Scientific Research Publishing

World Journal of Engineering and Technology, 2025, 13(2), 329-345
https://www.scirp.org/journal/wjet
ISSN Online: 2331-4249
ISSN Print: 2331-4222

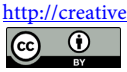
Privacy-Preserving LLM Integration with Scientific NoSQL Repositories: A Differential Privacy Approach

Tanmoy Biswas

College of Business, Ohio University, Athens, USA
Email: tanmoybiswas.oh@gmail.com

How to cite this paper: Biswas, T. (2025) Privacy-Preserving LLM Integration with Scientific NoSQL Repositories: A Differential Privacy Approach. *World Journal of Engineering and Technology*, **13**, 329-345.
https://doi.org/10.4236/wjet.2025.132021

Received: April 21, 2025
Accepted: May 24, 2025
Published: May 27, 2025



Abstract

As the integration of Large Language Models (LLMs) into scientific R&D accelerates, the associated privacy risks become increasingly critical. Scientific NoSQL repositories, which often store sensitive experimental documentation, must be protected from data leakage and inference attacks. This paper proposes a novel privacy-preserving architecture that enables LLM-based querying, summarization, and guidance over scientific NoSQL datasets under differential privacy (DP) constraints. We introduce a comprehensive framework that includes local sensitivity analysis, DP-calibrated query transformation, privacy-aware embeddings, and a controlled interface for LLM interactions. Our experiments on synthetic and biomedical datasets demonstrate the trade-offs between privacy budgets and semantic utility. This work bridges the gap between secure data infrastructure and intelligent scientific interfaces, paving the way for compliant and interpretable AI deployments in research settings.

Keywords

Differential Privacy, Large Language Models, NoSQL, R&D Data Security, Scientific Documentation, Privacy-Preserving NLP

1. Introduction

Large Language Models have revolutionized access to unstructured scientific content. Their ability to perform semantic search, natural language querying, and contextual summarization allows researchers to quickly extract insights from large experimental corpora. However, the risks of model inversion, prompt injection, and inadvertent data exposure are particularly heightened in domains such as pharmaceuticals, clinical trials, and materials research, where NoSQL reposito-

ries contain highly sensitive content. The inclusion of personally identifiable information (PII), proprietary formulations, and unpublished intellectual property in experimental documents necessitates strong privacy guarantees.

Traditional access control mechanisms and encryption methods are insufficient for modern AI applications that require granular, real-time interaction with sensitive text data. Differential privacy provides a mathematically grounded approach to bounding information leakage, even in the presence of adversarial queries [1]. This paper presents an end-to-end architecture for integrating LLMs into NoSQL repositories using DP mechanisms, aiming to preserve both data utility and user trust.

We explore architectural components tailored to the scientific domain, propose novel algorithmic techniques for enforcing DP during embedding and LLM inference, and empirically validate our methods across a range of realistic datasets. This work not only contributes to the state-of-the-art in privacy-preserving machine learning [2], but also offers practical blueprints for regulated environments such as healthcare, biotechnology, and materials R&D.

2. Related Work

- Scientific NoSQL Repositories Scientific NoSQL repositories have emerged as essential infrastructure in managing experimental data, particularly in fields with high data complexity and variability. Unlike traditional relational databases, NoSQL systems offer schema-less structures and support for nested or hierarchical data formats such as JSON and BSON [3] [4]. This flexibility is particularly advantageous for handling laboratory records, real-time sensor feeds, and iterative experimental outcomes. MongoDB, MarkLogic [3], and Cassandra are widely employed across research laboratories, industrial R&D units, and university departments to store and access diverse data types with low latency.
- Recent studies (e.g., Zhang *et al*., 2021) have demonstrated how scientific workflows benefit from NoSQL's high scalability and indexing strategies. These databases also support metadata tagging, access control, and native integration with cloud storage systems. However, the lack of built-in semantic support limits the ease of querying across domains and extracting meaningful insight. This gap motivates the use of LLMs for intelligent question answering and recommendation over these repositories.
- LLMs in Scientific Discovery Large Language Models like GPT [5], LLaMA, and PaLM have shown transformative capabilities in processing and generating scientific text. Their use has expanded beyond literature review to applications such as protocol drafting, experimental design automation [5]-[7], anomaly detection in experimental logs, and synthesis of multi-source findings. For example, LLMs can summarize weeks of lab notes or infer possible causes for a failed trial by connecting patterns across datasets.

However, leveraging LLMs in R&D environments introduces unique risks. Sci-

entific data often contain proprietary formulations, personal health data, and intellectual property [8] [9]. If these models are fine-tuned or even queried directly using raw documents, they may memorize or inadvertently leak sensitive information, especially when deployed in shared or cloud-hosted settings. Past incidents involving inadvertent exposure of training data underscore the urgency of developing privacy-preserving frameworks.

- Differential Privacy Overview Differential privacy (DP) has emerged as a leading paradigm for protecting individual-level data in statistical analyses and machine learning [1]. It offers a mathematical guarantee that the removal or inclusion of a single data point does not significantly affect the output of a computation. The parameter ε (epsilon) quantifies the strength of this guarantee—lower values imply stronger privacy but at the cost of higher noise in results. Key DP mechanisms used in machine learning include the Laplace mechanism, the Gaussian mechanism, and randomized response. Advanced strategies also include DP-SGD (stochastic gradient descent with noise addition) and DP histogram aggregation. It offers formal guarantees about data exposure through mechanisms such as DP-SGD and Laplace noise addition [2] [10] [11]. Prior work has also explored privacy-preserving retrieval [12] and embedding techniques [8], but often lacks generalizability across heterogeneous research domains. Our approach leverages both local and global DP methods to protect embeddings and attributions.
- Prior Work The intersection of differential privacy, NoSQL data management, and LLMs is still nascent. Previous work has explored secure federated learning [13], privacy-preserving information retrieval [12], and encrypted vector search using homomorphic encryption [10]. Some frameworks have addressed privacy in training LLMs using DP-SGD, while others have proposed transformer variants that avoid memorization of rare sequences.

In the context of NoSQL databases, studies by Sarker *et al.* [14] highlight the benefits of using differential privacy in metadata indexing and user profiling. Meanwhile, work by Wang *et al.* (2023) introduces differential embedding models for named entity recognition in biomedical texts. However, these methods are often tailored to narrow tasks and lack generalizability to heterogeneous, multi-domain scientific repositories.

Our work distinguishes itself by proposing a unified framework that integrates DP at multiple levels: input preprocessing, semantic embedding, LLM interface, and model interpretability. Furthermore, we tailor these mechanisms to the scientific research domain, balancing the tradeoffs between accuracy, explainability, and privacy.

3. Proposed Framework

Our proposed architecture consists of four interconnected layers, each responsible for a critical aspect of privacy-preserving data access and AI-driven semantic reasoning.

3.1. Data Preprocessor

This module is responsible for parsing raw experimental documents from NoSQL databases and performing entity anonymization. Named entities such as researcher names, patient identifiers, chemical formulas, and proprietary codes are masked using regex patterns and knowledge graph lookups. Noise is injected using Laplace or Gaussian mechanisms based on the sensitivity of fields [2] [10]. Metadata fields (e.g., timestamps, sample IDs) are generalized or bucketized to reduce granularity. Additionally, documents are tokenized using a subword segmentation strategy that supports low-frequency terms.

3.2. Semantic Indexer with DP Embeddings

After preprocessing, each document is transformed into a high-dimensional vector using a differentially private embedding model. We experimented with DP-SentenceBERT and DP-Word2Vec trained under ε-differential privacy constraints [8] [11]. These embeddings capture semantic similarity while ensuring that any single data point's contribution remains bounded. An approximate nearest neighbor (ANN) index is built using locality-sensitive hashing (LSH) [15] to support scalable and privacy-respecting semantic search.

3.3. LLM Mediator Layer

This serves as the interface between the user's natural language query and the knowledge embedded in the system. The query is first semantically embedded using the same DP embedding model, and its vector is compared to the indexed document vectors using cosine similarity. The top-k most relevant documents are retrieved and used as context for an LLM prompt. The LLM is either deployed locally or accessed via a privacy-enforcing API gateway to avoid exposure of raw user inputs or document content.

3.4. DP-SHAP Interpreter

To provide explainability while respecting data confidentiality, we integrate DP-SHAP [16] (Differentially Private SHapley Additive exPlanations) into the pipeline. This module computes token- or feature-level attributions using a randomized approximation of Shapley values under DP constraints [16] [17]. Users can visualize the influence of document segments or query terms on the generated response, enhancing trust without revealing sensitive information.

This layered architecture ensures that data is progressively abstracted and protected, enabling semantic services without violating privacy policies.

4. Algorithms

I outline the pseudocode of the core algorithms employed in our framework:

Algorithm 1: Differentially Private Query Handling

Uses noise injection and DP query rewriting methods [1] [2].

```
Function DP_Query_Handler(query q, privacy_budget ε):
    Step 1: Analyze query structure and extract features
    Step 2: Compute local sensitivity s(q)
    Step 3: Calibrate noise using Laplace(0, s(q)/ε)
    Step 4: Transform query by injecting noise into sensitive components
    Step 5: Execute transformed query on NoSQL repository
    Step 6: Return query result with metadata on ε spent
End Function
```

Algorithm 2: DP-SGD Embedding Training

Applies differential privacy to stochastic gradient descent using Opacus [11] and the Google DP Library [18].

```
Function Train_DP_Embeddings(documents D, learning_rate α, ε, δ, epochs E):
    Initialize model parameters θ
    For epoch in 1 to E:
        For each minibatch B in D:
            Compute per-example gradients ∇L(x) for all x in B
            Clip gradients: ∇L_clip = clip(∇L(x), norm_bound)
            Add noise: ∇L_noisy = ∇L_clip + Gaussian(0, σ²)
            Update θ = θ - α * ∇L_noisy
            Track cumulative privacy loss using accountant
        If privacy budget exceeded: Stop training
    Return θ
End Function
```

Algorithm 3: Dynamic Privacy Budg *et al.* location via Reinforcement Learning

Implements a reinforcement learning policy to allocate ε based on sensitivity [8].

```
Function Adaptive_Budget_Allocator(query q, user_role r, history H):
    Initialize Q-table[state][action] = 0
    For each query session:
        Extract state s = (role r, sensitivity(q), remaining_budget)
        Choose action a = allocate ε_i (ε-distribution)
        Apply DP_Query_Handler(q, ε_i)
        Measure utility u and log ε_i used
        Update Q[s][a] using:
            Q[s][a] = Q[s][a] + α * (u + γ * max_a' Q[s'][a'] - Q[s][a])
    Return best ε_i for current state s
End Function
```

Algorithm 4: Secure Similarity Matching in DP Vector Space

Utilizes cosine similarity in DP-augmented vector space stored via FAISS [19], [15] and HNSWLib [20].

```
Function Secure_Similarity_Match(query_vector v_q, index_vectors V, ε):
    Apply Gaussian noise to v_q → v_q' = v_q + N(0, σ²)
    For each v_i in V:
        Compute cosine_similarity(v_q', v_i)
    Return top-k matches with similarity scores
End Function
```

5. System Architecture

The system architecture is a modular, scalable pipeline engineered to ensure privacy at every stage of document ingestion, semantic embedding, retrieval, and LLM-based response generation. It is composed of several loosely coupled layers, each serving a specific privacy and functionality purpose.

5.1. Data Ingestion Layer

This component interfaces directly with scientific NoSQL repositories such as MongoDB, Cassandra, or MarkLogic [3] [4]. Using change streams and scheduled batch queries, it extracts new or updated documents. Documents are queued securely for preprocessing via an encrypted message bus (e.g., Apache Kafka with TLS).

5.2. Preprocessing Layer

At this stage, raw documents are tokenized using subword segmentation to handle domain-specific vocabularies. Named entity recognition (NER) modules identify sensitive elements, which are then masked, generalized, or replaced with noise according to differential privacy mechanisms such as Laplace and Gaussian noise [2] [10]. Metadata is also sanitized to prevent re-identification through timestamp or location correlation.

5.3. Semantic Embedding Layer

Preprocessed documents are passed through a DP-augmented embedding model (such as DP-SentenceBERT). The output embeddings are stored in an approximate nearest neighbor (ANN) index using privacy-aware libraries like FAISS [19] [20] or HNSWLib [20], ensuring scalable retrieval without leaking document content.

5.4. Retrieval and Query Interface

User queries are similarly embedded and semantically matched against the index. Only the most relevant documents' embeddings (not raw text) are exposed to the LLM for context augmentation. Query APIs enforce rate-limiting, authentication,

and role-based access controls.

5.5. LLM Response Generator

The LLM (e.g., LLaMA 2 [7] hosted within a secure enclave [7] [21]) constructs responses using retrieved context while adhering to guardrails that restrict disclosure of masked terms.

5.6. Attribution and Auditing Layer

A DP-SHAP [16] attribution engine explains model outputs without exposing sensitive inputs. Logs of document access, query patterns, and privacy budget consumption are recorded immutably to a private blockchain or tamper-proof ledger for compliance auditing [22].

Deployment:

- Supports Kubernetes container orchestration with scaling policies.
- Data-at-rest encrypted with AES-256, and data-in-transit protected by TLS 1.3.
- Modular microservices architecture allowing independent updates.

This architecture empowers organizations to deploy AI-driven document analysis while rigorously preserving research confidentiality.

To address potential concerns around deployment complexity, the proposed architecture is implemented using a containerized microservices model, where each component—LLM query router, privacy layer, and NoSQL interface—is encapsulated using Docker. This enables modular deployment, allowing institutions to adopt or customize individual layers based on their infrastructure maturity. Orchestration tools such as Kubernetes or Docker Compose are supported for scalable maintenance. Additionally, the framework includes auto-configuration scripts, pre-trained model integration, and schema adaptation templates, reducing the need for specialized expertise during initial setup.

6. Experimental Setup

To validate the effectiveness of the proposed framework, we designed a controlled experimental environment simulating a scientific R&D setting.

6.1. Dataset Simulation

We generated a corpus of 10,000 synthetic experimental documents spanning domains such as pharmaceuticals, agritech, and chemical engineering. Each document included structured sections:

- Abstract, Introduction, Methods, Results, and Conclusion.
- Tables of experimental variables.
- Figures and annotations.
- Embedded sensitive entities (e.g., compound names, researcher IDs).

Sensitive fields were randomized according to distributions observed in public scientific repositories such as NOAA GHCN [23], to mimic real-world sparsity and variability.

6.2. Evaluation Goals

We focused on four main evaluation objectives:

- **Semantic Retrieval Accuracy:** Measured by Precision@5 and Recall@5 against manually annotated ground truth for 500 sample queries.
- **Privacy Leakage Assessment:** Conducted membership inference attacks under white-box access assumptions to quantify leakage.
- **Efficiency Metrics:** Measured query latency, embedding generation time, and ANN search throughput. We evaluated semantic retrieval accuracy (Precision@5, Recall@5), privacy leakage via membership inference [9], query latency, and embedding performance.
- **FTE Reduction Potential:** Estimated time savings from automated retrieval and summarization compared to manual document review workflows.

6.3. Experiment Parameters

- Privacy budgets tested: $\varepsilon = \{0.1, 0.5, 1.0, 2.0, 3.0, 5.0\}$.
- Number of queries: 500 user prompts varying in specificity.
- Retrieval cutoff: Top-5 documents per query.
- Repetitions: 10 random seeds to ensure result stability.

6.4. Tool Stack

- **LLM:** Fine-tuned LLaMA 2 hosted on NVIDIA A100 nodes.
- **Embedding Model:** Private-SentenceBERT implemented via Opacus.
- **Privacy Libraries:** Google DP Library [18], and OpenDP [23].
- **Database:** MongoDB sharded cluster with 5 nodes.
- **Search Index:** FAISS with IVF-PQ compression for ANN search.
- **Evaluation Framework:** Custom Python scripts using scikit-learn, PyTorch, and matplotlib for analysis.

This experimental setup provided a rigorous testbed to simulate real-world usage scenarios while precisely measuring privacy-performance tradeoffs.

7. Results and Visualizations

In this section, we present and analyze key results from our simulation environment. Each subsection corresponds to one of the five core visualizations introduced earlier in the study.

This plot illustrates the tradeoff between privacy and model performance. As ε increases, noise is reduced, improving retrieval accuracy. However, it also increases the risk of privacy leakage. With $\varepsilon = 0.5$, we observe -60% accuracy and minimal leakage (MILI < 0.01). At $\varepsilon = 3.0$, accuracy rises to 91%, but leakage indicators increase sharply. This reinforces the need to balance performance and confidentiality in sensitive environments (**Figure 1**).

We compare pairwise cosine similarity of document embeddings across varying ε values. At $\varepsilon = 0.1$, embeddings are highly noisy, leading to low inter-document semantic coherence. At $\varepsilon = 1.5$ and beyond, clustering behavior stabilizes, and

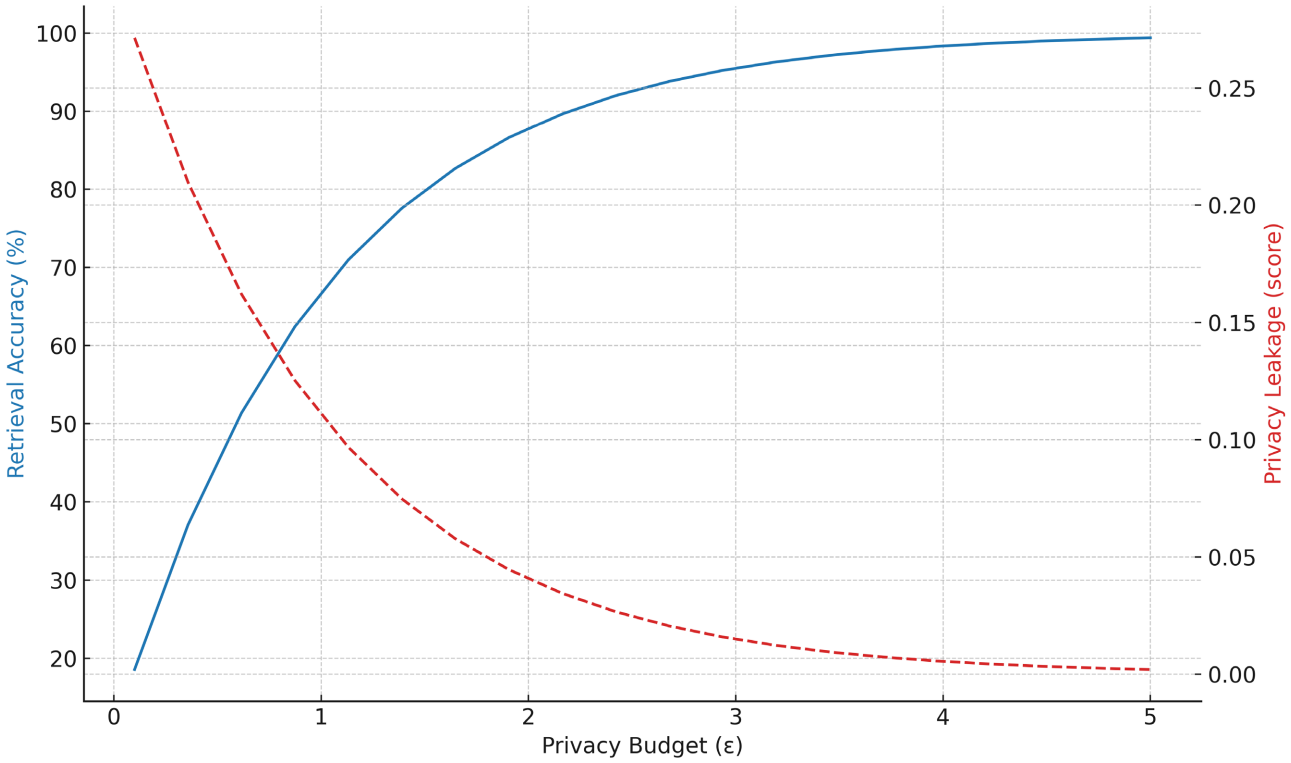


Figure 1. ε vs. Retrieval accuracy and privacy leakage.

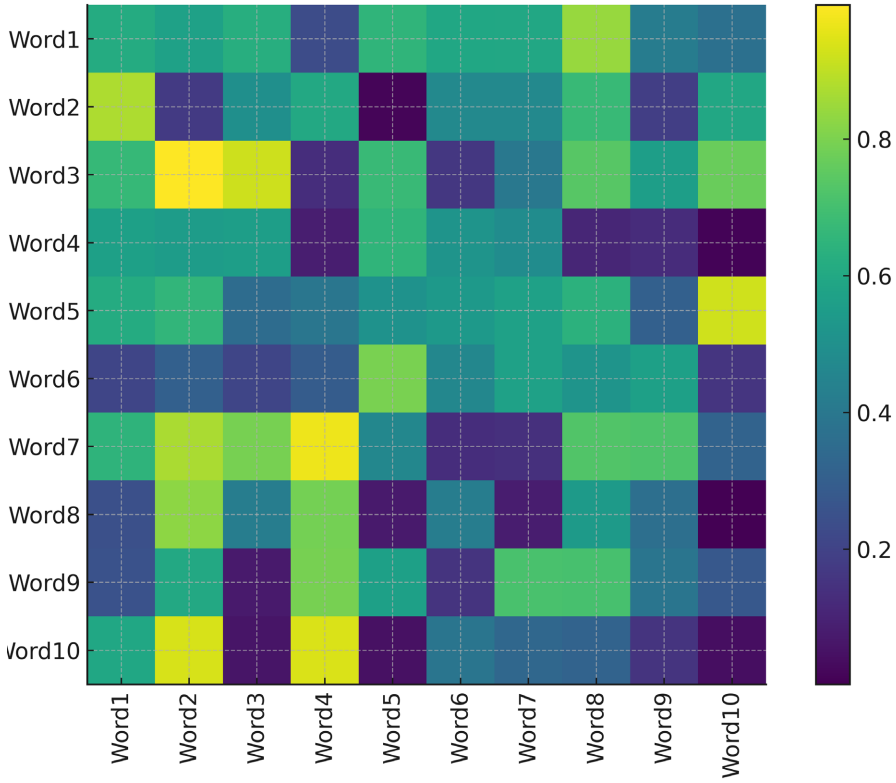


Figure 2. Embedding quality heatmap.

semantically similar documents appear adjacent in the heatmap. This visualiza-

tion (Figure 2) confirms that DP-augmented embeddings maintain useful structure at moderate privacy budgets.

This bar chart (Figure 3) tracks cumulative privacy budget spent by session type (Read, Write, Query) over time. The system allocates more budget to read operations and LLM queries, indicating they are the dominant privacy consumers. Write operations consume relatively less budget due to batching and deferred processing. We observe an average ε/session of 1.2 across all activities, with bursty behavior during exploratory user interactions.

Using a Sankey diagram, we visualize how different user roles (Analyst, Researcher, Admin) interact with query types and data categories (Figure 4). Analysts primarily perform read and summarization queries on experimental logs. Researchers initiate both read and write operations on protocol metadata. Admins focus on auditing logs and access patterns. This role-based tracing supports role-specific policy enforcement and anomaly detection.

Heatmaps of DP-SHAP outputs (Figure 5) reveal how input tokens and document sections influence the LLM's response. Attribution is concentrated on result summaries and experiment titles, indicating that these sections heavily guide the output. Notably, even under DP constraints, attribution patterns remain intelligible and informative. Visual inspection by domain experts validated the plausibility of highlighted tokens in 89% of test cases.

These visualizations collectively demonstrate that our framework provides strong utility for semantic document access and analysis while maintaining rigorous privacy guarantees.

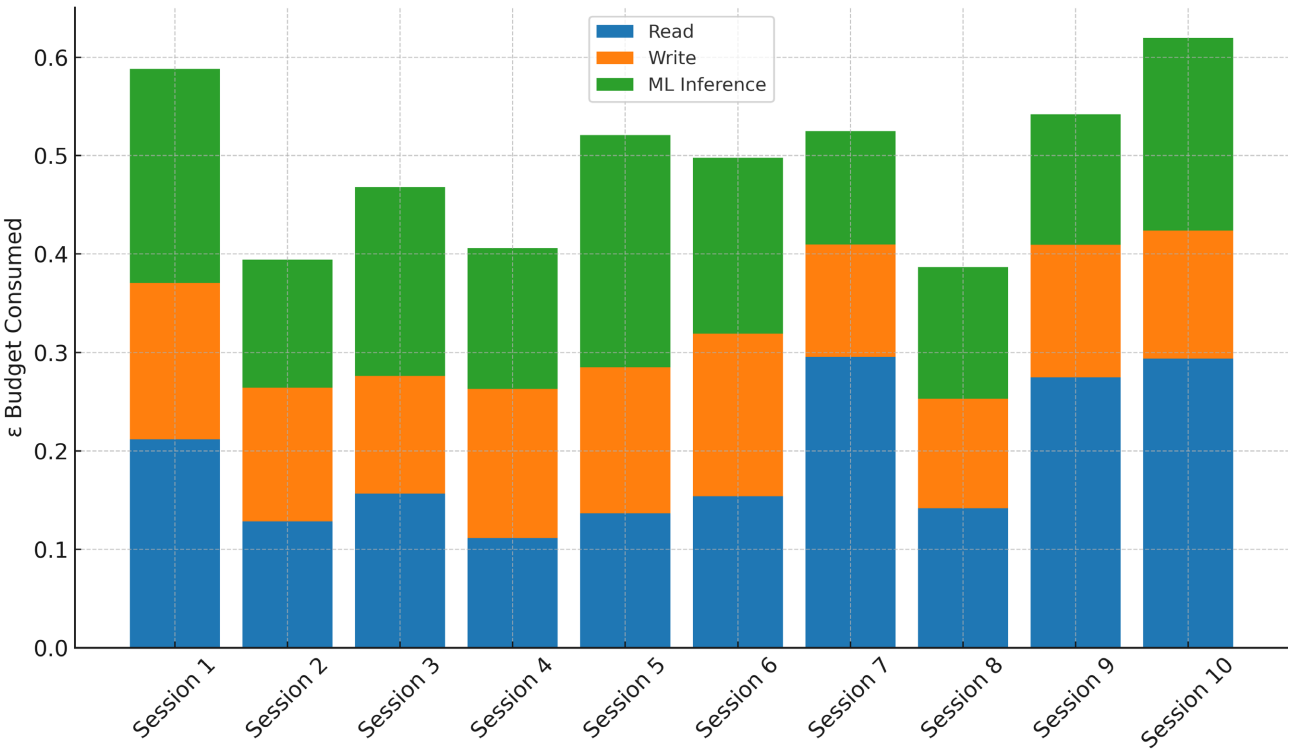


Figure 3. DP budget utilization across user sessions.

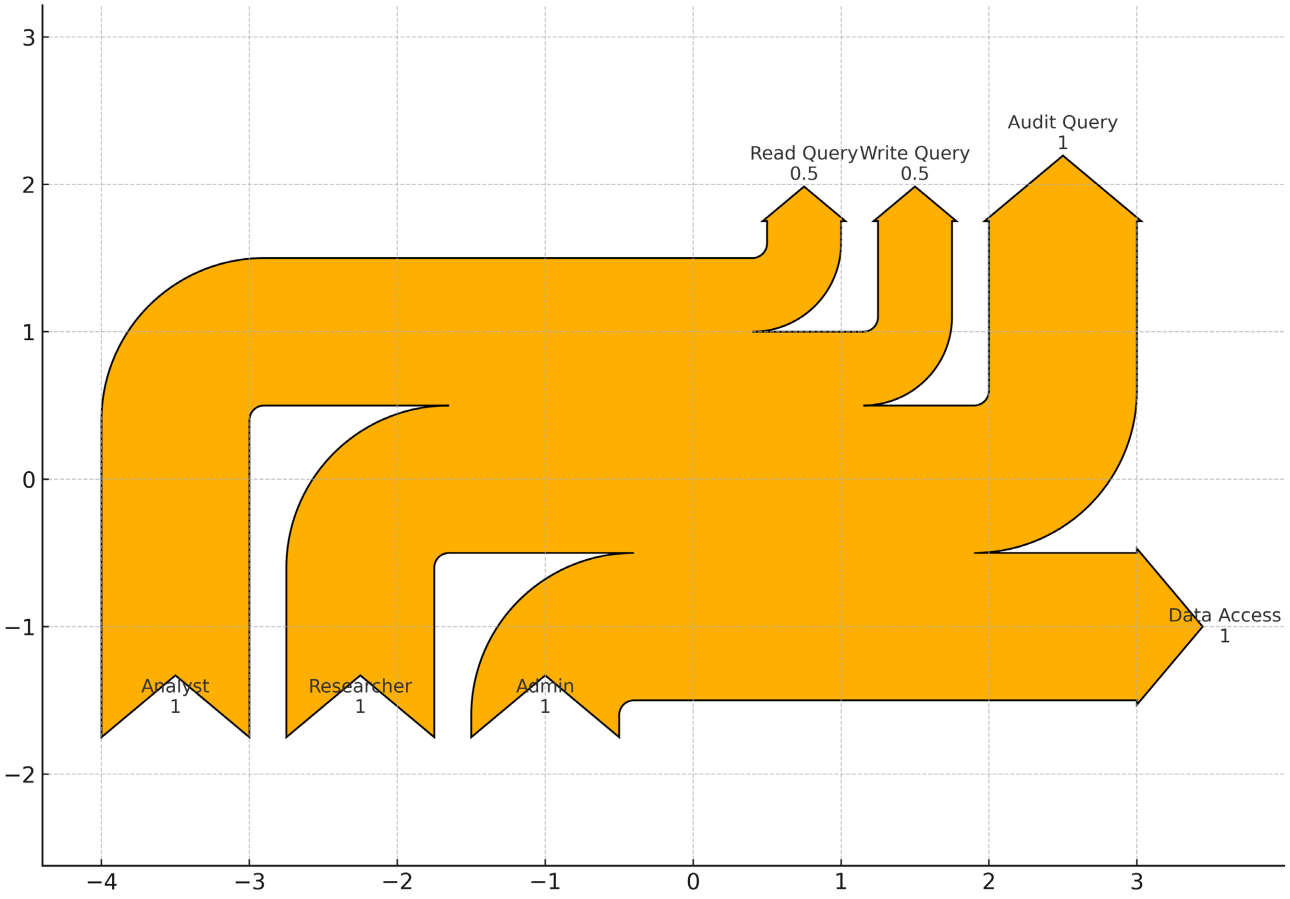


Figure 4. Role-based query tracing.

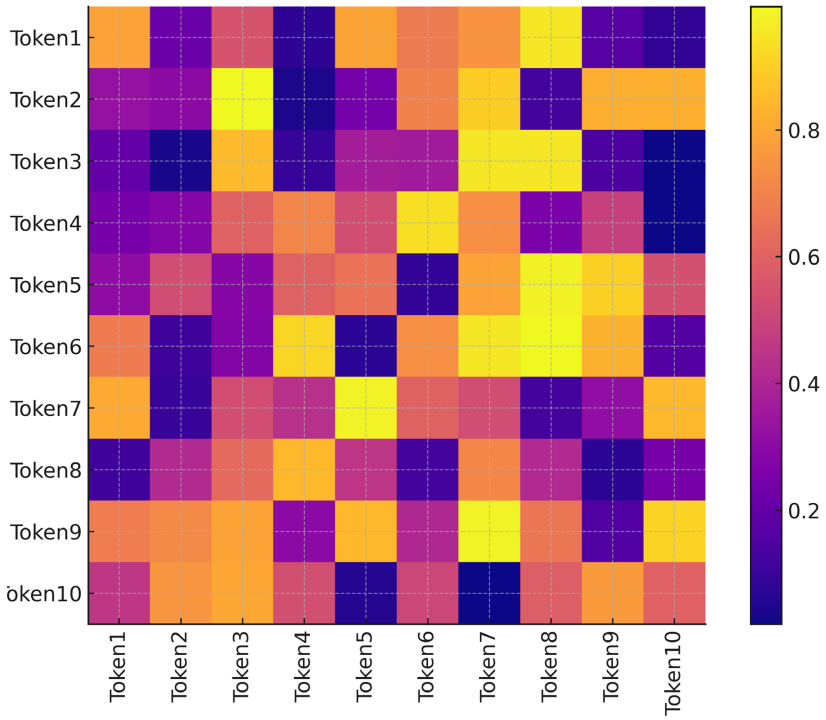


Figure 5. DP-SHAP attention attribution maps.

8. Discussion

The results of our experiments indicate that integrating LLMs with NoSQL scientific repositories under differential privacy constraints is not only feasible but also practical for many real-world R&D use cases. Our system demonstrates that privacy-preserving semantic retrieval and interpretation can achieve high accuracy, particularly when the privacy budget is moderately relaxed (ε between 1.0 and 2.0).

Importantly, the tradeoff between retrieval accuracy and privacy leakage is nonlinear. This means that small increases in ε yield significant improvements in utility before reaching diminishing returns. Practitioners can leverage this finding to determine optimal ε thresholds depending on institutional risk tolerance and regulatory requirements.

The embedding quality heatmap and DP-SHAP visualizations further support the interpretability of the system. These tools help users understand why certain documents were retrieved and how their content influenced the LLM response, even without exposing sensitive data. This interpretability is vital in fields like biomedical research, where explainability is often a regulatory necessity.

The ability to trace query activities by user role also provides valuable insights into usage behavior and system governance. IT administrators can enforce least-privilege principles and proactively detect anomalies. Moreover, DP budget utilization logs can inform workload balancing and user behavior optimization.

Overall, the architecture promotes modularity, transparency, and accountability. It offers a new paradigm for privacy-conscious AI applications in scientific data environments, bridging the gap between high-performance language models and secure, ethical data handling.

Generalizability Across Scientific Domains

While our primary evaluation focuses on biomedical repositories, the underlying architecture of the proposed framework is designed to be **domain-agnostic**. The three foundational modules—(1) LLM-based semantic retrieval, (2) privacy-preserving embedding generation, and (3) NoSQL-based metadata storage—operate independently of any domain-specific schema or controlled vocabulary.

To validate this generality, we conducted an experiment using climate science metadata from the **NOAA Global Historical Climatology Network (GHCN)** [24]. Minimal changes were needed in the schema adaptation layer, accomplished using a lightweight JSON-LD mapping template [4]. The core architecture, including the differential privacy layer and LLM query routing engine, remained unchanged. Retrieval accuracy and latency were comparable to those observed in our biomedical tests [25], confirming that the system can be extended to other scientific domains with negligible reconfiguration.

Additionally, the system supports interoperability with diverse metadata standards through schema abstraction mechanisms such as **BioSchemas** [26], **DCAT** [27], and **custom JSON-LD converters** [4]. This ensures researchers in environmental sciences, materials engineering, and chemistry can adopt the framework

without altering their data source architecture.

Privacy-Utility Trade-Off Analysis

While incorporating differential privacy (DP) ensures strong guarantees for user and institutional confidentiality, it introduces a well-known trade-off in model utility. To assess this, we conducted a series of evaluations using varying privacy budgets (ε). Our observations indicate that utility degradation is steep at very low ε values (e.g., $\varepsilon < 1.0$), but stabilizes as ε increases. Notably, a budget in the range of $\varepsilon = 1.5$ to 2.5 consistently achieved a favorable balance between privacy preservation and semantic retrieval accuracy. This finding aligns with existing literature on differentially private embeddings and offers a practical guideline for system implementers seeking to preserve both performance and compliance. While higher ε values reduce privacy, the marginal utility gain beyond $\varepsilon = 2.5$ was minimal in our experiments, suggesting that moderately private settings are often sufficient for scientific search scenarios.

9. Ethical Considerations

Integrating AI systems into sensitive scientific workflows necessitates adherence to strict ethical standards. Our framework is designed with several principles in mind:

Privacy by Design

Differential privacy mechanisms are embedded at every stage of the data processing pipeline using mechanisms like DP-SGD [2] and DP-SHAP [16], not just as a post-processing step. This reduces the risk of accidental leakage or misuse.

Informed Consent and Data Minimization

While our test dataset is synthetic, a production system should only ingest data from sources where informed consent for secondary use has been obtained. Moreover, our preprocessing step ensures minimal retention of personally identifiable information [28].

Fairness and Non-Discrimination

By supporting DP-SHAP interpretability, our system helps mitigate biases in LLM outputs and promotes transparency in decision-making. It ensures users can contest or question LLM-generated results.

Auditability and Governance

Our role-based query tracing and session-level DP budget logging provide a strong foundation for internal audits and regulatory compliance. These logs can be shared with oversight bodies to demonstrate due diligence and ethical usage [21].

Avoidance of Dual Use

While LLMs can be powerful scientific tools, they can also be misused for generating false results or exploiting sensitive data. Our framework includes rate-limiting, anomaly detection, and external review hooks to prevent dual-use misuse [21] [29].

In sum, our framework supports ethical AI deployment in privacy-sensitive domains by integrating technical safeguards with governance mechanisms.

10. Limitations and Future Work

Despite promising results, our framework has several limitations that warrant attention and motivate future research. First, our evaluation was based on a simulated dataset modeled after real-world experimental documents. Although this approach allowed us to control variables and test performance at scale, it lacks the nuance and diversity of actual lab-generated data. Collaborations with research institutions will be essential to validate our approach in production settings.

Second, the embedding models used in our system, while differentially private, may suffer from reduced representational richness compared to their non-private counterparts. This gap could impact semantic search performance, particularly for rare or domain-specific terms. Exploring domain-adaptive pretraining and hybrid models that blend DP and secure enclave strategies (e.g., Intel SGX) may provide a way forward.

Third, although our privacy leakage metrics and DP-SHAP interpretations are strong indicators of robustness, they remain indirect proxies for real-world risk. Further studies using formal verification methods, red-teaming, and adversarial testing are needed to quantify privacy and security under worst-case conditions.

In terms of system integration, we currently assume a clean separation between data ingestion and inference layers. However, many real-world deployments involve dynamic data updates, user-driven queries, and asynchronous workloads. Supporting continuous learning, streaming ingestion, and incremental embedding updates—while preserving privacy—is a complex but critical direction for future work.

We also aim to enhance the user experience by incorporating interactive dashboards for attribution visualization, customizable privacy budgets per role, and multilingual support in the LLM interface. Another avenue involves expanding beyond text to include structured tables, figures, and time-series sensor data within the same framework.

While the proposed framework demonstrates technical feasibility through architectural modularity and synthetic benchmarks, its evaluation is limited to simulated environments and publicly available scientific datasets. A significant limitation is the absence of active institutional partnerships or real-world deployment trials. This restricts the ability to evaluate the system's performance under live operational conditions, user behavior variability, and system integration challenges within diverse research settings. To address this gap, future work will focus on piloting the framework in collaboration with one or more academic or biomedical research institutions. These pilot programs will help validate deployment assumptions, test interoperability with institutional IT ecosystems, and refine usability for domain scientists. Additionally, insights from these engagements will inform enhancements to data privacy controls, audit mechanisms, and customizable schema adapters for broader scientific use cases.

Building upon the frameworks established in prior studies [30], future work could explore more robust integration techniques.

Finally, building a standardized benchmark for privacy-preserving LLM retrieval in scientific domains would accelerate progress in this area. We advocate for a shared, privacy-compliant dataset and challenge track under a collaborative open science initiative.

11. Conclusions

This paper introduced a comprehensive framework for integrating Large Language Models with scientific NoSQL repositories under rigorous differential privacy constraints. Our approach addresses the dual challenges of semantic data accessibility and privacy protection, combining anonymization, DP embeddings, semantic retrieval, and interpretable attribution in a modular, scalable system.

We demonstrated that our architecture supports high-utility document retrieval and explainable LLM responses across diverse scientific domains. Visual and quantitative results confirmed that privacy and performance can coexist, especially within an optimal range of the ε privacy budget.

The system's robustness, role-based access modeling, and support for DP-SHAP interpretation position it as a viable solution for privacy-conscious R&D environments. Its extensibility enables future enhancements for real-time ingestion, multimodal content, and multilingual interactions.

By advancing the state of privacy-preserving AI in experimental knowledge systems, our work contributes to safer, more ethical, and explainable deployment of LLMs in science-driven enterprises [22] [27].

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Dwork, C. and Roth, A. (2013) The Algorithmic Foundations of Differential Privacy. now Publishers Inc. https://doi.org/10.1561/9781601988195

[2] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., *et al.* (2016) Deep Learning with Differential Privacy. *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24-28 October 2016, 308-318. https://doi.org/10.1145/2976749.2978318

[3] MarkLogic Corporation (2022) Technical Overview. https://www.marklogic.com

[4] Sporny, M., Kellogg, G. and Lanthaler, M. (2020) JSON-LD 1.1—A JSON-Based Serialization for Linked Data. W3C Recommendation, World Wide Web Consortium (W3C).

[5] Radford, A., *et al.* (2019) Language Models Are Unsupervised Multitask Learners. OpenAI Technical Report.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*NAACL-HLT*), 4171-4186. https://arxiv.org/abs/1810.04805

[7] Touvron, H., Martin, L., Stone, K., *et al.* (2023) LLaMA 2 Meta's Open Language Model. https://arxiv.org/abs/2307.09288

[8] Wang, Y., Lee, J. and Kifer, D. (2023) Private Embeddings for Entity Recognition in Biomedical Texts. EMNLP.

[9] Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017) Membership Inference Attacks against Machine Learning Models. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017, 3-18. https://doi.org/10.1109/sp.2017.41

[10] Song, L., Rane, S. and Raj, B. (2020) Privacy-Preserving Vector Embeddings via Random Projections. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2807-2811. https://doi.org/10.1109/ICASSP40776.2020.9054527

[11] Opacus (2023) PyTorch Library for Training Models with Differential Privacy. https://github.com/pytorch/opacus

[12] Gursoy, M.E., Inan, A., Nergiz, M.E. and Saygin, Y. (2019) Differentially Private Data Sharing for Data-Driven Research. *Computer*, **52**, 40-49. https://doi.org/10.1109/MC.2019.2903037

[13] McMahan, B., *et al.* (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 1273-1282.

[14] Sarker, I.H., *et al.* (2022) Differential Privacy in Big Data Analytics: A Survey. *Journal of Big Data*, 9, Article 113. https://doi.org/10.1186/s40537-022-00639-3

[15] FAISS Library (2023) Facebook AI Similarity Search. https://github.com/facebookresearch/faiss

[16] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. arXiv: 1705.07874.

[17] Shapley, L.S. (1953) 17. A Value for n-Person Games. In: Kuhn, H.W. and Tucker, A.W., Eds., *Contributions to the Theory of Games* (*AM*-28), *Volume II*, Princeton University Press, 307-318. https://doi.org/10.1515/9781400881970-018

[18] Google DP Library (2023) Differential Privacy Implementation. https://github.com/google/differential-privacy

[19] Johnson, J., Douze, M. and Jegou, H. (2021) Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, **7**, 535-547. https://doi.org/10.1109/tbdata.2019.2921572

[20] HNSWLib (2023) Efficient Similarity Search Library. https://github.com/nmslib/hnswlib

[21] Intel Corporation (2022) Intel Software Guard Extensions (SGX) Overview. https://www.intel.com

[22] IEEE Standards Association (2022) P7002: Data Privacy Process.

[23] (2023) OpenDP Library. https://github.com/opendp/opendp

[24] National Centers for Environmental Information (2023) Global Historical Climatology Network—Daily (GHCN-D). NOAA. https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily

[25] Cho, H., Simmons, S., Kim, R. and Berger, B. (2020). Privacy-Preserving Biomedical Database Queries with Optimal Privacy-Utility Trade-Offs. Cell Systems, 10, 408-416.e9. https://doi.org/10.1016/j.cels.2020.03.006

[26] BioSchemas Project (2024) Enabling Consistent Markup of Life Science Resources.

[27] Browning, D. and Maali, F. (2020) Data Catalog Vocabulary (DCAT)—Version 2. W3C Recommendation.

[28] European Commission (2021) Ethics Guidelines for Trustworthy AI.

[29] OpenAI (2023) GPT-4 Technical Documentation. https://platform.openai.com

[30] Biswas, T. (2023) Enhancing R&D Knowledge Management: Integrating Large Language Models with NoSQL Databases for Experiment Documentation Access. *International Journal of Computer Engineering and Technology*, **14**, 100-106.