

Privacy-Preserving Federated Retrieval-Augmented Generation for Distributed Knowledge Systems

Nichola Hannigan

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
nichola.work@binghamton.edu

Marcio Giron

Department of Computer Science, George Mason University, Fairfax, VA, USA.
marcio.work@gmu.edu

Yeni Zedillo

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
yeniz@buffalo.edu

Lazzaro Nonis

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
nonis515@ku.edu

Abstract

Privacy-preserving federated retrieval-augmented generation has emerged as a critical systems direction for distributed knowledge environments in which organizations and communities need the reasoning power of large generative models without centralizing sensitive documents, proprietary indexes, or user interaction logs. This paper examines the architectural, infrastructural, governance, and policy dimensions of federated retrieval-augmented generation as a socio-technical system rather than a narrow model optimization problem. It argues that the central design challenge is not merely protecting individual model updates, but coordinating heterogeneous retrieval sources, heterogeneous privacy guarantees, heterogeneous institutional rules, and heterogeneous trust boundaries across a distributed knowledge fabric. The discussion analyzes secure aggregation, differential privacy, encrypted retrieval, local indexing, semantic fusion, and distributed evaluation as interdependent components whose trade-offs shape robustness, fairness, sustainability, and accountability. It further considers how governance mechanisms, auditability, model cards, datasheets, and policy constraints must be embedded into the system lifecycle. The paper concludes that privacy-preserving federated retrieval-augmented generation can support more equitable and resilient knowledge systems only if privacy is treated as an infrastructural property, retrieval as a governed common-pool resource, and evaluation as a multi-stakeholder process rather than a centralized benchmark exercise.

Keywords

privacy-preserving federated learning; retrieval-augmented generation; distributed knowledge systems; differential privacy; secure aggregation; trustworthy artificial intelligence; socio-technical infrastructure; governance.

1. Introduction

Distributed knowledge systems increasingly confront a structural contradiction. The most capable generative artificial intelligence systems depend on broad access to diverse, current, and domain-specific information, yet the most valuable information is often held by hospitals, government agencies, research libraries, industrial laboratories, community archives, and regional institutions that cannot or should not transfer it to a central operator. Retrieval-augmented generation offers a partial response by allowing a language model to consult external sources at inference time, thereby improving factuality and domain adaptation without fully internalizing all knowledge into model parameters [5]. Federated learning offers another partial response by allowing multiple parties to train or adapt models without sharing raw data, using local computation and aggregated updates as the primary learning signal [1]. Privacy-preserving federated retrieval-augmented generation combines these trajectories but also intensifies their unresolved tensions. It asks how a distributed system can retrieve, rank, fuse, and generate from knowledge that remains legally, institutionally, and technically decentralized while preserving privacy, security, fairness, and operational viability.

The significance of this problem extends beyond technical novelty. Distributed knowledge systems are embedded in public infrastructures, professional workflows, and civic institutions. A federated retrieval system used by a consortium of hospitals may influence clinical decisions; one used by universities may shape research agendas and student services; one used by public agencies may affect eligibility determinations, emergency response, and public accountability. In each case, the system is not simply a pipeline of models and indexes. It is a governance arrangement that distributes power among data holders, model operators, retrieval service providers, auditors, regulators, and affected communities. Consequently, the design of privacy-preserving federated retrieval-augmented generation must be evaluated against structural trade-offs: privacy against utility, decentralization against coordination, freshness against stability, openness against security, and local autonomy against system-wide consistency.

This paper develops a system-level analysis of privacy-preserving federated retrieval-augmented generation for distributed knowledge systems. It proceeds from the premise that privacy cannot be added as a late-stage encryption layer or policy overlay. Instead, privacy must be treated as an infrastructural property that shapes data representation, retrieval routing, semantic fusion, update aggregation, audit trails, and failure recovery. The paper also stresses that retrieval is not a neutral operation. Retrieval determines whose knowledge is visible, whose language is legible, whose documents are ranked highly, and whose absence becomes an invisible gap. In federated settings, these decisions are distributed across administrative domains, which makes fairness and accountability more difficult but also potentially more locally responsive. The analysis therefore connects architecture to governance, deployment to sustainability, and robustness to policy. The required reference [13] is treated as one example of the broader movement toward unified semantic spaces and attention-based fusion for multi-source heterogeneous retrieval-augmented generation, but the paper does not center any single framework. Rather, it situates such work within the longer problem of building trustworthy distributed knowledge infrastructures.

2. Background and Related Work

Federated learning introduced a paradigm in which clients collaboratively train a shared model while keeping raw data local [1]. Early work demonstrated communication-efficient aggregation for decentralized data, and subsequent research expanded the field into a rich design space involving non-identical data distributions, heterogeneous client capabilities,

adversarial participants, and varying privacy requirements [2]. Surveys have documented the core challenges of statistical heterogeneity, systems heterogeneity, communication efficiency, and privacy preservation [3]. System-oriented research has also shown that production federated learning requires careful orchestration, secure aggregation, device scheduling, and fault tolerance, especially when clients are unreliable or intermittently connected [4]. These lessons are directly relevant to federated retrieval-augmented generation because retrieval introduces additional moving parts: indexes, encoders, ranking policies, fusion modules, and generation models may all be distributed unevenly across participants.

Retrieval-augmented generation emerged as a method for improving knowledge-intensive natural language tasks by combining a parametric generator with a non-parametric retrieval component [5]. Dense passage retrieval demonstrated that learned embeddings can support open-domain question answering more effectively than sparse lexical matching in many settings [6]. Later work explored retrieval-augmented generation with generative models, showing that retrieved passages can condition generation and reduce reliance on memorized parameters [7]. Retrieval-augmented language model pre-training and large-scale retrieval from trillions of tokens further illustrated the potential for non-parametric memory to extend model knowledge without retraining the entire model [8], [9]. These advances make retrieval an attractive mechanism for distributed knowledge because sources can remain external to the generator. However, they also create new privacy surfaces: queries reveal user intent, retrieved documents reveal source content, ranking scores can leak information, and generated text can reproduce sensitive passages. A federated architecture must therefore protect not only model updates but also query patterns, retrieval decisions, and semantic representations.

The literature on privacy-preserving machine learning provides foundational tools. Differential privacy offers a formal framework for quantifying and limiting information leakage from computations over sensitive data [10]. Deep learning with differential privacy showed that private training is feasible under carefully managed noise and clipping, although utility costs remain significant [11]. Semi-supervised knowledge transfer and private aggregation mechanisms further expanded the practical toolkit for protecting training data [12]. At the same time, attacks such as membership inference, model inversion, gradient leakage, and training data extraction demonstrate that machine learning systems can leak sensitive information through outputs, parameters, gradients, and generated content [15], [16], [17], [18]. Security surveys of federated learning emphasize that privacy and security are related but distinct: encryption and secure aggregation can protect updates from an honest-but-curious server, but they do not automatically prevent poisoning, inference, or governance failures [19]. These insights frame privacy-preserving federated retrieval-augmented generation as a multi-layer problem rather than a single cryptographic guarantee.

3. Architectural Foundations of Privacy-Preserving Federated Retrieval-Augmented Generation

A privacy-preserving federated retrieval-augmented generation architecture must coordinate at least five layers: local data and indexing, local or delegated retrieval, privacy-preserving representation and aggregation, semantic fusion across heterogeneous sources, and generation with policy-aware output controls. In a fully decentralized design, each participant maintains a local index over its own documents and exposes only a constrained retrieval interface. Queries may be encrypted or obfuscated, and retrieved passages may be summarized, redacted, or transformed before they enter a shared generation process. In a partially

centralized design, a coordinator may manage routing and fusion while participants retain control over documents and update submission. The choice between these models is not merely technical. It determines who can observe query patterns, who can audit ranking behavior, who bears the cost of index maintenance, and who can exclude or include particular knowledge domains.

Secure aggregation and differential privacy are often proposed as core safeguards. Secure aggregation prevents the coordinating server from seeing individual model updates, while differential privacy limits what can be inferred from aggregate statistics or released outputs [10], [11]. Yet retrieval complicates both mechanisms. Model updates are not the only sensitive artifact; retrieval queries can reveal a user's condition, research interest, legal question, or political concern. Retrieved passages can reveal source content even if the underlying document remains local. Attention weights and fusion scores can expose which sources influenced a response. A robust architecture therefore needs privacy accounting across the entire inference and adaptation path, not only during training. It must consider whether local indexes are encrypted, whether embeddings are invertible, whether cached queries persist, and whether generated citations create a side channel that identifies a particular repository.

Heterogeneous sources create a further architectural challenge. Different participants may use different document formats, languages, metadata standards, ontologies, and access-control policies. A unified semantic space can help align representations across these differences, while attention-based fusion can weigh retrieved evidence according to relevance, reliability, and context [13]. However, unification also introduces risks. If all sources are mapped into a shared embedding space, the embedding itself may become a centralized knowledge asset that reveals semantic relationships across protected domains. If fusion is centralized, the fusion module may become a point of concentrated power and a target for attack. If fusion is fully local, system-wide coherence may suffer. The design must therefore balance shared semantics with local sovereignty, using techniques such as federated embedding alignment, policy-aware routing, and privacy-preserving ranking. The goal is not to eliminate heterogeneity but to govern it without forcing premature convergence.

Generation is the final and often most visible layer. A generator that receives retrieved evidence must decide how to synthesize, attribute, and qualify that evidence. In a distributed knowledge system, generation must respect source-specific licensing, consent, embargoes, and confidentiality requirements. It must also avoid hallucinating consensus where sources disagree and avoid amplifying dominant sources simply because they are easier to retrieve. Policy-aware generation can enforce constraints at inference time, but such enforcement requires interpretable provenance and auditable decision logs. The architecture must therefore support controlled disclosure, citation fidelity, and contestability. Without these properties, privacy-preserving federated retrieval-augmented generation may protect raw data while producing outputs that are still unfair, unaccountable, or unsafe.

4. Distributed Knowledge Systems and Retrieval Infrastructure

Distributed knowledge systems are not simply federated databases. They are institutional arrangements in which knowledge is produced, validated, contested, and maintained by communities with different incentives and obligations. A hospital's clinical notes, a university's research repository, a government agency's case files, and a community archive's oral histories differ not only in format but in governance. Retrieval-augmented generation can make these collections more accessible, but it can also flatten their distinct logics into a

universal query interface. The infrastructure challenge is to preserve context while enabling cross-domain retrieval. This requires metadata standards, provenance tracking, consent management, access-control languages, and durable identifiers that survive institutional change. It also requires retrieval protocols that can express not only relevance but also legitimacy, jurisdiction, sensitivity, and recency.

From an infrastructure perspective, retrieval is a latency-sensitive and resource-intensive operation. Dense retrieval involves embedding queries and documents, searching large vector indexes, and ranking candidate passages [6]. Large-scale retrieval may require distributed index shards, caching layers, and hardware accelerators [9]. In a federated setting, these resources are unevenly distributed. Large institutions may operate high-performance retrieval clusters, while small clinics, local libraries, or community organizations may have limited connectivity and compute. If the architecture assumes always-on, high-bandwidth participation, it will systematically exclude less resourced participants, thereby reproducing existing inequalities in knowledge production. Sustainable federated retrieval must therefore support asynchronous participation, compressed representations, intermittent connectivity, and tiered service levels. It must also avoid shifting hidden costs onto the weakest nodes in the network.

Privacy-preserving federated retrieval also faces the problem of index freshness and drift. Knowledge is not static. Clinical guidelines change, laws are amended, scientific consensus evolves, and community histories are revised. A distributed system must propagate updates without forcing participants to reveal their full indexes or continuously expose query traffic. Federated indexing and update protocols can help, but they raise questions about versioning, conflict resolution, and authority. Who decides which version of a guideline is current? How are contradictory sources represented? How are deletions and retractions handled across caches? These are not only technical questions. They are governance questions about epistemic authority. A system that silently favors the most recently updated source may disadvantage communities that document knowledge in slower cycles. A system that preserves all versions may overwhelm users with conflicting evidence. The infrastructure must therefore make temporal and jurisdictional context visible rather than hidden behind a single relevance score.

The socio-technical dimension of retrieval infrastructure is especially visible in public-interest applications. In public health, federated retrieval could allow hospitals and clinics to share evidence without exposing patient records, but it could also enable insurers or employers to infer sensitive patterns if query controls are weak. In education, a federated research assistant could help students access distributed archives, but it could also reproduce paywalls, language bias, and institutional gatekeeping. In government, federated retrieval could improve interagency coordination, but it could also create new surveillance pathways if audit and purpose limitation are inadequate. These examples show that retrieval infrastructure is never neutral. It encodes policy choices about who may ask what, from whom, under which conditions, and with what accountability.

5. Privacy, Security, and Trust Boundaries

Privacy in federated retrieval-augmented generation must be understood across multiple trust boundaries. The first boundary is between the user and the local device or institution. The second is between a participant and the coordinator or aggregation service. The third is between participants who may be curious, competitive, or malicious. The fourth is between the system and external observers who may monitor network traffic, side channels, or public

outputs. The fifth is between the system and future uses of derived data, including logs, embeddings, and fine-tuned adapters. A single privacy mechanism cannot address all these boundaries. Differential privacy may protect aggregate statistics but not query confidentiality. Secure aggregation may protect update confidentiality but not retrieval content. Encryption may protect data in transit and at rest but not inference-time leakage through generated text. A defense-in-depth strategy is necessary, but it must be coordinated so that protections do not conflict or create false assurances.

Attack surfaces in this setting are broad. Membership inference can reveal whether a particular document or individual was used in training or retrieval [17]. Model inversion can reconstruct representative features of sensitive data from model behavior [16]. Gradient leakage can recover training samples from shared updates in some federated configurations [18]. Training data extraction can cause generative models to reproduce memorized content [15]. Indirect prompt injection can manipulate retrieval-augmented systems by inserting adversarial content into documents that are later retrieved and followed by the generator [14]. These attacks are not hypothetical; they reflect the adversarial nature of distributed knowledge environments. In a federated retrieval system, an attacker may control a participant, a document repository, a network route, or a user account. The system must therefore assume that some retrieved evidence may be poisoned, some updates may be malicious, and some queries may be designed to probe protected information.

Trust boundaries also shape the usability of privacy guarantees. If privacy protections are so strong that they prevent meaningful retrieval, participants will route around the system or share data through informal channels. If protections are weak or opaque, participants may refuse to join. A workable design must make privacy trade-offs legible to stakeholders, not merely technically compliant. This requires privacy budgets that are understandable, audit logs that are meaningful, and governance processes that allow participants to contest or withdraw. It also requires clear separation between personal data, institutional data, and public knowledge. A community archive may permit broad retrieval of public histories but restrict culturally sensitive materials. A hospital may permit aggregate clinical retrieval but prohibit patient-level inference. A government agency may permit interagency retrieval under statutory purpose limitations. The architecture must represent these distinctions without reducing them to a single binary flag.

Finally, trust in federated retrieval-augmented generation depends on verifiability. Participants need assurance that aggregation was performed correctly, that privacy budgets were respected, and that retrieval policies were enforced. Cryptographic proofs, secure enclaves, transparency logs, and independent audits can contribute, but they must be embedded in a governance framework that assigns responsibility and remedies harm. Verification is not only a technical property; it is a social relationship supported by institutions. Without meaningful verification, privacy claims become reputational assertions. With excessive verification overhead, participation becomes impractical. The design challenge is to create proportionate trust mechanisms that match the sensitivity of the knowledge domain and the risk of the application.

6. Governance, Policy, and Institutional Implications

Governance is the connective tissue of privacy-preserving federated retrieval-augmented generation. Because the system spans multiple institutions, no single actor can legitimately define all rules. Governance must address membership, purpose limitation, data minimization, retention, access control, dispute resolution, and benefit sharing. It must also define how

decisions about retrieval inclusion and exclusion are made. If a source is omitted from the federated index, its knowledge becomes invisible to the system. If a source is over-weighted, it may dominate generated answers. These are political and epistemic decisions, not merely engineering parameters. Effective governance requires representative participation from data holders, domain experts, affected communities, and system operators. It also requires mechanisms for appeal when outputs cause harm or when sources are misrepresented.

Policy frameworks add external constraints. Data protection laws may require purpose limitation, data subject rights, and cross-border transfer safeguards. Sector-specific regulations may impose additional requirements in health, finance, education, and public administration. Competition policy may scrutinize centralized control over retrieval infrastructure. Intellectual property law may restrict the use of copyrighted materials in retrieval and generation. National security rules may limit certain knowledge flows. A federated architecture can help satisfy some of these constraints by keeping data local, but it can also complicate compliance because data may be processed across jurisdictions at inference time. Policymakers and system designers must therefore collaborate on interoperable standards for consent, provenance, audit, and accountability. The system should be able to demonstrate compliance without requiring the disclosure of protected data.

Documentation practices are part of governance infrastructure. Datasheets for datasets and model cards for model reporting provide structured ways to communicate intended use, limitations, composition, and ethical considerations [21], [22]. In a federated retrieval system, documentation must extend to indexes, retrieval policies, fusion models, and aggregation protocols. Each participant may have different documentation obligations, and the coordinator may need to synthesize them into a system-level transparency report. Documentation alone does not guarantee fairness, but it creates the conditions for scrutiny. It also helps new participants understand what they are joining and helps auditors trace how a particular output was produced. Without documentation, governance becomes reactive and opaque.

Fairness and accountability must also be understood institutionally. A federated system may be technically neutral while producing unequal outcomes because participating institutions differ in data quality, language coverage, and historical representation. A model trained or adapted on distributed data may perform worse for communities whose knowledge is underrepresented or whose documents are not digitized. Retrieval may favor well-indexed sources over oral or embodied knowledge. Governance must therefore include affirmative measures to support underrepresented sources, not only non-discrimination rules. It must also allocate resources for community review, translation, and metadata creation. Otherwise, privacy-preserving federation may protect existing inequalities by making them harder to see. The policy implication is that privacy and fairness are not separate goals; they are linked through the distribution of infrastructural capacity and epistemic authority.

7. Robustness, Fairness, and Sustainability

Robustness in federated retrieval-augmented generation refers to the system's ability to maintain reliable performance under adversarial, heterogeneous, and changing conditions. Adversarial robustness includes resistance to poisoned documents, malicious updates, prompt injection, and denial-of-service attacks [14], [19]. Statistical robustness includes resilience to non-identical data distributions, concept drift, and source dropout [2], [3]. Operational robustness includes graceful degradation when participants disconnect, when indexes become stale, or when a retrieval service fails. These dimensions are interconnected. A system that is robust to malicious updates but brittle to source dropout may still fail in practice. A system

that performs well under stable conditions may collapse when a major participant leaves or when regulatory rules change. Robustness must therefore be designed as a portfolio of mechanisms, including redundancy, reputation, anomaly detection, fallback retrieval, and human escalation.

Fairness in this context has distributive, procedural, and representational dimensions. Distributive fairness concerns whether the benefits and burdens of the system are shared equitably. Procedural fairness concerns whether participants and affected communities have a voice in governance. Representational fairness concerns whether retrieved and generated knowledge reflects diverse perspectives without stereotyping or erasing minority viewpoints. Differential privacy can interact with fairness in complex ways; noise may disproportionately affect small subgroups, and aggregation may obscure disparities [10]. Federated learning can improve privacy but may also reinforce bias if local data are skewed [2]. Retrieval can amplify bias if ranking models favor dominant sources [6]. A fair system must therefore combine technical metrics with participatory evaluation. It must allow communities to define what fair treatment means in their context and to contest outputs that violate those expectations.

Sustainability has infrastructural, environmental, and institutional dimensions. Infrastructural sustainability requires that the system remain maintainable over time, with clear ownership, funding, and update pathways. Environmental sustainability requires attention to the energy and carbon costs of distributed indexing, embedding, retrieval, and generation. A federated architecture may reduce some centralization costs but increase communication and redundant computation. Institutional sustainability requires that participation remains worthwhile for all parties, including those with fewer resources. If only large institutions can afford to maintain high-quality indexes and privacy infrastructure, the federation will become a cover for centralization. Sustainable design therefore favors modularity, open standards, low-resource participation modes, and shared maintenance responsibilities. It also requires long-term governance commitments rather than short-term pilot funding.

The tension between robustness and openness is particularly acute. Open retrieval systems can be more transparent and contestable, but they are also easier to poison or exploit. Closed systems may be more secure, but they concentrate power and reduce accountability. A federated approach can balance these extremes by allowing multiple retrieval paths, diverse curation policies, and layered access controls. However, federation does not automatically resolve the tension. It redistributes it. The design must decide which decisions are local, which are shared, and which are delegated to independent auditors or regulators. These decisions should be explicit, revisable, and aligned with the values of the knowledge communities involved.

8. Evaluation, Deployment, and Operational Considerations

Evaluation of privacy-preserving federated retrieval-augmented generation cannot rely solely on centralized benchmarks. Standard accuracy, latency, and throughput metrics remain important, but they must be complemented by privacy leakage tests, fairness audits, robustness stress tests, and governance reviews. Evaluation must also account for the distributed nature of the system. A model may perform well on a global benchmark while failing for a small language community or a specialized clinical domain. Retrieval may be fast on average while being inaccessible to participants with low bandwidth. Privacy may hold under one threat model while failing under another. A multi-level evaluation framework should therefore combine local validation, cross-site validation, adversarial simulation, and

participatory assessment. It should also document negative results and failure modes, not only aggregate improvements [22].

Deployment introduces operational constraints that are often invisible in research prototypes. Identity management, key rotation, secure aggregation orchestration, version control, index synchronization, and incident response must all be maintained in production [4]. Participants may have different security postures, patching schedules, and legal obligations. The coordinator may need to support multiple jurisdictions and languages. Users may expect low-latency responses even when retrieval requires cross-institutional queries. These pressures can lead to shortcuts that weaken privacy or fairness. For example, caching retrieved passages for speed may create new data retention risks. Logging queries for debugging may create surveillance capacity. Fine-tuning on retrieved content may embed sensitive information into model parameters. Deployment planning must therefore treat privacy, fairness, and security as continuous operational functions rather than one-time design choices.

Incident response is especially important. In a distributed knowledge system, a breach may occur at one participant but affect the entire federation. A poisoned document may influence many generated outputs before it is detected. A compromised key may expose aggregated updates. A biased ranking model may systematically disadvantage a group. The system needs protocols for detection, containment, notification, remediation, and public accountability. These protocols must be agreed upon before deployment and tested through exercises. They must also respect the autonomy of participants while enabling coordinated action. Without incident response, trust will erode quickly after the first visible failure. With well-designed response mechanisms, failures can become opportunities for learning and improvement.

Operational sustainability also depends on user-facing design. Users need to understand when a response is based on retrieved evidence, which sources contributed, what limitations apply, and how to challenge an output. Explanations should be accessible without exposing sensitive source content. Confidence indicators should be calibrated and honest. Citation mechanisms should distinguish between direct quotation, paraphrase, and synthesis. In federated settings, provenance may be partial or protected, which makes transparency more difficult but no less important. The system should communicate uncertainty and jurisdictional limits rather than presenting a single authoritative answer. This is particularly important in high-stakes domains where users may otherwise over-trust the system.

9. Future Directions and Research Agenda

Several research directions follow from this analysis. First, privacy accounting must become end-to-end. Future work should develop methods that track privacy loss across retrieval, fusion, generation, and adaptation, not only during federated training [10], [11]. Second, federated retrieval requires new protocols for secure and efficient index operations. Research is needed on encrypted search, private set intersection, secure ranking, and distributed vector search that can operate under realistic resource constraints. Third, semantic fusion across heterogeneous sources must be designed with sovereignty in mind. Unified semantic spaces and attention-based fusion can improve cross-domain retrieval [13], but they must be accompanied by mechanisms that preserve local meaning, licensing, and access control. Fourth, evaluation must become participatory and multi-stakeholder. Benchmarks should be co-designed with affected communities and should measure not only utility but also privacy, fairness, robustness, and contestability.

Fifth, governance research should examine how federated knowledge systems can be governed as common-pool resources. This includes studying membership rules, conflict resolution, benefit sharing, and accountability across jurisdictions [20], [21]. Sixth, policy research should explore how existing data protection, intellectual property, and sectoral regulations apply to retrieval-augmented generation in federated settings. Regulators may need new tools for auditing distributed systems without requiring centralized access to sensitive data. Seventh, sustainability research should quantify the environmental and institutional costs of different federation architectures. A system that is privacy-preserving but environmentally wasteful or institutionally fragile is not truly sustainable. Eighth, security research should continue to investigate adversarial retrieval, prompt injection, and data extraction in federated contexts [14], [15]. As retrieval systems become more capable, attackers will develop more sophisticated methods for manipulating or extracting knowledge.

A final direction concerns human oversight. Federated retrieval-augmented generation should support meaningful human judgment rather than replacing it. This means designing interfaces that reveal provenance, uncertainty, and disagreement. It means creating escalation paths for high-stakes decisions. It means allowing domain experts and affected communities to correct, annotate, or contest retrieved knowledge. Human oversight is not a substitute for technical safeguards, but it is essential for handling the ambiguity and context sensitivity that distributed knowledge systems inevitably encounter. The most robust system will be one that combines strong privacy engineering with strong institutions and engaged communities.

10. Conclusion

Privacy-preserving federated retrieval-augmented generation offers a promising path for distributed knowledge systems that need the capabilities of generative artificial intelligence without centralizing sensitive data. However, the promise depends on treating the problem as a socio-technical design challenge rather than a narrow model or cryptography problem. The architecture must protect queries, indexes, updates, retrieved content, and generated outputs across multiple trust boundaries. It must support heterogeneous sources, uneven resources, and conflicting jurisdictional rules. It must embed governance, documentation, audit, and contestability into the system lifecycle. It must evaluate success through privacy, fairness, robustness, sustainability, and accountability, not only accuracy or latency. The required reference [13] illustrates one technical direction toward unified semantic spaces and attention-based fusion, but technical fusion alone cannot resolve the deeper tensions of authority, representation, and power. The future of federated retrieval-augmented generation will therefore depend on interdisciplinary collaboration among computer scientists, engineers, legal scholars, policymakers, domain experts, and affected communities. If designed with care, it can contribute to knowledge infrastructures that are more private, more equitable, more resilient, and more democratic than either centralized platforms or isolated local systems.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR.
2. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>

3. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50–60. <https://doi.org/10.1109/MSP.2020.2975749>
4. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... Roselander, J. (2019). Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1, 374–388.
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
6. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
7. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 874–880). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.eacl-main.74>
8. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 3929–3938). PMLR.
9. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206–2240). PMLR.
10. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
11. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308–318). ACM. <https://doi.org/10.1145/2976749.2978318>
12. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*.
13. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In *2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 591–597). IEEE.
14. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (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* (pp. 79–90). ACM. <https://doi.org/10.1145/3605764.3623985>

15. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... Raffel, C. (2021). Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (pp. 2633–2650). USENIX Association.
16. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322–1333). ACM. <https://doi.org/10.1145/2810103.2813677>
17. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (pp. 3–18). IEEE. <https://doi.org/10.1109/SP.2017.41>
18. Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 14774–14784.
19. Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115, 619–640. <https://doi.org/10.1016/j.future.2020.10.007>
20. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. <https://fairmlbook.org>
21. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
22. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM. <https://doi.org/10.1145/3287560.3287596>