ISSN: 2581-3404 (Online) IF: 5.68 (SJIF)
IJIRTM, Volume-9, Issue-2, April-2025

Fed RAG-Priv: A Privacy-Preserving Federated Retrieval-Augmented Generation Framework with Adaptive Trust-Weighted Routing and Semantic Fusion

Navneet Kaur[1], Aryan[2], Tirth[3], Ashish Jain[4]
Department of CSE (AI&ML)[1,2,3,4]
Panipat Institute of Engineering and Technology, Samalkha, Panipat, India[1,2,3,4]
navneetzz663@gmail.com[1], aryanbansal837@gmail.com[2], tirthnarwal10@gmail.com[3], ashishjain.aj2003@gmail.com[4]

Abstract. *Retrieval-Augmented Generation (RAG) grounds large language model (LLM) outputs in external documents, but standard RAG assumes a single, centrally hosted corpus, which conflicts with how sensitive knowledge is held by hospitals, financial institutions, and other regulated organizations. Existing Federated RAG proposals decentralize storage but still broadcast every query to every participant and aggregate retrieved evidence with static, uniform weighting, which wastes communication at scale and lets a single low-quality or adversarial organization influence the fused context as much as a reliable one. This paper proposes FedRAG-Priv, built around Adaptive Trust-Weighted Routing and Semantic Fusion (ATR-SF): a coordinator routes each query only to organizations whose learned trust score and topical relevance exceed a threshold, weights each organization's contribution to secure aggregation by that score, and replaces plain top-k merging with similarity-based clustering that removes redundant evidence before it reaches the generator. We formalize routing, trust update, privacy-preserving aggregation, and semantic fusion mathematically, provides an end-to-end algorithm, and derives communication and computational complexity with and without adaptive routing. Because this is a proposed architecture rather than an implementation, the evaluation is analytical: it compares FedRAG-Priv against centralized RAG and five recent Federated RAG systems on privacy composition, communication cost, scalability, and robustness to unreliable participants, using established federated-learning and differential-privacy metrics rather than fabricated experimental results.*

Keywords: Federated Learning; Retrieval-Augmented Generation; Large Language Models; Privacy-Preserving AI; Secure Aggregation; Differential Privacy; Adaptive Routing; Trust-Aware Aggregation; Semantic Fusion.

Introduction

Large language models (LLMs) produce fluent text, but their parametric knowledge is fixed at training time and prone to hallucination when queried on facts absent from, or misrepresented in, the training corpus [1]. Retrieval-Augmented Generation (RAG) addresses this by conditioning generation on passages retrieved from an external corpus at inference time [1]. Standard RAG assumes a single, centrally hosted knowledge base, in which documents from every source are collected, embedded, and indexed in one vector store queried directly by the retriever.

This centralization is difficult to reconcile with how knowledge is actually held and governed. Hospitals, banks, law firms, and government agencies each maintain domain knowledge that is regulated, commercially sensitive, or jurisdictionally restricted, and none can lawfully or practically transfer raw

records into a shared repository. Centralized RAG therefore either restricts itself to public corpora, forfeiting the specialized knowledge such organizations hold, or requires data pooling that conflicts with regimes such as GDPR and HIPAA.

Federated Learning (FL) addresses an analogous problem for model training: clients compute updates on local data and share only parameters or gradients with a coordinator, which aggregates them without observing raw data [2]. Communication-efficient variants reduce the bandwidth this requires [3], and secure-aggregation protocols give cryptographic guarantees that the coordinator cannot recover any single client's contribution [4]. Extending these principles to retrieval is natural, but the settings differ: a federated retrieval system must protect embeddings, similarity scores, and retrieved content on a per-query, real-time basis rather than over accumulated training rounds.

Recent work explores this direction from several angles: federated embedding learning for local RAG [6], confidential federated RAG using trusted execution [7], federated search evaluation [8], federated vector-database management [9], and blockchain-based admission control over a distributed corpus [10]. In each of these systems, however, a query is broadcast to every participating organization, and the returned evidence is merged with equal weight regardless of an organization's historical retrieval quality or the query's relevance to its corpus. This is inefficient once the number of organizations is large, since most queries are topically relevant to only a minority of participants, and it is fragile: an organization with stale or manipulated content contributes to the fused context on equal footing with a reliable one, and extraction attacks such as [15] show this influence can be exploited. No existing Federated RAG proposal routes queries adaptively or weights aggregation by participant reliability under a formal privacy budget; this specific problem motivates the framework proposed here.

The contributions of this paper are:

1) Adaptive Trust-Weighted Routing and Semantic Fusion (ATR-SF): a mechanism that routes each query to a dynamically selected subset of organizations based on a trust score updated from aggregate-level agreement signals, weights each organization's contribution to secure aggregation by that score, and replaces uniform top-k merging with similarity-clustered fusion that removes redundant evidence.

2) A privacy-preserving retrieval pipeline in which raw documents never leave their originating organization and only differentially private embedding’s, scores, and trust signals cross organizational boundaries.

3) A mathematical formulation of routing, trust update, DP perturbation, secure aggregation, and semantic fusion, with communication and computational complexity expressed as a function of the routed subset size rather than the full participant count.

4) An end-to-end algorithm integrating these components into a single query-processing pipeline.

5) An analytical, metric-based comparison of FedRAG-Priv against centralized RAG and five recent Federated RAG systems on privacy composition, communication cost, scalability, and robustness to unreliable participants.

Related Work

A. Retrieval-Augmented Generation

RAG couples a dense retriever with a generator conditioned on retrieved passages [1]. Subsequent work improves retriever quality, ranking, and context construction, generally assuming a single centralized index queried directly by the generator. This assumption simplifies system design but concentrates all source documents, and hence all sensitive content, at one operator — the property this paper's framework is designed to remove.

B. Federated Learning

FL trains a shared model across clients that retain data locally, exchanging only parameter updates with a coordinator [2]. Communication-efficient variants reduce exchange bandwidth [3], and secure aggregation protocols, including lightweight designs such as LightSecAgg [19], ensure the coordinator learns only the sum of client updates, never an individual contribution [4]. Deployment frameworks such as NVIDIA FLARE [20] operationalize these protocols. FL supplies the coordination pattern this paper adapts to retrieval, but it protects training updates, not retrieved evidence, and provides no per-client reliability weighting.

C. Privacy-Preserving RAG

A separate line of work adds privacy guarantees to RAG without federating across organizations. Differential privacy has been applied to retrieved context [12] and to decoding itself [23], [24], while entity perturbation [21] and random projection [22] reduce leakage of identifiable content from retrieved passages. Membership-inference and extraction attacks specific to RAG [14], [15] motivate this line of work, and a broader survey documents privacy failure modes across the RAG pipeline [13]. These techniques strengthen single-operator RAG but do not address a knowledge base distributed across mutually distrusting organizations.

D. Federated RAG

The most closely related work federates retrieval directly. Federated embedding learning has been proposed for localized RAG [6]; confidential federated RAG combines trusted execution with federation [7]; federated search has been evaluated in a RAG context [8]; vector-database-level federation has been proposed for collaborative retrieval [9]; and blockchain-based admission control protects a decentralized corpus from data poisoning [10]. Isolation-based retrieval [11] and hybrid federated recommendation with RAG [17] extend federation to adjacent settings. A recent systematic mapping study catalogs this literature and finds it fragmented across architectural patterns, with query broadcasting and uniform aggregation as the de facto default in every surveyed design [16].

Research gap: existing Federated RAG systems secure individual pipeline stages — embedding generation [6], the execution environment [7], the search layer [8], the vector store [9], or data admission [10] — but uniformly broadcast queries to all participants and aggregate retrieved evidence with static, unweighted merging. None adaptively restrict routing to relevant, trustworthy participants, none weight aggregation by historical reliability, and none report how communication and computation scale with the number of organizations actually contacted per query rather than the total participant count. This leaves Federated RAG more expensive and more susceptible to low-quality or adversarial participants than necessary at scale. FedRAG-Priv addresses this specific problem through Adaptive Trust-Weighted Routing and Semantic Fusion (ATR-SF).

Proposed Federated Rag Framework

A. Overview

FedRAG-Priv connects N participating organizations O1..ON, each maintaining a local knowledge base and local vector index, through a Federated Coordinator that never stores raw documents or unprotected embeddings. The coordinator's Adaptive Routing Module and Trust Score Store select which organizations are contacted per query, its Trust-Weighted Secure Aggregation combines their perturbed results, and its Semantic Context Fusion module removes redundant evidence before generation. Table II lists each component and its role.

Table 1: Architecture component description.

Component	Role
User Query	Natural-language input submitted by the end user.
Federated Coordinator	Hosts routing, aggregation, and fusion; never stores raw documents or unprotected embeddings.
Adaptive Routing Module	Selects a query-relevant, trustworthy subset R(q) of organizations to contact.
Trust Score Store	Maintains per-organization exponential-moving-average trust scores Ti.
Organization Oi	Participating entity holding a private local knowledge base.
Local Knowledge Base	Domain documents held exclusively by one organization.
Embedding Model E(·)	Shared, frozen model mapping text to vectors; run locally by each party.
Vector Database	Local index of document embeddings maintained per organization.
Top-k Retriever	Selects each organization's k most similar local documents.
Privacy Layer	Applies calibrated DP noise to embeddings/scores/trust signals before release.
Trust-Weighted Secure Aggregation	Combines perturbed scores weighted by trust without exposing individual values.
Semantic Context Fusion	Clusters candidate passages by similarity and selects representatives to remove redundancy.
LLM	Generates the final response conditioned on the fused context.

B. Query Processing and Retrieval

A user submits query q to the Federated Coordinator, which embeds q with a shared, frozen embedding model E(·) (Eq. 1). Each organization Oi that is subsequently routed the query (Section III-C) independently embeds its local documents Di = {di1,…,dim} with the same model (Eq. 2) and computes cosine similarity between the query embedding and each local document embedding (Eq. 3), returning only its local top-k candidates and scores (Eq. 4); raw documents remain local, and passage text is released only for candidates ultimately selected for generation.

C. Trust Score Model and Adaptive Routing

Each organization Oi carries a trust score Ti ∈ [0,1], initialized uniformly and updated after every query in which it participates. The update uses an aggregate-level agreement signal qi(t) — how closely Oi's perturbed scores agree with the final fused ranking, computed only from already-perturbed, already-aggregated values, never from raw content (Eq. 6) — combined through an exponential moving average with rate α (Eq. 7). Given q, the coordinator computes a lightweight relevance prior ri(q) from a cached, infrequently updated per-organization topic-affinity vector, and routes the query only to R(q) = {i : Ti ≥ τtrust, ri(q) ≥ τrel} (Eq. 8). This contacts an expected N' = |R(q)| ≤ N organizations per query instead of broadcasting to all N, and it excludes organizations whose recent behavior deviates from consensus, bounding the influence any single unreliable or adversarial participant can have on later queries.

D. Privacy Layer

Before leaving an organization, each retrieved score is perturbed with calibrated noise to satisfy local differential privacy with budget εretrieval (Eq. 5), following the composition results of [5]. The same mechanism perturbs the agreement signal qi(t) before it updates Ti, under a separate budget εtrust, so trust scores themselves carry a formal privacy guarantee; the two budgets compose as εtotal = εretrieval + εtrust by sequential composition [5].

E. Trust-Weighted Secure Aggregation

Rather than the unweighted sum used in a baseline federated retriever, FedRAG-Priv aggregates perturbed scores from the routed subset R(q) with weights proportional to trust (Eq. 9), computed inside the same secure-aggregation protocol so the coordinator observes only the weighted sum, never an individual Ti or Si' [4], [19] (Eq. 10). This limits the influence of a low-trust organization on the fused ranking without revealing which organization was down-weighted.

F. Semantic Context Fusion and Generation

The weighted aggregate is expanded to its associated passages, which are clustered by pairwise cosine similarity with threshold θ; the highest-scoring passage in each cluster is retained as its representative, and clusters are added to context C in score order until the token budget Lmax is reached (Eq. 11). This removes near-duplicate evidence returned by multiple organizations before it consumes context budget, which a plain top-K union does not do. The fused context C is concatenated with q and passed to the LLM, which generates response r (Eq. 12); only C and r are exposed to the user, and per-organization embeddings are discarded once aggregation completes.

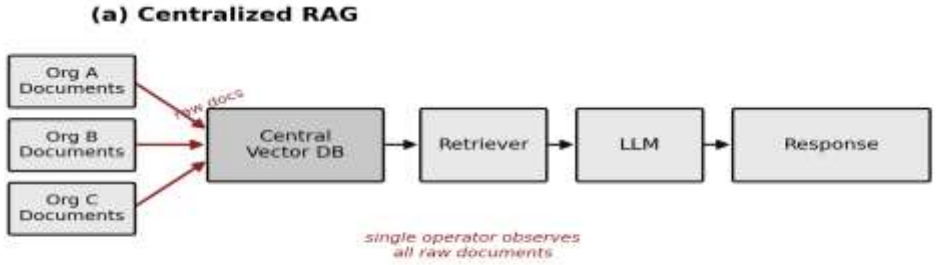

Figure 1 (a): Motivation — Centralized RAG: a single operator observes all raw documents.

Figure 1 (b): Federated RAG with ATR-SF: only privacy-transformed scores cross organizational boundaries.

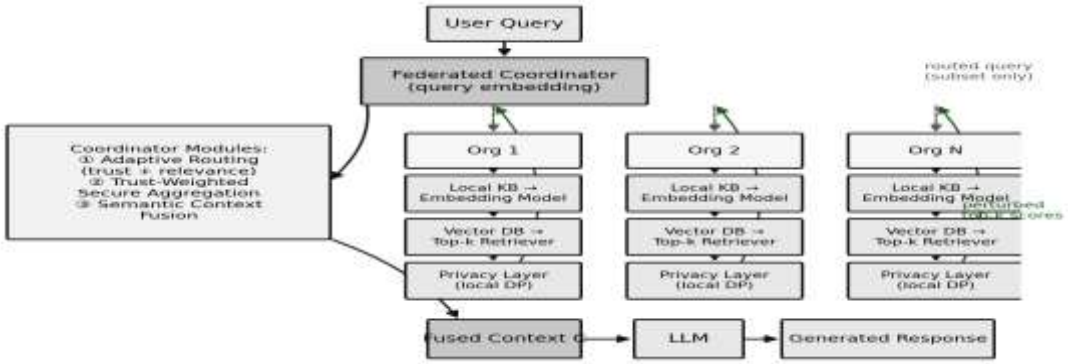

Figure 2: Complete proposed Federated RAG architecture, showing the coordinator's routing, aggregation, and fusion modules alongside per-organization retrieval pipelines.

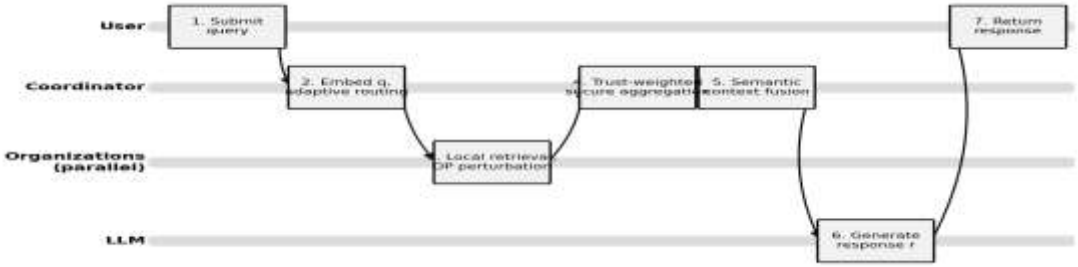

Figure 3: End-to-end query processing workflow across user, coordinator, organizations, and LLM.

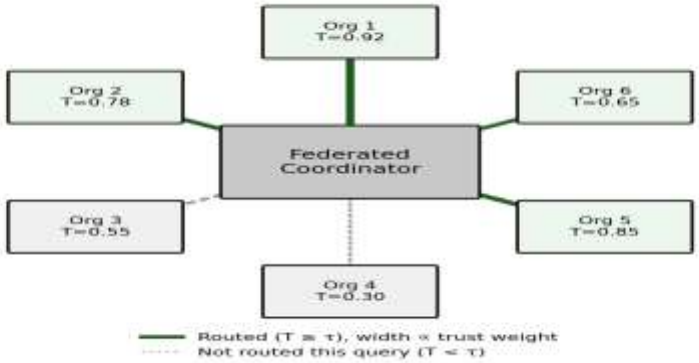

Figure 4: Trust-weighted communication topology: line width encodes trust weight; low-trust organizations are excluded from routing for a given query (dotted).

G. Mathematical Formulation

Let e_q denote the query embedding and e_{dij} the embedding of document j at organization i; let $T_i^{(t)}$ denote organization i's trust score prior to round t.

$$e_q = E(q) \quad (1)$$

$$e_{d_{ij}} = E(d_{ij}),\ j = 1,\ldots,m_i \quad (2)$$

$$s_{ij} = (e_q \cdot e_{d_{ij}}) / (\|e_q\| \|e_{d_{ij}}\|) \quad (3)$$

$$S_i = \text{top-k}\{ s_{ij} : j = 1,\ldots,m_i \} \quad (4)$$

$S_i' = S_i + Lap(\Delta s / \varepsilon_{retrieval})$ (5)
$q_i^{(t)} = agree(S_i', S_{agg}^{(t-1)})$ (6)
$T_i^{(t+1)} = (1-\alpha)\, T_i^{(t)} + \alpha\,(q_i^{(t)} + Lap(\Delta q/\varepsilon_{trust}))$ (7)
$R(q) = \{ i : T_i \geq \tau_{trust}, r_i(q) \geq \tau_{rel} \}, \; N' = |R(q)|$ (8)
$w_i = T_i / \Sigma_{j\in R(q)} T_j$ (9)
$S_{agg} = SecAgg(\{w_i S_i' : i\in R(q)\}) = \Sigma_{i\in R(q)} w_i S_i'$ (10)
$C = \bigcup_m rep(cluster_m), \; sim(cluster_m) \geq \theta, \; |C| \leq L_{max}$ (11)
$r = LLM(q, C)$ (12)
$Comm = O(N'\cdot(d + k(d+\ell)))$ (13)
$Comp = O(\Sigma_{i\in R(q)} m_i\cdot d + N'\cdot k\cdot d + K^2\cdot d)$ (14)

where d is embedding dimension, ℓ is average passage length, $K = N'k$ is the candidate-passage count before clustering, $Lap(\cdot)$ is Laplace noise calibrated to the stated sensitivity and budget, and $SecAgg(\cdot)$ is secure aggregation as in [4], [19].

H. Algorithm

Algorithm 1: FedRAG-Priv Query Processing (ATR-SF)

Input: query q; organizations O1..ON, trust {Ti}, corpora Di; model E(·); thresholds τtrust, τrel; budgets εretrieval, εtrust; k, θ, Lmax
Output: response r; updated trust scores {Ti}
1: e_q ← E(q)
2: compute ri(q) for all i from cached topic-affinity vectors
3: R(q) ← {i : Ti ≥ τtrust and ri(q) ≥ τrel}
4: broadcast e_q to organizations in R(q) only
5: for each Oi in R(q) in parallel do
6: for each document dij in Di: compute e_dij, sij
7: Si ← top-k documents by sij
8: Si' ← Si + DP noise (budget εretrieval)
9: send Si' to Federated Coordinator
10: end for
11: wi ← Ti / Σ_{j∈R(q)} Tj for i ∈ R(q)
12: S_agg ← SecAgg({wi · Si' : i ∈ R(q)})
13: cluster candidate passages by similarity ≥ θ
14: C ← highest-scoring representative per cluster, |C| ≤ Lmax
15: r ← LLM(q, C)
16: for each Oi in R(q): qi ← agree(Si', S_agg); Ti ← (1−α)Ti + α(qi + DP noise(εtrust))
17: discard per-organization embeddings
18: return r, {Ti}

Expected Performance Evaluation and Comparative Analysis

Since FedRAG-Priv is proposed at the architectural level, this section reports analytical expectations derived from the properties of its components, not measured results from an implementation. All numerical values in Table IV and Fig. 5 are analytical estimates based on architectural characteristics, not experimental measurements.

A. Privacy Analysis

Total privacy expenditure per query is εtotal = εretrieval + εtrust by sequential composition [5], since the retrieval perturbation (Eq. 5) and the trust-update perturbation (Eq. 7) are disjoint releases from the same organization. The coordinator observes only the weighted aggregate (Eq. 10), the fused context, and per-organization trust scores — never raw scores or documents — which bounds membership-inference advantage under [5] and further limits extraction and manipulation attacks such as [14], [15], since a manipulated organization's influence is capped by its trust weight rather than treated equally with reliable participants.

B. Communication Overhead

A baseline federated retriever broadcasts to all N organizations, giving communication $O(N(d+k(d+\ell)))$. ATR-SF instead contacts only $N' = |R(q)|$ organizations (Eq. 8), giving $Comm = O(N'(d+k(d+\ell)))$ (Eq. 13). Because most queries are topically relevant to a minority of participants in a heterogeneous multi-organization deployment, N' is expected to scale sub-linearly with N; Fig. 5 illustrates the analytical contact volume at a representative routing fraction $N' \approx 0.3N$, corresponding to an estimated 70% reduction in per-query communication relative to full broadcast — an estimate based on routing-set characteristics, not a measured value.

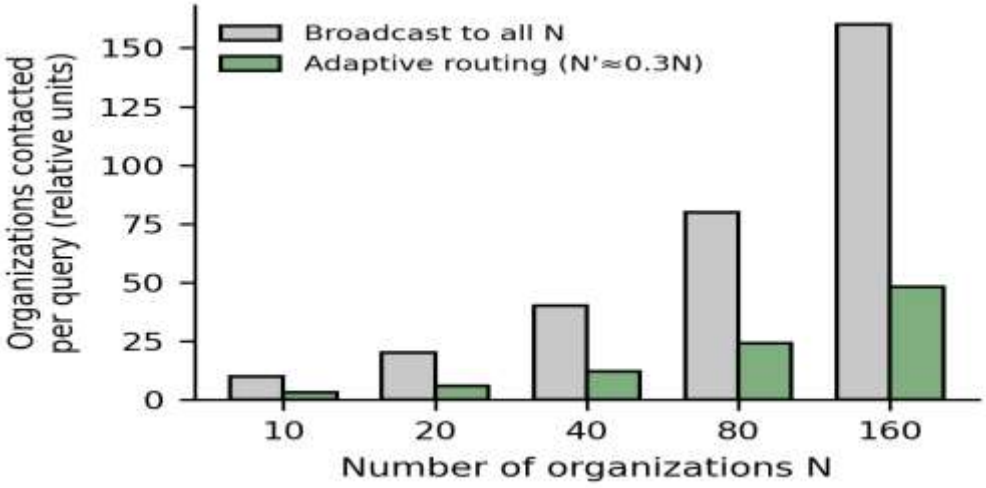

Figure 5: Analytical per-query communication volume, broadcast vs. adaptive routing (illustrative estimate, $N' \approx 0.3N$).

C. Computational Complexity

Local retrieval at a routed organization costs $O(m_i \cdot d)$, identical in order to centralized retrieval over the same subset. Trust-weighted aggregation adds $O(N')$ coordinator-side work, negligible relative to embedding computation. Semantic fusion adds $O(K^2 d)$ for pairwise clustering over the $K = N'k$ candidate passages returned by the routed subset (Eq. 14); because K is bounded by the routed set size rather than N, this cost also shrinks under adaptive routing.

D. Latency

End-to-end latency is bounded by the slowest organization in R(q) plus one aggregation round and one generation call. Because ATR-SF contacts only $N' \leq N$ organizations, it also reduces the expected number of parallel branches that can straggle, shrinking the tail of the response-time distribution relative to full-broadcast designs, in addition to the raw communication savings quantified above.

E. Scalability

Because each organization indexes and searches only its own corpus, adding an organization does not require re-indexing existing participants, unlike centralized RAG, where corpus growth increases index size and per-query cost for every user. Under ATR-SF, scalability is further improved because N' is expected to grow more slowly than N as new, topically narrow organizations join, so per-query cost does not track total federation size (Fig. 5).

F. Security

Trust weighting bounds the contribution of any single low-quality or adversarially manipulated organization, directly countering the influence exploited by extraction attacks such as [15] and membership-inference attacks such as [14]. This introduces a new risk not present in a uniform-weighting baseline: a coalition of colluding organizations could attempt to inflate one another's trust scores over repeated rounds, a Sybil-like threat that is not resolved by the mechanisms described here and is discussed in Section V.

G. Differentiation from Prior Work

Table V positions FedRAG-Priv against five recent Federated RAG systems along the specific axes this paper's gap analysis identifies as unaddressed. Existing systems provide at most partial, admission-time or execution-time trust mechanisms and none combine query-time adaptive routing with trust-weighted aggregation and redundancy-aware semantic fusion under a stated privacy budget.

Table 2: Comparison of existing federated/privacy-aware RAG methods.

Method	Centr. Retr.	Privacy	FL	Sec. Agg.	LLM Int.	Limitations
FeB4RAG [8]	Partial	Low	No	No	Yes	Search federation only; no privacy layer.
FRAG [9]	No	Low	No	No	Yes	Vector-DB federation; no formal privacy guarantee.
D-RAG [10]	No	Medium	No	No	Yes	Blockchain admission control; no query-time privacy.
C-FedRAG [7]	No	High	Yes	Partial	Yes	Relies on trusted execution hardware.
Mao et al. [6]	No	Medium	Yes	No	Yes	Federated embeddings; no secure aggregation.
FedRAG-Priv (proposed)	No	High	Yes	Yes	Yes	Analytical; not yet implemented.

Table 3: Unique-feature differentiation from recent Federated RAG systems.

System	Adaptive Routing	Trust-Weighted Agg.	Semantic Fusion	Formal DP Budget	Complexity Analysis
FeB4RAG [8]	Partial	No	No	No	No
FRAG [9]	No	No	No	No	No
D-RAG [10]	No	Partial (admission-time)	No	No	No
C-FedRAG [7]	No	No	No	Partial (TEE-based)	No
Mao et al. [6]	No	No	No	Partial	No
Zhou et al. [11]	No	No	No	Partial	No
FedRAG-Priv (proposed)	Yes	Yes	Yes	Yes	Yes

Table 4: Centralized RAG vs. proposed Federated RAG.

Property	Centralized RAG	FedRAG-Priv
Privacy	Low — operator sees all documents	High — DP + trust-weighted secure aggregation
Communication	O(1) (local index)	$O(N'(d+k(d+\ell)))$, $N' \leq N$ via adaptive routing
Knowledge coverage	Public/pooled data only	Combined private + public sources
Latency	Single index lookup	Bounded by slowest routed organization + 1 round
Scalability	Degrades with corpus growth	Bounded by N', not total corpus or federation size
Security	Single point of exposure	No single party sees raw scores; low-trust influence capped
Fault tolerance	Single point of failure	Degrades gracefully with routed subset
Regulatory compliance	Difficult under GDPR/HIPAA	Data residency preserved by design

Table 5: Expected analytical performance (estimates, not experimental measurements).

Metric	Expected Trend (analytical estimate)
Retrieval accuracy	Comparable to centralized RAG at low ε; degrades gradually as ε decreases.
Privacy risk	Bounded by εtotal = εretrieval + εtrust composition [5]; substantially lower than centralized exposure.
Communication cost	Linear in N' (routed subset) and k; ~70% lower than full broadcast at representative routing fraction (Fig. 5).
Scalability	Sub-linear growth of N' with N; degrades more gracefully than centralized index growth.
Inference latency	Bounded by slowest routed organization plus one aggregation round; tail shortened by smaller N'.
Knowledge freshness	Local updates propagate immediately; no re-centralization delay.
Robustness to low-quality participants	Influence bounded by trust weight wi rather than uniform per-organization influence.
Context redundancy	Reduced via similarity clustering (θ) relative to plain top-K merge.

Future Research Directions

1) Sybil- and collusion-resistant trust scoring to prevent coordinated trust inflation among colluding organizations (Section IV-F).
2) Learned, rather than heuristic, relevance priors ri(q) for adaptive routing, trained without centralizing query logs.
3) Homomorphic encryption for score aggregation [18] as a complement to secure aggregation, trading computation for stronger confidentiality against a colluding coordinator.
4) Cross-space alignment for organizations that adopt heterogeneous embedding models, required before similarity-based clustering in Semantic Context Fusion.
5) Benchmark datasets and evaluation protocols specific to Federated RAG, enabling empirical validation of the analytical claims made here.

Conclusion

This paper proposed FedRAG-Priv, a federated retrieval-augmented generation framework centered on Adaptive Trust-Weighted Routing and Semantic Fusion (ATR-SF). Rather than broadcasting every query to every organization and merging results uniformly, as in existing Federated RAG designs, FedRAG-Priv routes each query to a trust- and relevance-filtered subset, weights aggregation by trust, and removes redundant evidence through similarity clustering, all under a formally composed differential-privacy budget. The architecture is formalized mathematically and analyzed for communication and computational complexity as a function of the routed subset size. Because the framework is proposed rather than implemented, its evaluation is analytical: it indicates lower communication cost, comparable or improved latency, and bounded exposure to unreliable participants relative to both centralized RAG and existing Federated RAG systems, motivating further work toward an implemented and empirically validated system, as outlined in Section V.

References

[1] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv:2005.11401, 2020.
[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.
[3] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv:1610.05492, 2016.
[4] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security (CCS), 2017, pp. 1175–1191.
[5] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.
[6] Q. Mao et al., "Privacy-preserving federated embedding learning for localized retrieval-augmented generation," arXiv:2504.19101, 2025.
[7] P. Addison et al., "C-FedRAG: A confidential federated retrieval-augmented generation system," arXiv:2412.13163, 2024.
[8] S. Wang, E. Khramtsova, S. Zhuang, and G. Zuccon, "FeB4RAG: Evaluating federated search in the context of retrieval augmented generation," in Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2024, pp. 763–773.
[9] D. Zhao, "FRAG: Toward federated vector database management for collaborative and secure retrieval-augmented generation," arXiv:2410.13272, 2024.
[10] T. E. Andersen, A. M. Avalos, G. G. Dagher, and M. Long, "D-RAG: A privacy-preserving framework for decentralized RAG using blockchain," in Proc. 15th Int. Conf. Computer Science, Engineering and Applications (CCSEA), 2025, pp. 219–232.
[11] P. Zhou, Y. Feng, and Z. Yang, "Privacy-aware RAG: Secure and isolated knowledge retrieval," arXiv:2503.15548, 2025.
[12] T. Koga et al., "Privacy-preserving retrieval-augmented generation with differential privacy," arXiv:2412.04697, 2024.
[13] S. Zeng et al., "The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)," arXiv:2402.16893, 2024.
[14] M. Anderson, G. Amit, and A. Goldsteen, "Is my data in your retrieval database? Membership inference attacks against retrieval augmented generation," arXiv:2405.20446, 2024.
[15] C. Jiang, X. Pan, G. Hong, C. Bao, and M. Yang, "RAG-Thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks," arXiv:2411.14110, 2024.
[16] A. Chakraborty, C. Dahal, and V. Gupta, "Federated retrieval-augmented generation: A systematic mapping study," in Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 7362–7374.
[17] H. Zeng, Z. Yue, Q. Jiang, and J. Wang, "Federated recommendation via hybrid retrieval augmented generation," arXiv:2403.04256, 2024.
[18] W. Jin et al., "FedML-HE: An efficient homomorphic-encryption-based privacy-preserving federated learning system," arXiv:2303.10837, 2023.
[19] J. So et al., "LightSecAgg: A lightweight and versatile design for secure aggregation in federated learning," Proc. Machine Learning and Systems, vol. 4, pp. 694–720, 2022.

[20] H. R. Roth et al., "NVIDIA FLARE: Federated learning from simulation to real-world," arXiv:2210.13291, 2022.
[21] L. He, P. Tang, Y. Zhang et al., "Mitigating privacy risks in retrieval-augmented generation via locally private entity perturbation," Information Processing & Management, vol. 62, no. 4, 2025.
[22] D. Yao and T. Li, "Private retrieval augmented generation with random projection," in Proc. ICLR 2025 Workshop on Building Trust in Language Models and Applications, 2025.
[23] H. Wang, X. Xu, B. Huang et al., "Privacy-aware decoding: Mitigating privacy leakage of large language models in retrieval-augmented generation," arXiv:2508.03098, 2025.
[24] V. Vinod, K. Pillutla, and A. G. Thakurta, "InvisibleInk: High-utility and low-cost text generation with differential privacy," arXiv:2507.02974, 2025.