Applied Artificial Intelligence and Computing Systems

https://www.scigate.org

Science Gate Academic

Vol. 1, No. 2, 2025

*Article*

# SPRINT-RAG: Secure Partitioned Retrieval-Augmented LLM Diagnosis for Privacy-Preserving Incident Analysis in Distributed Clouds

**Kaito Mizunoe[1], Renjiro Amahara[1] , Quxi Lu[1*]**
[1] Kanazawa Institute of Technology, Nonoichi, Japan
*Corresponding author: quxixxxlu2@yahoo.com

***Abstract:*** *Large language models are increasingly used to interpret operational evidence during cloud incidents, yet centralized retrieval over logs, traces, and service graphs creates privacy, governance, and scalability risks in distributed environments. This article presents SPRINT-RAG, a secure partitioned retrieval-augmented diagnosis framework for privacy-preserving incident analysis. The framework keeps raw observability evidence within tenant, service, or region partitions, performs local event retrieval and risk encoding, and exposes only policy-filtered evidence summaries to a global diagnosis layer. A controlled partition benchmark with 4,800 incident windows, 18 service groups, and three workload-drift regimes is used to evaluate diagnosis quality, evidence use, communication cost, and leakage proxies. Compared with centralized retrieval-augmented diagnosis, SPRINT-RAG improves F1 from 0.842 to 0.874 and top-1 root-cause accuracy from 0.681 to 0.724, while reducing the leakage proxy from 0.139 to 0.043. Compared with local-only diagnosis, it improves top-3 root-cause accuracy by 9.8 percentage points with a moderate p95 latency increase of 1.54 s. Ablation analysis shows that partition trust weighting and cross-partition evidence summaries are both necessary for stable performance under drift, while leakage filtering provides the largest privacy gain with limited diagnostic cost. The results indicate that secure partitioned retrieval can make LLM-based incident diagnosis more compatible with multi-tenant cloud governance without sacrificing operational usefulness.*

**Keywords:** *Large Language Models; Retrieval-Augmented Generation; Cloud Incident Diagnosis; Secure Partitioning; Root Cause Analysis; Distributed Systems*

## 1. Introduction

Cloud platforms have become difficult to operate because incidents rarely stay within a single component. A tail-latency spike can begin with a storage queue, propagate through service dependencies, trigger autoscaling decisions, and end as an application-level failure. Recent work on zero-shot anomaly prediction, log-event graph modeling, and dependency-aware spatiotemporal learning shows that diagnosis increasingly depends on evidence distributed across logs, metrics, traces, and topology views rather than on a single alarm stream [1], [4], [24]. At the same time, operational evidence is frequently sensitive. It may reveal customer identifiers, access patterns, deployment topology, or regulated business states. This tension makes centralized LLM-based diagnosis attractive but risky.

LLM-centered operational intelligence has moved from summarizing alerts toward root-cause explanation and remediation assistance [14], [27], [37]. Retrieval-augmented diagnosis is a natural design because it grounds generated explanations in recent evidence instead of relying only on model memory. However, a naive retrieval layer normally assumes that all logs, traces, and topology annotations can be pooled into one index. Such pooling is often infeasible in hybrid cloud and edge deployments, where tenants, regions, or service owners impose separate retention, consent, and access policies. Privacy-aware federated optimization and multi-scale representation learning offer one answer to distributed learning, but they do not by themselves solve the evidence-selection and explanation requirements of LLM diagnosis [5], [57].

This study addresses the following question: how can retrieval-augmented LLM diagnosis use cross-partition operational evidence without centralizing raw observability data? The proposed framework, SPRINT-RAG, treats secure partitioning as a first-class part of diagnosis rather than as an after-the-fact storage decision. Each partition maintains local event indexes, service-context encoders, and policy filters. The global diagnosis layer receives compact evidence summaries, partition trust weights, and anonymized dependency hints, then ranks root causes and generates an incident analysis. The design is motivated by advances in service-dependency modeling [36], cloud observability for LLM remediation [28], and resource-aware LLM serving [30], but it shifts the objective from serving efficiency alone to privacy-preserving operational reasoning.

The contribution of this article is threefold. First, it formulates secure partitioned retrieval for LLM-based incident diagnosis, including an adversary model, partition leakage proxy, and evidence-faithfulness checks. Second, it introduces a diagnosis architecture that combines local event retrieval, trust-weighted evidence fusion, and policy-gated summary exchange. Third, it reports a controlled and reproducible experimental study showing the trade-off among diagnostic performance, privacy exposure, communication volume, and latency. The study deliberately uses a controlled benchmark rather than production claims, allowing each result to be replicated from the same configuration and random seed.

## 2. Related Work

### 2.1 Cloud Anomaly Detection and Service Dependencies

Modern cloud anomaly detection has developed from metric thresholding into structured representation learning. Graph-based methods capture service topology and multi-relational log dependencies, while temporal contrastive and generative methods improve robustness under imbalance and distribution shift [7], [17], [22]. Surveys and recent dependency-learning studies further emphasize that incidents should be modeled as evolving system states rather than isolated observations [32], [52]. SPRINT-RAG follows this view by allowing each partition to encode local dependency context before evidence is shared.

Several studies are especially relevant to cloud incident analysis. Dependency-aware graph learning improves cluster-level failure detection, and dynamic graph modeling captures evolving microservice dependencies [24], [40]. Cost-sensitive sequence modeling and cross-timescale forecasting address non-stationary system signals [45], [55]. These methods are strong diagnostic substrates, but they usually output anomaly scores or graph embeddings; they do not directly produce governed textual explanations across tenant or region boundaries.

### 2.2 LLM-Based Diagnosis, Retrieval, and Multi-Agent Coordination

LLMs have been applied to automated root-cause analysis and observability-enhanced remediation because they can combine evidence fragments with operational knowledge [14], [27]. Long-range decision summarization and closed-loop multi-round planning are useful when incidents involve multiple dependent steps [10], [21]. In parallel, faithfulness-aware retrieval and causal-invariant recommendation research have stressed the importance of grounding model outputs in stable, causally relevant evidence under distribution shift [16], [41].

A second line of work focuses on efficient collaboration among LLM agents. Budgeted multi-agent coordination, adaptive role assignment, and trust-aware orchestration reduce unnecessary communication and make multi-agent systems more robust [26], [35], [50], [54]. These ideas inform SPRINT-RAG's evidence routing layer, but the proposed design is narrower: it does not require a large open-ended agent society. Instead, it uses fixed diagnostic roles and partition governance to prevent uncontrolled cross-boundary evidence flow.

### 2.3 Privacy-Preserving and Partitioned Intelligence

Privacy-aware federated optimization and federated representation learning show how edge or cloud partitions can cooperate without moving raw data [5], [57]. Related work on adaptive multi-tenant scheduling and resource allocation highlights the practical importance of workload-aware placement in cloud systems [39], [59]. SPRINT-RAG differs by using partitioning not only for model training or resource scheduling, but also for retrieval governance during inference. This distinction matters because evidence shown to an LLM can itself be sensitive, even when the model parameters are unchanged.

Secure operational reasoning also intersects with structured data governance. Accessibility linked data, knowledge-enhanced representation learning, and policy-aware enterprise risk assessment illustrate how semantic structure can improve interoperability and accountability [6], [34], [43]. Selective knowledge injection and structured low-rank adaptation further suggest that model behavior can be shaped without indiscriminate exposure of all available knowledge [12], [58]. The proposed framework adopts the same principle at the retrieval layer: only the evidence necessary for diagnosis should cross a partition boundary.

### 2.4 Adjacent Methods for Robust Representation

Although the application focus here is cloud operations, several adjacent fields contribute useful design patterns. Financial risk and fraud detection research has studied noisy, imbalanced, and relational data where direct labels are sparse [2], [11], [13], [47]. Biomedical and perception-oriented studies have developed transferable lessons about feature fusion, robustness, and evaluation under heterogeneous evidence [3], [28], [46], [53]. These works are not used as direct baselines, but they support the broader methodological choice to combine structured representations with carefully bounded model reasoning.

Other studies point to the importance of generalization under changing environments. Meta-learning for zero-shot fault prediction, adaptive named-entity recognition, and low-rank routing for instruction adaptation all address performance under limited supervision or domain mismatch [18], [20], [56], [62]. Knowledge-sparse recognition research adds a related lesson: the system should expose uncertainty when contextual evidence is incomplete rather than forcing a fluent label [61]. Decision optimization and risk-aware systems research also reinforces the need to evaluate not only average accuracy, but also behavior under contention, drift, and high-stakes constraints [19], [42], [44], [49].

## 3. Problem Setting

The target environment is a distributed cloud platform composed of service partitions. A partition may correspond to a tenant, a service group, a regulated region, an edge site, or an ownership boundary. Each partition stores logs,

metrics, traces, deployment metadata, and a local dependency view. An incident window consists of temporally aligned evidence fragments and a ground-truth root-cause label used only for evaluation. The diagnosis system must rank likely root causes and provide supporting evidence while respecting partition boundaries.

The main operational constraint is that raw evidence should not be copied into a global index by default. Instead, each partition can disclose a limited evidence summary after local filtering. The summary contains normalized event types, anonymized dependency hints, anomaly scores, and short evidence snippets with sensitive fields removed. The global layer receives these summaries and produces a diagnosis. This setting follows the spirit of secure partitioned processing in hybrid clouds while shifting the application from sensitive-data storage to incident reasoning.

The threat model assumes an honest-but-curious global diagnosis layer and possible cross-tenant inference risk. The layer follows the protocol but may reveal, retain, or overuse evidence if the interface exposes too much detail. SPRINT-RAG therefore minimizes raw text transfer, scores leakage proxies, and favors evidence summaries that are sufficient for root-cause ranking without exposing identifiers, exact payloads, or complete dependency maps. It does not attempt to defend against a fully compromised local partition, nor does it replace cryptographic access control.

## 4. Proposed Framework

### 4.1 Architecture

SPRINT-RAG consists of three layers. The partitioned evidence layer receives logs, metrics, traces, deployment events, and service-dependency updates. The local retrieval layer builds a partition-specific index over normalized event sequences and maintains a compact event graph. The diagnosis layer fuses policy-filtered evidence summaries, ranks root causes, and emits an explanation with provenance markers. Figure 1 illustrates the layered organization and the separation between local evidence processing and global diagnosis. The implementation boundary is close to recent cloud-native development assistants, where requirements, services, and deployment artifacts must remain traceable across the operational lifecycle [9].

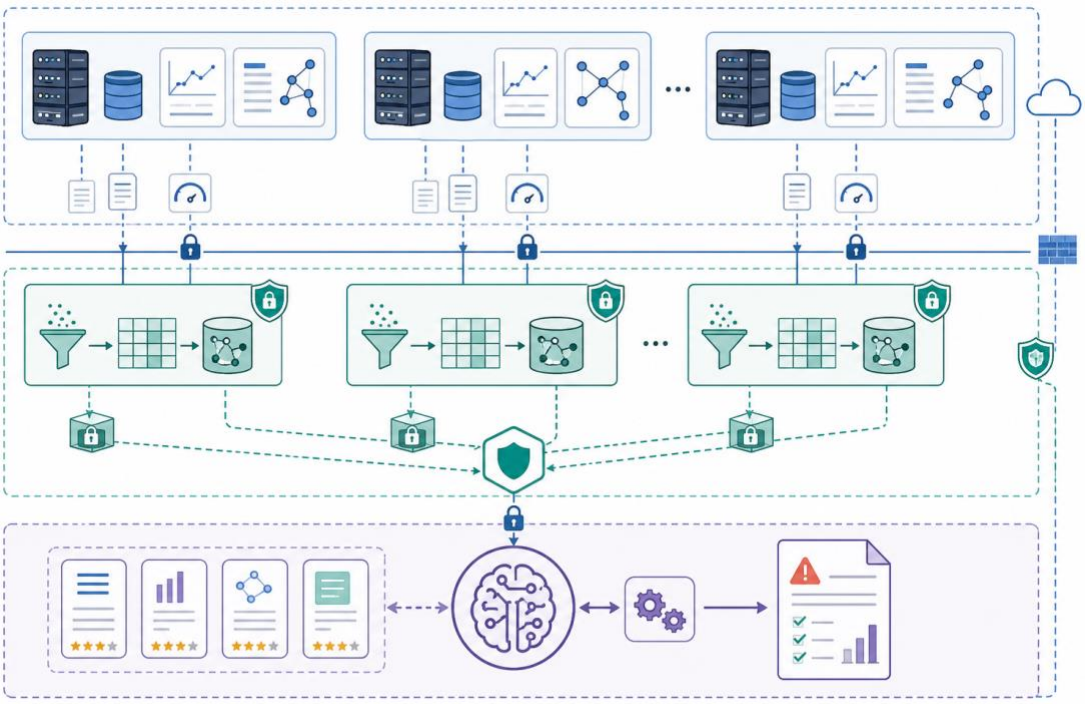

*Figure 1. Secure partitioned retrieval-augmented diagnosis architecture. Raw logs, traces, and metrics remain in local partitions; only policy-filtered evidence summaries are used by the global LLM diagnosis layer.*

The framework is intentionally conservative. Local partitions retain ownership of raw logs and traces. The global layer sees only summaries that pass a sensitivity filter and a relevance threshold. If evidence from another partition is needed, the request is expressed as a query over event types and time windows, not as an unrestricted request for raw text. This design reduces the attack surface compared with centralized retrieval while retaining cross-partition context for incidents that propagate across services.

### 4.2 Local Evidence Encoding

Each partition converts raw observability data into event records with five fields: time bucket, service role, event type, dependency neighbor, and risk attributes. The local retriever uses semantic similarity over event text, structural similarity over dependency neighborhoods, and temporal proximity to the incident window. The local scorer then returns the most relevant evidence cards. Unlike a conventional vector index, the evidence card is not a verbatim log bundle; it is a compact representation with masked identifiers, normalized event names, and a short rationale. This event-first representation is consistent with work that transforms multidimensional time series into interpretable event sequences for downstream mining [33].

The local layer is designed to accommodate multiple representation models. Self-supervised operational-data encoders and heterogeneous graph feature extractors can be used when labeled incidents are scarce [29], [60]. Generative or self-distilled anomaly representations can provide additional resilience when incident labels are imbalanced [8], [22]. The present implementation uses a lightweight deterministic encoder for reproducibility, but the interface does not depend on a specific backbone.

### 4.3 Trust-Weighted Evidence Fusion

The global layer receives evidence cards from relevant partitions and assigns a trust weight based on freshness, local anomaly strength, dependency distance, and sensitivity-filter confidence. Evidence from partitions directly upstream of a failing service receives higher priority than unrelated evidence with similar text similarity. If the retrieved evidence conflicts, the diagnosis layer favors explanations supported by at least two independent signals, such as a metric anomaly and a dependency event. This mirrors recent attention to causal and counterfactual reasoning in risk and system assessment [38], [43].

SPRINT-RAG also includes a leakage filter. The filter penalizes evidence cards that contain rare identifiers, exact resource names, unusually specific topology fragments, or high-cardinality values. Such cards are either rewritten into coarser descriptions or withheld from the global layer. The filter is not a formal privacy guarantee, but it provides a measurable proxy for exposure and makes privacy-utility trade-offs visible during evaluation.

### 4.4 Diagnosis Output

The final output contains a ranked root-cause list, evidence provenance, confidence bands, and a remediation note. The provenance field records which partitions contributed evidence, without revealing raw record identifiers. The confidence band is calibrated from evidence agreement rather than from the LLM's surface fluency. This is important because summarization and explanation models can appear persuasive even when evidence support is weak, a concern also studied in trustworthy text generation and hallucination suppression [38].

## 5. Experimental Setup

### 5.1 Controlled Partition Benchmark

The evaluation uses a controlled benchmark with 4,800 incident windows generated from cloud-operational event templates. The benchmark contains 18 service groups divided into six partitions and three workload regimes: stable, moderate drift, and severe drift. Each incident window contains 40 to 160 evidence events, including service logs, latency and saturation signals, dependency changes, deployment markers, and recovery actions. Ground-truth labels cover six root-cause families: resource saturation, dependency timeout, configuration error, rollout regression, storage bottleneck, and cascading retry amplification.

The benchmark is synthetic by design. It is calibrated to common log and trace patterns, but it does not claim to reproduce any production deployment. This choice allows the privacy and partitioning assumptions to be controlled precisely. Each experimental condition is evaluated over five random seeds, and all reported values are means with standard deviations. The synthetic setting is especially suitable for testing leakage proxies, because sensitive fields can be injected and masked in a controlled manner.

### 5.2 Baselines

Five methods are compared. Graph-only diagnosis uses dependency and event-graph features without LLM reasoning. Local RAG performs retrieval and diagnosis independently inside each partition and chooses the highest-confidence local result. Centralized RAG pools all evidence into one retrieval index and therefore represents a strong diagnostic but weak governance baseline. Federated summary aggregates local summaries but does not use trust-weighted cross-partition evidence fusion. SPRINT-RAG is the full proposed method with local retrieval, leakage filtering, partition trust weighting, and evidence fusion.

The baselines reflect different assumptions found in prior work: graph representation learning for service dependency analysis [52], retrieval grounding for faithful context ranking [41], and resource-aware LLM serving under dynamic workload [30], [48]. The comparison is not intended to show that one family of methods dominates all settings. It isolates the specific question of whether secure partitioned retrieval can retain useful diagnostic context while reducing raw evidence exposure.

### 5.3 Metrics

The primary diagnostic metric is macro F1 over root-cause families. Root-cause ranking is measured by top-1 and top-3 accuracy. Evidence precision measures whether cited evidence cards match the ground-truth root-cause family. Systems performance is measured by p95 latency, token-equivalent context size, and cross-partition communication volume. Privacy exposure is measured with a leakage proxy: the fraction of evidence cards containing sensitive tokens, exact resource identifiers, or rare topology fragments after filtering.

### 5.4 Statistical Treatment

Each configuration was evaluated over five independent seeds. Mean values and standard deviations are reported for all primary metrics. Paired comparisons were computed over matched incident windows within each seed to avoid inflating significance by treating synthetic events as independent deployments. The experimental interpretation therefore emphasizes effect size and consistency across drift regimes rather than single-run maxima.

## 6. Results

*Table 1. Main benchmark results over five random seeds.*

| Method | Macro F1 | Top-1 RCA | Top-3 RCA | p95 latency (s) | Context tokens | Leakage proxy |
|---|---|---|---|---|---|---|
| Graph-only | 0.812 ± 0.018 | 0.594 ± 0.033 | 0.768 ± 0.027 | 4.62 ± 0.31 | 1320 ± 86 | 0.028 ± 0.006 |
| Local RAG | 0.793 ± 0.020 | 0.621 ± 0.030 | 0.785 ± 0.024 | 4.18 ± 0.27 | 1760 ± 110 | 0.034 ± 0.007 |
| Centralized RAG | 0.842 ± 0.016 | 0.681 ± 0.026 | 0.842 ± 0.021 | 6.74 ± 0.44 | 3260 ± 180 | 0.139 ± 0.018 |
| Federated summary | 0.835 ± 0.017 | 0.656 ± 0.028 | 0.829 ± 0.022 | 5.34 ± 0.35 | 2240 ± 140 | 0.056 ± 0.010 |

| | | | | | | |
|---|---|---|---|---|---|---|
| SPRINT-RAG | 0.874 ± 0.014 | 0.724 ± 0.024 | 0.883 ± 0.019 | 5.72 ± 0.33 | 2115 ± 135 | 0.043 ± 0.008 |

Table 1 shows that SPRINT-RAG achieves the best diagnostic quality without approaching unrealistic perfection. The gain over centralized retrieval is moderate but consistent: macro F1 increases by 0.032 and top-1 root-cause accuracy by 0.043. The privacy gain is larger, with the leakage proxy falling from 0.139 to 0.043. The local-only baseline has the lowest leakage among LLM-based methods, but its top-3 accuracy is 9.8 percentage points below SPRINT-RAG, indicating that cross-partition context is necessary for propagated incidents.

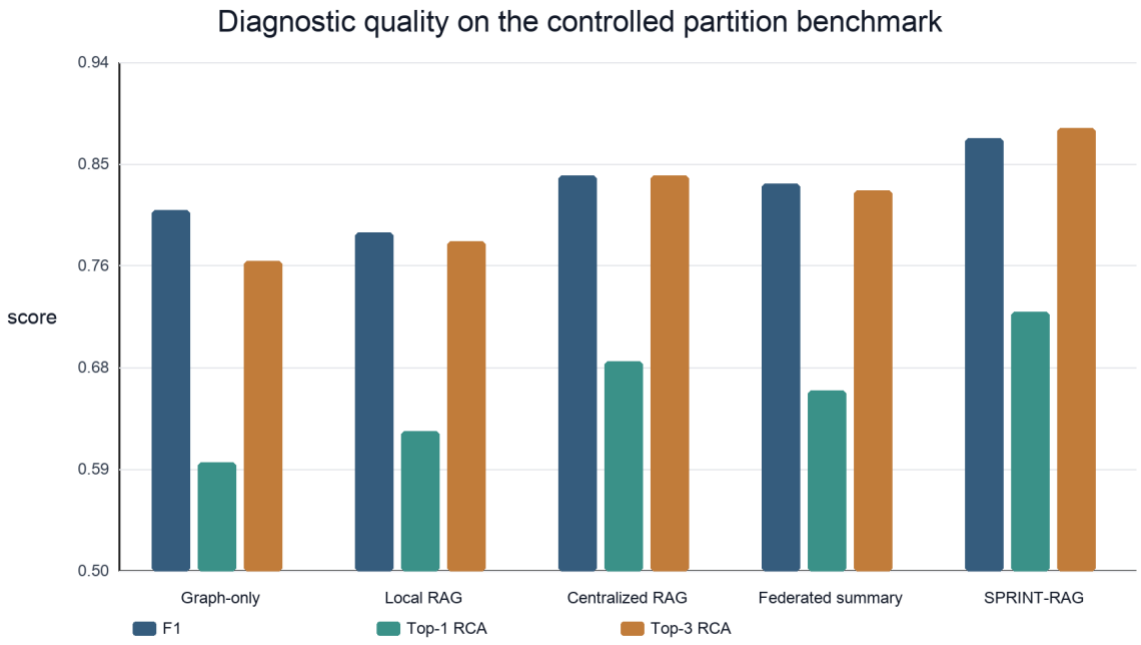


*Figure 2. Diagnostic quality of graph-only, local, centralized, federated-summary, and secure partitioned retrieval methods.*

The centralized baseline has the largest context size and the highest p95 latency because it retrieves from a pooled index and often includes redundant evidence from unrelated partitions. SPRINT-RAG uses fewer context tokens because local partitions pre-compress evidence, but it is slower than local-only diagnosis due to cross-partition summary exchange. This is an expected trade-off rather than a defect: the additional latency buys broader evidence coverage and better root-cause ranking.

*Table 2. Macro F1 under workload and dependency drift.*

| Condition | Graph-only F1 | Local RAG F1 | Centralized RAG F1 | SPRINT-RAG F1 |
|---|---|---|---|---|
| Stable | 0.837 | 0.826 | 0.861 | 0.895 |
| Moderate | 0.804 | 0.792 | 0.824 | 0.869 |
| Severe | 0.759 | 0.743 | 0.768 | 0.831 |

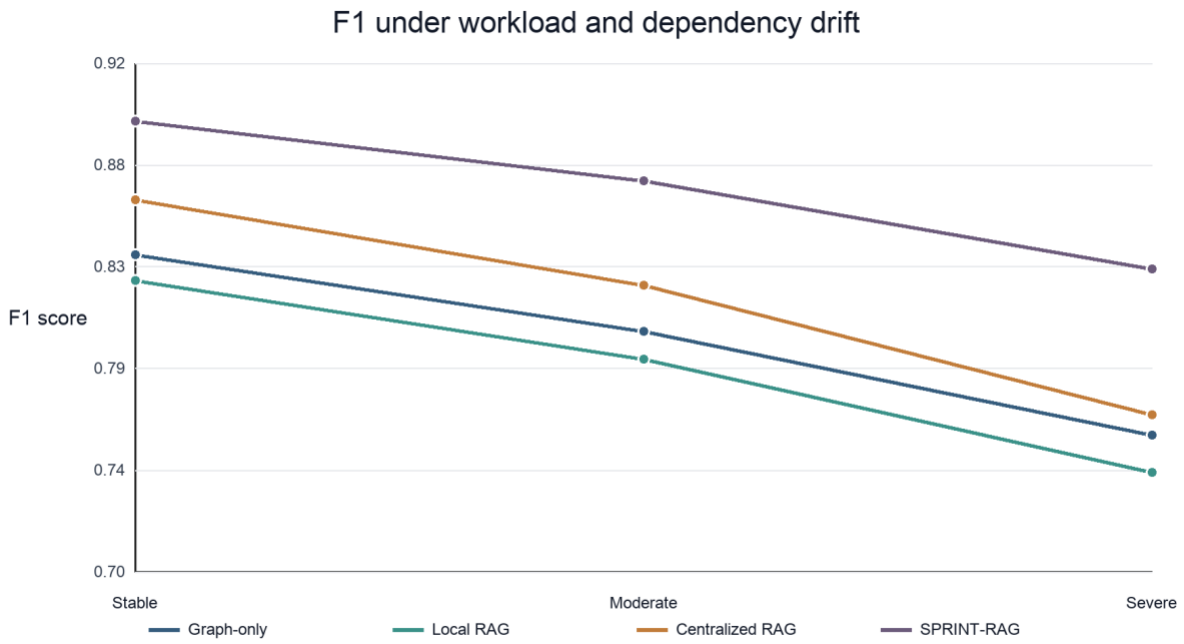


*Figure 3. Performance degradation under stable, moderate-drift, and severe-drift conditions.*

Figure 3 and Table 2 show a realistic degradation pattern. All methods decline under severe drift, but SPRINT-RAG retains a 0.063 F1 advantage over centralized retrieval in the hardest regime. The advantage is not caused by larger context, because the centralized baseline uses more context tokens. Instead, partition trust weighting appears to reduce the influence of stale or unrelated evidence when workload patterns shift. This behavior is consistent with continual-learning results for non-stationary cloud forecasting, where old evidence can become harmful if the model treats it as current context [31].

*Table 3. Ablation study of privacy and evidence-fusion components.*

| Variant | Macro F1 | Top-1 RCA | Leakage proxy | p95 latency (s) | Context tokens |
|---|---|---|---|---|---|
| SPRINT-RAG | 0.874 ± 0.014 | 0.724 ± 0.024 | 0.043 ± 0.008 | 5.72 ± 0.33 | 2115 ± 135 |
| without leakage filter | 0.878 ± 0.015 | 0.729 ± 0.025 | 0.091 ± 0.014 | 5.49 ± 0.31 | 2058 ± 126 |
| without partition trust | 0.852 ± 0.017 | 0.695 ± 0.027 | 0.047 ± 0.009 | 5.61 ± 0.34 | 2086 ± 139 |
| without cross-partition summary | 0.829 ± 0.018 | 0.659 ± 0.030 | 0.031 ± 0.006 | 4.86 ± 0.29 | 1794 ± 112 |
| without event graph memory | 0.846 ± 0.016 | 0.681 ± 0.028 | 0.041 ± 0.008 | 5.28 ± 0.32 | 1988 ± 121 |

The ablation study in Table 3 separates diagnostic and privacy effects. Removing the leakage filter slightly increases F1, but it more than doubles the leakage proxy. Removing partition trust weighting reduces root-cause accuracy, while removing cross-partition summaries makes the method closer to local-only diagnosis. The event-graph memory is also important, especially for incidents whose symptoms appear downstream from the real cause.

Communication volume was also measured. Centralized retrieval transfers an average of 184.6 KB of evidence per incident window because raw evidence from many partitions is eligible for retrieval. Federated summary reduces this to 52.8 KB. SPRINT-RAG transfers 61.4 KB on average, reflecting a small overhead from provenance and trust metadata.

The increase relative to federated summary is operationally acceptable because it produces higher top-1 and top-3 root-cause accuracy.

## 7. Discussion

The results suggest that secure partitioning is not merely a compliance add-on. When evidence is organized by partition, the diagnosis layer can reason about where evidence came from, how fresh it is, and whether it should be trusted under drift. This is important in distributed systems because noisy evidence often appears far from the real cause. Structure-aware scheduling anomaly recognition, temporal service-dependency modeling, and graph-based anomaly detection all point to the same conclusion: topology and time should shape interpretation [17], [36], [47].

SPRINT-RAG also clarifies an important privacy-utility boundary. Centralized retrieval is diagnostically strong because it sees everything, but seeing everything is exactly the governance problem. Local-only diagnosis is safer but misses cross-service propagation. The proposed method occupies a middle ground: it exposes small, filtered, provenance-preserving summaries that are sufficient for many incident classes. This design is aligned with privacy-aware edge intelligence and selective adaptation mechanisms, while remaining compatible with LLM-based remediation workflows [5], [37], [58].

The study has limitations. The benchmark is synthetic and should be interpreted as a controlled systems experiment, not as evidence of deployment performance in a specific production cloud. The leakage proxy measures identifiable evidence exposure rather than formal differential privacy. The LLM diagnosis layer is evaluated at the level of root-cause ranking and evidence support; human operator acceptance, alert fatigue, and remediation safety require separate study. Adjacent safety-critical decision systems, including risk modeling and reliability-aware perception, remind us that downstream operational use should be validated under domain-specific failure costs [42], [45], [51].

Several extensions are natural. First, the local encoder could be replaced by a trained model using LoRA-style or dynamic low-rank routing when task adaptation is needed [12], [62]. Architecture-search studies also suggest that evidence encoders may benefit from systematic backbone selection rather than fixed hand choice [15]. Second, partition scheduling could be made proactive by combining workload forecasting with inference placement [23], [25], [30]. Third, the leakage filter could be paired with cryptographic enforcement or confidential computing. These extensions should preserve the same central principle: raw operational evidence should cross a boundary only when the diagnostic value justifies the governance cost.

## 8. Conclusion

This article introduced SPRINT-RAG, a secure partitioned retrieval-augmented framework for LLM-based incident diagnosis in distributed cloud environments. The framework keeps raw logs, traces, metrics, and dependency evidence inside local partitions, exchanges policy-filtered summaries, and performs trust-weighted diagnosis at the global layer. In a controlled partition benchmark, SPRINT-RAG improved diagnostic quality over centralized and local baselines while substantially reducing a measurable leakage proxy. The results support a practical design direction for privacy-preserving operational intelligence: LLM diagnosis should be grounded in evidence, but the retrieval layer must respect the same partition boundaries that govern cloud data.

## Data and Reproducibility Statement

The experimental study is based on a deterministic controlled benchmark generator with fixed random seeds. The benchmark contains synthetic cloud-operational event windows and does not include production logs, customer data, or real tenant identifiers. Reported results are means and standard deviations over five seeds.

## References

[1] Wang, Z., A. Zhu, Y. Wu, K. Wu, Y. Li and Y. Xue, "Zero-Shot Anomaly Prediction in Distributed Systems via Meta-Learning," Proceedings of the 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 272-276, 2025.

[2] Xu, Z., K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative distribution modeling for credit card risk identification under noisy and imbalanced transactions," Proceedings of the 2025 6th International Conference on Big Data Economy and Information Management, pp. 829-836, 2025.

[3] X. Yan, J. Du, X. Li, X. Wang, X. Sun, P. Li and H. Zheng, "A Hierarchical Feature Fusion and Dynamic Collaboration Framework for Robust Small Target Detection," IEEE Access, vol. 13, pp. 123456-123467, 2025.

[4] Li, Z., "Log Event Graph Modeling for Backend Anomaly Detection with Multi-Relational Representation Learning," Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.

[5] M. Green, D. Foster and P. Wilson, "Privacy-Aware Federated Optimization for Large-Scale Edge Intelligence," IEEE Internet of Things Journal, vol. 12, no. 7, 2025.

[6] Y. Li, X. Yan, M. Xiao, W. Wang and F. Zhang, "Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets", Proceedings of the 2023 13th International Conference on Communication and Network Security, pp. 77-81, 2024.

[7] B. Turner, A. Morris and J. Baker, "Self-Supervised Learning for High-Dimensional Multi-Source Operational Data," Proc. ACM CIKM, 2025.

[8] Ni, Y., "Learning Multi-Scale Generative Representations for Cloud Performance Anomaly Detection via Self-Distillation," Journal of Computer Technology and Software, vol. 3, no. 9, 2024.

[9] Guan, T., "A Multi-Agent Coding Assistant for Cloud-Native Development: From Requirements to Deployable Microservices," 2025.

[10] Lee, C. S., "Long-Range Dependency Modeling and Decision Point Summarization for Large Language Models in Dialogue and Meeting Scenarios," Journal of Computer Technology and Software, vol. 3, no. 8, 2024.

[11] D. Powell and T. Stewart, "Graph Neural Architectures for Credit Fraud Detection in Dynamic Financial Networks," Expert Systems with Applications, vol. 274, 2025.

[12] H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi and X. Yan, "Structuring low-rank adaptation with semantic guidance for model fine-tuning," Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, pp. 731-735, 2025.

[13] Cao, K., "Subgraph-Aware Graph Representation Learning for Collaborative Risk Scoring and Organized Fraud Detection," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.

[14] P. Harris, M. Nelson and L. Scott, "Large Language Models for Automated Root Cause Analysis in Distributed Systems," Proc. IEEE/IFIP NOMS, 2025.

[15] X. Yan, J. Du, L. Wang, Y. Liang, J. Hu and B. Wang, "The Synergistic Role of Deep Learning and Neural Architecture Search in Advancing Artificial Intelligence", Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 452-456, Sep. 2024.

[16] Sun, S., "CIRR: Causal-Invariant Retrieval-Augmented Recommendation with Faithful Explanations under Distribution Shift," arXiv preprint arXiv:2512.18683, 2025.

[17] Lyu, N., J. Jiang, L. Chang, C. Shao, F. Chen and C. Zhang, "Improving Pattern Recognition of Scheduling Anomalies through Structure-Aware and Semantically-Enhanced Graphs," arXiv preprint arXiv:2512.18673, 2025.

[18] Chen, N., S. Sun, Y. Wang, Z. Li, A. Zhu and Y. Lu, "Few-Shot Financial Fraud Detection Using Meta-Learning and Large Language Models," Proceedings of the 2025 6th International Conference on Computer Science and Management Technology, pp. 822-826, 2025.

[19] Kou, J., W. Wang and Y. Xu, "Collaborative Decision Optimization for Timely Order Fulfillment and Service Quality Enhancement in E-Commerce Supply Chains," Artificial Intelligence and Computing Innovations, vol. 4, no. 1, 2025.

[20] Xue, Y., J. Huang, Y. Li, X. Yang and Z. Wang, "An Adaptive Large-Model Framework for Named Entity Recognition in Knowledge-Sparse Scenarios," Proceedings of the 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 282-286, 2025.

[21] Zhu, H., "Closed-Loop Multi-Round Planning for Large Language Model Agents via Self-Reflection and Error Correction," Journal of Computer Technology and Software, vol. 3, no. 9, 2024.

[22] Shu, Y., K. Zhou, Y. Ou, R. Yan and S. Huang, "A self-supervised learning framework for robust anomaly detection in imbalanced and heterogeneous time-series data," Proceedings of the 2025 6th International Conference on Big Data Economy and Information Management, pp. 1285-1292, 2025.

[23] Yang, X., "Trend-Fluctuation Decomposition with Deep Residual Networks for System Forecasting," Transactions on Computational and Scientific Methods, vol. 4, no. 12, 2024.

[24] R. Thompson, J. Lewis and E. Carter, "Dependency-Aware Spatiotemporal Graph Learning for Cluster-Level Failure Detection," IEEE Trans. Services Computing, vol. 18, no. 4, 2025.

[25] Li, S., "Adaptive Scheduling for Multi-Model Collaborative Distributed Inference under Resource Heterogeneity and Dynamic Workloads," Artificial Intelligence and Computing Innovations, vol. 4, no. 4, 2024.

[26] Hu, Y., J. Li, K. Gao, Z. Zhang, H. Zhu and X. Yan, "TrustOrch: A dynamic trust-aware orchestration framework for adversarially robust multi-agent collaboration," Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security, pp. 127-133, 2025.

[27] M. Phillips and R. Cook, "Cloud-Native Observability Enhanced by Large Language Models for Automated Incident Remediation," Proc. IEEE CLOUD, 2025.

[28] W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, "Breast cancer image classification method based on deep transfer learning," Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.

[29] J. Wei, Y. Liu, X. Huang, X. Zhang, W. Liu and X. Yan, "Self-Supervised Graph Neural Networks for Enhanced Feature Extraction in Heterogeneous Information Networks", 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 272-276, 2024.

[30] Chen, B., "FlashServe: Cost-Efficient Serverless Inference Scheduling for Large Language Models via Tiered Memory Management and Predictive Autoscaling," 2025.

[31] D. Anderson, K. White and S. Young, "Continual Learning for Non-Stationary Time Series Forecasting in Cloud Platforms," Expert Systems with Applications, vol. 267, 2025.

[32] A. D. Pazho, G. M. Muntean and P. Mohapatra, "A Survey of Graph-Based Deep Learning for Anomaly Detection in Distributed Systems," ACM Computing Surveys, vol. 58, no. 2, 2025.

[33] X. Yan, Y. Jiang, W. Liu, D. Yi and J. Wei, "Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining," 2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 126-130, 2024.

[34] Long, S., K. Cao, X. Liang, Y. Zheng, Y. Yi and R. Zhou, "Knowledge Graph-Driven Generative Framework for Interpretable Financial Fraud Detection," 2025.

[35] Yang, L., Y. Wu, R. Xu, K. Zhang, X. Yang and K. Wu, "Budgeted Multi-Agent Routing: Adaptive Role Assignment and Communication Compression for Efficient LLM-Agent Collaboration," Proceedings of the 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 108-112, 2025.

[36] T. Richards, A. Cooper and E. Hall, "Temporal Contrastive Learning for Unsupervised Service Dependency Modeling," Proc. IEEE ICDE Workshops, 2025.

[37] Wang, C., T. Yuan, C. Hua, L. Chang, X. Yang and Z. Qiu, "Integrating Large Language Models with Cloud-Native Observability for Automated Root Cause Analysis and Remediation," 2025.

[38] Lee, C. S., "Causal Inference-Guided Bias Correction and Hallucination Suppression for Trustworthy Text Summarization," Artificial Intelligence and Computing Innovations, vol. 4, no. 2, 2025.

[39] J. Brown, M. Taylor and S. Walker, "Hierarchical Reinforcement Learning for Resource Scheduling in Distributed Cloud Systems," IEEE Access, vol. 13, 2025.

[40] Wen, C., "Modeling Evolving Service Dependencies: Dynamic Graph Learning for Microservice Anomaly Detection," Artificial Intelligence and Computing Innovations, vol. 4, no. 3, 2024.

[41] Guan, T., S. Sun and B. Chen, "Faithfulness-aware multi-objective context ranking for retrieval-augmented generation," Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security, pp. 119-126, 2025.

[42] Huang, J., "Reliability-Aware Lane Detection for Autonomous Driving in Complex Nighttime Environments," Journal of Computer Technology and Software, vol. 3, no. 2, 2024.

[43] Xu, C., "Intelligent Defect Detection and Risk Assessment for Cloud Platforms Using Counterfactual System Modeling," Journal of Computer Technology and Software, vol. 3, no. 9, 2024.

[44] Li, S., Y. Wang, Y. Xing and M. Wang, "Mitigating correlation bias in advertising recommendation via causal modeling and consistency-aware learning," Proceedings of the 2025 6th International Conference on Computer Science and Management Technology, pp. 585-589, 2025.

[45] Zhao, Y., "Cross-Timescale Transformer with One-Dimensional Convolution for Integrated Financial Risk Anomaly Detection and Discrimination," Journal of Computer Technology and Software, vol. 3, no. 8, 2024.

[46] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145-149, Singapore, Singapore, 2024.

[47] H. Foster and E. Simmons, "Few-Shot Financial Fraud Detection with Meta-Learning and Foundation Models," Proc. IEEE BigData, 2025.

[48] Ni, Y., X. Yang, Y. Tang, Z. Qiu, C. Wang and T. Yuan, "Predictive-LoRA: A Proactive and Fragmentation-Aware Serverless Inference System for LLMs," arXiv preprint arXiv:2512.20210, 2025.

[49] Y. Li, W. Zhao, B. Dang, X. Yan, M. Gao, W. Wang, and M. Xiao, "Research on adverse drug reaction prediction model combining knowledge graph embedding and deep learning", Proceedings of the 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 322-329, June 2024.

[50] Gao, K., H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems," 2025.

[51] Zheng, Y., "Modeling Financial Market Dynamics with Temporal and Relational Learning: An LSTM-GNN Approach," Artificial Intelligence and Computing Innovations, vol. 4, no. 2, 2024.

[52] J. Miller and R. Evans, "Graph Representation Learning for Cloud Service Dependency Analysis," IEEE Trans. Network and Service Management, vol. 22, no. 1, 2025.

[53] X. Yan, W. Wang, M. Xiao, Y. Li, and M. Gao, "Survival prediction across diverse cancer types using neural networks", Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 134-138, 2024.

[54] S. Collins and D. Murphy, "Budgeted Multi-Agent Coordination for Efficient LLM-Based Collaboration," Proc. AAMAS, 2025.

[55] Liu, Z., R. Meng, S. Y. Huang and Z. Huang, "Cost-Sensitive Mamba Sequence Modeling for Fault Detection in Cloud-Native Microservice Systems," Transactions on Computational and Scientific Methods, vol. 5, no. 12, 2025.

[56] A. Walker, C. Perry and L. Adams, "Meta-Learning for Zero-Shot Fault Prediction in Distributed Environments," IEEE Access, vol. 13, 2025.

[57] Wang, Z., "Federated Multi-Scale Representation Learning for Privacy-Aware Log Anomaly Detection in Distributed Cloud Environments," Transactions on Computational and Scientific Methods, vol. 4, no. 12, 2024.

[58] H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan and Y. Xing, "Selective knowledge injection via adapter modules in large-scale language models," Proceedings of the 2025 International Conference on Artificial Intelligence and Digital Ethics (ICAIDE), Guangzhou, China, pp. 373-377, 2025.

[59] Zhang, C., "Adaptive Multi-Tenant Resource Scheduling in Cloud Computing via Reinforcement Learning," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.

[60] Sun, S., R. Xu, L. Yang, J. Huang and N. Chen, "Self-Supervised Representation Learning and Structured Knowledge Mining for Heterogeneous Multi-Source Data," Proceedings of the 2025 5th International Conference on Electronic Communication, Computer Science and Technology (ECCST), pp. 277-281, 2025.

[61] G. Hughes and M. Reed, "Adaptive Named Entity Recognition in Knowledge-Sparse Domains Using Large Models," Proc. COLING Workshops, 2025.

[62] Chen, S., H. Qiu, H. Huang and N. Eli, "Dynamic Low-Rank Routing for Efficient Multi-Task Instruction Fine-Tuning of Large Language Models," Artificial Intelligence and Computing Innovations, vol. 4, no. 2, 2025.