Knowledge Access and Retrieval

Institute of Applied Artificial Intelligence and Robotics (IAAIR)
In collaboration with *TO64 Inc.*
Research Internship, Winter 2025

1 Project Overview

Current scientific RAG systems suffer from 10-30% citation hallucination rates and lack provenance guarantees for high-stakes domains such as clinical decision-making. While hybrid graph+vector architectures show promise, no production system exists that unifies literature, structured datasets, and ontologies with complete audit trails suitable for regulatory review. This internship builds that foundational infrastructure. This internship establishes a multi-modal knowledge fabric that unifies scientific literature, structured datasets, and ontologies using a hybrid graph+vector substrate. The goal is to deliver provenance-rich retrieval interfaces that support agentic orchestration and scientific discovery. We build on domain-adapted language models (e.g., SciBERT), scientific claim verification datasets (SciFact), hybrid RAG patterns, and graph reasoning over citation and biomedical knowledge graphs [8, 9, 12, 14–16].

2 Rationale

State-of-the-art scientific assistants increasingly rely on agentic workflows that plan, retrieve, attribute, and verify evidence before synthesis. Recent work demonstrates measurable gains from: (i) robust source attribution and citation alignment [1], (ii) knowledge-graph–augmented reasoning for hypothesis generation (SciAgents) [10, 11], and (iii) cross-modality retrieval for figures, tables, and domain images [15, 16]. In high-stakes domains (e.g., clinical guidance), RAG has shown improved accuracy and safety when combined with strong provenance controls [12]. A unified retrieval layer that fuses graphs and embeddings, grounded in open scholarly APIs (OpenAlex, Semantic Scholar) [5, 7], is therefore essential infrastructure for trustworthy, reproducible discovery.

3 Objectives

1. **Substrate**: Stand up vector and graph stores (e.g., Milvus/Weaviate + Neo4j); ingest literature, figures/tables, datasets, and ontologies; perform entity linking and citation-graph construction [6, 9].
2. **Interfaces**: Implement semantic search, subgraph (Cypher) queries, cross-modality retrieval, and entity grounding with calibrated confidence [15, 16].
3. **Provenance**: Record complete lineage (source, version, timestamp, transformations) and expose an attribution API aligned with agentic citation patterns [2].

4. **Integration**: Provide typed contracts for orchestrators/agents (LangChain/Haystack style RAG pipelines) with contract tests and evidence bundles [3, 4].

4 Core Research Questions

1. **Hybrid fusion**: When do graph-first vs. vector-first strategies maximize recall and faithfulness for scientific questions? How do we route adaptively [13]?
2. **Attribution**: What design choices (chunking, window recall, rerankers) most improve exact passage-level attribution without hurting answer quality [1]?
3. **Cross-modality**: How much do figure/table retrieval and visual-text alignment improve answerability on science-specific benchmarks [16]?
4. **Verification**: Can SciFact-style claim verification reduce hallucinations in the final synthesis loop [14]?

5 Expected Outcomes

- A running knowledge fabric (graph+vector) with reproducible ingestion pipelines and schemas.
- Retrieval APIs covering semantic, subgraph, and cross-modality search with confidence and evidence bundles.
- A provenance ledger (queryable lineage and audit trails) with export for reports and peer review.
- An evaluation report: nDCG/MRR, latency, cost, attribution fidelity, and error taxonomy across modalities.
- Final technical report, slide deck, and open-source artifacts (config, notebooks, seed datasets).

6 Baseline Guarantees and Risk Mitigation

Baseline guarantees. Even if hybrid fusion yields no measurable gains over strong baselines (e.g., standard vector RAG), we will:

- Deliver a well-documented ingestion stack (OpenAlex/Semantic Scholar APIs) with entity linking and citation graphs [5, 7].
- Provide an attribution layer and provenance ledger aligned to agentic citation patterns [2].
- Publish an error taxonomy and ablations on chunking, reranking, and modality mix [16].

Risks & mitigations.

- *API/data drift*: Pin API versions; snapshot seed corpora; add health checks.
- *Sparse ontologies*: Back off to weak linking (string/embedding), log uncertainties; use OpenAlex graph where available [6].
- *Attribution regressions*: Add contract tests that fail builds if citation spans are missing/misaligned [1].

7 Phaseline Outcomes (12 Weeks)

Week 1	Requirements, KPIs; pick stack (Neo4j + Milvus/Weaviate); schema drafts (works, authors, venues, entities).
Week 2	Ingestion v0 for literature (OpenAlex/Semantic Scholar); citation-graph construction; basic entity linking.
Week 3	Embedding selection (SciBERT + domain variants); index build; search baseline and latency budget.
Week 4	Ontology alignment; Cypher subgraph API; basic cross-modality extraction (figures/tables) [16].
Week 5	Hybrid fusion v0 (graph↔vector routing); reranker; early attribution checks.
Week 6	Provenance ledger v0; lineage endpoints; evidence bundle format.
Week 7	Evaluation sprint: nDCG@k/MRR, attribution fidelity, ablations; error taxonomy pass.
Week 8	Performance/cost optimisation; caching; async batchers.
Week 9	Integration with Orchestrator/Agents (LangChain/Haystack contracts) [3, 4].
Week 10	Cross-modality hardening; calibration of confidence; UX for evidence bundles.
Week 11	Security/access controls; docs; red-team tests on hallucination/attribution.
Week 12	Final demo; full report; reproducible artifacts (configs, notebooks, fixtures).

8 Evaluation Criteria

- **Retrieval quality**: nDCG@k, Recall@k, MRR on held-out scientific queries; SciMMIR subsets for figure/table sensitivity [16].
- **Attribution fidelity**: exact span match rate; citation coverage; wrong-source rate (lower is better) [1].
- **Answer quality**: human+LLM judging with *groundedness* checks; domain pilot where applicable (e.g., medical RAG setup [12]).
- **Efficiency**: p95 latency and cost/query under defined load; throughput and cache hit-rates.
- **Robustness**: drift tests (updated corpora), schema-evolution reindex time, and API health checks.

9 Required Skills and Resources

Skills. Strong Python; data engineering (ETL); graph databases (Neo4j/Cypher); vector DBs (Milvus/Weaviate/FAISS); retrieval and reranking; basic MLOps; scientific Python stack. Familiarity with LangChain/Haystack and evaluation design.

Resources. Cloud credits; managed Neo4j / Milvus (or self-hosted); access to OpenAlex and Semantic Scholar APIs; storage for snapshots; CI with contract tests; Grafana/Prometheus for monitoring.

10 Research vs Engineering Split

- **Research (60%)**: hybrid routing policies; attribution strategies; cross-modality retrieval; verification via SciFact-style pipelines; ablation studies and error taxonomy.
- **Engineering (40%)**: ingestion/ETL, schema design, API implementation, provenance ledger, performance, observability, CI/CD and security.

Notes on External Dependencies

We will prioritize open, rate-limited scholarly APIs (OpenAlex, Semantic Scholar) and reproducible snapshots; tutorials and contracts will follow widely used RAG frameworks [3–5, 7]. For agentic citation alignment we mirror recently published engineering patterns [1, 2].

References

[1] Introducing citations on the anthropic api. `https://www.anthropic.com/news/introducing-citations-api`, 2025.

[2] How we built our multi-agent research system (citationagent). `https://www.anthropic.com/engineering/multi-agent-research-system`, 2025.

[3] Haystack documentation: Introduction. `https://docs.haystack.deepset.ai/docs/intro`, 2025.

[4] Build a retrieval augmented generation (rag) app. `https://python.langchain.com/docs/tutorials/rag/`, 2025.

[5] Openalex api overview. `https://docs.openalex.org/how-to-use-the-api/api-overview`, 2025.

[6] Openalex api: Works. `https://docs.openalex.org/api-entities/works`, 2025.

[7] Semantic scholar academic graph api. `https://www.semanticscholar.org/product/api`, 2025.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv:1903.10676*, 2019. URL `https://arxiv.org/abs/1903.10676`.

[9] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(67), 2023. doi: 10.1038/s41597-023-01960-3. URL `https://www.nature.com/articles/s41597-023-01960-3`.

[10] Armin Ghafarollahi et al. Sciagents ai model drives hypothesis generation by harnessing multi-agent graph reasoning. *Advanced Materials (News/Feature)*, 2024. URL `https://advanced.onlinelibrary.wiley.com/doi/full/10.1002/adma.202413523`. Journal feature describing SciAgents.

[11] Armin Ghafarollahi et al. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv:2409.05556*, 2024. URL `https://arxiv.org/abs/2409.05556`.

[12] YH Ke et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 2025. URL https://www.nature.com/articles/s41746-025-01519-z.

[13] Aditya Nagori, Ricardo Accorsi Casonatto, Ayush Gautam, Abhinav Manikantha Sai Cheruvu, and Rishikesan Kamaleswaran. Open-source agentic hybrid rag framework for scientific literature review. *arXiv:2508.05660*, 2025. URL https://arxiv.org/abs/2508.05660.

[14] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *EMNLP*, 2020. URL https://aclanthology.org/2020.emnlp-main.609/.

[15] Yifei Wang, Yunrui Li, Lin Liu, Pengyu Hong, and Hao Xu. Advancing drug discovery with enhanced chemical understanding via asymmetric contrastive multimodal learning. *arXiv:2311.06456*, 2023. URL https://arxiv.org/abs/2311.06456. v6 (May 2025).

[16] Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhao Huang, et al. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv:2401.13478*, 2024. URL https://arxiv.org/abs/2401.13478.