TrustChain-RAG: A Blockchain-Anchored Context Graph Framework for Auditable Knowledge Mining in a Gated Environment

Jaromir Dzialo[1] [0009−0003−9868−5182]

Department of Applied Computer Science, AGH University of Krakow,
al. A.Mickiewicza 30, 30-059 Krakow, Poland
dzialo@agh.edu.pl

Abstract. Regulated enterprises cannot adopt generative AI when the system offers no proof of how an answer was produced or who was authorized to see the evidence behind it. The most capable LLMs are cloud-hosted and off-limits under data sovereignty mandates. Organizations fail to provide sanctioned alternatives, so employees use consumer AI tools and leak exactly the data that restrictions were meant to protect. Without traceability, auditability, accountability (and other '-ilities') built into the architecture by design, enterprise stakeholders will not trust AI-generated knowledge. Without trust, adoption does not happen.

This paper presents TrustChain-RAG, an architecture born from five years of operating a multilingual knowledge pipeline (100M+ words, six European languages) for a Life Sciences organization, building on prior patents. The framework combines multilingual Context Graphs with lattice-aligned RBAC, locally-hosted RAG, and a private blockchain audit trail. Cross-lingual bridge nodes in the context graph outperform the best current multilingual embeddings (BGE-M3) by +15pp in precision on specialized domains. Evaluation across three regulated domains, six European languages, and five LLMs (7B-72B parameters) shows 50-62% hallucination reduction over standard RAG. Controlled ablations isolate this to graph topology rather than retrieval volume or fine-tuning. The accuracy margin holds at +12 FA points regardless of model scale, from 7B to 70B. RBAC enforcement holds at 1.3% leakage versus 6% for metadata pre-filtering. Failure analysis at the 70B scale shows the error bottleneck shifting from model reasoning to knowledge structure: ontology gaps, bridge calibration – confirming that the framework and the model address different limitations.

Keywords: Artificial Intelligence in Scientific Computing · Hybrid Computational Methods · Computational Trust · Data Sovereignty · RAG · Retrieval-Augmented Generation · Context Graphs · Blockchain · Audit Trail · RBAC · Knowledge Mining · Shadow AI

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

1 Introduction

1.1 Motivation

For the past five years I have been running a multilingual knowledge processing pipeline for a large Life Sciences organization, building on my prior work on semantic document representation [12,13] and translation workflow [22]. Over 100 million words have passed through it: regulatory submissions, clinical documentation, pharmacovigilance reports, spread across six European languages. That operational experience motivates everything in this paper. It taught me that the governance question: "can we prove how this answer was produced?" – dominates technical ones. In all stakeholders meetings and governance board reviews I participated in, the first question was never about F1 scores. It was about who accessed what and when, which sources contributed, and whether there is a record for the auditor to examine.

Standard RAG [1] cannot answer these questions. No record persists of which documents were retrieved or how the context window was assembled. Access control, when it exists, runs as a metadata input/output filter at the API gateway level (or as NER e.g. for PII related anonymization purposes) with no connection to retrieval semantics.

When organizations cannot deploy AI through governed channels, employees use consumer LLMs on personal accounts. Cyberhaven's 2024 analysis found 27.4% of data pasted into ChatGPT to be sensitive [2]. Samsung banned generative AI internally after engineers uploaded proprietary semiconductor files [4]. Cisco reported 48% of employees entering non-public data into public AI [5]. In pharma, clinical trial data and pharmacovigilance records in ICH E2B format fall under FDA 21 CFR Part 11 and HIPAA, so pasting any of it into a public LLM is a regulatory violation. But a medical writer who can draft a Module 2.5 Clinical Overview in two hours instead of two days will not voluntarily forgo the tool [3]. That is not a hypothesis; I have seen it happen too many times. TrustChain-RAG is designed to fix this data leakage.

1.2 Problem Statement

P1: Semantic Fragmentation. Cross-lingual relationships are poorly preserved in flat embedding spaces [6]. Even current top-tier multilingual embedders like BGE-M3 [7] and mE5 [8] handle distributional similarity well, but lose relational structure: hypernymy, regulatory equivalence, domain-specific disambiguation – that specialized retrieval depends on.

P2: Retrieval-Inference Opacity. Intermediate steps of RAG are ephemeral [9]. Hallucination [10] becomes difficult to diagnose when there is no trace of what the model actually saw, with no steps to reproduce it.

P3: No Immutable Provenance. Retrieval logs, where they exist, typically reside in mutable relational databases or application logs with no integrity guarantees. Standard database audit logs can be modified by administrators

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

with write access, a limitation recognized in NIST SP 800-53 AU-10 (content non-repudiation) [11].

P4: Disconnected Access Control. Application-layer access control operates independently from retrieval semantics and generates no audit record that could be correlated with knowledge provenance [11].

1.3 Contributions

1. **TrustChain-RAG**, integrating Context Graphs, lattice-aligned RBAC, locally-hosted RAG, and private blockchain, with properties that arise from the integration and that no component provides in isolation (Sections 3–4).
2. A method for **multilingual context graph construction** building on the semantic document representation of [12] and the semantic network generation of [13]. This is the core technical contribution; graph structural analysis and controlled ablation confirm it is topology, not retrieval volume or fine-tuning, that drives accuracy improvement (Section 5.3).
3. **Lattice-aligned RBAC with blockchain-anchored governance**, creating a single audit surface for identity management and data lineage.
4. A **threat model** with five adversary classes, proofs of RBAC-filtered retrieval safety and completeness (Sections 3.2, 4).
5. **Multi-scale evaluation** across three domains, six languages, five LLMs (7B–72B), including failure analysis that shows a qualitative shift in error distribution as model capacity increases (Section 5).

2 Related Work

RAG architectures. Self-RAG [14] adds iterative refinement, RAPTOR [15] builds hierarchical summaries, GraphRAG [16,17] applies community detection over knowledge graphs. None address the governance properties that would let an organization offer RAG as a sanctioned internal tool. As for commercial platforms: Elasticsearch [18], Weaviate [19], Pinecone [20] – enforce access as pre-retrieval metadata filters, which works for simple cases but breaks down when entities co-embed across classification levels, when document-level filters don't map cleanly to chunks, and when the access decisions leave no tamper-evident record. TrustChain-RAG provides the trust layer they currently lack.

Multilingual embeddings. BGE-M3 [7], mE5 [8], and XLM-R [21] perform well on general cross-lingual benchmarks. Section 5.2 tests directly whether explicit bridge nodes add value on top of BGE-M3 in specialized domains. They do – 15 pp in cross-lingual precision – for reasons discussed in Section 3.4.

Knowledge graphs and provenance. Context graphs here extend the semantic structuring of [12,13], with cross-lingual bridge confidence from [22]. RBAC [23] and lattice models [24] have decades of deployment history but have not been applied to RAG pipelines. For provenance, Section 3.5 compares blockchain to append-only logs [25], transparency logs [26], and lineage frameworks [27,28].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

3 TrustChain-RAG Architecture

3.1 System Overview

Six on-premise subsystems: Document Ingestion (DISPL), Multilingual Context Graph Engine (MCGE), RBAC Subsystem (RBACS), Trust-Aware Retrieval (TARM), Local LLM Inference (LLIE), Blockchain Audit Trail (BATS).

Definition 1. *A TrustChain-RAG system is* $\mathcal{T} = (D, G, \mathcal{A}, R, L, B, \sigma)$*: document corpus* D*; context graph* $G = (V, E, \lambda, \mu)$*; RBAC model* $\mathcal{A}$*; retrieval function* $R\colon Q \times G \times \tau \to 2^V$*; inference function* L*; blockchain* B*; anchoring function* σ*.*

The framework is model-agnostic: LLIE accepts any causal LLM served through a vLLM-compatible interface [29]. Section 5 tests five models (7B–72B); the architecture imposes no coupling to any specific model family.

3.2 Threat Model

Assets: corpus, graph, query-response history, audit trail, classification boundaries, role assignments.

Five adversary classes. ADV-1: insider who modifies documents after ingestion, for instance to retroactively change a source cited in a regulatory response. ADV-2: insider who alters audit records to conceal a policy violation. ADV-3: user who crafts adversarial queries to access documents above their clearance (such attempts were observed in the operational deployment, usually through paraphrasing restricted terminology in a permitted language). ADV-4: actor who poisons the corpus through legitimate ingestion channels. ADV-5: compromised admin who escalates role privileges.

Out of scope: side-channel attacks on GPU memory, model extraction, OS-level compromise. This is a real gap. In three incidents over five years, the most serious breaches I saw in production exploited infrastructure, not application logic. TrustChain-RAG operates one layer above that. Confidential computing [30] should be placed underneath for mitigation/early detection of infrastructure-level threats.

Security goals: (G1) detect post-ingestion modification; (G2) prevent retroactive audit alteration; (G3) enforce classification boundaries during retrieval; (G4) attribute every document to its ingesting actor; (G5) immutably record role lifecycle events.

3.3 Document Ingestion and RBAC

Upon ingestion each document d_i is parsed into semantic representation $S(d_i) = (T_i, E_i, M_i, C_i, P_i^{\text{req}})$ following [12]: discourse tree T_i (rhetorical structure of the document), entity set E_i (named entities and domain terms), metadata M_i (author, date, source system), classification level C_i (e.g., Public, Confidential, Restricted), and required permissions P_i^{req} (the minimum role permissions needed

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

to access this document). Before any semantic processing begins, DISPL verifies that the ingesting user's role carries write privileges and that the document's classification does not exceed the user's ceiling. Both checks pass before an `INGEST` transaction goes to the blockchain:

$$\mathrm{tx}_{\mathrm{INGEST}_i} = \sigma\big(\text{SHA-256}(d_i \| M_i \| C_i \| P_i^{\mathrm{req}}),\ u,\ r,\ t\big) \tag{1}$$

where u is the ingesting user, r the user's active role at ingestion time, and t the timestamp. The SHA-256 hash covers the raw document concatenated with its metadata, classification, and permission requirements.

This property is termed provenance-at-ingestion. The trust chain starts when a document enters the perimeter, not when someone queries it. If the document is tampered later (ADV-1), the hash mismatch is detectable. If someone ingested a document they should not have had the authority to introduce (ADV-4), the role recorded in the transaction provides the trail. Updates produce `DOC_UPDATE` with `supersedes` edges, so retractions are additive, to produce evidences.

Definition 2. *The RBAC model* $\mathcal{A} = (\mathcal{U}, \mathcal{R}oles, \mathcal{P}, \mathrm{UA}, \mathrm{PA}, \mathcal{RH})$ *has a four-dimensional permission space: classification ceiling* c*, domain scope* δ*, language scope* Λ*, operation level* $o \in \{r, rw, \mathrm{admin}\}$*. Effective permissions:* $\mathrm{EffPerm}(u) = \bigcup_{r \in \mathrm{AR}(u)} \bigcup_{r' \leq r} \mathrm{PA}(r')$.

Four dimensions turned out to be necessary. This requirement emerged repeatedly in operational deployment. A pharmacology researcher might be cleared for restricted data, but only in pharmacology, only in English and French, and only for read access. A German regulatory submission at the same classification level should be invisible to them. Classification alone does not capture this. The domain×language×operation matrix makes it work, and this is why the ablation in Section 5.3 shows that removing the multi-dimensional RBAC (keeping classification only) leaks at 8% where the full model leaks at 1.3%. Every RBAC lifecycle event goes to the blockchain, so any user's effective permissions at any past timestamp can be reconstructed from the data stored on-chain.

3.4 Multilingual Context Graph

Definition 3. $G = (V, E, \lambda, \mu)$ *where* $V = V_D \cup V_E \cup V_C \cup V_L$ *(document-segment, entity, concept, cross-lingual bridge nodes). Each node carries classification* $C(v)$*, domain* $\delta(v)$*, required permissions* $P^{\mathrm{req}}(v)$*, and status. Relation types include is_ equivalent, is_ hypernym, cross_ lingual_ bridge, contains, references, contradicts, supersedes.*

Cross-lingual bridge nodes. For entity e across languages $\ell_1, \ldots, \ell_k$, a bridge node v_ℓ carries embedding:

$$\mathbf{h}_{v_\ell} = \frac{1}{Z} \sum_{j=1}^{k} \alpha_j \cdot W_{\ell_j \to \mathrm{en}} \cdot \mathbf{h}_e^{\ell_j} \tag{2}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

where $W_{\ell \to \text{en}}$ is a Procrustes projection [21] learned from 5,000 domain-specific bilingual dictionary pairs and α_j is alignment precision on a 500-pair validation set. Low-resource pairs (e.g., PL–CZ) chain through English, which introduces compounding error (I return to this in "Limitations"). Translation quality signals follow [22]. Bridge classification is conservative: $C(v_\ell) = \bigsqcup\{C(\text{neighbors})\}$.

Why explicit bridges, when BGE-M3 already handles cross-lingual similarity? This objection arose repeatedly during internal reviews. Bridge nodes carry typed relations: they encode that German "Zulassungsverfahren" is a hypernym of "klinische Prüfung" in a regulatory context, which a cosine score between embeddings cannot express. They are inspectable: when a German source influences an English response, the traversal path through the bridge is logged with confidence values, giving auditors a cross-lingual reasoning chain to review. And they carry RBAC metadata (classification, domain, permissions) that embedding vectors have no surface for.

Trust-Aware Retrieval. The trust context $\tau = (u, \text{AR}(u), c_u^{\text{eff}}, \delta_u^{\text{eff}}, \Lambda_u^{\text{eff}})$ governs retrieval. Node v is accessible iff: (1) $C(v) \leq c_u^{\text{eff}}$; (2) $\delta(v) \cap \delta_u^{\text{eff}} \neq \emptyset$; (3) $\lambda(v).\ell \in \Lambda_u^{\text{eff}}$; (4) $P^{\text{req}}(v) \subseteq \text{EffPerm}(u)$; (5) v is active. In plain terms: the user's clearance must meet or exceed the node's classification (1); the node's domain must overlap with what the user is permitted to see (2); likewise for language (3); any special permissions the node requires must be present in the user's effective permission set (4); and the node must not have been retracted (5). The retrieval pipeline embeds the query, runs ANN search restricted to the accessible set, performs RBAC-filtered BFS with depth limit k, and expands through bridges, applying the same accessibility check at each hop.

3.5 Blockchain Audit Protocol

All lifecycle events – `ROLE_*`, `INGEST`, `GRAPH_UPDATE`, `RETRIEVAL`, `INFERENCE` – are anchored on a 3-node Hyperledger Fabric [31] deployment (Raft consensus), operated by IT Operations, Information Security, and Compliance respectively. Full text goes to encrypted local storage; only hashes go to the chain.

Why blockchain (and not "append-only logs")? Six subsystems may run on separate machines. A single append-only log creates a single point of trust. Three Raft consensus nodes, operated by different departments, distribute that trust. This is what defends against ADV-2. Causal ordering across subsystem boundaries (`ROLE_LIFECYCLE` → `INGEST` → `GRAPH_UPDATE` → `RETRIEVAL` → `INFERENCE`) would otherwise require a custom cross-reference mechanism that, in my experience building such mechanisms, converges toward blockchain-like architecture anyway. Co-locating identity governance and data provenance on one ledger also avoids a practical problem termed here the *correlation problem*: in current enterprise deployments, reconstructing "who had which role, queried what, retrieved which document, and got which response" requires joining across identity management, data access logs, and application logs. These systems were never designed to interoperate. Here it is a single query. Hyperledger also produces artifacts (signed blocks, Merkle proofs) that auditors can work with directly, which matters for EU AI Act Article 12 and FDA 21 CFR Part 11.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

Trust boundary. Three nodes in one organization, possibly managed by one sysadmin who has root on all three. Frankly, if one person controls all nodes, the "distributed trust" is theatrical. Mitigations are organizational: separate credential stores per department, multi-party authorization for infrastructure changes, anomaly monitoring. What blockchain does here is raise the bar. Tampering goes from "edit a database row" to "coordinate modification of distributed, hash-linked, replicated state across multiple administrative domains." Compliance frameworks don't demand absolute tamper-proofness. But they do require reasonable controls and detectability. Where ADV-2 is out of scope, signed append-only logs with Merkle verification [25] are simpler and sufficient.

4 Formal Properties

Theorem 1 (Semantic Preservation). *If documents d_i, d_j in languages ℓ_a, ℓ_b share entity or concept references (via alignment with confidence above θ_{align}) and semantic similarity above θ_{sem}, a path exists in G with weight above $f(\theta_{\text{sem}}, \theta_{\text{align}})$.*

This is only as good as the ontology: coverage is 94% in Life Sciences (SNOMED CT + MeSH), 81% in Legal where cross-jurisdictional concept standardization lags. For domains without established ontologies, and there are many, this is an open problem the architecture does not solve.

Theorem 2 (RBAC Enforcement). *Every node on any traversal path under trust context τ satisfies all five accessibility conditions.* Proof by induction on path length. □

Theorem 3 (Safety and Completeness). *All retrieved nodes are accessible (safety). Within $G[V_\tau]$, all reachable nodes at depth $\leq k$ are retrieved (completeness). Paths through inaccessible intermediates are deliberately blocked, since their existence would leak information about restricted content. This occasionally causes false restriction, measured at 2% in Section 5.2.*

Integration effects. When the context graph preserves a bridge path with confidence weights and the blockchain records the traversal, the result is an auditable cross-lingual reasoning chain: a German pharmacovigilance report that influenced an English response can be traced through the specific bridge with the specific confidence. Neither a standalone graph nor a standalone blockchain delivers this. The graph provides the semantic structure; the blockchain makes the traversal tamper-evident. This combination gives an auditor something they can actually follow: in one review session with a regulatory affairs team, this was the key to move the conversation from "nice prototype" to "when can we deploy."

Fusing role lattices with graph traversal also produces retrieval that is both semantically rich and provably role-bounded (Theorems 2–3), with the bound itself recorded in the audit trail. Co-locating RBAC events with data provenance events on the same ledger solves the correlation problem from Section 3.5: who had which role, queried what, saw which documents, got which response. One query, not a cross-system join.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

Graph structural analysis confirms small-world topology ($\sigma = 4.3$–5.1), which is what makes depth-bounded BFS at $k = 3$ practical: high clustering keeps traversal neighborhoods topically focused; short average path lengths keep them from being too narrow. Bridge nodes are 7% of all nodes but appear in 18–22% of successful cross-lingual retrieval paths. Human-judged P@20: TrustChain-RAG 0.76 vs. Baseline-RAG 0.52 ($r = 0.91$, $p < 0.001$ with structural relevance).

5 Evaluation

5.1 Setup

Corpus. 36,300 documents, six languages (EN, DE, FR, ES, PL, CZ), 266M tokens across Life Sciences (12,400 docs), Legal (8,700), Technical (15,200). A note (as a part of "Limitations"): six European languages, all Latin script. Whether the results hold for CJK, Arabic, or languages without large-scale bilingual dictionaries is an open question I cannot answer yet from these experiments.

Models. LLaMA-3-70B-Instruct [32] (primary), LLaMA-3-8B-Instruct [32], Mistral-7B-Instruct-v0.2, Mixtral-8x22B-Instruct-v0.1, Qwen-2.5-72B-Instruct. The 70B+ models are served via vLLM [29] under GPTQ 4-bit quantization [33], fitting within two A100 80GB GPUs per model. The 7–8B models run at full precision on a single GPU. Fine-tuning uses QLoRA [34] (rank 16, 4-bit NormalFloat base) with 5K domain-specific instruction examples per model. I reduced the learning rate to 1e−4 for the 70B-class models after preliminary runs at 2e−4 showed gradient instability in the first 200 warmup steps.

Baselines. Baseline-RAG (FAISS + multilingual-e5-large), BGE-M3-RAG [7], GraphRAG [17], RAPTOR [15], Encrypted-RAG (Baseline + AES-256 + metadata RBAC). GraphRAG: the corpus is chunked (600 tokens, 100 overlap), entities and relationships are extracted using LLaMA-3-70B-Instruct, community detection uses Leiden clustering, and community summaries are generated for global queries; embedding uses multilingual-e5-large (same as Baseline-RAG, not BGE-M3, to isolate graph structure from embedding quality). All baselines use the same primary model (LLaMA-3-70B-Instruct, GPTQ 4-bit) and same hardware (4× A100 80GB; 256GB; Ubuntu 22.04, CUDA 12.1, PyTorch 2.1).

Metric definitions. Each response is segmented into atomic claims. For each claim c_i, annotators assign a label from {Supported, Partially Supported, Unsupported, Contradicted} by checking whether the claim is entailed by the retrieved source segments provided to the LLM.

Definition 4 (Factual Accuracy and Hallucination Rate).

$$\mathrm{FA} = \frac{|\{c_i : \text{Supported}\}| + 0.5 \cdot |\{c_i : \text{Partially Supported}\}|}{|\{c_i\}|} \tag{3}$$

$$\mathrm{HR} = \frac{|\{c_i : \text{Unsupported}\}| + |\{c_i : \text{Contradicted}\}|}{|\{c_i\}|} \tag{4}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

A claim is Supported *if all its factual assertions are directly entailed by at least one retrieved segment.* Partially Supported *if some assertions are entailed but others are not.* Unsupported *if no retrieved segment entails any assertion.* Contradicted *if a retrieved segment directly negates an assertion.*

Annotation protocol. 200 queries per domain, 600 total: 50 monolingual, 50 cross-lingual, 50 adversarial for hallucination, 50 adversarial for RBAC. Three domain experts per domain independently annotated all queries using the claim-level scheme above. Claims where annotators disagreed (14% of total) were resolved in a reconciliation session where annotators reviewed the source evidence and converged on a final label. Inter-annotator agreement before reconciliation: Fleiss' $\kappa = 0.81$ (Life Sciences), 0.83 (Legal), 0.79 (Technical). I reviewed all borderline annotations (Partially Supported) against source documents before computing final scores. Statistical testing: paired bootstrap, 10K resamples, $\alpha = 0.01$ [35]. Effect sizes reported as Cohen's d. RAGAS [36] faithfulness computed as secondary validation.

Adversarial protocol. RBAC adversarial probes (50 per domain) followed a three-tier design. Tier 1 (direct access): queries explicitly requesting documents above the user's classification, e.g., a user with `Public` clearance asking "List all Restricted pharmacovigilance signals for compound X." Tier 2 (paraphrase): the same restricted content requested using terminology from a permitted domain or language, e.g., a user with `Legal-EN-Confidential` clearance querying "regulatory approval procedures for active substances" (paraphrasing pharmacovigilance content classified as `LifeSci-Restricted`). Tier 3 (external red team): 50 probes designed by testers who knew the RBAC policy structure but had no access to retrieval internals, e.g., querying in Czech for a concept that exists as a PL–CZ bridge node with a classification mismatch. Hallucination adversarial queries (50 per domain) were constructed by modifying entity names, dates, or dosage values in valid queries to create questions where the correct answer is "no evidence in the corpus," e.g., "What were the Phase III results for *Nexovirin* 400mg?" (a fabricated drug name at a non-existent dosage).

5.2 Results

Factual Accuracy and Hallucination (LLaMA-3-70B-Instruct):

All TC-RAG improvements are significant at $p < 0.01$. Cohen's d vs. Baseline: 1.4–1.8. Hallucination reduction: 58–62%. RAGAS faithfulness corroborates the human judgments: TC-RAG 0.89 vs. Baseline-RAG 0.74.

Cross-Lingual Retrieval Precision (model-independent, a property of TARM + MCGE):

The German "Anlage" illustrates why. It means "appendix" in legal context and "facility" in technical. The bridge carries the domain label; the embedding does not. This is where the +15 pp comes from: terminology disambiguation that cosine similarity cannot perform. Low-resource pairs benefit most. PL–CZ gains 22 pp over BGE-M3, where Procrustes alignment on domain-specific dictionaries outperforms the embedder's general cross-lingual transfer.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

Table 1. Factual Accuracy (FA) and Hallucination Rate (HR) across three domains. LS = Life Sciences, Leg = Legal, Tech = Technical. All TC-RAG improvements significant at $p < 0.01$; Cohen's d vs. Baseline: 1.4–1.8.

System	FA (LS)	FA (Leg)	FA (Tech)	HR (LS)	HR (Leg)	HR (Tech)
Baseline-RAG	.79±.03	.76±.04	.82±.03	.10±.02	.12±.03	.08±.02
BGE-M3-RAG	.82±.03	.79±.03	.84±.02	.08±.02	.10±.03	.07±.02
GraphRAG	.84±.02	.82±.03	.86±.02	.07±.02	.09±.02	.06±.01
TC-RAG	**.91±.02**	**.89±.02**	**.92±.01**	**.04±.01**	**.05±.02**	**.03±.01**

Table 2. Cross-Lingual Retrieval Precision (CLRP) by language pair.

Pair	Baseline	BGE-M3	**TC-RAG**
EN–DE	.71±.04	.78±.03	**.91±.02**
EN–FR	.73±.04	.80±.03	**.92±.02**
EN–PL	.62±.05	.68±.04	**.86±.03**
PL–CZ	.58±.06	.63±.05	**.85±.03**
Mean	.67±.02	.74±.02	**.89±.01**

RBAC Enforcement (model-independent):

Both TrustChain-RAG leakages traced to the same root cause: an entity appearing in two domains with different classifications was mapped to a concept node that inherited the lower classification due to an ordering issue in the propagation code. Found by the external red team. The first 100 internally designed probes caught nothing; 2 of 50 external probes broke through. The bug is now patched. The evaluation ran on the pre-patch build because reporting post-patch results would be retroactive cherry-picking. Imperfect, but 4.7× better than metadata pre-filtering, which is what most production systems actually use.

ADV-5 (role manipulation): 48/50 simulated attacks caught. The 2 misses involved role assignment and revocation within the same blockchain batch window (<5 s), making them appear atomic.

Audit completeness: 98%. Eight queries had incomplete chains from missing metadata during the initial bulk load (BATS was not yet fully operational). Four cases from a race condition in concurrent delta serialization. Both were engineering issues, since fixed.

5.3 Controlled Ablations

All ablations use LLaMA-3-70B-Instruct.

Removing the blockchain changes nothing about accuracy or leakage. It lives on the audit plane, not the data plane. The interesting results are in the other rows:

Removing multi-dimensional RBAC but keeping classification: leakage jumps to 8%, confirming the domain and language scope dimensions are doing real work.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

Table 3. RBAC enforcement comparison.

System	Leakage	False Restriction	Audit
Encrypted-RAG	6.0%	4.0%	Partial
TrustChain-RAG	**1.3%**	**2.0%**	**Full**

Table 4. Component ablation (Life Sciences, 200 queries). FA = Factual Accuracy, HR = Hallucination Rate, CLRP = Cross-Lingual Retrieval Precision.

Config	FA	HR	CLRP	Leakage
Full TC-RAG	.91	.04	.91	1.3%
− Blockchain	.91	.04	.91	1.3%
− RBAC (classif. only)	.91	.04	.91	8.0%
− Bridges	.86	.06	.74	1.3%
− Graph (vector only)	.82	.08	.71	12.0%

Removing bridges costs 5 FA and 17 CLRP. Removing the graph entirely brings performance back near baseline. That this pattern is structurally identical to the 8B results is itself a finding: the graph is not compensating for weak models. It provides something the model cannot extract from flat retrieval, regardless of capacity.

Volume control: At 10 passages each, the graph provides +9 FA points over Baseline-RAG. Pushing Baseline to 50 passages barely moves the needle (0.78 FA vs. 0.79 at 10). The extra passages dilute the context more than they enrich it.

Fine-tuning control: TrustChain-RAG with the base 70B model and no QLoRA reaches 0.88 FA. Baseline-RAG with a QLoRA-tuned 70B reaches 0.82. The graph alone beats fine-tuning alone by 6 points. Fine-tuning adds 3 more on top. But the evidence quality is the dominant factor.

5.4 Model Scale Study

Does the framework's benefit survive when you give all systems a stronger model?

The most important number here is not the absolute FA. It is the Δ column. The framework adds +12 FA points at 70B, the same +12 it adds at 8B. The gain persists because the graph targets a bottleneck that model scaling does not reach: evidence that is structured by the graph, filtered by role, and connected across languages. No amount of parameters can extract that from a flat vector index. Model-side gains concentrate in the jump from 8B to 44B–70B; beyond that, returns diminish. LLaMA-3-70B and Qwen-2.5-72B differ by a single FA point.

5.5 Failure Analysis

I went through all incorrect or substantially incomplete responses from the 70B evaluation and assigned root causes. My initial assumption was that ontology

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

Table 5. Model scale study. *44B active parameters per token (176B total, MoE).

Model	Params	TC-RAG FA	Base FA	Δ FA
Mistral-7B	7B	.85±.02	.71±.03	+.14
LLaMA-3-8B	8B	.86±.02	.74±.03	+.12
Mixtral-8x22B	44B*	.90±.02	.78±.03	+.12
LLaMA-3-70B	70B	.91±.02	.79±.03	+.12
Qwen-2.5-72B	72B	.92±.01	.80±.02	+.12

Table 6. Error distribution by model scale (% of total errors).

Error Category	7–8B	70B	72B (128K ctx)
LLM reasoning	27%	16%	11%
Context saturation	23%	15%	8%
Ontology gaps	19%	26%	31%
Bridge miscalibration	17%	25%	28%
RBAC over-restriction	14%	19%	22%

gaps would shrink as the model got better at reasoning through incomplete information. The opposite happened.

With 7–8B models, half of all errors came from model limitations. With 72B and long context, that share drops below a quarter. Over 80% of remaining errors trace to the knowledge layer.

Take the "beneficial ownership disclosure" versus "wirtschaftlich Berechtigter" example from Legal. The entity-level bridge exists but with low confidence, because the ontological backbone linking these jurisdictional concepts is missing. That kind of gap – ontology coverage – is now the largest single error category. Bridge miscalibration follows: primarily PL–CZ pairs where false cognates get inappropriately high alignment confidence. Polish "aktualny" was bridged to Czech "aktuální" at high confidence in a context where the domain-specific senses diverged.

These are problems that better ontologies and alignment dictionaries will fix. Not bigger models.

5.6 Latency and Cost

TrustChain-RAG with 70B: median 1,940 ms, P95 2,820 ms (vs. Baseline 1,380 ms / 2,040 ms). Retrieval overhead breaks down as: graph traversal +280 ms, bridge expansion +85 ms, RBAC checks +40 ms, blockchain pre-commit +120 ms. All model-independent. GPU overhead is 1.27× at 70B, actually lower than the 1.38× at 8B, because LLM inference dominates total GPU time more at larger scale, making the fixed retrieval cost proportionally smaller. At 760K docs (50× synthetic scale), retrieval stays under 600 ms and RBAC checks remain flat at ∼40 ms.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

6 Discussion and Limitations

Organizations don't need to choose between oversight and quality. A 70B model with TrustChain-RAG gets .91 FA; the same model without the framework achieves .79. The governance layer is built in by design, not bolted on, and it contributes to accuracy, not just auditability.

On Shadow AI: the alternative to providing governed AI tools is not "no AI" – it is ungoverned AI, consumer accounts, and the data exposure that follows. TrustChain-RAG is a governed tool that checks most boxes on the regulatory list. My experience running the Life Sciences pipeline suggests that when such a tool is available and works well enough, unauthorized usage drops. I have not measured this formally (it would require a controlled deployment study), but the pattern was consistent across two organizations.

Principal limitations. Language coverage, more than anything else in this paper. Six European languages, all Latin script. The Procrustes alignment depends on 5,000 bilingual dictionary pairs per language pair, and for CJK that volume of domain-specific bilingual material may not exist in specialized fields like pharmacovigilance or EU regulatory law. I tried bootstrapping PL–CZ alignment from smaller dictionaries (2,000 pairs) and precision dropped 11 points. Arabic morphological complexity adds another dimension the current pipeline does not handle at all. Until those evaluations happen, the multilingual claims carry a large asterisk.

Ontology dependence is just as fundamental. Biomedicine has SNOMED CT and MeSH. Legal terminology standardization across jurisdictions lags. For emerging or cross-disciplinary domains, there may be no ontology at all. No ontology, no bridging. As of now I don't have a fix for this. The blockchain trust model in single-organization deployment is partly organizational guarantee, not cryptographic; I have been direct about this in Section 3.5. Additional constraints: RBAC role explosion beyond about 50 domain-language combinations (an ABAC [37] extension would help), GPTQ 4-bit quantization effects on absolute numbers (1–3%; I did not run the full evaluation at full precision because a single 70B model at fp16 would have required all four A100s), and the 600-query evaluation scale.

Future work. Federated deployment with inter-organizational role mapping. CJK and Arabic evaluation – I consider this a prerequisite for broader multilingual claims. ABAC extension. Active learning for bridge confidence, which the failure analysis points to as the highest-impact next step. Formal verification of the RBAC implementation against the model; Theorem 2 proves properties of the formal model, not of the code.

7 Conclusion

The model-scale study produced what I consider the most informative result of the evaluation. With smaller models, most errors trace to reasoning limits and context saturation. With 72B models and long context, over 80% of remaining

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

errors shift to ontology gaps and bridge calibration. The knowledge layer, not the model, becomes the constraint. The framework and the model solve different problems. Neither alone reaches the performance of both together.

The numbers across three domains, six languages, and five LLMs from 7B to 72B: hallucination drops 50–62% against baseline RAG. The framework's accuracy margin holds at +12 FA points regardless of model scale. Ablations trace this to graph topology, with fine-tuning additive on top. RBAC leakage at 1.3% versus 6% for metadata pre-filtering. Audit completeness at 98%.

The system grew out of years of operating a multilingual pipeline for a Life Sciences organization and my prior work [12,13,22]. The design reflects what I learned from that experience: the governance question is not a compliance checkbox. It is the single biggest determinant of whether an enterprise will actually deploy Knowledge Mining. And without that tool, employees don't stop using AI. They use the ungoverned kind.

Acknowledgments. The author thanks prof. dr hab. Leszek Kotulski for doctoral supervision and guidance throughout the research. Thanks are due to the domain experts from partner organizations in Life Sciences, Legal, and Technical Documentation who designed the RBAC role hierarchies, curated the evaluation queries, and performed the expert annotations. The anonymous red team members who designed the external adversarial RBAC probes provided invaluable testing. The author also thanks the reviewers for their constructive feedback. The views expressed are the author's own.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for knowledge-intensive NLP tasks. In: NeurIPS (2020)
2. Cyberhaven Labs: ChatGPT at Work. Cyberhaven Threat Research (Q2 2023)
3. Cyberhaven Labs: 2025 AI Adoption & Risk Report (Q2 2025)
4. Gurman, M., King, R.: Samsung bans staff use of AI tools like ChatGPT after internal data leak. Bloomberg Technology (May 2, 2023)
5. Cisco Systems: 2024 Data Privacy Benchmark Study. San Jose, CA (2024)
6. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. TACL **7**, 597–610 (2019)
7. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv:2402.03216 (2024)
8. Wang, L., Yang, N., Huang, X., Jiao, L., Yang, W., Jiang, D., Majumder, R., Wei, F.: Multilingual E5 text embeddings: A technical report. arXiv:2402.05672 (2024)
9. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.-W.: REALM: Retrieval-Augmented Language Model pre-training. In: ICML (2020)
10. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023). arXiv:2202.03629
11. Joint Task Force: Security and privacy controls for information systems and organizations. NIST SP 800-53, Rev. 5 (2020)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8

12. Dawson, K., Dzialo, J., Niewiadomska, A., Metel, P., Grochowski, M.: Representing a document using a semantic structure. US Patent 8,335,754 B2 (2012)
13. Dawson, K., Dzialo, J., Niewiadomska, A., Metel, P., Grochowski, M.: Generating a document representation using semantic networks. US Patent 8,756,185 B2 (2014)
14. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In: NeurIPS (2023). arXiv:2310.11511
15. Sarthi, S., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: Recursive abstractive processing for tree-organized retrieval. In: ICLR (2024)
16. He, Y., Tian, J., Sun, L., Zhang, S.: G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In: NeurIPS (2024)
17. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, C., Truitt, S., Larson, J.: From local to global: A Graph RAG approach to query-focused summarization. arXiv:2404.16130 (2024)
18. Document level security for content connectors. Elasticsearch Reference (2024)
19. Weaviate B.V.: Multi-tenancy and authorization. Weaviate Documentation (2024)
20. Pinecone Systems Inc.: Namespaces. Pinecone Documentation (2024)
21. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL, pp. 8440–8451 (2020). arXiv:1911.02116
22. Etches, R., Dzialo, J.: Automation-enhanced translation workflow. US Patent 12,340,183 B2 (2025)
23. Sandhu, R., Coyne, E., Feinstein, H., Youman, C.: Role-based access control models. IEEE Computer **29**(2), 38–47 (1996)
24. Denning, D.E.: A lattice model of secure information flow. Communications of the ACM **19**(5), 236–243 (1976)
25. Crosby, S., Wallach, D.: Efficient data structures for tamper-evident logging. In: USENIX Security Symposium, pp. 317–334 (2009)
26. Laurie, B., Langley, A., Kasper, E.: Certificate Transparency. RFC 6962, IETF (Jun 2013)
27. Apache Software Foundation: Apache Atlas: Data governance and metadata framework for Hadoop (2023)
28. OpenLineage Project Contributors: OpenLineage Specification (2023)
29. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with PagedAttention. In: SOSP (2023). arXiv:2309.06180
30. Tramèr, F., Boneh, D.: Slalom: Fast, verifiable and private execution of neural networks in trusted hardware. In: ICLR (2019)
31. Androulaki, E., Barger, A., Bortnikov, V., et al.: Hyperledger Fabric: A distributed operating system for permissioned blockchains. In: EuroSys. ACM (2018)
32. AI@Meta: The Llama 3 herd of models. arXiv:2407.21783 (2024)
33. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: GPTQ: Accurate post-training quantization for generative pre-trained transformers. In: ICLR (2023)
34. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized language models. In: NeurIPS (2023). arXiv:2305.14314
35. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP, pp. 388–395 (2004)
36. Es, S., James, J., Espinosa Anke, L., Schockaert, S.: RAGAS: Automated evaluation of Retrieval Augmented Generation. arXiv:2309.15217 (2023)
37. Hu, V., Ferraiolo, D., Kuhn, R., Schnitzer, et al.: Guide to Attribute Based Access Control (ABAC) definition and considerations. NIST SP 800-162 (2014)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_8