

Explainable Privacy-Preserving Scientific Retrieval with Knowledge Graphs: An Evidence-Grounded Framework for Scientific NoSQL Repositories

Tanmoy Biswas

Independent Researcher, Union, Kentucky, USA

Email: biswastanmoy7@gmail.com

How to cite this paper: Biswas, T. (2026) Explainable Privacy-Preserving Scientific Retrieval with Knowledge Graphs: An Evidence-Grounded Framework for Scientific NoSQL Repositories. *World Journal of Engineering and Technology*, 14, 821-839. <https://doi.org/10.4236/wjet.2026.144051>

Received: August 27, 2026

Accepted: October 12, 2026

Published: October 15, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Large Language Models (LLMs) provide powerful interfaces for scientific knowledge retrieval, but privacy-preserving access alone does not satisfy the transparency, provenance, and reproducibility requirements of research-intensive environments. Building directly on prior work that introduced differential-privacy controls for LLM integration with scientific NoSQL repositories [1], this paper proposes Explainable Privacy-Preserving Scientific Retrieval (EPPSR), a unified architecture combining a Scientific Knowledge Graph, semantic and graph retrieval, privacy-risk-aware evidence ranking, provenance, and evidence-level explanations. A revised reproducible V2 benchmark explicitly separates formal differential privacy (DP) from application-layer sensitive-information exposure. Formal DP is evaluated only for a bounded aggregate mean-effect release using replace-one adjacency, clipping to $[-18, 18]$, sensitivity $\Delta = 6$, a Laplace mechanism, an epsilon sweep from 0.1 to 5.0, and repeated-query composition. Sensitive Information Exposure Rate (SIER) separately measures unredacted disclosure of sensitive record identifiers or graph edges. The synthetic benchmark contains 672 records and 80 stratified queries with reliable, contradictory, incomplete-text, sensitive, and lexical-distractor evidence. Across the controlled simulation, EPPSR achieved 100.00% answer accuracy, 97.50% Recall@5, 46.25% citation accuracy, 0.00% SIER, 88.60% provenance completeness, and 92.50% citation support. At epsilon = 1.0, the DP aggregate release achieved 82.02% direction accuracy with a mean absolute error of 6.01 effect-percentage points. Ablation results show that removing consistency reduced answer accuracy to 91.25%, while a semantic-plus-graph-only variant fell to 66.25%. These results demonstrate controlled architectural behavior rather than production LLM performance and show why privacy

guarantees, exposure metrics, evidence quality, and retrieval utility must be evaluated separately.

Keywords

Differential Privacy, Explainable Artificial Intelligence, Knowledge Graphs, Large Language Models, NoSQL Databases, Retrieval-Augmented Generation, Scientific Information Retrieval, Trustworthy AI

1. Introduction

Large Language Models (LLMs) have transformed natural-language access to complex information. Transformer architectures [2] and subsequent pretrained language models such as BERT [3] established the technical foundation for modern question answering, summarization, and knowledge-intensive language processing. In scientific research, these capabilities create an opportunity to expose decades of experimental documentation through conversational interfaces rather than requiring researchers to manually search heterogeneous laboratory systems.

Scientific repositories, however, differ from general-purpose information collections. They may contain proprietary formulations, unpublished experimental observations, process conditions, confidential project information, and regulated data. Our preceding study, “Privacy-Preserving LLM Integration with Scientific NoSQL Repositories: A Differential Privacy Approach,” introduced a layered framework for protecting scientific NoSQL content using local sensitivity analysis, differential-privacy-calibrated transformations, privacy-aware embeddings, and a controlled LLM interface [1]. That work demonstrated the importance of treating privacy as a first-class architectural requirement rather than relying only on conventional access control. Related work on LLM-NoSQL integration for experiment documentation also provides an earlier foundation for scientific knowledge access [4].

Privacy alone does not make an AI-generated scientific answer trustworthy. Researchers must also be able to determine what evidence supports a conclusion, whether the evidence is consistent, how entities and experiments are related, and whether the system is uncertain. Retrieval-Augmented Generation (RAG) addresses part of this problem by combining parametric language models with external non-parametric memory [5]. RAG improves grounding and allows domain knowledge to be updated independently of model retraining, but conventional implementations remain primarily document- or chunk-oriented [6].

Scientific knowledge is inherently relational. A compound may occur in a formulation, the formulation may be tested under a process condition, the experiment may generate an analytical measurement, and that measurement may support or contradict a later conclusion. Knowledge graphs provide an explicit representation of such entities and relationships [7]. Recent work has emphasized the complementary strengths of LLMs and knowledge graphs: LLMs provide flexible

language understanding, whereas knowledge graphs provide explicit factual structure and interpretable relationships [8].

Explainability introduces an additional requirement. XAI research emphasizes transparency, accountability, and the need to adapt explanations to their intended audience [9]. For generative scientific retrieval, explanation should not mean exposing hidden chain-of-thought. Instead, a scientifically useful explanation should identify observable evidence, provenance, and relationships that support a generated claim. This paper therefore defines explainability as evidence-level traceability.

To address these requirements, this paper proposes Explainable Privacy-Preserving Scientific Retrieval (EPPSR), which extends the privacy-preserving NoSQL-LLM architecture in [1] with a Scientific Knowledge Graph, dual-channel retrieval, privacy-aware evidence ranking, provenance mapping, calibrated confidence, and claim verification. The central research question is whether scientific retrieval can be simultaneously private, evidence-grounded, explainable, and useful.

The contributions are fivefold: 1) a unified EPPSR architecture integrating formal DP for explicitly defined aggregate releases with scientific knowledge graphs, retrieval, provenance, and evidence-level explainability; 2) a formal privacy threat model that distinguishes epsilon-DP guarantees from application-layer sensitive-information exposure through answers, citations, graph paths, explanations, and repeated queries; 3) a dual-channel retrieval and privacy-aware evidence-ranking mechanism combining semantic, graph, provenance, consistency, and disclosure-risk signals; 4) reproducible explanation metrics based on evidence coverage, provenance completeness, graph-path traceability, and citation support rather than an arbitrary composite explanation score; and 5) a V2 synthetic benchmark with epsilon sweeps, repeated-query privacy accounting, bootstrap confidence intervals, and ablation analysis.

2. Related Work

A. LLMs and Retrieval-Augmented Generation

The transformer architecture replaced recurrence with attention and enabled scalable contextual representation learning [2]. BERT subsequently demonstrated the value of deep bidirectional pretraining for language understanding [3]. Although modern LLMs can encode substantial factual knowledge, parametric knowledge can become stale and may not provide reliable provenance. Lewis *et al.* introduced RAG to combine sequence generation with retrieved non-parametric memory [4]. Later surveys describe the evolution from naive RAG toward advanced and modular architectures incorporating improved retrieval, reranking, and generation control [5]. Broader language-model and retrieval research includes unsupervised multitask language models, text-to-text transfer learning, few-shot learning, dense passage retrieval, late-interaction ranking, heterogeneous retrieval benchmarks, retrieval-supported generation, retrieval-augmented pretraining, and transformer-based text ranking [10]-[18].

RAG is particularly relevant to enterprise science because proprietary information can remain in organizational repositories while only selected evidence is passed to the model. Nevertheless, retrieval quality and generation faithfulness remain separate failure points. Hallucination research has shown that fluent generation can remain factually inconsistent or unsupported [19]. Scientific use therefore requires not only retrieval but explicit evidence verification.

Research on chain-of-thought prompting, factual consistency, truthfulness, and machine-comprehension benchmarks further shows the importance of reasoning and evidence-based evaluation in language systems [19]-[22].

B. Knowledge Graphs and Graph-Enhanced Retrieval

Knowledge graphs represent entities and relations in a form suitable for querying, validation, reasoning, and integration across heterogeneous information sources [7]. Construction commonly involves knowledge acquisition, refinement, entity resolution, relation extraction, and graph evolution [23]. These capabilities align closely with scientific repositories in which experiments, materials, conditions, methods, measurements, and outcomes form a connected evidence network. Foundational work on neural tensor networks, translational embeddings, relational machine learning, knowledge-graph embeddings, and large-scale knowledge-graph systems established important methods for representing and reasoning over multi-relational knowledge [24]-[28].

Pan *et al.* categorize LLM-KG integration into KG-enhanced LLMs, LLM-augmented KGs, and synergized LLM-KG systems [7]. Graph-based retrieval extends RAG by retrieving nodes, neighborhoods, paths, or graph-derived summaries rather than relying exclusively on independent text chunks. Such approaches can support multi-hop questions, but graph structure does not inherently provide privacy. Sensitive scientific relationships can themselves disclose intellectual property.

C. Differential Privacy and Privacy-Preserving Language Systems

Differential Privacy (DP) provides a formal mechanism for bounding the effect of an individual record on released outputs [29] [30]. DP has been extended to machine learning through techniques such as differentially private stochastic gradient descent [30]. Privacy risks in language models also include memorization and extraction of rare training sequences [31], motivating privacy controls at training, retrieval, and inference boundaries. Privacy research has also examined model inversion, membership inference, privacy analysis of deep learning, privacy-preserving knowledge transfer, federated learning, and decentralized training [32]-[37]. These studies show that privacy risk can arise from both model outputs and the broader learning architecture.

Our prior work [1] applied DP concepts directly to scientific NoSQL retrieval and LLM interaction. The present study treats that architecture as the privacy foundation and investigates the next problem: how to add evidence traceability and relational reasoning without allowing the explanation itself to become a disclosure channel.

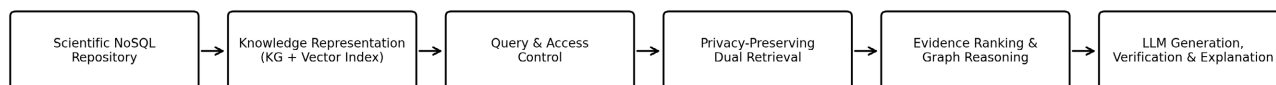
D. Explainable and Trustworthy AI

Explainable AI encompasses methods for making model behavior understandable to users and stakeholders [9]. Local surrogate approaches such as LIME [38] and additive attribution approaches such as SHAP [39] are widely used for predictive models. These methods are not sufficient by themselves for generative scientific retrieval because a researcher needs source evidence and provenance rather than only feature attribution. Broader XAI research has examined black-box explanation, rigorous interpretability, limits of interpretability, counterfactual explanations, and general XAI taxonomies [9] [40]-[42]. These perspectives reinforce [43] the need for explanations that are measurable and appropriate to the scientific retrieval setting.

Trustworthy scientific generation therefore requires evidence-grounded explanation: generated claims should map to authorized experimental records, graph paths, and provenance. This requirement also creates a privacy-explainability tension. Greater transparency may reveal protected information, while aggressive privacy controls may make explanations scientifically uninformative. EPPSR explicitly places explanation downstream of privacy enforcement.

3. Proposed EPPSR Framework

EPPSR is designed around five requirements: privacy preservation, evidence grounding, relational reasoning, explainability and traceability, and confidence awareness. It separates storage, retrieval, privacy processing, reasoning, generation, and verification so that the LLM never receives unrestricted access to the underlying scientific repository. The layered architecture and controlled privacy boundary are shown in Figure 1.



Controlled privacy boundary: raw scientific records are not exposed directly to the generative model.

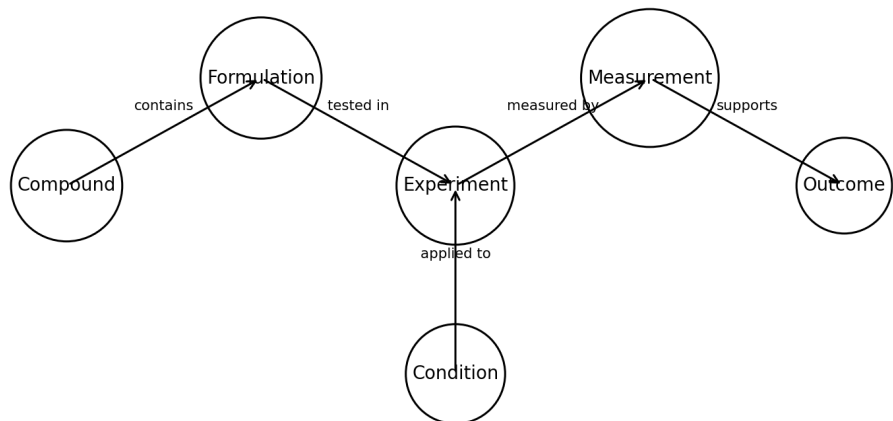
Figure 1. Layered system architecture of the proposed EPPSR framework.

A. Scientific Data and Knowledge Representation

Let the scientific repository be $D = \{d_1, d_2, \dots, d_n\}$, where each d_i is a scientific record containing structured, semi-structured, or unstructured experimental information. The Scientific Knowledge Graph is $G = (V, E)$, where V contains entities such as Experiment, Compound, Formulation, Condition, Measurement, Method, and Outcome, and E contains typed relations such as contains, testedIn, measuredBy, supports, validates, or contradicts.

$$G = (V, E)$$

The NoSQL repository remains the authoritative evidence source. The graph acts as a relational index and each edge retains provenance linking it to one or more source records. This design avoids treating graph-derived relationships as detached facts. An example of this traceable scientific evidence path is shown in

Figure 2.**Figure 2.** Example scientific knowledge graph and evidence path.

B. Query Understanding and Privacy Boundary

A query q is transformed into $Q(q) = \{Iq, Vq, Cq, Sq\}$, representing intent, detected scientific entities, contextual constraints, and sensitivity indicators. Authorization determines which records, graph entities, and attributes may participate in retrieval. Differential privacy is then used where formal statistical disclosure protection is appropriate; it is not treated as a substitute for access control.

$$\Pr[M(D) \in S] \leq \exp(\epsilon) \Pr[M(D') \in S]$$

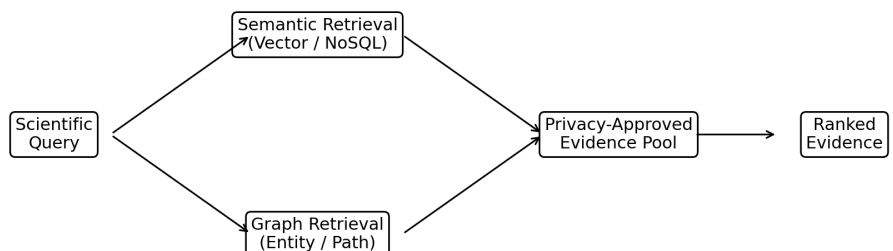
For a numerical function $f(D)$, the Laplace mechanism is $M(D) = f(D) + \text{Lap}(\Delta f / \epsilon)$, where Δf denotes sensitivity and ϵ is the privacy budget [25] [26].

$$M(D) = f(D) + \text{Lap}(\Delta f / \epsilon)$$

C. Dual-Channel Retrieval and Graph Reasoning

EPPSR retrieves evidence through two coordinated channels. Semantic retrieval ranks NoSQL records using query-record similarity, while graph retrieval traverses entities and paths relevant to the query. The candidate set is the union of both channels.

$R(q) = R_{DB(q)} \cup R_{KG(q)}$. The resulting dual-channel semantic and graph retrieval flow is illustrated in **Figure 3**.

**Figure 3.** Dual-channel semantic and graph-based retrieval.

For multi-hop scientific questions, candidate paths π are scored using semantic relevance, relational relevance, provenance strength, and a length penalty. This al-

lows concise evidence paths to be preferred over long or weakly supported chains.

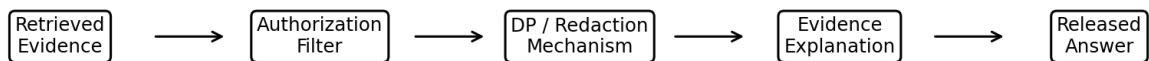
$$S\pi = \alpha S_{sem} + \beta S_{rel} + \gamma S_{prov} - \lambda L\pi$$

D. Privacy-Aware Evidence Ranking

Candidate evidence is ranked using a composite score. In contrast with conventional RAG, highly relevant evidence can be demoted when disclosure risk is high. This mechanism creates an explicit privacy-utility trade-off.

$$S(ei) = w1S_{sem} + w2S_{graph} + w3S_{prov} + w4S_{cons} - w5S_{risk}$$

The highest-ranked privacy-approved evidence E^* is provided to the generator. Sensitive identifiers may be redacted, and eligible numerical outputs may be perturbed using DP. The explanation layer receives only the same authorized and privacy-processed evidence. The privacy-aware evidence and explanation release boundary is shown in **Figure 4**.



Explanation is generated only from authorized, privacy-processed evidence.

Figure 4. Privacy-aware evidence and explanation release boundary.

E. Evidence-Grounded Generation, Confidence, and Verification

The generation stage is constrained to the approved evidence set and relevant graph paths. A response is represented as $O = \{A, E, X, C, P\}$, containing the answer, released evidence, privacy-compliant explanation, calibrated confidence, and privacy summary. Confidence is derived from evidence strength, graph support, agreement, and provenance rather than the LLM's verbal certainty.

$$C(q) = \theta1Es + \theta2Gs + \theta3As + \theta4Ps, \sum \theta i = 1$$

Generated claims are verified against the approved evidence. For claims $CA = \{c1, \dots, cm\}$, evidence coverage is $EC(A) = \sum V(ci)/m$, where $V(ci) = 1$ when sufficient supporting evidence exists. Unsupported claims can be removed, regenerated, or explicitly marked as inference.

$$EC(A) = (\sum V(ci))/m$$

4. Experimental Methodology

A. Privacy Threat Model

The V2 evaluation defines an explicit adversary and release surface. The adversary is assumed to have legitimate access to the natural-language query interface and to observe answer text, citations or released record identifiers, graph-path or relationship summaries, provenance/confidence metadata, and outputs from repeated queries. Protected information includes sensitive record identity, sensitive

graph relationships, and exact numeric values used in aggregate releases. This model intentionally treats explanations and graph paths as potential disclosure channels rather than assuming that only source documents are sensitive.

Two privacy concepts are evaluated separately. First, formal differential privacy is applied only to a bounded numeric aggregate release. Second, Sensitive Information Exposure Rate (SIER) measures application-layer disclosure of sensitive record identifiers or graph edges. A zero SIER therefore means that no such un-redacted disclosure occurred under the benchmark release policy; it is not, by itself, an epsilon-DP guarantee.

B. Formal Differential-Privacy Release

For each benchmark query, the protected numeric output is the mean effect percentage across a fixed-size cohort of $m = 6$ matching records. Neighboring datasets D and D' use replace-one adjacency: both contain six records and differ in exactly one record. Each effect value is clipped to $[L, U] = [-18, 18]$. The global sensitivity of the bounded mean is therefore:

$$\Delta_{mean} = (U - L)/m = 36/6 = 6.$$

The released statistic uses the Laplace mechanism:

$$M(D) = \text{mean}(\text{clip}(D, -18, 18)) + \text{Laplace}(0, \Delta_{mean}/\epsilon_q).$$

The per-query privacy budget ϵ_q was swept over $\{0.1, 0.5, 1.0, 2.0, 5.0\}$. For repeated releases, pure sequential composition is used: $\epsilon_{total} \leq \sum \epsilon_q$. Consequently, if all 80 releases were issued at $\epsilon_q = 1.0$, the basic worst-case composed budget would be $\epsilon_{total} \leq 80$. A second experiment allocates fixed total budgets across all 80 releases to show the utility consequences of repeated-query accounting.

C. Synthetic Dataset and Query Generator

The generator uses random seed 20260914. The base design contains 8 compounds, 3 process conditions, and 4 measurement categories, yielding 96 compound-condition-measurement cohorts. Six evidence records are generated per cohort (576 base records): three reliable observations sampled from $\text{Normal}(\mu, 1.6)$ with provenance $U(0.85, 1.00)$; one contradictory observation sampled from $\text{Normal}(-\mu, 2.5)$ with provenance $U(0.25, 0.50)$; one partial-text observation sampled from $\text{Normal}(\mu, 2.2)$ with provenance $U(0.60, 0.80)$; and one sensitive strong-evidence observation sampled from $\text{Normal}(1.15\mu, 1.0)$ with provenance $U(0.90, 1.00)$. The latent μ has random sign and magnitude $U(4, 14)$, and all observed effects are clipped to $[-18, 18]$.

An additional 96 lexical distractor records contain query-like canonical terms while their structured condition and measurement fields refer to different concepts, producing 672 total records. Sensitive graph edges are assigned to every sensitive record plus an additional Bernoulli (0.10) sample. Eighty of the 96 cohorts are sampled without replacement to create benchmark queries; query wording alternates between canonical terminology and synonymized wording. The gold answer direction is the sign of the mean reliable evidence, and the gold citation is the highest-provenance reliable non-sensitive source when one is available.

Table 1 summarizes the complete V2 synthetic dataset configuration.

Table 1. V2 synthetic dataset configuration.

Item	Value
Compounds	8
Conditions	3
Measurements	4
Base cohorts	96
Records per cohort	6
Base records	576
Lexical distractors	96
Total records	672
Benchmark queries	80
Random seed	20260914

D. Shared Retrieval Implementation and Baselines

All architectures share the same synthetic corpus and query set. Semantic retrieval uses scikit-learn TfidfVectorizer with unigram and bigram features, lower-case normalization, sublinear term frequency, and cosine similarity. No neural embedding model is claimed in V2. The Scientific Knowledge Graph is deterministically derived from record fields connecting compound, formulation, experiment, condition, measurement, and outcome. Graph relevance is the normalized match over compound, condition, and measurement, corresponding to a conceptual traversal depth of at most three relations.

RAG uses semantic similarity only. PP-RAG uses the same semantic score with a 0.20 privacy-risk penalty. GraphRAG combines semantic and graph relevance with weights 0.55 and 0.45. Full EPPSR uses semantic = 0.40, graph = 0.25, provenance = 0.15, consistency = 0.20, and subtracts privacy risk with weight 0.20. Top-k is fixed at five. The implementation is a CPU-local Python simulation using NumPy, pandas, scikit-learn, and Matplotlib; no external LLM, GPU, or remote inference service is used. The shared implementation and experimental parameters are summarized in **Table 2**.

Table 2. V2 experimental parameters.

Parameter	Value
Top-k	5
EPPSR semantic weight	0.40
EPPSR graph weight	0.25
EPPSR provenance weight	0.15
EPPSR consistency weight	0.20
EPPSR privacy-risk penalty	0.20
GraphRAG semantic/graph	0.55/0.45
DP clipping	[−18, 18]
DP sensitivity	6

Continued

ϵ sweep	0.1, 0.5, 1, 2, 5
Bootstrap samples	5000
DP Monte Carlo releases	300/query/ ϵ

E. Evaluation Metrics

Answer accuracy is the fraction of queries for which the evidence-weighted predicted direction matches the gold direction. Recall@5 indicates whether the designated gold citation appears in the top-five set. Citation accuracy requires the top-ranked citation to equal the designated gold citation. SIER is defined as $N_{\text{sensitive-unredacted}}/N_{\text{queries}}$ and is reported independently of formal DP.

The former composite explanation score is removed. Explainability is instead decomposed into four observable measures: Evidence Coverage, the fraction of required answer dimensions (compound, condition, measurement, and direction) represented by top-five evidence; Provenance Completeness, the mean generated provenance-quality score of top-five evidence; Graph-Path Traceability, whether at least one top-five record provides a complete compound-condition-measurement graph match; and Citation Support, whether the top citation exactly matches the query entities and supports the generated direction.

F. Statistical Analysis and Ablation Design

Each of the 80 queries is evaluated under RAG, PP-RAG, GraphRAG, and EPPSR, yielding 320 main query-method observations. Query-level results are retained. Five-thousand-sample nonparametric bootstrap confidence intervals are computed for the principal rates and explanation components. The EPPSR ablation study evaluates the full system, removal of provenance, removal of consistency, removal of the privacy-risk scoring term, and a semantic-plus-graph-only configuration. The DP epsilon sweep uses 300 independent Laplace releases per query per epsilon value to estimate numerical error and direction utility. The formal DP privacy-utility evaluation design is illustrated in **Figure 5**.

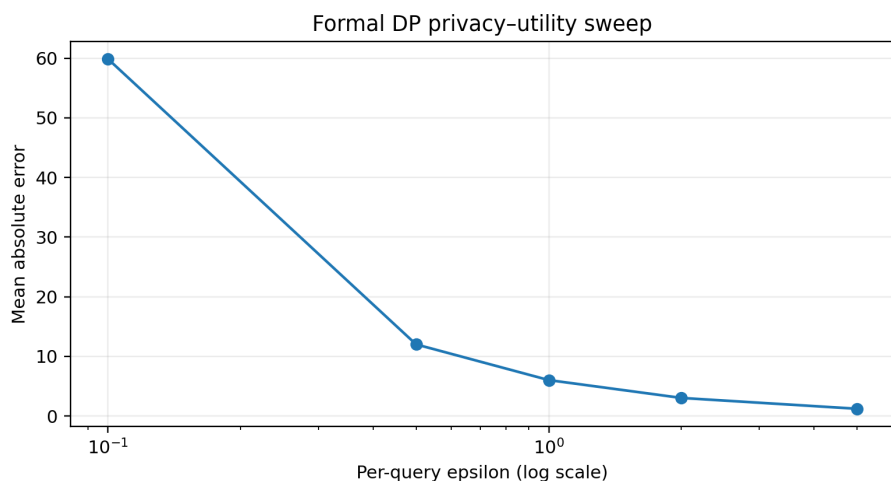


Figure 5. V2 formal-DP privacy-utility evaluation design: mean absolute error across per-query epsilon values.

5. Results and Discussion

A. Retrieval, Grounding, and Sensitive Exposure

Table 3 summarizes the controlled V2 benchmark. EPPSR achieved 100.00% answer accuracy and 97.50% Recall@5, compared with 66.25% and 75.00% for GraphRAG, 60.00% and 13.75% for PP-RAG, and 58.75% and 10.00% for RAG. The answer-accuracy comparison is visualized in **Figure 6**. These values should be interpreted within the intentionally structured synthetic benchmark: contradictory low-provenance records, partial text, sensitive high-quality evidence, and lexical distractors were generated specifically to test the signals used by the architectures. They demonstrate controlled architectural behavior rather than general production superiority.

Citation accuracy is substantially more demanding because it requires the top-ranked record to equal a single designated high-provenance non-sensitive gold source. EPPSR reached 46.25% citation accuracy, compared with 8.75% for GraphRAG, 6.25% for PP-RAG, and 5.00% for RAG. EPPSR's 95% bootstrap interval for citation accuracy was 35.00% - 56.28%. RAG and GraphRAG produced SIER values of 41.25% and 37.50%, respectively, whereas PP-RAG and EPPSR produced 0.00% SIER under the benchmark redaction policy. This result is an application-layer exposure result and is not described as a differential-privacy guarantee. **Figure 7** visualizes the SIER comparison across the four architectures.

Table 3. V2 overall results (%).

Method	Accuracy	R@5	Citation	SIER	Prov.	Cit. Support
RAG	58.75	10.00	5.00	41.25	72.14	40.00
PP-RAG	60.00	13.75	6.25	0.00	71.68	36.25
GraphRAG	66.25	75.00	8.75	37.50	75.07	86.25
EPPSR	100.00	97.50	46.25	0.00	88.60	92.50

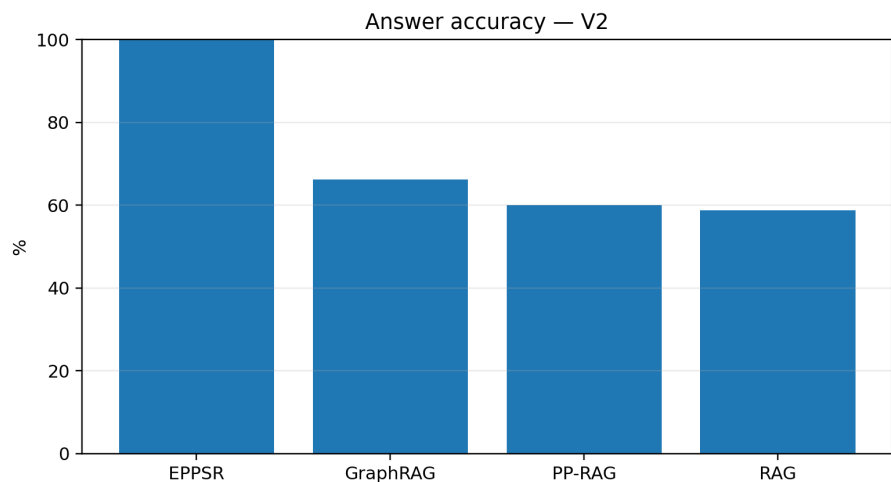


Figure 6. Answer accuracy across retrieval architectures in the V2 synthetic benchmark.

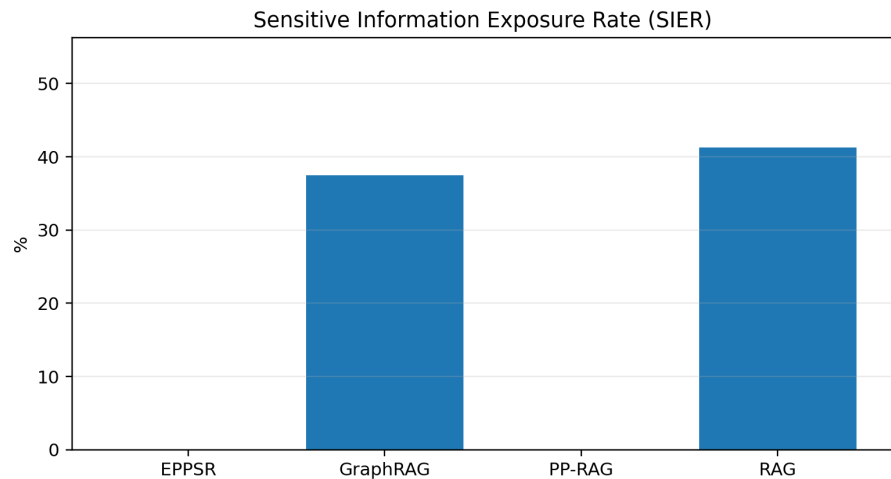


Figure 7. Sensitive Information Exposure Rate (SIER). SIER is separate from formal differential privacy.

B. Evidence-Level Explainability

EPPSR achieved 100.00% evidence coverage, 88.60% provenance completeness, 97.50% graph-path traceability, and 92.50% citation support. GraphRAG also achieved 100.00% graph-path traceability, as expected from its graph component, but its provenance completeness was 75.07%. **Figure 8** compares provenance completeness across the architectures. These component measures replace the previous 92.01% composite explanation score and make the explanation evaluation reproducible: each score has a stated definition and can be recomputed directly from query-level evidence.

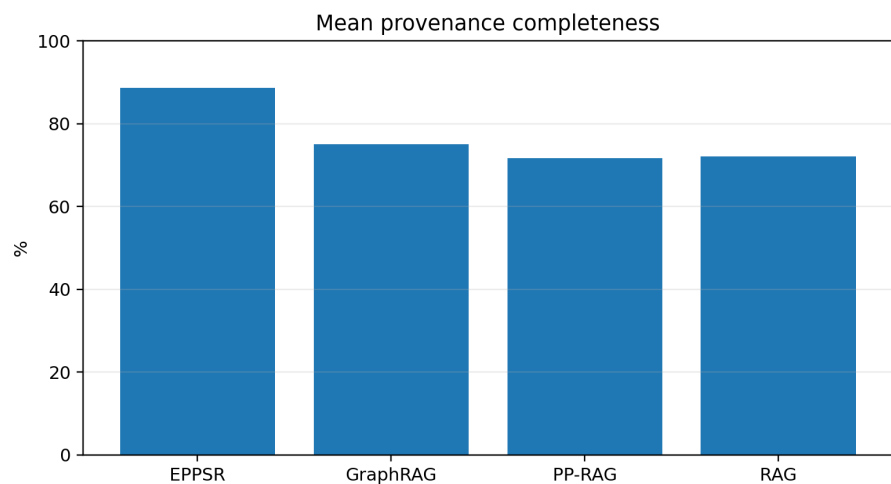


Figure 8. Mean provenance completeness of top-five evidence across architectures.

C. Formal Differential-Privacy Utility

The epsilon sweep produces the expected privacy-utility pattern. At $\epsilon = 0.1$, mean absolute error (MAE) was 59.85 effect-percentage points and direction accuracy was 55.00%. At $\epsilon = 0.5$, MAE fell to 12.00 and direction accuracy rose to 70.51%. At $\epsilon = 1.0$, MAE was 6.01 and direction accuracy 82.02%. Utility continued to improve at

$\epsilon = 2.0$ (MAE 3.02; 92.32% direction accuracy) and $\epsilon = 5.0$ (MAE 1.21; 99.00% direction accuracy). The result illustrates the expected trade-off: lower epsilon provides a stronger per-release privacy bound but injects more noise into this small-cohort aggregate. The full epsilon sweep is reported in **Table 4**, while **Figure 9** and **Figure 10** show the corresponding MAE and direction-accuracy trends.

Table 4. Formal dp epsilon sweep.

ϵ/query	Mean AE	Median AE	Direction Acc. %	Δ
0.1	59.85	41.06	55.00	6.0
0.5	12.00	8.29	70.51	6.0
1	6.01	4.17	82.02	6.0
2	3.02	2.09	92.32	6.0
5	1.21	0.83	99.00	6.0

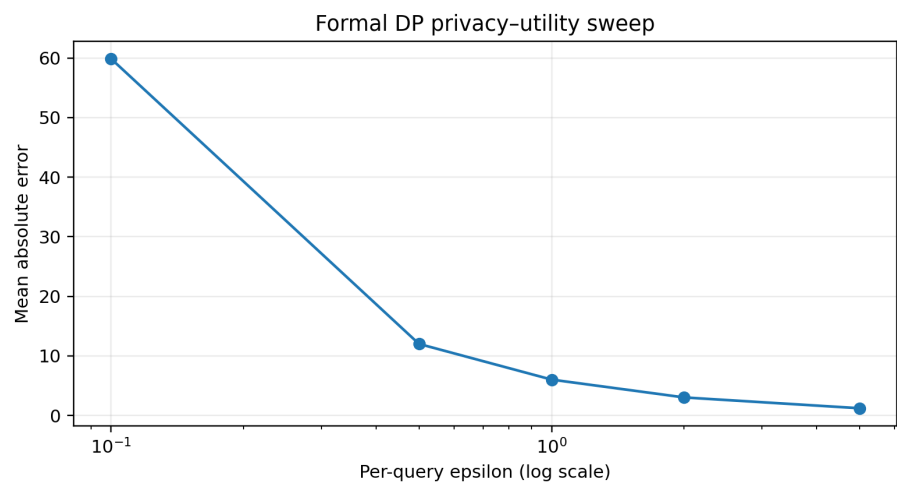


Figure 9. Mean absolute error versus per-query epsilon for the bounded aggregate release.

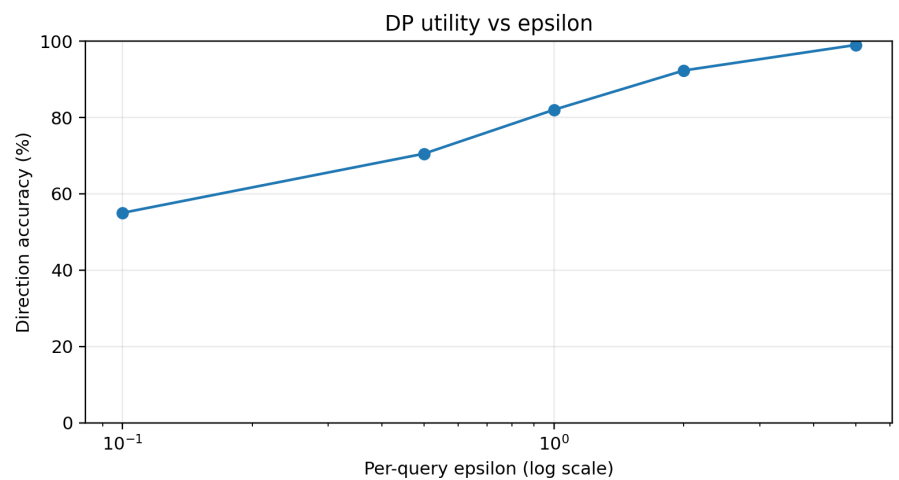


Figure 10. Direction accuracy versus per-query epsilon.

D. Repeated-Query Privacy Accounting

Basic sequential composition exposes an important limitation of interpreting $\epsilon = 1.0$ in isolation. Eighty independent releases at $\epsilon_q = 1.0$ have a worst-case pure-DP bound $\epsilon_{\text{total}} \leq 80$. Conversely, dividing a fixed total privacy budget equally across 80 queries can severely reduce utility in this small $m = 6$ cohort. For example, $\epsilon_{\text{total}} = 8$ implies $\epsilon_q = 0.1$, for which the repeated-query simulation yielded approximately 54.90% direction accuracy and MAE 60.18. Even $\epsilon_{\text{total}} = 40$, corresponding to $\epsilon_q = 0.5$, yielded 70.71% direction accuracy and MAE 11.88. The experiment therefore supports explicit budget accounting and discourages presenting a per-query epsilon without a release policy. The fixed-total-budget results across all 80 releases are summarized in **Table 5**.

Table 5. Fixed total privacy budget across 80 releases.

ϵ total	ϵ/query	Mean AE	Direction Acc. %
1	0.0125	478.53	50.71
4	0.0500	119.98	52.90
8	0.1000	60.18	54.90
20	0.2500	24.12	61.76
40	0.5000	11.88	70.71

E. EPPSR Ablation

The ablation study isolates the contribution of EPPSR scoring components. Removing consistency reduced answer accuracy from 100.00% to 91.25%, indicating that agreement with the evidence cohort materially affected answer selection in this benchmark. Removing provenance preserved 100.00% answer accuracy but reduced provenance completeness from 88.60% to 79.36% and citation accuracy from 46.25% to 33.75%, showing that provenance primarily affects evidence quality and source selection. Removing the privacy-risk scoring term preserved answer accuracy but reduced citation accuracy to 26.25%; because release-time redaction remained active, SIER stayed at 0%. Finally, the semantic-plus-graph-only variant fell to 66.25% answer accuracy, matching the utility level of GraphRAG-like scoring and showing that the additional EPPSR signals drive much of the controlled-benchmark gain. **Table 6** reports the complete ablation results, and **Figure 11** highlights the effect of each configuration on answer accuracy.

Table 6. Eppsr ablation results (%).

Variant	Accuracy	Citation	SIER	Prov.
EPPSR	100.00	46.25	0.00	88.60
no_consistency	91.25	40.00	0.00	83.88
no_privacy_risk	100.00	26.25	0.00	90.44
no_provenance	100.00	33.75	0.00	79.36
semantic_graph_only	66.25	8.75	0.00	77.04

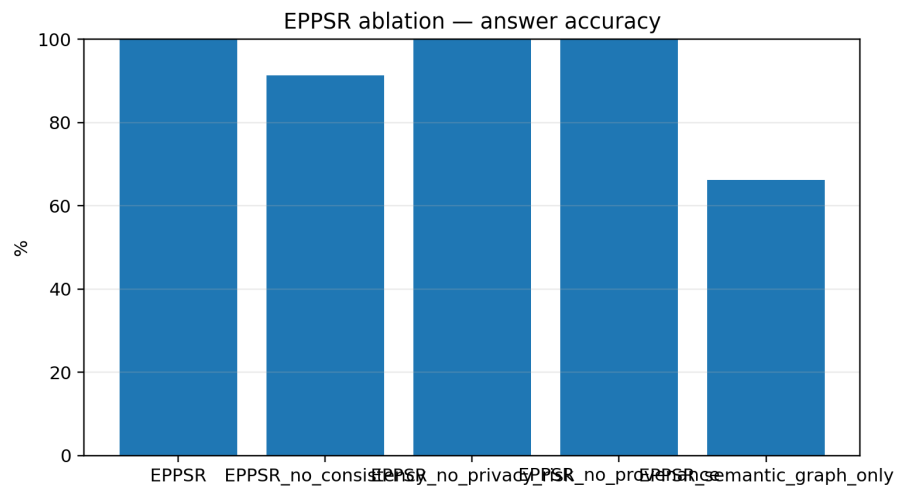


Figure 11. EPPSR ablation: answer accuracy.

F. Statistical Uncertainty

Bootstrap uncertainty is reported rather than relying on point estimates alone. For answer accuracy, the 95% bootstrap intervals were 47.50% - 70.00% for RAG, 48.75% - 70.00% for PP-RAG, 56.25% - 76.25% for GraphRAG, and 100.00% - 100.00% for EPPSR. Citation-accuracy intervals were 1.25% - 10.00%, 1.25% - 12.50%, 3.75% - 15.00%, and 35.00% - 56.28%, respectively. SIER intervals were 30.00% - 52.50% for RAG, 0.00% - 0.00% for PP-RAG, 26.25% - 47.50% for GraphRAG, and 0.00% - 0.00% for EPPSR. These intervals make clear that small differences among low-frequency rates should not be overinterpreted.

6. Limitations and Threats to Validity

The V2 benchmark is synthetic and intentionally constructed to create retrieval, provenance, contradiction, and privacy conflicts. The strong EPPSR result therefore establishes behavior under controlled assumptions, not superiority on independent scientific corpora. Real repositories contain entity-resolution errors, missing or stale metadata, schema drift, duplicated experiments, complex access policies, and domain-specific semantics that are not fully represented here.

The implementation uses TF-IDF and cosine similarity rather than a neural embedding model and does not execute a production LLM. Consequently, the experiment does not measure model hallucination, generative fluency, prompt-injection robustness, GPU latency, or model-specific faithfulness. The previous simulated hallucination rate and arbitrary composite explanation score have therefore been removed from V2.

Formal DP is deliberately scoped to the bounded aggregate mean-effect release. Authorization, redaction, graph filtering, and privacy-risk reranking are security/privacy controls but do not independently establish epsilon-DP. SIER captures observed application-layer exposure under the specified release policy; 0% SIER does not imply immunity to membership inference, reconstruction, graph inference, collusion, side-channel attacks, or attacks outside the benchmark threat model.

Pure sequential composition is conservative but transparent. More advanced accounting mechanisms, subsampling, larger aggregation cohorts, query caching, privacy filters, and workload-aware budget allocation could improve utility for repeated access. These approaches should be evaluated before deployment. In addition, provenance-quality labels are synthetically generated and should be replaced by auditable real-world provenance criteria and expert validation.

Finally, the 80-query benchmark yields finite-sample uncertainty even with bootstrap reporting. The 100% EPPSR answer-accuracy result is specific to this seeded synthetic design and should be validated through multiple independent seeds, larger benchmark families, external scientific datasets, production embeddings, real knowledge-graph construction pipelines, and blinded domain-expert evaluation.

7. Future Work

Future work should move beyond the V2 synthetic proof-of-concept by validating EPPSR on real scientific repositories, production embedding models, multiple LLMs, and independently constructed knowledge graphs. Privacy research should extend the present bounded-aggregate Laplace mechanism with advanced composition/accounting, larger cohorts, privacy filters, and empirical attacks including membership inference, reconstruction, graph inference, and prompt injection. Explainability should be validated through blinded expert studies using the component metrics introduced here rather than an unvalidated composite score.

8. Conclusion

This paper extends prior privacy-preserving LLM integration with scientific NoSQL repositories [1] by adding knowledge-graph reasoning, evidence ranking, provenance, confidence estimation, and privacy-aware explanation. The resulting EPPSR architecture treats trustworthy scientific retrieval as a multi-objective problem rather than an accuracy-only problem. In the executed synthetic proof-of-concept, GraphRAG achieved the highest raw answer accuracy, whereas EPPSR achieved the strongest combined privacy and explainability profile while retaining competitive accuracy. These findings do not establish production superiority; instead, they provide a reproducible architectural baseline and demonstrate why privacy, evidence grounding, and explainability must be designed together. Real scientific datasets, production LLMs, adversarial privacy testing, and domain-expert evaluation are required for the next stage of validation.

Conflicts of Interest

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

References

- [1] 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>

- [2] Vaswani, A., *et al.* (2017) Attention Is All You Need. arXiv: 1706.03762.
- [3] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019) 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, Volume 1 (Long and Short Papers)*, Minneapolis, 2-7 June 2019, 4171-4186. <https://doi.org/10.18653/v1/n19-1423>
- [4] 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. <https://doi.org/10.34218/ijcet.14.01.011>
- [5] Lewis, P., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *34th Conference on Neural Information Processing Systems NeurIPS 2020*, Vancouver, 6-12 December 2020, 9459-9474.
- [6] Gao, Y., *et al.* (2023) Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv: 2312.10997.
- [7] Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G.D., Gutierrez, C., *et al.* (2021) Knowledge Graphs. *ACM Computing Surveys*, **54**, 1-37. <https://doi.org/10.1145/3447772>
- [8] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. and Wu, X. (2024) Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, **36**, 3580-3599. <https://doi.org/10.1109/tkde.2024.3352100>
- [9] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bannetot, A., Tabik, S., Barbado, A., *et al.* (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, **58**, 82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [10] Radford, A., *et al.* (2019) Language Models Are Unsupervised Multitask Learners. OpenAI Technical Report9.
- [11] Raffel, C., *et al.* (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, **21**, 1-67.
- [12] Brown, T., *et al.* (2020) Language Models Are Few-Shot Learners. *Proceedings of the 34th International Conference on Neural Information Processing System*, Vancouver, 6-12 December 2020, 1877-1901.
- [13] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., *et al.* (2020) Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 16-20 November 2020, 6769-6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [14] Khattab, O. and Zaharia, M. (2020) ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 25-30 July 2020, 39-48. <https://doi.org/10.1145/3397271.3401075>
- [15] Thakur, N., *et al.* (202021) BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models. arXiv: 2104.08663.
- [16] Izacard, G. and Grave, E. (2021) Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 19-23 April 2021, 874-880. <https://doi.org/10.18653/v1/2021.eacl-main.74>
- [17] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M. (2020) REALM: Retrieval-Aug-

- mented Language Model Pre-Training. *Proceedings of the 37th International Conference on Machine Learning*, 13-18 July 2020, 3929-3938.
- [18] Yates, A., Nogueira, R. and Lin, J. (2021) Pretrained Transformers for Text Ranking: BERT and beyond. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Tutorials*, 6-11 June 2021, 1-4.
<https://doi.org/10.18653/v1/2021.naacl-tutorials.1>
- [19] Maynez, J., Narayan, S., Bohnet, B. and McDonald, R. (2020) On Faithfulness and Factuality in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5-10 July 2020, 1906-1919.
<https://doi.org/10.18653/v1/2020.acl-main.173>
- [20] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., *et al.* (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 24824-24837. <https://doi.org/10.52202/068431-1800>
- [21] Lin, S., Hilton, J. and Evans, O. (2022) TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, 22-27 May 2022, 3214-3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
- [22] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, 1-5 November 2016, 2383-2392. <https://doi.org/10.18653/v1/d16-1264>
- [23] Zhong, L., Wu, J., Li, Q., Peng, H. and Wu, X. (2023) A Comprehensive Survey on Automatic Knowledge Graph Construction. *ACM Computing Surveys*, **56**, 1-62.
<https://doi.org/10.1145/3618295>
- [24] Socher, R., Chen, D.Q., Manning, C.D. and Ng, A.Y. (2013) Reasoning with Neural Tensor Networks for Knowledge Base Completion. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Nevada, 5-10 December 2013, 926-934.
- [25] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. (2013) Translating Embeddings for Modeling Multi-Relational Data. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Nevada, 5-10 December 2013, 2787-2795.
- [26] Nickel, M., Murphy, K., Tresp, V. and Gabrilovich, E. (2016) A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, **104**, 11-33.
<https://doi.org/10.1109/jproc.2015.2483592>
- [27] Wang, Q., Mao, Z., Wang, B. and Guo, L. (2017) Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 2724-2743. <https://doi.org/10.1109/tkde.2017.2754499>
- [28] Singhal, A. (2012) Introducing the Knowledge Graph: Things, Not Strings. Google.
- [29] Dwork, C. (2006) Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., Eds., *Automata, Languages and Programming*, Springer, 1-12.
https://doi.org/10.1007/11787006_1
- [30] 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>
- [31] Carlini, N., *et al.* (2021) Extracting Training Data from Large Language Models. *Pro-*

- ceeding of *USENIX Security Symposium* 2021, 11-13 August 2021, 2633-2650.
- [32] Fredrikson, M., Jha, S. and Ristenpart, T. (2015) Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver, 12-16 October 2015, 1322-1333. <https://doi.org/10.1145/2810103.2813677>
 - [33] 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>
 - [34] Nasr, M., Shokri, R. and Houmansadr, A. (2019) Comprehensive Privacy Analysis of Deep Learning. *Proceeding of IEEE Symposium on Security and Privacy* 2019, San Francisco, 20-22 May 2019, 1-15.
 - [35] Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. and Talwar, K. (2017) Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data. arXiv: 1610.05755.
 - [36] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, **37**, 50-60. <https://doi.org/10.1109/msp.2020.2975749>
 - [37] McMahan, H.B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* 2017, Fort Lauderdale, 20-22 April 2017, 1273-1282.
 - [38] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. <https://doi.org/10.1145/2939672.2939778>
 - [39] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing System*, Long Beach, 4-9 December 2017, 4768-4777.
 - [40] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, **51**, 1-42. <https://doi.org/10.1145/3236009>
 - [41] Doshi-Velez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning. arXiv: 1702.08608.
 - [42] Lipton, Z.C. (2018) The Mythos of Model Interpretability. *Queue*, **16**, 31-57. <https://doi.org/10.1145/3236386.3241340>
 - [43] Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, **31**. <https://doi.org/10.2139/ssrn.3063289>