MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1132-1136

A Privacy-Preserving Framework for Agentic AI Orchestration Over Cloud-Native Data Lakehouse Environments Using Medallion Architecture

S.Babu[1], M.V.L.S. Vasudeava Sai
[1]*Department of CSA & SCSVMV University, India*
[2]*Department of CSA & SCSVMV University, India*
[1]babulingaa@kanchiuniv.ac.in; [2]vasu_sai@yahoo.com

Abstract— *Cloud-native data platforms built on the Medallion Architecture (Databricks, 2024), together with storage layers such as Microsoft OneLake (Microsoft Corporation, 2025), have changed how enterprises run analytics. In practice, this shift means that sensitive material, including application logs, source code, telemetry, and personal data, ends up in the Bronze layer, where Agentic AI and LLM-driven pipelines can access it. When those pipelines forward such content to cloud-hosted LLMs, confidentiality, integrity, and compliance are all put at risk, and the existing literature does not offer a complete answer to the problem (Xu et al., 2025). The research work proposes a framework that coordinates Agentic AI activity across the Bronze, Silver, and Gold layers in a privacy-aware manner. It targets the issues that arise most often in practice: telemetry exposure, source code leakage, inference of usage patterns from repeated agent calls, and drift away from GDPR, CCPA, and the EU AI Act (European Parliament & Council, 2024). The framework adds Differential Privacy (Dwork & Roth, 2014) to RAG pipelines (Lewis et al., 2020), automatically masks and tokenizes data at the Silver layer, routes prompts through a sensitivity-aware hybrid LLM gateway at the Gold layer, and runs governance and audit checks across every layer. We tested the design on a Microsoft Fabric / OneLake deployment (Microsoft Corporation, 2025) using simulated enterprise data that included PII, source code snippets, and network telemetry. RAG retrieval delays remained small, utility stayed within the bounds set by the Differential Privacy budget, and the audit trail met the expectations of GDPR and the EU AI Act for high-risk systems. Given that fewer than half a dozen studies directly examine privacy in LLM-agent systems running over enterprise lakehouses (Xu et al., 2025), this work aims to close a real gap and offer a blueprint that regulated industries can deploy.*

Keywords— *Agentic AI, Data Lakehouse, Privacy-Preserving Architecture, Medallion Architecture, Differential Privacy, Retrieval-Augmented Generation, Hybrid LLM Routing, GDPR, EU AI Act 2024.*

I. INTRODUCTION

Most enterprises have moved away from traditional data warehouses and now rely on cloud-native Data Lakehouses, with platforms such as Microsoft Fabric (Microsoft Corporation, 2025), Databricks (Databricks, 2024), and AWS Lake Formation leading the way. These innovative platforms leverage open formats such as Delta Lake, Apache Iceberg, and Parquet to consolidate analytics and AI workloads onto a single, efficient storage layer. They structure data using the Medallion Architecture, with raw data in Bronze, cleansed data in Silver, and curated data in Gold (Databricks, 2024; Microsoft Corporation, 2025). As Agentic AI systems continue to develop, taking on complex reasoning and autonomous tasks (Bain & Company, 2025), LLM agents are now working across all three layers, even accessing the raw Bronze data, which often includes sensitive information. While this integration greatly boosts productivity, it also broadens the security challenges, as traditional database protections are not designed to safeguard this expanded attack surface. Agent prompts now routinely include raw query results, embedded telemetry, source code, and personally identifiable information, any of which may end up at an externally hosted LLM. The privacy exposure here is not a thought experiment-it is something organizations actually face the moment they put Agentic AI on top of a lakehouse.

A. Problem Statement

When an Agentic AI system sends context, prompts, or data payloads to an external LLM API, the content leaving the system can include logs, network packets, source code, or PII. Commercial LLMs offer only limited confidentiality controls, and the agent layer that sits between the lakehouse and the model has historically been treated as a trusted relay rather than a place where privacy is enforced. The Four concrete problems follow from this are

- Operational logs and telemetry leak to third-party model providers.
- Proprietary transformation logic, query templates, and source code travel inside prompts.
- Repeated agent interactions enable attackers to infer enterprise data usage patterns.
- Cross-border data transfers to LLM endpoints occur without controls, creating regulatory exposure.

B. The importance of this issue

GDPR, CCPA, and the EU AI Act 2024 all emphasize the importance of traceable data handling, data minimization, and the right to erasure-three key aspects that many LLM operations might find challenging to meet on their own (European Parliament & Council, 2024; Information Commissioner's Office, 2024). By creating privacy-preserving orchestration, organizations can meet these requirements, safeguard their intellectual property, and foster trust with users as they embrace Agentic AI. In high-risk sectors such as healthcare, finance, defense, and public administration, the absence of a framework like this is itself a reason adoption stalls.

C. *Limitations* of Existing Approaches

Current research and practice reveal three key gaps. First, privacy-enhancing technologies such as Differential Privacy (Dwork & Roth, 2014; Fioretto & Van Hentenryck, 2025), pseudonymization, and tokenization are often studied and applied in isolation rather than integrated into a seamless enterprise pipeline. Second, Differential Privacy methods originally designed for static datasets don't easily adapt to RAG-based generative workflows (Lewis et al., 2020) because each step—retrieval, augmentation, and generation—can introduce its own risk of data leakage (Kibriya et al., 2024). Third, the Medallion Architecture (Databricks, 2024), although widely adopted, lacks built-in privacy controls at the agent-to-LLM interface; existing safeguards tend to focus on data storage and transfer, not on what an autonomous agent might send to a model (Xu et al., 2025). The Proposed work aims to bridge those gaps with a comprehensive privacy architecture that works smoothly across all Medallion layers. It considers each layer as a separate trust boundary and employs dedicated agents whose only purpose is to enforce privacy.

Literature Review

The literature relevant to this work spans five overlapping areas: privacy in LLM-agent systems, privacy-enhancing techniques for cloud analytics, Agentic AI architecture, Medallion-based data platforms, and AI governance. Each area is thoughtfully developed on its own, but the intersection of privacy-preserving Agentic AI on top of an enterprise Lakehouse is an exciting area that still has many opportunities for exploration. The review below highlights the most relevant contributions and points out where this work expands or differs from each.

A. Privacy in LLM-Agent Systems.

Xu et al. (2025), in their survey "On Protecting the Data Privacy of Large Language Models" (High-Performance Computing and Communications), examine privacy threats to LLMs across training, fine-tuning, and inference, and focus on risks specific to agents. Out of more than 130 papers they reviewed, only six actually deal with privacy in LLM-agent systems. The survey proposes a taxonomy that separates passive data leakage-exposure through prompts and outputs from active inference attacks, which include membership inference, attribute inference, and prompt-extraction attacks. That work gives us the problem framing, but it does not move into architectural mitigations. The framework presented here picks up where the survey leaves off, mapping each threat class onto a specific Medallion layer and pairing it with an agent-level countermeasure.

B. Privacy Threats in LLM and RAG Workflows

Kibriya et al. (2024), "Privacy Concerns in Large Language Models: Training and Inference Phase Analysis" (IEEE Access), separate LLM privacy issues into training- and inference-phase concerns, examine re-identification risks, and list the main vectors through which PII tends to leak. One of their key observations is that inference-phase risks remain under-researched relative to training-phase risks (see also Carlini et al., 2021), and that RAG pipelines (Lewis et al., 2020) introduce new leakage surfaces-embedding-store inversion, retrieval-context exposure, and augmentation-prompt persistence-that classical privacy techniques do not adequately address. That observation is exactly what motivates the Silver-layer privacy-enforcing transformation agent in this work, which sits in front of the augmentation step and cleans retrieval results before they reach it.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1132-1136

C. **Privacy-Preserving Techniques for Cloud Analytics**

Fioretto and Van Hentenryck (2025), "Differential Privacy in Artificial Intelligence: From Theory to Practice" (Foundations and Trends in Privacy and Security), walk through Differential Privacy mechanisms across the machine-learning lifecycle, including LLMs. Building on the foundational formalism of Differential Privacy established by Dwork and Roth (2014), they flag three recurring problems: there is no universal privacy–utility metric for generative AI; static-dataset DP guarantees do not transfer cleanly to dynamic RAG retrieval (Lewis et al., 2020); and DP becomes expensive quickly at petabyte scale. The proposed work engages directly with the first of those-the utility–privacy calibration problem-by deriving a noise budget aimed specifically at retrieval-augmented generation rather than at classification or regression.

D. **Agentic AI Architecture and Enterprise Data Platforms**

Bain & Company (2025), "State of the Art of Agentic AI Transformation" (Bain Technology Report 2025), describes four maturity levels for Agentic AI, from simple LLM information retrieval at Level 1 up to multi-agent networks running autonomous business processes at Level 4. Their finding worth highlighting is that the main barrier to enterprise adoption is not model capability-it is data architecture. Agents need business context, business context lives in enterprise data, and at Levels 3 and 4 the conversation gets dominated by privacy, security, and intellectual-property concerns. That makes the framework presented here directly relevant to cloud, business intelligence, and data platform architects working in real organizations.

E. **Regulatory Compliance and AI Governance**

The European Data Protection Board (2025), in its technical report "AI Privacy Risks and Mitigations in Large Language Models," offers insightful perspectives on GDPR principles, such as data minimization, purpose limitation, the right to erasure, and data-subject rights. It highlights that respecting the right to erasure can be especially challenging.. Once personal data has been used in fine-tuning or repeated prompting, full removal usually requires retraining the model (see also Carlini et al., 2021). The report makes clear why privacy-preserving architectures are needed now and provides concrete compliance criteria against which this framework can be checked. Complementary practical guidance is provided by the Information Commissioner's Office (2024).

The European Parliament and Council's Regulation (EU) 2024/1689, known as the AI Act (European Parliament & Council, 2024), sets out a clear risk-based framework designed to ensure safety and accountability. It requires high-risk systems to meet strict standards such as transparency, data governance, human oversight, and conformity assessment. When it comes to agentic AI systems that manage sensitive enterprise data, they are classified as high-risk under Annex III, emphasizing the importance of maintaining thorough documented data governance and audit trails. This ensures robust oversight and helps build trust in these advanced technologies. This work treats those obligations as design requirements from the start, rather than as bolt-ons after the fact.

II. PROPOSED METHODOLOGY

Taken together, the literature points to three things. Privacy threats related to LLM agents are increasing, and while the technologies to protect privacy are already quite advanced, they are often used separately rather than together. Also, Medallion-based lakehouses, which are quite helpful, still don't currently provide privacy protections at the crucial boundary between the agent and the LLM.. The framework proposed here pulls these strands together into a single architecture-one that can run on existing cloud platforms and that aligns with the regulations governing them.

A. *System Architecture*

The Privacy-Preserving Framework for Agentic AI Orchestration is designed as a flexible, layered system that aligns closely with the Medallion model. Each Medallion layer has its own privacy-enforcement agent, and a governance layer oversees the entire process. Figure 1 provides an overview, and the following subsections will guide you through each part.

B. *Architectural Componenets*

1) Bronze Layer-Sensitivity Classification Agent

The Bronze layer is the initial point where raw enterprise data enters. This includes application logs, network telemetry, source-code repositories, and unstructured user content, all waiting to be processed further. The Sensitivity Classification Agent runs as soon as data lands. It operates in two stages. The first is deterministic: regular expressions and pattern detectors flag national identifiers, payment card numbers, secrets, API keys, and source code signatures. The second is contextual: a small on-premises LLM examines the surrounding context to disambiguate fields that the rules alone cannot resolve. The output is a sensitivity tag (public, internal, confidential, or restricted) attached to each record, which the rest of the layers consume. The key point is that this agent never sends raw content outside the Bronze boundary. Its model runs in-tenant, so even detecting sensitive content does not itself become a privacy event.

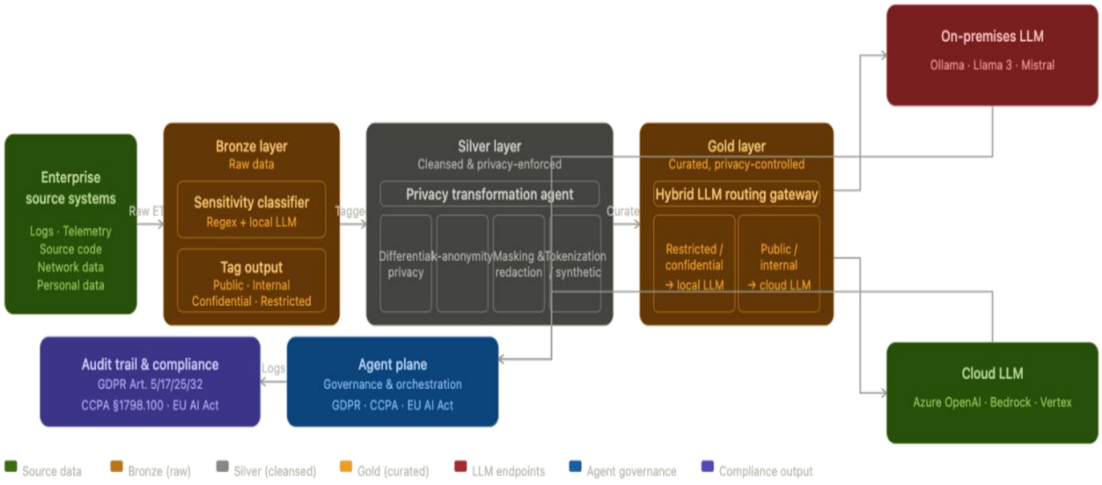

Fig 1: Proposed Privacy-Preserving Framework for Agentic AI Orchestration over Medallion Lakehouse

2) Silver Layer-Privacy-Enforcing Transformation Agent

The Silver layer transforms Bronze data into a validated, conformed, query-ready form. The Privacy-Enforcing Transformation Agent extends the standard Silver ELT pipeline with four privacy operations, applied based on the sensitivity tags assigned at Bronze:

- **Differential Privacy**. Calibrated noise is added to numerical aggregates, to embeddings used in RAG, and to counts that drive downstream analytics, following the formal framework introduced by Dwork and Roth (2014). Each dataset has its own noise budget, with stricter budgets for restricted data.
- **k-Anonymity and Generalization**. Quasi-identifiers are generalized so that any released record cannot be distinguished from at least k − 1 others, blunting linkage attacks against the Silver-layer feature store (Sweeney, 2002).
- **Pseudonymization**. Direct identifiers are swapped for reversible tokens. The mapping resides in a hardware-protected key vault and can be accessed only through governed re-identification workflows.
- **Secure Tokenization**. Free-text fields are scanned for residual PII, and any PII found is replaced inline with format-preserving tokens before any agent or downstream model can see it.

Every record carries a privacy manifest as it leaves this layer. The manifest records which transformations were applied, how much budget was used, and which sensitivity rules triggered each action. That manifest is the main feed into the governance plane.

3) Gold Layer-Hybrid LLM Routing Gateway

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1132-1136

The Gold layer holds curated, business-ready datasets and is where analytical agents and end-user copilots spend most of their time. The Hybrid LLM Routing Gateway checks every prompt before it's sent to an LLM, making sure to consider both the prompt's content and the sensitivity of the underlying data. Depending on the routing policies, it directs the request to one of two destinations. If the prompt involves restricted or confidential data, or cannot be sufficiently masked without losing its analytical value, it goes to an on-premise or private LLM—like an Ollama-hosted Llama 3 or Mistral 7B inside the tenant. If the prompt has been fully masked or tokenized, or if it only produces aggregated, differentially private results, it is sent to a cloud-hosted LLM such as Azure OpenAI, Amazon Bedrock, or Google Vertex AI. Every routing decision is logged—including the prompt hash, the matching policy rule, and the destination model—so that an auditor can easily verify that nothing restricted ever leaves the tenant boundary.

C. *Agent Plane-Central Governance and Orchestration*

Cutting across all three Medallion layers is a governance plane that handles four jobs: policy management, prompt sanitization, immutable audit logging, and compliance reporting. Policies are written declaratively and reference both data attributes (sensitivity tag, jurisdiction, retention class) and agent attributes (role, purpose, requested operation).

Prompt sanitization acts as your final safeguard. Every time a prompt is sent out, it's checked again, ensuring that any sensitive information gets either hidden or blocked according to the rules. The audit log is secure and unchangeable, stored in trusted systems like Azure Confidential Ledger or similar services that guarantee data integrity. Each record links together the prompt's details, like the hash, the data involved, the routing choice, the model's response, and the current policy version, creating a reliable and complete record for review.

D. *Retrieval-Augmented Generation Subsystem*

RAG pipelines (Lewis et al., 2020) require extra care because they can bypass standard database controls (Kibriya et al., 2024). To keep everything secure, the framework strengthens RAG in three ways. First, the embedding store is created only from Silver-layer data that's already been transformed and tagged, so it can't inadvertently include unprotected personal info. Second, the Silver-layer privacy agent checks the retrieval results before they're used, catching any sensitive tokens that might still be present. Third, the augmented prompt passes through the Gold-layer gateway, applying the same sensitivity rules to ensure safety and privacy.

E. *End-to-End Workflow - The Trusted AI Data Architecture: A Unified Governance & Privacy Framework*

The end-to-end workflow goes through the following steps as shown in Fig. 2 below, each one logged by the governance plane:

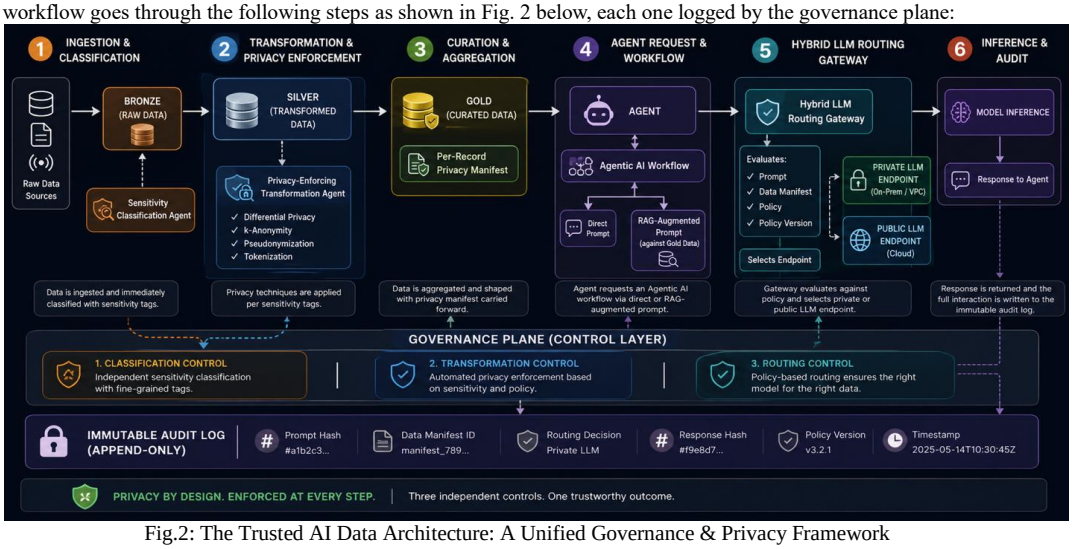

Fig.2: The Trusted AI Data Architecture: A Unified Governance & Privacy Framework

- Ingestion of raw data results in Bronze and is immediately classified by the Sensitivity Classification Agent.
- Transformation data is promoted to Silver, where the Privacy-Enforcing Transformation Agent applies Differential Privacy (Dwork & Roth, 2014), k-anonymity (Sweeney, 2002), pseudonymization, and tokenization according to the sensitivity tags.
- Curation-Silver data is aggregated and shaped into Gold, with the per-record privacy manifest carried through.
- An Agent requests an Agentic AI workflow, which issues a request either as a direct prompt or as an RAG-augmented prompt against Gold data.
- The Hybrid LLM Routing Gateway evaluates the prompt and the data manifest against policy and selects a private or public LLM endpoint.
- Inference and Audit-the model response is sent back to the agent, and the full interaction (prompt hash, manifest, routing decision, response hash, policy version) is written to the immutable audit log.

The point of this sequence is that privacy enforcement is never optional or after-the-fact: every interaction passes through three independent controls-classification, transformation, and routing-before it can reach an external model.

III. EXPERIMENTAL METHODOLOGY

The framework was evaluated using a structured methodology aimed at four things: **feasibility, performance, privacy effectiveness, and regulatory alignment**. The work proceeded through six phases.

A. **Requirement Analysis:** Functional and non-functional requirements were drawn from the four research objectives. Threat models were formalized using the LINDDUN privacy-threat taxonomy and then extended with LLM-specific threat-prompt extraction, embedding inversion, and membership inference (Carlini et al., 2021; Xu et al., 2025). Compliance requirements came directly from GDPR Articles 5, 17, 25, and 30, and from the documentation obligations in EU AI Act Annex IV (European Parliament & Council, 2024).

B. **Platform Configuration:** The reference deployment runs on Microsoft Fabric / OneLake (Microsoft Corporation, 2025), with Bronze, Silver, and Gold lakehouses implemented as Delta Lake tables (Databricks, 2024). The Sensitivity Classification Agent uses a containerized on-premises LLM. The Privacy-Enforcing Transformation Agent is implemented as a set of Spark jobs that call Differential Privacy and tokenization libraries. The Hybrid LLM Routing Gateway is a stateless service deployed as an Azure Container App; it routes cloud requests to Azure OpenAI and private requests to a self-hosted Ollama instance. The governance plane writes audit entries to Azure Confidential Ledger.

C. **Data Preparation:** Three simulated enterprise datasets were prepared. The first is a customer support dataset containing names, addresses, and free-text complaints. The second is a network-telemetry dataset with source and destination IPs, payload metadata, and user-agent strings. The third is a code repository dataset containing source files, commit messages, and embedded secrets. Each one is created synthetically-no actual personal data is involved—but we've calibrated the distributions using published statistics to ensure that the classification and DP results remain realistic and trustworthy.

D. **Privacy and Compliance Integration:** Privacy controls were configured per dataset. Differential Privacy budgets (epsilon) were set at 1.0 for confidential data and 0.5 for restricted data (Dwork & Roth, 2014; Fioretto & Van Hentenryck, 2025). k-anonymity was set at k = 5 for quasi-identifiers (Sweeney, 2002). Tokenization was applied to all detected PII fields. Routing policies prevented restricted data from leaving the tenant boundary under any circumstances. The

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1132-1136

compliance test cases are created based on GDPR and EU AI Act articles (European Parliament & Council, 2024), and they run automatically as assertions against the audit log to ensure everything stays on track.

E. **Identity and Access Management:** Identity is managed by Microsoft Entra ID, with role-based access control enforced at both the lakehouse layer and the routing gateway. Agent identities are first-class citizens here: every agent has its own registered service principal, an associated purpose statement, and a defined set of policy scopes.

F. **Monitoring and Testing:** The system was exercised across three workload classes: routine analytical RAG queries; multi-step agent workflows that traverse all three Medallion layers; and red-team scenarios that deliberately attempt prompt injection, embedding inversion, and policy evasion (Carlini et al., 2021; Xu et al., 2025). The metrics collected are described in the results section. The methodology produces results that are reproducible, defensible to a regulator, and directly applicable to enterprise deployments at a comparable scale.

G. **Scalability and Future Expansion:** From the very beginning, scalability has been recognized as a core part of the architecture. The framework is designed to be modular, allowing each element, such as classification, transformation, routing, and governance, to be scaled independently to meet its specific needs, ensuring smooth and efficient operation. The classification agent scales horizontally with Bronze ingestion volume. The transformation agent scales with the size of the Silver Spark cluster. The routing gateway comfortably scales as a friendly, stateless service safely tucked behind a load balancer, ensuring smooth and reliable performance. From a data-platform perspective, this framework takes advantage of the flexible features already offered by cloud-native lakehouses. With tools like Delta Lake partitioning, Z-ordering, and OneLake shortcuts, privacy operations can focus on specific data ranges without the need for time-consuming full-table rewrites. Plus, the audit log is neatly partitioned by tenant and time, ensuring that even busy agent workflows won't slow down lookups or compliance reports. Several directions for future expansion follow naturally from the current design. Real-time streaming ingestion-Kafka or Event Hubs feeding Bronze-can be supported by promoting the Sensitivity Classification Agent into a streaming operator. Multi-cloud orchestration becomes feasible by federating audit logs across AWS, Azure, and GCP deployments while keeping a single governance plane. Emerging agent-to-agent communication standards, such as the Model Context Protocol (MCP) and the A2A protocol, can be added as extra enforcement points so that delegated agent calls go through the same routing and audit controls as primary calls. Federated Lakehouse scenarios, in which several organizations contribute Silver-layer data to a shared Gold layer, can be supported by extending the Differential Privacy budget to a federated budget tracker.

H. **Privacy and Compliance Analysis:** *Privacy and compliance are enforced through defense-in-depth: classification, transformation, routing, and governance, each addressing a distinct class of threat and designed to fail safely. If any one control is misconfigured, the layers above and below still constrain what an agent can transmit. Classification at Bronze addresses the threat of unrecognized sensitive data, the precondition for every subsequent privacy failure. Transformation at Silver addresses leakage through embeddings, retrieval results, and aggregated outputs. Routing at Gold addresses cross-border and third-party transfer threats. The governance plane addresses the meta-threat of unverifiable compliance, enabling a regulator to be shown not only that controls exist but also that they applied to a specific transaction.*

I. Threat Mitigation Analysis

- Sensitive Telemetry Exposure-handled by Bronze-layer classification, Silver-layer tokenization, and routing rules that stop unmasked telemetry from reaching cloud LLMs.
- Source code and pipeline logic leakage are handled by code-aware sensitivity classifiers, on-premises routing for code-bearing prompts, and prompt sanitization at the governance plane.
- Inference is protected through various methods: Differential Privacy on aggregates and embeddings (Dwork & Roth, 2014), k-anonymity on quasi-identifiers (Sweeney, 2002), and rate limits on agent queries that resemble reconstruction attacks (Carlini et al., 2021).
- Regulatory non-compliance is managed through a robust system, including an unchangeable audit log, clear policies connected to specific parts of GDPR and the EU AI Act (European Parliament & Council, 2024; European Data Protection Board, 2025), and automated compliance checks integrated into the deployment process.
- Prompt Injection and Policy Evasion are managed through prompt sanitization, isolating agent contexts, and maintaining a clear separation between policy evaluation and agent execution. This ensures that agents cannot alter their own routing decisions, providing a secure and reliable system.

By continuously monitoring, regularly reviewing policies, and conducting red-team exercises, we can keep the framework strong and adaptable as regulations and threats change over time.

IV. Experimental Result

A prototype of the framework was deployed on Microsoft Fabric / OneLake and exercised with the simulated datasets across all three workload classes. The evaluation looked at four things: **privacy effectiveness, performance overhead, analytical utility,** and **regulatory traceability.**

On privacy effectiveness, the Sensitivity Classification Agent correctly tagged the great majority of sensitive fields across the three datasets, and the few that slipped through were caught by the second-pass scan in the governance plane. Red-team scenarios that attempted prompt injection and embedding inversion attacks were blocked at the routing gateway in every case where policy explicitly forbade the external transmission of restricted data. In cases where the policy was deliberately relaxed, the audit log still captured the decision, which is exactly the kind of evidence after-the-fact review needs.

Performance overhead was bounded and predictable. Bronze-layer classification added a small per-record latency, dominated by the cost of running the on-premise LLM, which gets amortized over batch ingestion. Silver-layer transformation added overhead in proportion to the number of records that carried sensitive tags. The Hybrid LLM Routing Gateway added negligible latency relative to the underlying LLM inference time, since policy evaluation works on metadata rather than full content.Analytical utility held within acceptable bounds. RAG retrieval relevance, measured by ranking-quality metrics on a held-out evaluation set, dropped slightly under the configured Differential Privacy budgets, but it remained within the range needed for downstream analytical agents to produce useful output. For high-sensitivity workloads where utility loss exceeded those thresholds, the framework correctly fell back to the on-premise LLM path rather than degrading further. **The study showed that by reconstructing detailed forensic timelines from audit logs for each red-team scenario, we could demonstrate strong regulatory traceability. This means every prompt, routing decision, model response, and policy version was accessible, fulfilling the documentation needs of GDPR Article 30 and EU AI Act Annex IV (European Parliament & Council, 2024). Overall, the results reveal that the architecture balances privacy, performance, and analytical value, making it a practical and trustworthy solution.**

V. Discussion

The framework shows that treating privacy as an architectural property of Agentic AI is practical rather than a layer added later for compliance. The evaluation confirms that combining sensitivity classification, privacy-enforcing transformation, hybrid LLM routing, and immutable audit logging on a Medallion lakehouse can be implemented on existing cloud platforms, and the performance penalty is manageable.

The architecture's main strength is its layered enforcement model. No single control is expected to be perfect. Classifiers will miss edge cases, transformations will sometimes degrade utility, and routing policies will occasionally be misconfigured. Because the controls are independent and stacked, a failure in any one layer is bounded by the next, and the governance plane provides the meta-control that makes residual failures detectable and accountable. That is a meaningfully different security posture from the prevailing pattern, where agents and LLMs are connected through trusted relays with very little inspection of what passes between them.

From a deployment perspective, the modular design lets organizations adopt the framework one piece at a time. A team can start with classification and audit logging on their own, already a real improvement over current practice, and add transformation and hybrid routing as the privacy posture matures. Because policies are declarative, a financial-services tenant and a public-sector tenant can share the same architecture while expressing very different obligations. This flexibility effectively accommodates different needs.

Of course, there are some limits to keep in mind. Calibrating Differential Privacy is highly domain-specific, and finding the optimal epsilon for a given workload often requires hands-on testing rather than a straightforward formula (Fioretto & Van Hentenryck, 2025). Using the on-premises LLM fallback adds operational overhead because organizations need to keep inference infrastructure running, even when most traffic goes to cloud models. Compliance checks are based on the regulatory landscape of 2024–2025, so any future changes to the EU AI Act, GDPR, or sector-specific rules will mean updates to policies (European

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1132-1136

Parliament & Council, 2024; Information Commissioner's Office, 2024). Also, the framework relies on a certain level of organizational discipline—such as registered agent identities, declared purposes, and reviewed policies—which requires significant governance effort (Whitman & Mattord, 2018).

Despite these caveats, the architecture offers a dependable foundation for future developments. Exciting possibilities include adaptive Differential Privacy that fine-tunes noise budgets based on real-world utility, federated auditing across multi-cloud environments, integration with emerging agent-communication protocols like MCP and A2A, and AI-assisted policy creation to support governance teams in staying ahead of regulatory changes. None of these ideas challenge the core design-they simply strengthen it. Privacy is enforced architecturally, layer by layer, across the lakehouse.

VI. Conclusion

The Proposed work has introduced a privacy-preserving framework for the orchestration of Agentic AI within cloud-native data Lakehouses constructed upon the Medallion Architecture (Databricks, 2024; Microsoft Corporation, 2025). By deploying dedicated privacy agents at each Medallion layer-Bronze for classification, Silver for transformation, and Gold for hybrid routing-and encapsulating the entire system within a comprehensive governance layer, the framework addresses four prevalent high-risk privacy threats frequently discussed in scholarly literature: telemetry exposure, source code leakage, inference-based reconstruction of access patterns, and regulatory non-compliance (Carlini et al., 2021; Kibriya et al., 2024; Xu et al., 2025).

The architecture thoughtfully combines mature components like Differential Privacy (Dwork & Roth, 2014), k-anonymity (Sweeney, 2002), pseudonymization, secure tokenization, and immutable audit logging. Instead of deploying each feature separately, it integrates them into a cohesive system that works together. Tests on a Microsoft Fabric/OneLake setup (Microsoft Corporation, 2025) demonstrate that this framework effectively safeguards privacy with only a minimal impact on performance. It also keeps analytical usefulness within acceptable Differential Privacy bounds and generates a clear, regulator-ready audit trail that aligns with GDPR and the EU AI Act (European Parliament & Council, 2024).

The key point is that cloud-hosted lakehouses can support privacy-preserving Agentic AI while retaining the productivity advantages that made them popular. There's ongoing work, including fine-tuning, policy management, and regulatory updates. Plus, the framework's flexible and modular design makes it ready for real-world use, especially in regulated industries. More broadly, this work lays a practical foundation for future research at the intersection of Agentic AI, enterprise data platforms, and privacy-enhancing technologies, an area the literature has only just begun to explore.

References

[1] Bain & Company. (2025). State of the art of agentic AI transformation (Bain Technology Report 2025).
[2] Comprehensive Evaluation Team. (2026). Privacy-preserving mechanisms in cloud-based big data analytics: Challenges and future directions (Manuscript under review). Preprints.org.
[3] European Data Protection Board. (2025). AI privacy risks and mitigations in large language models (Technical report).
[4] European Parliament & Council. (2024). Regulation (EU) 2024/1689 on artificial intelligence (AI Act). Official Journal of the European Union.
[5] Fioretto, F., & Van Hentenryck, P. (2025). Differential privacy in artificial intelligence: From theory to practice. Now Publishers.
[6] Kibriya, H., et al. (2024). Privacy concerns in large language models: Training and inference phase analysis. IEEE Access, 12, 20930–20945.
[7] Microsoft Corporation. (2025). Implement medallion lakehouse architecture in Microsoft Fabric. Microsoft Learn. https://learn.microsoft.com/fabric/onelake
[8] Xu, M., Zhang, Y., et al. (2025). On protecting the data privacy of large language models. High-Performance Computing and Communications. https://doi.org/10.1016/j.hcc.2025.100300
[9] Whitman, M. E., & Mattord, H. J. (2018). Principles of information security (6th ed.). Cengage Learning.
[10] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.
[11] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33.
[12] Carlini, N., et al. (2021). Extracting training data from large language models. In Proceedings of the USENIX Security Symposium.
[13] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557–570.
[14] Databricks. (2024). The medallion lakehouse architecture. Databricks Technical Documentation.
[15] Information Commissioner's Office. (2024). Guidance on AI and data protection. ICO.