Privacy & Security

Federated RAG must secure both retrieval and generation. *C-FedRAG* confines the entire pipeline to SGX enclaves, shielding queries and documents from untrusted servers. *FRAG* encrypts vector search with IND-CPA–secure homomorphic kNN, avoiding raw-index exposure, while *FedE4RAG* performs local augmentation at inference to minimise data movement. These techniques uphold regulations such as "right-to-be-forgotten" but add latency and can hamper cross-silo reasoning; future work should formalise threat models and quantify the privacy–utility frontier.

4.2 Model Adaptation & Knowledge Management

Maintaining knowledge without a central store is equally demanding. *FedE4RAG* trains dense retrievers federatively, preserving locality, while adapter-style updates and federated embedding learning cut communication. Hybrid planners like *RAGRoute* select local or remote answers when indices drift. Yet index synchronisation, conflict resolution, and continual adaptation across heterogeneous silos remain open. CRDT-based distributed indexing and meta-learned retriever personalisation are promising, but benchmarks must cover timestamped document shifts, bandwidth ceilings, and real-world constraints.

5 Applications

Federated RAG is particularly suited for domains where data sensitivity, personalization, and access to localized knowledge are paramount. In healthcare, it enables clinical question answering systems that draw on hospital-specific knowledge bases while ensuring that sensitive patient data remains on-premise, complying with strict privacy regulations. In financial services, it facilitates regulation-aware summarization and reporting by incorporating institution-specific documents and compliance guidelines without exposing confidential data externally. Education, enterprise knowledge management, and legal technology also present high-impact opportunities. In education, federated RAG can power personalized tutoring systems that align with local curricula and student learning profiles. Enterprise use cases include internal document retrieval and report generation without requiring the centralization of proprietary con-

tent. Legal applications range from case-specific document analysis to retrieval-augmented reasoning over firm-confidential corpora.

These examples illustrate the broad utility of federated RAG in supporting *privacy-preserving*, *context-aware*, and *domain-specific* language generation across distributed settings.

6 Challenges, Gaps and Trends

In this section, we discuss the overarching challenges, research gaps, and emerging trends for federated RAG. Table 4 summarizes how our research questions (RQ1–RQ4) align with specific challenge areas identified from the 18 primary studies.

Research Question	Key Insights / Challenge Areas
RQ1	System heterogeneity in data and clients; need for scalable federated RAG architectures.
RQ2	Privacy and security trade-offs; mechanisms for trust and compliance in federated setups.
RQ3	Retrieval and generation efficiency; balancing quality against resource costs.
RQ4	Evaluation gaps and future directions, such as benchmarks and domain-specific adaptations.

Table 4: Mapping of RQ1–RQ4 to key insights and challenge areas in federated RAG.

Design Patterns and Practical Guidelines

Beyond high-level challenges, recent studies also surface recurring *design patterns* that can guide implementers in practice:

- **Retrieval.** Systems employ either selective query routing (e.g., RAGRoute) to reduce redundant cross-silo traffic, or encrypted kNN search (e.g., FRAG) to ensure confidentiality at the cost of higher latency.

- **Generation.** Approaches range from centralized generation, which improves consistency but raises privacy concerns, to fully client-side generation, which preserves data sovereignty but increases coordination complexity. Hybrid designs (e.g., GPT-FedRec) combine global knowledge with local personalization.

- **Personalization.** Adapter tuning and federated embedding alignment (e.g., FedE4RAG) allow lightweight personalization without overwhelming communication budgets, making them attractive for heterogeneous client settings.

- **Trade-offs.** Every design choice involves balancing privacy, latency, and accuracy. For example, secure enclaves (C-FedRAG) provide strong protection but add overhead; selective routing reduces cost but may risk coverage. These trade-offs highlight the importance of aligning system design with domain-specific regulatory and operational requirements.

Taken together, these patterns provide actionable guidance: implementers can view Federated RAG architectures not as ad hoc designs but as configurable combinations of retrieval, generation, and personalization strategies, each with well-understood implications for scalability and compliance.

System Heterogeneity, Scalability, and Performance Trade-offs – Heterogeneous clients introduce latency, uneven resource use, and noisy retrieval results. To address this, recent work explores selective query routing (e.g., *RAGRoute*), benchmark-driven source selection strategies FEB4RAG, and encrypted caching with probabilistic gating *FRAG*. Together, these methods illustrate how different design choices balance recall, privacy, and latency, highlighting the inherent trade-offs in federated RAG. Overall, such advances directly target the architectural and efficiency concerns that must be solved for scalable deployment (RQ1).

Privacy and Security Considerations – Federated RAG must guarantee data confidentiality throughout both retrieval and generation, making privacy a core concern (RQ2). Traditional RAG pipelines risk information leakage through exposed queries or retrieved documents. Recent systems address this through trusted execution and encryption. *C-FedRAG*, for example, runs indexing and generation entirely inside secure enclaves, protecting sensitive content even when orchestrated by untrusted servers. *FRAG* further strengthens guarantees via homomorphic encryption, enabling encrypted vector search without exposing embeddings or indices—achieving IND-CPA security with acceptable overhead. Compliance with data governance (e.g., right-to-be-forgotten) motivates designs that avoid raw data centralization. *FedE4RAG* exemplifies this by performing local augmentation at inference, ensuring data minimization. However,

privacy-preserving techniques often introduce latency and limit cross-silo reasoning, raising the challenge of balancing security with system utility. Future work must formalize threat models and quantify trade-offs between confidentiality, accuracy, and responsiveness in real-world deployments.

Architectural Maturity – Retrieving and updating knowledge without centralized storage is a central challenge for federated RAG systems (RQ3). *FedE4RAG* proposes client-side retriever training via federated learning, improving generalization while respecting data locality. To mitigate resource constraints, current approaches favor parameter-efficient methods such as adapter tuning or federated embedding learning. Hybrid pipelines like *RAGRoute* arbitrate between federated and RAG responses based on confidence, improving robustness when document indices are stale. However, synchronizing indices, resolving data conflicts, and enabling continual adaptation across diverse silos remain open problems. For domain-specific systems, e.g., hospital QA linked to EHR silos, federated RAG must reconcile personalization with evolving knowledge, potentially through distributed indexing, meta-learning, and lightweight model updates. Future work should explore CRDT-based index synchronization and meta-learned retriever personalization to enable continual, privacy-preserving adaptation across heterogeneous silos. Benchmarking these techniques under realistic, timestamped document shifts and resource constraints remains an open evaluation priority.

Evaluation and Benchmarking – Federated RAG is a new topic, hence evaluation frameworks (RQ4) are lacking. Early studies reused QA criteria or imitated federated environments, making comparisons impossible. Federated retrieval and creation require standardized datasets and metrics to account for client relevance, privacy constraints, and network costs. This gap is being filled by recent work: *FeB4RAG* provides a dataset for federated RAG search, allowing for realistic multi-silo evaluation of retrieval algorithms. FedRAG may leverage the MIRAGE benchmark (used by RAGRoute) and a new MIMIC-IV-based clinical QA dataset to evaluate systems in certain areas. Still, evaluating issues exist. Relevance across clients requires reinterpreting IR measurements like recall or MRR. Current measurements make it difficult to measure a system's privacy-preserving quality (utility attained for a given level of privacy). Real-world deployment studies are rare; most are lab-based. Common evaluation protocols, such as federated versions of popular QA tasks and defined success criteria (accuracy vs. bandwidth vs. privacy), will advance the industry. Community-driven toolkits (e.g., Flower FedRAG demonstrations) and open-source frameworks are lowering experimental barriers, which are beneficial. Future work should introduce privacy–utility Pareto benchmarks that vary attacker models and bandwidth ceilings, alongside live-stream leaderboards that track adaptation latency and cost as documents drift over time.

6.1 Outlook and Future Opportunities

Federated RAG research is rapidly evolving, addressing challenges in privacy, scalability, and personalization. Key opportunities include reducing cross-silo latency via neural routing and CRDT-based index synchronization, strengthening privacy through Pareto benchmarks that jointly track attacker models and bandwidth ceilings, and enabling continual adaptation with meta-learned retriever personalization and tiny-ML adapters scheduled to respect heterogeneous devices. Lifelong evaluation resources such as live-stream leaderboards and federated versions of popular QA tasks will clarify trade-offs among accuracy, privacy, and cost as documents drift. Domain-specific optimizations for healthcare, finance, and multi-enterprise settings remain crucial, and community toolkits are lowering experimental barriers. Overall, federated RAG offers a compelling path toward trustworthy, knowledge-grounded AI that operates securely in decentralized environments. Future research could empirically investigate and quantify these identified architectural trade-offs across diverse real-world federated contexts, thereby enabling more precise architectural recommendations tailored explicitly to domain-specific operational constraints.

7 Conclusion

Federated Retrieval-Augmented Generation (RAG) has rapidly moved from concept to practice by pairing privacy-preserving training with knowledge-grounded generation. Our systematic mapping of 18 primary studies charts this evolution, classifying architectures, contributions, and domains, and highlighting a 2024 research surge driven by large-scale language-model adoption and regulatory pressure for confidential data handling.

Where it already helps – Early deployments

demonstrate clear value in compliance-sensitive settings: *C-FedRAG* answers clinical questions while keeping electronic health-record data in-hospital; *FedE4RAG* powers enterprise knowledge assistants without centralizing proprietary documents; and *RAGRoute* cuts 75% of redundant cross-silo queries, lowering latency for finance workloads. Across these use cases, federated RAG improves factual grounding, reduces hallucinations, and upholds data-sovereignty requirements that vanilla RAG or centralized LLM serving cannot meet.

What has been solved and what has not – Secure enclaves, homomorphic encryption, and differential-privacy mechanisms now offer end-to-end confidentiality; neural routing and probabilistic caching tame bandwidth costs; and federated embedding learning supplies parameter-efficient personalization. Yet four hurdles persist:

- *Scalable, CRDT-style index synchronization to keep silos consistent*,
- *Meta-learned retriever adaptation that survives concept drift*,
- *Privacy–utility evaluation protocols that expose trade-offs under realistic attacker models*, and
- *Live benchmarks that measure cost and latency as documents evolve*.

Why this mapping study matters –Federated Retrieval-Augmented Generation has rapidly progressed from early conceptual proposals to a diverse body of system designs, benchmarks, and evaluation resources. Our systematic mapping of 18 primary studies charts this trajectory, identifying common architectural patterns, key challenges, and emerging opportunities across domains such as healthcare, finance, and enterprise knowledge management.

In moving beyond description, this paper evolves from a static mapping to a *field-shaping foundation*: it combines structured analysis with empirical metrics, introduces theoretical perspectives through a unified objective function, and distills implementation guidance into actionable design patterns. Together, these contributions provide both researchers and practitioners with a clearer roadmap for advancing Federated RAG toward scalable, trustworthy, and domain-ready deployment.

Limitations

As with any systematic mapping study, our work is subject to certain limitations that should be acknowledged when interpreting its findings.

External Validity A precisely defined corpus of 18 main papers from 2020–2025 was used to consolidate Federated RAG research trends and conclusions. We specifically reviewed retrieval-augmented generation in federated or privacy-preserving systems across NLP, IR, ML, and security to assure external validity. We used backward reference tracing and cross-domain search tactics, although relevant research, especially those in proprietary industries or less often indexed venues, may have been missed. Current research trends are mostly from Western and East Asian academic and corporate ecosystems and may not fully represent worldwide research efforts.

Construct Validity Federated RAG is new and its borders are changing across communities. Despite following a tight inclusion process and Kitchenham's mapping study principles, different definitions of a Federated RAG system (e.g., static knowledge bases vs. dynamic retrieval) may generate conceptual diversity. Federated learning system publications sometimes used retrieval or augmentation processes without mentioning RAG models. We chose inclusivity, yet conceptual ambiguity may affect our categories' completeness or framing. The focus on peer-reviewed and preprint literature ignored grey literature, technical documentation, and non-archival systems like deployed prototypes.

Conclusion Validity Our results stem from a small set of studies, so generalizations should be cautious; the 2024 surge may be a short-term spike. Evolving definitions and tools mean our taxonomy is a snapshot of a rapidly changing landscape.

Other Limitations Despite a comprehensive search, some studies using unconventional terms or industry-only contributions may have been missed, limiting reproducibility.

Acknowledgements

We thank the Complex Data Analysis and Reasoning Lab at Arizona State University for computational support, the anonymous reviewers for their thoughtful feedback, and our lab cat, Coco, for ensuring our professor maintained just the right mix of sanity and chaos during deadlines.

References

Parker Addison, Minh-Tuan H. Nguyen, Tomislav Medan, Jinali Shah, Mohammad T. Manzari, Brendan McElrone, Laksh Lalwani, Aboli More, Smita Sharma, Holger R. Roth, Isaac Yang, et al. 2024. C-FedRAG: A confidential federated retrieval-augmented generation system.

Keith Bonawitz, Vladimir Ivanov, Benjamin Kreuter, Alessandro Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1175–1191.

Tom B Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901.

Tonmoy Debnath, Md Nurul Absar Siddiky, Muhammad Enayetur Rahman, Prosenjit Das, and Antu Kumar Guha. 2025. A comprehensive survey of prompt engineering techniques in large language models. *TechRxiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Rachid Guerraoui, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, and Martijn de Vos. 2025. Efficient federated search for retrieval-augmented generation. In *Proceedings of the 5th Workshop on Machine Learning and Systems*, pages 74–81.

Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. 2022. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pages 691–707. Springer.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Beomyeol Jeon, S M Ferdous, Muntasir Raihan Rahman, and Anwar Walid. 2021. Privacy-preserving decentralized aggregation for federated learning. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6.

Emily Jiang. 2024. *Clinical Question-Answering over Distributed EHR Data*. Ph.D. thesis, Massachusetts Institute of Technology.

Jincheol Jung, Hongju Jeong, and Eui-Nam Huh. 2025. Federated learning and rag integration: A scalable approach for medical large language models. *Preprint*, arXiv:2412.13720.

Peter Kairouz, H. Brendan McMahan, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210.

Barbara Kitchenham, David Budgen, and Pearl Brereton. 2011. Mapping study methodology. *Empirical Software Engineering*, 16(6):791–810.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Yuxiang Ku, Wen-tau Chen, Guillaume Bouchard, Douwe Kiela, et al. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Ilia Kulikov, Marjan Ghazvininejad, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 9459–9474.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, et al. 2021. Pre-train prompt tune: Towards generalizing to unseen tasks. In *Proceedings of ACL-IJCNLP*.

Qianren Mao, Qili Zhang, Hanwen Hao, Zhentao Han, Runhua Xu, Weifeng Jiang, Qi Hu, Zhijun Chen, Tyler Zhou, Bo Li, Yangqiu Song, Jin Dong, Jianxin Li, and Philip S. Yu. 2025. Privacy-preserving federated embedding learning for localized retrieval-augmented generation. *Preprint*, arXiv:2504.19101.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hao Peng, Haoran Li, Yangqiu Song, Vincent Zheng, and Jianxin Li. 2021. Differentially private federated knowledge graphs embedding. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1416–1425.

Alec Radford, Jeffrey Wu, Rewon Child, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of NAACL-HLT*.

Parshin Shojaee, Sai Sree Harsha, Dan Luo, Akash Maharaj, Tong Yu, and Yunyao Li. 2025. Federated retrieval augmented generation for multi-product question answering. *Preprint*, arXiv:2501.14998.

Milad Shokouhi and Luo Si. 2011. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *Preprint*, arXiv:2401.13256.

Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, and Guido Zuccon. 2024b. Feb4rag: Evaluating federated search in the context of retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 763–773.

Yebo Wu, Chunlin Tian, Jingguang Li, He Sun, Kahou Tam, Li Li, and Chengzhong Xu. 2025. A survey on federated fine-tuning of large language models. *arXiv preprint arXiv:2503.12016*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Runhua Xu, Nathalie Baracaldo, Yi Zhou, Ali Anwar, Swanand Kadhe, and Heiko Ludwig. 2022. Detrust-fl: Privacy-preserving federated learning in decentralized trust setting. In *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*, pages 417–426. IEEE.

Huimin Zeng, Zhenrui Yue, Qian Jiang, and Dong Wang. 2024. Federated recommendation via hybrid retrieval augmented generation. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8078–8087. IEEE.

Dongfang Zhao. 2024. FRAG: Toward federated vector database management for collaborative and secure retrieval-augmented generation. *arXiv preprint arXiv:2410.13272*.

A Extended Resources and Comparative Synthesis

A.1 Benchmarks and Toolkits

To complement the taxonomy and trends discussed in the main paper, we provide expanded descriptions of benchmarks and toolkits that now support Federated RAG evaluation and deployment:

- **C-FedRAG** (Addison et al., 2024): A confidential QA system executed entirely in Trusted Execution Environments (SGX), demonstrating strong privacy guarantees in clinical domains.
- **RAGRoute** (Guerraoui et al., 2025): A query routing framework that reduces redundant cross-silo queries by up to 75% while maintaining accuracy.
- **FRAG** (Zhao, 2024): A homomorphic-encryption-based kNN retrieval system offering IND-CPA security with latency within 3–5$\times$ of plaintext baselines.
- **GPT-FedRec** (Zeng et al., 2024): A hybrid recommendation model combining collaborative filtering with RAG, achieving +5–7% personalization hit-rate.
- **FedE4RAG** (Mao et al., 2025): A modular federated dense retriever framework with homomorphic encryption and distillation for end-to-end privacy.
- **UniMS-RAG** (Wang et al., 2024a): A unified multi-source RAG system integrating client-side retrieval with a shared generator for adaptive latency control.
- **MIRAGE** (Xiong et al., 2024): A clinical QA benchmark designed for federated hospital data, formalizing evaluation in healthcare contexts.
- **MKP-QA** (Shojaee et al., 2025): A benchmark for enterprise multilingual QA across product silos, supporting federated evaluation in industry.
- **FeB4RAG** (Wang et al., 2024b): A federated benchmark spanning 16 BEIR-derived domains to study routing strategies and retrieval efficiency.
- **RAGAS** (Es et al., 2024): An evaluation toolkit introducing hallucination, factuality, and robustness metrics tailored to federated RAG pipelines.
- **FedRAG** (Addison et al., 2024): A distributed EHR QA system that introduces hierarchical retrieval across hospitals while preserving patient privacy.

- **RAG** (Wang et al., 2024a): A medical large language model integrated with RAG, demonstrating gains over FL-only baselines.
- **DP-FedKGE** (Peng et al., 2021): A framework for differentially private federated knowledge graph embeddings, enabling reasoning with entity-level privacy.
- **PPDA** (Jeon et al., 2021): A lightweight decentralized aggregation protocol that ensures privacy without relying on DP or HE.
- **DeTrustFL** (Xu et al., 2022): A cryptographic trust consensus mechanism preventing disaggregation attacks in federated setups.
- **FedX** (Han et al., 2022): An unsupervised federated representation learning system using cross-client knowledge distillation.
- **FL–LLM surveys** (Wu et al., 2025): A survey of federated LLM fine-tuning methods, with explicit discussion of RAG integration.
- **FRAD (Zhao)** (Zhao, 2024): A general-purpose federated architecture demonstrating cryptographic retrieval under multi-domain constraints.

Together, these resources provide the beginnings of a shared evaluation layer, enabling more consistent comparisons across systems and domains.

A.2 Full Comparative Synthesis of Surveyed Studies

Table 5 extends the synthesis presented in Section 3.1 by aligning each of the 18 Federated RAG studies (2020–2025) with their reported benefits and outcomes. This consolidated view highlights where empirical improvements have been quantified, where cryptographic and privacy guarantees dominate, and where new benchmarks or toolkits have been introduced.

B Comparative Synthesis of Federated Studies

In addition to the qualitative taxonomy presented in Section3.1, we provide a consolidated comparative synthesis of all 18 Federated RAG studies surveyed between 2020–2025 (Table5). This extended table aligns each system with its reported benefit and empirical or conceptual outcome, thereby offering a more granular view of how diverse contributions map to performance, privacy, efficiency, and evaluation. For instance, system-level advances such as C-FedRAG and RAGRoute quantify measurable improvements in accuracy and communication efficiency, while cryptographic frameworks like FRAG demonstrate the latency trade-offs inherent in strong privacy guarantees. Benchmarking and evaluation initiatives, including FeB4RAG, MIRAGE, and RAGAS, highlight the field's growing emphasis on standardized protocols, whereas systems like FedE4RAG and UniMS-RAG focus on architectural integration. Early foundational efforts (e.g., DP-FedKGE, PPDA, DeTrustFL) continue to shape privacy-preserving mechanisms, while recent domain-specific benchmarks (e.g., MKP-QA) expand application horizons. Taken together, this synthesis complements the main text by grounding the survey's classifications in concrete evidence and by illustrating the breadth of technical strategies underpinning Federated RAG research.

System + Reported Benefit	Outcome / Contribution
C-FedRAG (Addison et al., 2024)	
Clinical QA in trusted enclaves	+12.7% QA accuracy (59.8→72.5) using SGX
RAGRoute (Guerraoui et al., 2025)	
Adaptive query routing across silos	–75% redundant queries with 72% accuracy
FRAG (Zhao, 2024)	
Encrypted kNN retrieval (HE-based)	3–5× latency compared to plaintext
GPT-FedRec (Zeng et al., 2024)	
Personalized recommendation	+5–7% hit rate improvement
FedE4RAG (Mao et al., 2025)	
Federated dense retrievers + HE distillation	End-to-end privacy-preserving retrievers
UniMS-RAG (Wang et al., 2024a)	
Multi-source retrieval + shared generator	Unified retrieval-control, adaptive latency
MIRAGE (Xiong et al., 2024)	
Clinical QA benchmark	First federated benchmark for clinical QA
MKP-QA (Shojaee et al., 2025)	
Enterprise multilingual QA benchmark	Cross-silo multilingual evaluation
FeB4RAG (Wang et al., 2024b)	
Federated BEIR benchmark	Evaluation across 16 BEIR-derived domains
RAGAS (Es et al., 2024)	
Evaluation toolkit for hallucination/consistency	Metrics for hallucination, evidence quality, robustness
FedRAG (Addison et al., 2024)	
Distributed EHR QA	Hierarchical hospital retrieval pipeline
RAG (Wang et al., 2024a)	
Medical LLM + RAG integration	Outperforms FL-only baselines on clinical QA
DP-FedKGE (Peng et al., 2021)	
Differentially private embeddings	Entity-level DP protection for KG embeddings
PPDA (Jeon et al., 2021)	
Lightweight decentralized aggregation	Privacy without DP/HE overhead
DeTrustFL (Xu et al., 2022)	
Cryptographic trust consensus	Prevents disaggregation attacks in FL
FedX (Han et al., 2022)	
Cross-client knowledge distillation	Unsupervised federated representation learning
FL–LLM surveys (Wu et al., 2025)	
Survey of FL-LLM fine-tuning	Overview, references to RAG methods
FRAD (Zhao) (Zhao, 2024)	
General multi-domain architecture	Cryptographic constraints, scalable retrieval

Table 5: Comparative synthesis of 18 Federated RAG studies (2020–2025). Each system (bold, first line) is paired with its reported benefit (indented line) and the corresponding outcome or contribution.