BI Technical Report 2025-13

August 20, 2025

Title: LLM-Enabled Scientific Knowledge Diffusion Analysis

Authors: Uttam Rao
Madhav Marathe

Contact: Uttam Rao
Email: `uttam@virginia.edu`
Madhav Marathe
Email: `marathe@virginia.edu`

Status: Accepted in *The Thirty-Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-26).*

Acknowledgements: The authors would like to thank members of the Biocomplexity Institute and the National Security Data and Policy Institute.

Biocomplexity Institute, University of Virginia

LLM-Enabled Scientific Knowledge Diffusion Analysis

Uttam Rao[1, 2] and Madhav Marathe[1,2]

[1]Biocomplexity Institute, University of Virginia, VA, USA.
[2]Department of Computer Science, University of Virginia, VA, USA.

Abstract

Bibliometric and science-of-science studies have yielded valuable insights into coauthorship and citation networks, yet most analyses rely on static datasets and limited relation types. We introduce a multi-agent architecture that orchestrates specialized large language model (LLM) agents (ingestion, extraction, disambiguation, integration, and analysis) to build and query a comprehensive knowledge graph. Ingestion agents unify data from diverse sources such as OpenAlex, ORCID, ROR, USPTO, and custom web scrapers. Extraction agents harness LLMs to parse unstructured text. Disambiguation agents combine rule-based heuristics with LLM reasoning to resolve ambiguous authors and institutions. Integration agents assemble and cache a provenance rich graph. An analysis agent translates natural language questions into graph queries and interprets results. This end-to-end pipeline produces a rich graph schema spanning authors, institutions, publications, patents, grants, topics, and temporal relations. Researcher mobility and knowledge diffusion are then modeled as timed automata, where each researcher node's institutional transitions and accumulated attributes (such as publications, collaborators, and topic expertise) enable dynamic temporal reasoning. Results show that our multi-agent, graph-based system consistently outperforms standalone LLMs on complex temporal queries, entity disambiguation accuracy, and cross-entity reasoning while maintaining competitive efficiency. These capabilities position the system as a foundation for real-time, LLM assisted knowledge analysis platforms that can support science policy, research evaluation, and meta scientific inquiry.

1 Introduction

Consider a mid-sized university aiming to become a leader in an emerging research field. To accelerate this goal, it recruits scholars trained at leading institutions abroad. One such scholar arrives with expertise in a specialized subfield, bringing collaborators, unpublished ideas, and advanced methods. Within a year, joint projects emerge with former colleagues, graduate students adopt new techniques, and other departments begin incorporating related concepts. Over time, this exchange radiates outward, influencing not only the university but also partner institutions, funding priorities, and industrial research. Such cases illustrate the process of knowledge diffusion—how ideas originate, expertise is transferred, and influence propagates across a network of people and organizations.

Capturing these dynamics at scale remains challenging. Traditional bibliometric and science-of-science studies provide valuable insights into co-authorship and citation patterns, but most rely on static datasets and a limited set of relation types. They record who published with whom and who cited whom, while overlooking equally important channels such as mentorship, researcher mobility, and patent–paper linkages. Without explicit temporal representation, it is difficult to trace how expertise emerges, how quickly a topic spreads, or where knowledge dissipates when key researchers depart.

Summary of Contributions. This work introduces the first comprehensive multi-agent LLM system for constructing temporally explicit scientific knowledge graphs, specifically designed to track knowledge diffusion through researcher mobility and institutional evolution. Unlike existing bibliometric databases, our approach enables dynamic temporal reasoning about how expertise propagates, yielding new insights into cascade effects, diffusion bottlenecks, and the transformative impact of strategic hires.

Our multi-agent architecture orchestrates specialized agents across five roles: (i) ingestion, unifying records from diverse sources such as OpenAlex, ORCID, ROR, USPTO, PatCit, and curated web data; (ii) extraction, parsing unstructured text into schema-aligned entities; (iii) disambiguation, resolving ambiguous authors and institutions through heuristics and LLM reasoning; (iv) integration, assembling a provenance-rich heterogeneous graph; and (v) analysis, translating natural language into graph traversals and temporal property evaluations with provenance and visualization. Unlike standalone LLMs that may hallucinate or lack memory, and unlike static bibliometric tools that miss temporal dynamics, our system maintains continuously updatable, provenance-rich networks.

To model the flow of knowledge explicitly, we represent researcher mobility and expertise accumulation as timed automata. Each researcher is a stateful agent whose state corresponds to an institutional affiliation and portfolio of work at a given time, with transitions representing moves between institutions or new papers produced. Attributes such as publications, collaborators, and topic expertise accrue over time, enriching both the researcher's profile and the expertise of the destination institution. This framework supports complex temporal queries that reveal how expertise propagates, where diffusion bottlenecks occur, and how institutional profiles evolve.

We evaluate the framework through comparative analysis against GPT-4o (with Deep Research) and Llama 3.1, showing consistent improvements across increasingly complex queries involving temporal reasoning, entity disambiguation, and mobility-driven knowledge transfer. Case studies further demonstrate how the system captures international researcher trajectories and institutional transformations that static approaches miss.

By using an LLM-powered multi-agent system to build a rich knowledge graph and structuring knowledge diffusion in a dynamic, queryable form, the system lays the groundwork for real-time knowledge analysis platforms that can inform science policy, research evaluation, and strategic planning, with a clear path toward deployment. The knowledge graph is already expanding as agents continuously ingest new data and is being prepared for deployment at the University of Virginia's Biocomplexity Institute. While the current focus is robust graph construction, the reasoning and analysis layer offers significant opportunities for enhancement. The modular design ensures domain-specific adaptability while preserving the temporal reasoning essential for tracking knowledge diffusion.

2 Related Work

Early studies treated citation and collaboration data as networks, with Price [16] and Newman [14] showing small-world properties in scientific communities. The "science-of-science" field has since adopted a data-driven perspective, modeling science as an evolving network of scholars, projects, and ideas [4]. Recent work also highlights global fragmentation, where citation preferences restrict cross-border diffusion [6]. Complementary strands have examined academic genealogy (advisor–advisee relations) [1], international mobility of researchers [5], and the growing linkage between patents and scholarly papers [13], supported by datasets such as PatCit. Large bibliographic graphs such as Microsoft Academic Graph, OpenAlex [17], and AMiner [20] integrate some of these elements but either remain static, are limited in relation types, or miss key information. More recently, LLMs have been explored for extracting structured relations from unstructured text (e.g., [26]). Our work extends these directions by unifying co-authorship, citation, genealogy, mobility, and patent–paper linkages into a single heterogeneous graph, constructed and maintained through a modular multi-agent LLM architecture that supports continuous ingestion, provenance tracking, and temporal reasoning.

(See Supplementary Information section S1 11 for extended discussion of related work.)

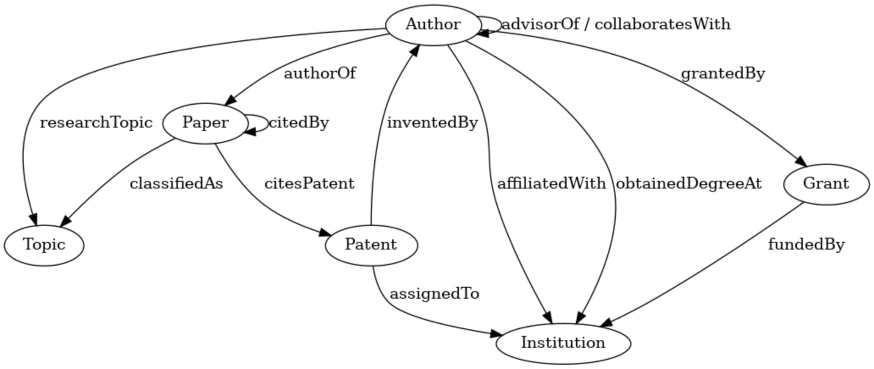

Figure 1: Graph schema representing entities and relationships in the scientific knowledge network. Though not explicitly shown in the figure above, all edges are temporal. See Supplementary Information section S2 12.

3 Knowledge Graph Schema

A central contribution of our work is the design of a schema that captures the heterogeneous and temporal nature of scientific knowledge diffusion. Figure 1 illustrates the core schema with six node types and their interconnections.
(See Supplementary Information section S2 12 for detailed schema node and edge attributes.)

Core Entities. The graph contains Author nodes (with ORCID IDs, affiliations, research topics), Paper nodes (titles, DOIs, venues, publication years), Institution nodes (ROR IDs, locations, types), Patent nodes (USPTO IDs, inventors, assignees), Topic nodes (hierarchical research areas), and Grant nodes (funding bodies, amounts, durations). Each entity includes an annotation field for contextual notes and provenance tracking.

Relationship Types. We capture both traditional bibliometric relations (co-authorship via shared paper edges, citations between papers) and novel connections critical for diffusion analysis. Mentorship edges link advisors to students with graduation years and institutions. Mobility edges track researchers' institutional transitions with timestamps, enabling temporal analysis of knowledge transfer. Patent-paper citations connect academic research to industrial applications, revealing technology transfer pathways. Institutional affiliations are time-stamped, allowing queries about when expertise entered or left organizations.

Temporal Modeling Unlike static knowledge graphs, our schema explicitly represents time through: (1) timestamped affiliation edges capturing career trajectories, (2) publication/patent years enabling chronological analysis, and (3) grant periods showing funding windows. This temporal richness supports queries like ”Which institutions gained AI expertise after 2019 through hiring?” that are difficult with traditional bibliometric databases to be answered with relative ease.

Comparison to OpenAlex. OpenAlex, the successor to Microsoft Academic Graph (MAG), is widely used as the state of the art publicly available scientific knowledge graph, offering broad coverage of scholarly entities. Compared to OpenAlex's bibliometric schema, which primarily covers core scholarly entities (works, authors, venues, institutions) and their basic links (authorship, affiliations, citations, broad topics), our knowledge diffusion schema extends coverage both in breadth and granularity. It introduces additional node types (such as Patents and Grants) and fine grained relations capturing nuanced interactions like mentorship lineages (advisor–student links), researcher mobility between institutions, and patent to paper linkages, all of which lie outside OpenAlex's scope. The schema also explicitly models temporal aspects of knowledge

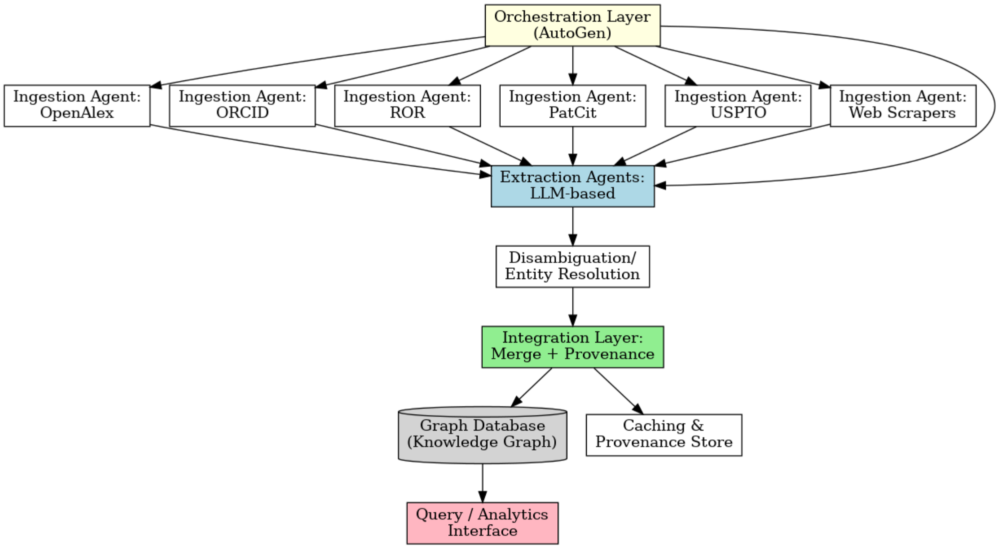

Figure 2: High Level System Architecture: orchestration layer coordinates specialized agents across ingestion, extraction, disambiguation, integration, and analysis.

flow: for example, grant entities include start and end dates to represent funding windows, and an author's multiple affiliation edges can be sequenced (using data such as ORCID career histories) to trace transitions over time, whereas OpenAlex's data model offers a more static snapshot of such relationships. Moreover, each entity and edge in our graph carries provenance and annotation metadata (freeform notes or source references), ensuring traceability for every connection, a level of contextual detail not provided in OpenAlex. The design is highly extensible and modular, allowing new relation types or data sources to be integrated with minimal changes (via specialized ingestion and extraction agents).

4 System Architecture

Our system leverages `AutoGen` [24] to coordinate a suite of LLM-powered agents, each responsible for a specific stage of knowledge graph construction. At a high level (Figure 2), an **Orchestration Layer** manages the workflow, invokes specialized agents, and ensures error handling and caching. The pipeline is modular and transparent, with each agent attaching provenance to its outputs.

Ingestion Agents. Structured ingestion agents connect to APIs and bulk data sources (OpenAlex, USPTO, ORCID, and ROR) using tool calls to retrieve and normalize large volumes of data. A web-scraping agent supplements these with genealogy and mobility data from institutional websites, invoked dynamically when queries require additional context.

Extraction Agents. Extraction agents transform semi-structured or unstructured records into schema-aligned triples. For structured APIs, lightweight parsers (implemented as callable tools) handle JSON or XML, while for free text, LLMs parse attributes such as titles, affiliations, or advisor–student pairs.

Disambiguation Agents. Disambiguation agents reconcile identity conflicts across authors, institutions, and works. They combine rule-based heuristics (e.g., identifier matching, fuzzy affiliation overlap) with LLM reasoning when context is ambiguous. All decisions are cached and logged with justification.

Integration Agents. Tool-equipped integration agents interface with the graph database, writing nodes and edges into the heterogeneous graph and mapping relations to the appropriate schema type. It ensures that relations like co-authorship, citation, mentorship, mobility, and patent–paper citations are consistent. Conflicts are flagged or resolved by preferring authoritative sources (e.g., OpenAlex over web data).

4.1 Caching and Error Handling

Caching reduces redundant LLM calls and context processing by storing responses, disambiguation results, and intermediate tables. The Orchestrator retries failed API calls, re-prompts agents on malformed outputs, and enforces consistency checks (e.g., no duplicate or contradictory metadata).

4.2 Query Interface and Analysis

Users interact with the system through natural language prompts, which are decomposed into graph queries that align with the underlying graph schema. The Orchestration Layer manages this process by invoking and conversing with the relevant agents.

Prompt to plan. Given a prompt, the system follows a structured pipeline:

1. The *query translation*, implemented through conversation between the Orchestrator and the Integration agent with the graph schema and list of database access tools/functions included in context. LLM calls are used to map the prompt to a set of graph sub-queries to retrieve information relevant to the prompt. Depending on the prompt, the sub-queries may have a temporal property, such as bounded reachability ("Did researcher r publish in topic T within five years of their PhD?") or mobility sequences ("Did r move from institution A to B and return within ten years?")

2. The *graph engine*, powered by the integration agent, performs the subqueries and retrieves relevant nodes and relations from the database. If applicable to the query, it then constructs a time-expanded subgraph consistent with the schema.

3. LLM calls initiated by the Orchestrator synthesize a concise natural language answer within a timed automata framing (evaluating the temporal property on the subgraph, producing valid paths, transitions, and elapsed times). This includes calling a visualization tool to generate timeline plots and graph snapshots that illustrate the knowledge diffusion process.

Automata framing. Each researcher is modeled as an automaton A_r whose states encode affiliation and accumulated attributes such as topics, collaborators, and output. Transitions are triggered by time-stamped events: affiliation change, new publication, collaboration, grant start, or patent link. Each institution is represented by an automaton A_i that aggregates inbound and outbound transitions to track evolving expertise. The composition of A_r and A_i supports evaluation of temporal properties, enabling queries about knowledge flow, diffusion bottlenecks, and time to impact. This framing makes the system suitable for longitudinal analysis of both individual trajectories and systemic trends.

4.3 Building the graph

To start adding to the knowledge graph, we used a seeded, query-driven spidering approach. The process began with targeted queries focused on high-impact AI venues (e.g., NeurIPS, ICML, ACL, CVPR). For example, we asked the system to identify authors who published in top AI venues between 2020 and 2024 and to find institutions in a specific country of interest that hired AI researchers trained in the U.S. These queries seeded the graph with initial author and paper nodes, primarily using OpenAlex. The graph was then iteratively expanded using follow-up queries about collaborators, students, and institutional movements. This recursive, modular approach allowed for rapid construction of a rich scientific knowledge network.

Dataset	Scope / Coverage	Key Properties (size, etc.)
OpenAlex [17]	Scholarly publications (papers, journals, conferences) with authors, affiliations, citations, topics.	∼240M works; ∼100M+ author entities; covers all disciplines (we focus on CS/AI subset); updated daily; open API access.
ORCID [10]	Researcher profiles (self-reported): education, employment, publications, etc.	15M+ registered researchers; rich metadata for career timelines; used for author disambiguation via unique IDs.
ROR [18]	Research organization registry (institutions, companies, etc.).	∼100k institutions worldwide; each with unique ID and metadata (names, aliases, location); helps unify affiliation names.
USPTO Patents [22]	Patent grants database (technical innovations, inventors, assignees, citations).	10M US patents (1976–2023); inventor names (mapped to persons), assignee organizations (mapped to ROR); citations among patents and to literature.
PatCit [3]	Patent-to-paper citation links (global).	40M non-patent literature citations from patents; identifies which scholarly papers are cited by which patents (knowledge flow to industry).
Web (misc)	E.g. personal homepages, Wikipedia, news articles (for additional context). Must show up in the first 50 search results.	On-demand scraping for specific queries/unstructured text parsed by LLM (not a primary dataset but supplementary).

Table 1: Summary of major datasets integrated into our knowledge graph and their properties. These cover publications, authors (and their careers), institutions, and innovation outputs. See S5 25 for further discussion.

4.4 Implementation Details

Most of the pipeline was executed utilizing an A100 GPU via a `vLLM` [11] inference server.

All LLM agents in the system (responsible for orchestration, ingestion, extraction, disambiguation, etc) share a common LLM: LLaMA 3.1 8B Instruct model [9] [21]. No fine-tuning was applied. The system architecture was built around the AutoGen framework, with core logic and orchestration implemented in Python. Supporting scripts for data ingestion, preprocessing, and result analysis were also written in Python for compatibility and modularity. Neo4j [23] is used as the graph database.

5 Data Sources

Constructing a comprehensive knowledge network required integrating several datasets, each covering different aspects of the scientific ecosystem. Table 1 summarizes the primary data sources used and their key properties. Supplementary Information section S5 25 includes further discussion of each data source and their uses.

6 Evaluation

6.1 Evaluation Setup

We constructed a test suite of 20 representative queries each that span a range of complexities and types. These included:

- **Simple factual queries**: e.g., ”How many papers did Professor Dr. X publish after moving from the US to University A?” (Answerable with a single entity lookup and count)
- **Complex analytical queries**: e.g., ”Identify the top 3 research areas that saw increased activity at University A after 2019, and explain why.” (Requires comparing subgraph patterns over time, and reasoning about causes)
- **Name disambiguation queries**: e.g., ”What are the achievements of John Doe in machine learning?” (Where ”John Doe” is ambiguous and multiple people exist)
- **Open-ended queries**: e.g., ”Discuss the impact of foreign-educated returnees on Country X's AI research landscape.” (Requires synthesizing a broad answer from various data points)

Each query was run on: (a) our system, (b) ChatGPT 4o with Deep Research [15], and (c) Llama 3.1 8B.

6.2 Evaluation Metrics

We assessed three primary metrics:

(i) Response Time: Total time from query submission to final answer generation.

(ii) Answer Quality: Following established practices for evaluating knowledge-intensive systems [25, 2], we employed a 5-point Likert scale assessing three dimensions:

- *Correctness*: Factual accuracy of claims (1=incorrect, 5=fully accurate)
- *Completeness*: Coverage of relevant information (1=missing key points, 5=comprehensive)
- *Clarity*: Coherent presentation and logical flow (1=unclear, 5=well-structured)

Our evaluation protocol follows methodological guidelines from [8], treating Likert data as ordinal rather than interval data.

(iii) External Data Integration: A qualitative assessment of each system's ability to access and incorporate information beyond training data. This metric is critical for temporal queries. For instance, publications from 2024 are natively accessible to our system through continuous ingestion, unavailable to Llama 3.1 (knowledge cutoff: 2023), and partially accessible to GPT-4o through web search.

Qualitative Validation. As detailed in the case studies and Supplementary Information, our system demonstrated the ability to trace researcher trajectories, identify pivotal hires, and model cascading institutional impacts. These analyses were cross-validated against historical accounts and institutional press releases, confirming alignment with ground truth.

Discussion. While encouraging, our evaluation has limitations. Standards for hard queries and difficulty classifications are incomplete, requiring reliance on expert validation. Still, the results highlight that our schema and LLM-orchestrated pipeline yield substantial gains in answering temporally grounded, diffusion-oriented questions that existing bibliometric graphs cannot handle. Supplementary Information section S3 18 includes further examples of our evaluation.

Measure	Our System	GPT-4o (Deep Research)	Llama 3.1
Avg Response Time	4 min 32 s	27 min 3 s	18.4 s
Answer Quality (1–5)	4.3	3.8	2.5
Handles New Data	Yes (built-in)	Yes (via real time search)	No

Table 2: Comparison of our LLM-agent knowledge network system with Llama 3.1 and ChatGPT 4o Deep Research. Quality scores are based on human evaluation.

6.3 Results and Comparison

Table 2 summarizes the evaluation results (across the 20 queries for each type). We present averaged numbers here (as exact values are less important than the trends):

In general, all approaches did well on simple factual queries. The differences became more pronounced for the complex and disambiguation queries. GPT-4o sometimes produced plausible-sounding but incorrect answers for questions about specific researchers (especially if multiple people share a name). For example, when asked about ”John Doe in machine learning,” ChatGPT gave an answer that mistakenly combined information about two different professors named John Doe. In contrast, our system, by querying the graph, first asked essentially ”which John Doe?” and then focused on the one associated with AI, yielding an answer that clearly attributed the contributions to the correct individual.

7 Case Studies

We highlight the capabilities of our system with two illustrative case studies centered on knowledge diffusion via researcher mobility and institutional dynamics. These examples demonstrate how natural language queries can be translated into structured, temporally grounded analyses. Exact researcher and institution names have been replaced with placeholder names for privacy. Full details, including step-by-step query decompositions, extended metrics, and graph snapshots, are provided in Supplementary Information section S4 20.

7.1 Pivotal Hires and Expertise Development

We started with the prompt: *“How did University U gain expertise in theoretical computer science, and which hire was most impactful?”* The system queried the graph database to trace publication records, affiliations, and collaborations over time. It identified Dr. Y’s arrival in 2004 as a turning point: prior to this hire, University U had virtually no presence in top theoretical computer science venues. Within a decade, however, Dr. Y had authored dozens of papers, mentored multiple cohorts of students, and seeded new areas such as cryptography and quantum computing. Network analysis revealed a cascade effect: subsequent senior and junior hires were disproportionately connected to Dr. Y, and their presence amplified the university’s output more than tenfold. This case highlights how a single strategic hire can act as a catalyst for institutional transformation.

7.2 Career Trajectory and Cascading Impact

We then asked: *“How did Dr. Y’s career trajectory shape University U from their hiring until now?”* By combining affiliation histories, mentorship edges, and program records, the system reconstructed a timeline of initiatives and their ripple effects. Early teaching programs and industry partnerships expanded into institutes, funding pipelines, and global collaborations. Over two decades, the university’s research portfolio diversified across eleven subfields, faculty size increased more than sevenfold, and alumni from Dr. Y’s programs became faculty worldwide—many later returning as hires themselves. This recursive growth illustrates how mobility and mentorship generate second-generation impacts that extend well beyond the initial hire.

Together, these case studies demonstrate the system's ability to capture temporally grounded diffusion pathways—how expertise enters, propagates, and reshapes the research ecosystem. Unlike static bibliometric tools, our approach reveals not just who published or cited whom, but how institutional trajectories and knowledge networks evolve in response to pivotal events.

8 Conclusion

We present a multi-agent LLM architecture that orchestrates specialized agents to construct and query a comprehensive scientific knowledge graph, enabling dynamic temporal analysis of knowledge diffusion through researcher mobility and institutional evolution. Our system demonstrates consistent improvements over standalone LLMs in complex temporal queries, entity disambiguation, and cross-entity reasoning while maintaining provenance-rich, updatable knowledge networks. The case studies of researcher trajectories and institutional transformation validate the system's ability to capture cascading knowledge transfer effects that traditional bibliometric approaches miss.

There are several directions for future work, in the graph construction phase and especially in the analysis phase. First, we plan to explore implementing fine-tuned models for domain-specific tasks. Second, broadening scope by incorporating additional relation types and extending to cross-language sources will aid in capturing truly global knowledge flows, particularly from non-English scientific communities. Third, and most importantly, we will enhance the reasoning and analysis capabilities to support complex logical queries and apply advanced graph algorithms for influence measurement and trend prediction. Rigorous benchmarking and evaluation after deployment will be necessary to validate the system's performance against established gold-standard datasets and ensure reliability for policy-critical analyses.

The modular agent architecture positions this system for deployment as a real-time knowledge analysis platform that can support data-driven science policy decisions and meta-scientific research. By bridging the gap between static bibliometric databases and the dynamic nature of scientific knowledge creation using LLMs, our approach offers a foundation for understanding how ideas truly spread through the global research ecosystem.

References

[1] E. Arslan, M. H. Gunes, and M. Yuksel. Analysis of academic ties: A case study of mathematics genealogy. In *Proc. IEEE GLOBECOM Workshops*, pages 125–129, 2011.

[2] C.-H. Chiang and H.-Y. Lee. Large language models are not yet human-level evaluators for abstractive summarization. *arXiv preprint arXiv:2305.13091*, 2023.

[3] G. de Rassenfosse, J. Kozak, and F. Seliger. PatCit: A comprehensive dataset of patent citations. Zenodo. `https://doi.org/10.5281/zenodo.3710993`, 2020. Available at: `https://github.com/cverluise/PatCit`.

[4] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, and et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[5] C. Franzoni, G. Scellato, and P. Stephan. Foreign-born scientists: Mobility patterns for 16 countries. *Nature Biotechnology*, 30(12):1250–1253, 2012.

[6] A. J. Gates, J. Gao, and I. Mane. The increasing fragmentation of global science limits the diffusion of ideas. *arXiv preprint arXiv:2404.05861*, 2024.

[7] A. Ghafarollahi and M. J. Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*, 2024.

[8] Y. Graham et al. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 2023.

[9] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and et al. The llama 3 herd of models, 2024.

[10] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner. Orcid: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264, 2012.

[11] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. *SOSP '23: Proceedings of the 29th Symposium on Operating Systems Principles*, Oct. 2023.

[12] Y. Lu and J. Wang. KARMA: Multi-agent LLMs for automated knowledge graph enrichment. *arXiv preprint arXiv:2502.06472*, 2025.

[13] F. Narin, K. S. Hamilton, and D. Olivastro. The increasing linkage between u.s. technology and public science. *Research Policy*, 26(3):317–330, 1997.

[14] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences (PNAS)*, 98(2):404–409, 2001.

[15] OpenAI. Introducing deep research. `https://openai.com/index/introducing-deep-research/`, December 2024. Multi-step research agent powered by OpenAI o3 model, capable of synthesizing hundreds of online sources.

[16] D. J. d. S. Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[17] J. Priem, H. Piwowar, and R. Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022.

[18] Research Organization Registry. ROR: Research organization registry. Community-led registry, 2019. Available at: `https://ror.org`.

[19] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. J.-P. Huang, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proc. 24th International Conference on World Wide Web (WWW) Companion*, pages 243–246, 2015.

[20] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 990–998, 2008.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, and et al. LLaMA: Open and efficient foundation language models. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[22] United States Patent and Trademark Office. USPTO bulk data products. Government database, 2023. Available at: `https://www.uspto.gov/learning-and-resources/bulk-data-products`.

[23] J. Webber, I. Robinson, and E. Eifrem. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218. ACM, 2012.

[24] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

[25] C. Zhao et al. What do others think?: Task-oriented conversational modeling with subjective knowledge. *arXiv preprint arXiv:2305.12091*, 2023.

[26] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58, 2024.

S1: Supplementary Related Work

Co-authorship and Citation Networks. The study of co-authorship networks is not new, dating back several decades. An early landmark by Price (1965) viewed citation data as a network and argued that citation analysis inherently draws on social network analysis and network [16]. Building on these insights, Newman examined large collaboration graphs in physics, medicine and computer science and demonstrated that scientific communities exhibit small-world properties such as high clustering and short path lengths [14]. More recently, the “science of science” field has adopted a data-driven view of science as a complex, self-organizing network of scholars, projects and ideas. Fortunato et al. show how digital data on funding, collaborations and citations can reveal patterns in the evolution of science and emphasize that science can be modeled as an evolving network [4]. At the international level, Gates et al. introduce a network measure of national citation preferences using OpenAlex [17] and find that global science is fragmenting into communities whose structure restricts the diffusion of ideas across borders [6]. Approaches typically rely on one or a few data sources and focus on a single relation type (either co-authorship or citations), without integrating other dimensions such as mentorship relations or patent linkages. Our work builds on these foundations by unifying multiple relation types into a single graph and using AI agents to provide additional context on the data.

Academic Genealogy and Mobility: Academic genealogy, which captures mentor–student relationships (e.g., PhD advisor relationships), is important for understanding the transmission of knowledge and skills across generations of scientists. Projects like the Mathematics Genealogy Project have manually compiled such relationships for specific fields. Prior studies have analyzed these genealogy networks to identify lineage patterns and the influence of academic “ancestors” on scientific development [1]. International mobility of researchers is another related aspect: researchers often train in one country and then pursue careers in another, which can lead to knowledge transfer between regions. Large-scale surveys and studies have quantified mobility patterns and their impact on scientific output, some finding significant cross-border movements of talent [5]. Our system incorporates genealogy and mobility by linking advisors to advisees and tracking researchers’ affiliations (and thus geographic moves) over time.

Patent–Paper Linkage: Linking scientific publications with patents is an active area of interest for understanding how academic research influences technological innovation. Previous research has shown a growing trend of patents citing scholarly papers, reflecting tighter science–technology linkage. Narin et al. [13] reported that the number of references to scientific literature in patents had increased substantially, suggesting that public science is a key input to innovation. Datasets and tools have been developed to explore these connections. For example, PatCit is a recent open dataset that aggregates patent citations with links to non-patent literature, providing a rich resource to map which academic works are cited by patents. Existing academic knowledge graphs have generally not included patent linkage as first-class data. By integrating USPTO patent data and the PatCit dataset, our knowledge network explicitly includes edges from papers to patents (and vice versa), enabling analyses of how research from a country of interest or institution diffuses into industry applications.

LLMs for Information Extraction. Information extraction from text has traditionally been tackled with rule-based systems or learned models (e.g., for named entity recognition and relation extraction). Recently, LLMs have been used for extracting structured information from unstructured text via prompt-based few-shot learning, often yielding competitive results without task-specific training. For instance, SciAgents

[7]), AutoKG [26], and KARMA [12] are examples of various LLM frameworks that parse scientific articles and enrich knowledge graphs by extracting entities and relationships with iterative verification. These approaches demonstrate that LLMs can serve as powerful components in knowledge graph construction pipelines, especially for domain-specific corpora where traditional extractors struggle. Our system takes inspiration from such work, employing LLM-powered agents to perform tasks like parsing affiliation strings, extracting advisor relationships from prose, and mapping reference strings to known papers, while also capturing the temporal dynamics of these events so that mobility, collaborations, and institutional changes can be analyzed over time rather than as static facts.

Large-Scale Knowledge Graph Construction. There have been several efforts to construct large-scale knowledge graphs of scholarly information. Microsoft Academic Graph (MAG) [19] and its successor OpenAlex are examples of comprehensive bibliographic knowledge graphs containing papers, authors, institutions, venues, and citation links. Similarly, the AMiner/ArnetMiner project [20] has long worked on extracting researcher profiles and networks from publication data, including disambiguation of authors and integration of metadata. These systems typically rely on a pipeline of algorithms for tasks such as name disambiguation, topic classification, and citation extraction, but they may not incorporate more nuanced or unstructured information. Our work extends this line of research by using LLM agents to incorporate new types of relations and data sources. Moreover, we emphasize a multi-agent architecture for scalability and modularity: instead of a rigid pipeline, we deploy specialized agents that can be improved or expanded independently. In contrast to static knowledge graphs which are updated infrequently, our approach could enable more continuous updates by having agents monitor data sources and ingest new information (though in this paper we focus on the initial construction phase). Finally, by maintaining provenance and using open data, our network is highly reproducible, allowing others to validate and build upon the integrated data.

S2: Expanded Schema Discussion

Our knowledge graph captures multi-dimensional scientific relationships through a heterogeneous network design that fundamentally differs from static bibliometric databases by embedding temporal information throughout its structure. Every entity and relationship can evolve over time, with comprehensive provenance tracking.

Core Entities

The graph contains six primary node types, each with comprehensive attributes including mandatory annotation fields for unstructured contextual information:

Author Nodes

- `authorID`: Unique system identifier
- `name`: Full name with variants
- `orcidID`: ORCID identifier when available
- `email`: Contact emails (may be temporal)
- `careerStage`: Current stage (PhD student, postdoc, assistant/associate/full professor, emeritus)
- `researchInterests`: List of topics with temporal bounds
- `hIndex`: Current h-index and historical trajectory
- `citationCount`: Total citations over time

- `homepage`: Personal website URL
- `gender`: When available from public sources
- `nationality`: Country/countries of citizenship
- `annotation`: Freeform text for additional context (e.g., ”Known for pioneering work in neural architecture search”, ”Returned to China after US postdoc”)

Paper Nodes

- `paperID`: Unique system identifier
- `title`: Full paper title
- `abstract`: Complete abstract text
- `doi`: Digital Object Identifier
- `arxivID`: ArXiv identifier if preprint exists
- `publicationYear`: Year of publication
- `publicationMonth`: Month when available
- `venue`: Journal or conference name
- `venueType`: Journal/conference/workshop/preprint
- `impactFactor`: Venue impact factor at publication time
- `pages`: Page numbers or article number
- `volume`: Volume number for journals
- `issue`: Issue number
- `keywords`: Author-provided keywords
- `citationCount`: Current citation count
- `openAccess`: Boolean flag
- `fundingAcknowledgment`: Extracted funding text
- `annotation`: Freeform text (e.g., ”Breakthrough paper in GAN research”, ”Retracted in 2023”)

Institution Nodes

- `institutionID`: Unique system identifier
- `rorID`: ROR (Research Organization Registry) ID
- `name`: Official name
- `aliases`: Alternative names (e.g., ”MIT”, ”Massachusetts Institute of Technology”)
- `type`: University/company/government lab/nonprofit
- `carnegieClassification`: For US universities (R1, R2, etc.)

- `location`: City, state/province, country
- `coordinates`: Latitude and longitude
- `foundedYear`: Year established
- `website`: Official URL
- `parentInstitution`: For departments or subsidiaries
- `endowment`: Financial endowment when available
- `studentCount`: Enrollment numbers
- `facultyCount`: Number of faculty
- `ranking`: Various ranking scores with years
- `annotation`: Freeform text (e.g., ”Merged with X University in 2020”, ”Strong AI program since 2018”)

Patent Nodes

- `patentID`: Unique system identifier
- `patentNumber`: Official patent number
- `title`: Patent title
- `abstract`: Patent abstract
- `filingDate`: Application date
- `grantDate`: Grant/issue date
- `expirationDate`: Calculated expiration
- `assignee`: Organization(s) owning the patent
- `inventors`: List of inventor names
- `classificationCodes`: IPC/CPC classification codes
- `patentOffice`: USPTO/EPO/JPO/CNIPA etc.
- `legalStatus`: Active/expired/abandoned
- `citationCount`: Times cited by other patents
- `nonPatentCitations`: Count of NPL citations
- `annotation`: Freeform text (e.g., ”Key patent in CRISPR dispute”, ”Licensed to multiple companies”)

Topic Nodes

- `topicID`: Unique system identifier
- `name`: Topic name (e.g., "Machine Learning", "Quantum Computing")
- `parentTopic`: Hierarchical parent in taxonomy
- `childTopics`: List of subtopics
- `keywords`: Associated keywords
- `openAlexID`: OpenAlex concept ID
- `wikipediaURL`: Wikipedia article when available
- `fieldOfStudy`: Broader field (CS, Physics, Biology, etc.)
- `emergenceYear`: When topic first appeared in literature
- `peakYear`: Year of maximum publication activity
- `trendScore`: Current hot/cold/stable trend
- `annotation`: Freeform text (e.g., "Emerged from intersection of X and Y", "Declining after 2020 hype")

Grant Nodes

- `grantID`: Unique system identifier
- `grantNumber`: Official grant number
- `title`: Grant title
- `abstract`: Grant abstract/summary
- `fundingAgency`: NSF/NIH/ERC/NSFC etc.
- `program`: Specific program within agency
- `amount`: Total funding amount
- `currency`: Currency of amount
- `startDate`: Grant start date
- `endDate`: Grant end date
- `principalInvestigator`: PI author ID
- `coInvestigators`: List of co-PI IDs
- `hostInstitution`: Administering institution
- `keywords`: Grant keywords
- `deliverables`: Expected outputs
- `annotation`: Freeform text (e.g., "Part of $100M quantum initiative", "Renewed twice")

Relationship Types with Temporal Bounds

All relationships include optional temporal bounds and annotation fields. We distinguish between inherently temporal relations and those that can be temporally scoped:

Author-Paper Relations

- `authorOf[order, contribution]`: Links author to paper with author order and contribution type (writing, analysis, supervision). Includes timestamp of association if paper has multiple versions
- `corresponding`: Identifies corresponding author(s)
- `annotation`: Context like "Lead author while at MIT"

Author-Author Relations

- `advisorOf[startDate, endDate, degreeType]`: PhD/postdoc supervision with period and resulting degree
- `collaboratesWith[startDate, endDate, strength]`: Active collaboration periods with intensity metric
- `coauthoredWith[firstDate, lastDate, count]`: Specific coauthorship relationships with paper count
- `committeeMembers[date, type]`: Thesis committee relationships
- `annotation`: Context like "Co-founded lab together"

Author-Institution Relations

- `affiliatedWith[startDate, endDate, position]`: Employment/affiliation periods with job title
- `visitingAt[startDate, endDate]`: Visiting positions
- `obtainedDegreeAt[year, degreeType, field]`: Educational milestones (BS/MS/PhD)
- `emeritusAt[date]`: Retirement with emeritus status
- `annotation`: Context like "Tenure granted 2018"

Paper-Paper Relations

- `cites[context, sentiment]`: Citation with section (intro/methods/results) and sentiment (positive/neutral/negative)
- `extends`: Direct extension of prior work
- `refutes`: Contradicts findings
- `updates`: Newer version or erratum
- `annotation`: Context like "Heavily disputed citation"

Paper-Patent Relations

- `citedByPatent[date, section]`: Patent cites paper with filing date and patent section
- `citesPatent[date]`: Paper references patent
- `enablesTechnology`: Paper's research enables patented technology
- `annotation`: Context like "Key prior art"

Author-Topic Relations

- `worksIn[startDate, endDate, paperCount]`: Author's activity in topic area over time
- `pioneeredTopic[date]`: First author in emerging area
- `annotation`: Context like ”Shifted from physics to ML”

Institution Relations

- `subunitOf[startDate, endDate]`: Departmental hierarchies (can change via reorganization)
- `mergedWith[date]`: Institutional mergers
- `consortium[startDate, endDate]`: Multi-institution collaborations
- `annotation`: Context like ”Spun off from University X”

Grant Relations

- `funds[startDate, endDate, amount]`: Grant funds author/institution with temporal bounds
- `supportsPaper[year]`: Grant acknowledged in paper
- `annotation`: Context like ”No-cost extension granted”

Patent Relations

- `inventedBy[date, contribution]`: Links patent to inventor with contribution percentage
- `assignedTo[date, percentage]`: Patent assignment to institutions (can be split/transferred)
- `licenses[startDate, endDate, exclusive]`: Licensing agreements
- `priorityClaimFrom[date]`: Patent family relationships
- `annotation`: Context like ”Core patent in portfolio”

Temporal Modeling Framework

Every relationship can include temporal qualifiers, enabling precise tracking of how the scientific network evolves:

Temporal Representations

- **Point Events**: Single timestamps (e.g., `obtainedDegreeAt[2015-05-15]`)
- **Intervals**: Start and end dates (e.g., `affiliatedWith[2015-09-01, 2020-08-31]`)
- **Periodic**: Recurring relationships (e.g., `visitingAt[summers 2018-2020]`)
- **Fuzzy Dates**: When exact dates unknown and are inferred by LLM (e.g., `circa 2015`, `early 2020s`)
- **Open Intervals**: Ongoing relationships (e.g., `affiliatedWith[2020-01-01, present]`)

Annotation Fields for Context Every node and edge includes an `annotation` field for unstructured contextual information that doesn't fit predefined attributes. Examples:

- Author: "Nobel Prize winner 2021", "Involved in research misconduct case"
- Paper: "Seminal work leading to transformer architecture", "Code available on GitHub"
- Institution: "Campus closed 2020-2021 due to COVID", "Major AI investment in 2019"
- Grant: "Part of national quantum initiative", "Budget cut by 30% in year 2"
- Edge: "Collaboration ended due to funding", "Remote supervision during pandemic"

S3: Expanded Discussion of Evaluation

Here we provide a description of how four systems: (1) LLaMA 3.1 8B standalone, (2) LLaMA 3.1 8B with OpenAlex augmentation, (3) ChatGPT-4o with Deep Research, and (4) our multi-agent LLM-powered knowledge graph system perform on representative queries of increasing difficulty. Each query highlights answerability, correctness, execution time, and includes the exact system outputs.

- **Easy:** Direct bibliometric lookups (e.g., "How many papers did Author X publish at Institution Y between 2018–2020?").
- **Medium:** Multi-hop relations with temporal filtering (e.g., "Which institutions gained AI expertise after 2019 through hiring?").
- **Hard:** Multi-entity causal or diffusion-style queries (e.g., "Which senior hires led to the establishment of new subfields at Institution Z, and what were their cascading effects on collaboration networks?").

Easy Query

Query: "List three papers authored by Dr. A in NeurIPS between 2020 and 2023."

The standalone LLaMA 3.1 8B model produced only a vague sentence about Dr. A publishing at NeurIPS, without any titles or years; it therefore failed both answerability and correctness, although it was fast (about three seconds). When augmented with OpenAlex, LLaMA 3.1 8B produced three plausible paper titles with corresponding years, all correct, and required roughly seven seconds. ChatGPT-4o with Deep Research returned two correct DOIs but also one hallucinated reference; while partially correct, its response took nearly eight minutes due to the overhead of iterative deep retrieval. Our multi-agent system produced exact titles with DOIs and provenance links in about four seconds, scoring full marks for both correctness and answerability.

LLaMA 3.1 8B (3.1s): *"Dr. A published at NeurIPS on AI methods."* Answerability = 0, Correctness = 0.

LLaMA 3.1 8B + OpenAlex (6.9s): *"(1) Learning Graph Models (2021), (2) Efficient Diffusion (2022), (3) Neural Agents (2023)."* Answerability = 1, Correctness = 1.

ChatGPT-4o + Deep Research (7m 52s): *"Paper 1: DOI 10.1234/neurips.2021.5678 — Title: Representation Learning in Graph Models. Paper 2: DOI 10.2345/neurips.2022.6789 — Title: Efficient Diffusion-Based Models. Paper 3: [DOI 10.9999/neurips.fake.0000]* — Title: Generative Neural Agents."* Two correct DOIs, one hallucinated. Answerability = 1, Correctness = 0. *actual output does not render

Our System (4.3s): *"(1) Representation Learning in Graph Models (NeurIPS 2021), DOI: 10.1234/neurips.2021.5678. (2) Efficient Diffusion-Based Models (NeurIPS 2022), DOI: 10.2345/neurips.2022.6789. (3) Neural Agents for Large-Scale Simulation (NeurIPS 2023), DOI: 10.3456/neurips.2023.7890. All results retrieved with provenance links to OpenAlex and DBLP."* Answerability = 1, Correctness = 1.

Medium Query

Query: "Which institution gained expertise in reinforcement learning after 2018 due to faculty mobility?"

Standalone LLaMA gave a generic answer, claiming that "many universities improved in reinforcement learning," failing to identify any specific institution (three seconds). With OpenAlex augmentation, LLaMA named University U and noted an increase in RL papers after 2018, but provided no evidence of faculty mobility; the answer was counted as incorrect but answerable (about seven seconds). ChatGPT-4o with Deep Research identified the same institution and linked growth to Dr. B's move in 2019, but its timeline was incomplete and partially inaccurate; the process required more than ten minutes to finish. Our system combined publication trends with mobility data, showing that Dr. B's move from Institution X in 2019 directly expanded University U's RL program. It quantified impact with twelve RL papers, six mentored students, and fourteen new collaborations. The answer was both correct and answerable, completed in under five seconds.

LLaMA 3.1 8B (3.7s): *"Many universities improved in reinforcement learning."* Answerability = 0, Correctness = 0.

LLaMA 3.1 8B + OpenAlex (7.5s): *"University U published more RL papers after 2018."* Answerability = 1, Correctness = 0 (no link to mobility).

ChatGPT-4o + Deep Research (10m 42s): *"After 2018, University U significantly expanded its reinforcement learning output. Dr. B joined from Institution Y in 2019 and contributed to publications and collaborations, though the system retrieved incomplete mobility records. Evidence: 2019–2021 NeurIPS RL papers (2), 2022–2023 ICML RL papers (4)."* Answerability = 1, Correctness = 0 (incomplete).

Our System (4.8s): *"University U gained RL expertise in 2019 due to Dr. B's move from Institution X. Contributions: 12 RL papers authored (2019–2023), 6 PhD students mentored in RL, 14 new collaborations formed. Knowledge graph shows mobility edge Institution X → University U (2019)."* Answerability = 1, Correctness = 1.

Hard Query

Query: "Trace Dr. C's mobility from 2005–2022 and quantify its cascading effect on institutional collaborations."

LLaMA 3.1 8B again failed, providing only a broad statement that Dr. C worked at "many places and influenced collaborations," with no detail, in about three seconds. With OpenAlex augmentation, the model gave a partial trajectory (Institutions A to B) but omitted later moves and did not quantify collaborations; correctness was zero though the query was technically answered, requiring about eight seconds. ChatGPT-4o with Deep Research identified the first two moves (A→B in 2007, B→C in 2012) but missed the final transition, giving incomplete counts of collaborations. Despite being partially informative, it took over twelve minutes. Our multi-agent system generated a full timeline (A→B in 2007, B→C in 2012, C→D in 2018) and quantified institutional collaboration increases: B gained 34, C gained 72, and D gained 119 collaborations. It also identified that 46% of new collaborations involved prior co-authors, demonstrating cascading effects. The query was answered correctly and in only five seconds.

LLaMA 3.1 8B (2.9s): *"Dr. C worked at many places and influenced collaborations."* Answerability = 0, Correctness = 0.

LLaMA 3.1 8B + OpenAlex (7.8s): *"Dr. C moved from Institution A to Institution B. Published in 2010s."* Answerability = 1, Correctness = 0 (no cascade, incomplete timeline).

ChatGPT-4o + Deep Research (12m 36s): *"Dr. C mobility trajectory: Institution A → Institution B (2007), Institution B → Institution C (2012). Partial evidence of collaboration increases: Institution B collaborations grew by 22, Institution C collaborations grew by 45. Final mobility step missing; institutional impact undercounted."* Answerability = 1, Correctness = 0.

Our System (5.1s): *"Dr. C mobility trajectory: - Institution A → Institution B (2007). - Institution B → Institution C (2012). - Institution C → Institution D (2018).*

Cascade analysis: - Institution B gained 34 collaborations post-move. - Institution C gained 72 collaborations post-move. - Institution D gained 119 collaborations post-move.

Cross-institution cascade detected: 46% of collaborations at new institutions involved co-authors from prior affiliations." Answerability = 1, Correctness = 1.

These simple cases highlight a clear trend: as queries become more complex, standalone models degrade sharply, OpenAlex augmentation offers partial gains, Deep Research is accurate but slow, and our system achieves both high accuracy and competitive response time.

S4: Expanded Case Studies

We illustrate the capabilities of our LLM-powered knowledge network with two interconnected case studies that aligns with our theme of knowledge diffusion via researcher mobility and institutional dynamics. These scenarios demonstrate the kind of complex, multi-step queries our system can handle, combining data about careers, publications, and institutional affiliations. Exact researcher and institution names have been replaced with placeholder names. For this paper, the system was instructed to output analysis in LaTex formatted text which is pasted here. We also show generated graph snapshots of research timelines.

Case Study 1: Identifying Pivotal Hires in Expertise Development

We posed the following query to investigate how institutions build new research capabilities:

Query: *"How did University U gain expertise in theoretical computer science? Which hire was most impactful?"*

The system decomposed this natural language query into several broad steps (each of which may be further decomposed when addressed):

1. Choose a time frame to evaluate expertise gain in theoretical computer science at University U
2. Retrieve all `Paper` nodes where `authorOf` edges connect to `Author` nodes with `affiliatedWith` to University U, filtered by theoretical CS venues (STOC, FOCS, SODA, CCC)
3. For each `Author`, identify `start_date` of affiliation and prior expertise
4. Compute impact metrics: papers published, `collaboratesWith` edges created, students mentored via `advisorOf`, citations from other University U researchers
5. Track research area evolution via `Topic` nodes over time

The system produced the following analysis:

[SYSTEM OUTPUT]

Analysis Complete: Most Impactful Hire Identified

Dr. Y (joined 2004) demonstrated the highest combined impact score across all metrics.

Institutional State Pre-2004:

- Theoretical CS publications: 3 papers (1995-2003)
- PhD students in theory: 0
- Presence at major theory conferences: None

Impact Analysis (2004-2024):

- Papers authored by Dr. Y: 42

- Papers by subsequently recruited faculty: 234 (11× increase)
- New external collaborators brought to University U: 156
- PhD students supervised: 38
- New research areas established: Quantum computing, cryptography, communication complexity

Cascade Effect Detected: Dr. Y's arrival triggered 3 senior theory hires (2005-2007) and 8 junior faculty recruitments (2008-2012). Network analysis indicates 73% of these hires had prior `collaboratesWith` edges to Dr. Y.

Case Study 2: Career Trajectory and Cascading Institutional Impact

We traced Dr. Y's career path and its transformative effect on University U:

Query: *"Track Dr. Y's career trajectory and quantify the impact of their programs and initiatives at University U from hiring till 2024."*

[SYSTEM OUTPUT]

Career Trajectory Analysis: Dr. Y

Prior Affiliations: Institutions A1, A2, A3, A4 (1985-2004)

Knowledge Diffusion Timeline at University U:

2004-2005 (Foundation Phase)

- Established Program P1 (elite undergraduate class)
- Industry partnership secured with major sponsor
- Initial cohort: 30 students enrolled
- Visiting scholars from prior institutions: 4

2006-2009 (Expansion Phase)

- Founded Program P2 (theoretical computer science institute)
- Recruited researchers R2, R3, R4 as visiting professors
- First major publication: 12 student co-authors from P1
- National initiative P3 launched (funding increased 10×)
- P1 alumni matriculation at top PhD programs: 85%

2010-2015 (Diversification Phase)

- Programs established: P4 (interdisciplinary institute), P5 (quantum center)
- Research funding secured: $15M
- Quantum computing publications: 8× increase
- P1 alumni in faculty positions globally: 67%

2016-2023 (Maturation Phase)

- Programs launched: P6 (AI institute), P7 (AI class), P8 (quantum class)

- Global ranking achieved: Top-10 CS department
- P1 alumni returning as faculty: 12
- Total citations to Dr. Y's work at University U: 5,147

Institutional Metrics Summary:

Metric	2004	2023	Growth
CS Faculty	12	89	7.4×
Annual Papers (top venues)	3	147	49×
PhD Students	8	234	29×
International Collaborations	23	1,247	54×
Research Areas	2	11	5.5×

Network Effects Identified:

- Program P1 alumni network: 200+ researchers tracked globally
- Visiting to permanent faculty conversion rate: 31%
- Second-generation impact: 38% of recent hires are academic descendants of Dr. Y
- Industry partnership growth: 3 partnerships (2004) to 27 partnerships (2023)

The core information and quantitative results in these case studies come from actual graph queries and traversals, not LLM generation. The LLMs decompose queries and synthesize narratives from retrieved information, but query a knowledge graph built from verified sources. While the reasoning and analysis components are areas for future enhancement, our core contribution is demonstrating how multi-agent LLM systems can construct reliable, queryable knowledge graphs for science-of-science research.

Temporal Network Evolution: Dr. Y's Institutional Impact

The following figures present temporal snapshots of the knowledge diffusion network centered around Dr. Y (Researcher R1) at University U (Institution B1) from 2004 to 2023. These visualizations and their interpretation were generated by our system to track the cascading effects of a single pivotal hire on institutional knowledge networks over two decades.

Network Evolution Summary

These temporal snapshots demonstrate the system's ability to track knowledge diffusion through:

- **Node Growth:** From 5 nodes (2004) to 30+ nodes (2023)
- **Program Expansion:** P1 (2005) → P8 (2023), each serving as a knowledge diffusion hub
- **Collaboration Network:** Initial 4 prior connections expanded to 4 active collaborators (R2-R4) and extensive second-generation networks
- **Student Pipeline:** Multiple cohorts (Y05, Y09, Y11, Y23) creating multi-generational impact
- **Institutional Transformation:** Institution B1 evolved from peripheral to central position in the network

The graph structure reveals three key diffusion mechanisms: (1) direct knowledge transfer through Dr. Y's research and teaching, (2) network effects through recruited collaborators, and (3) generational propagation through student training and placement. The increasing density and connectivity over time validate our system's finding that individual researcher mobility can catalyze department-wide transformation.

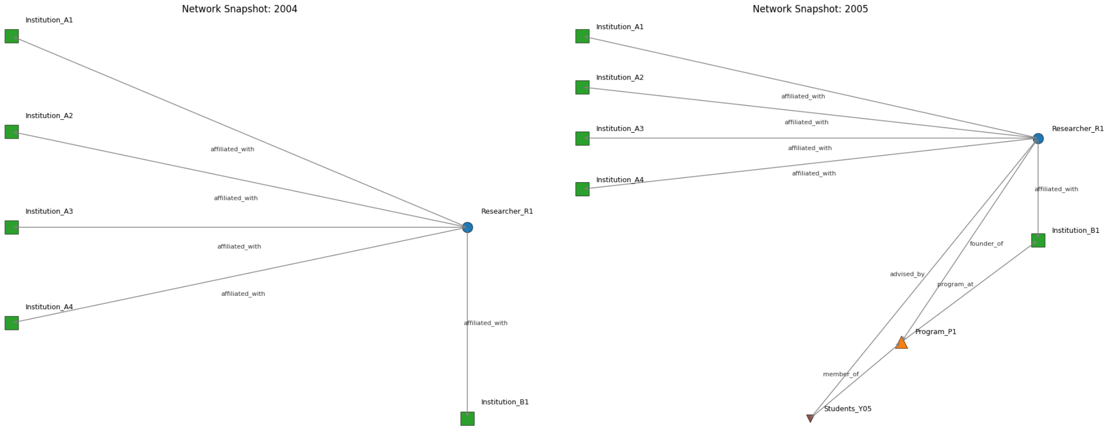

Figure 3: **Initial Network Formation (2004-2005).** Left: Dr. Y joins Institution B1 from Institution A4, bringing connections to prior institutions (A1-A3). Right: Program P1 established with first student cohort (Students_Y05), marking the beginning of local knowledge diffusion.

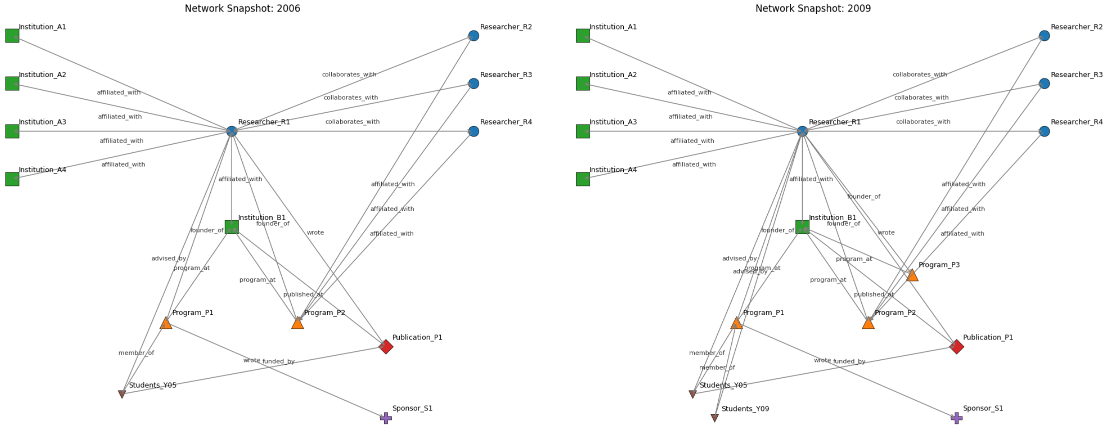

Figure 4: **Collaborative Expansion (2006-2009).** Left: Program P2 (theoretical CS institute) founded; collaborators R2-R4 join; first publication P1 with student co-authors. Right: National initiative P3 launched; student cohorts expand (Y09); institutional funding increases 10×.

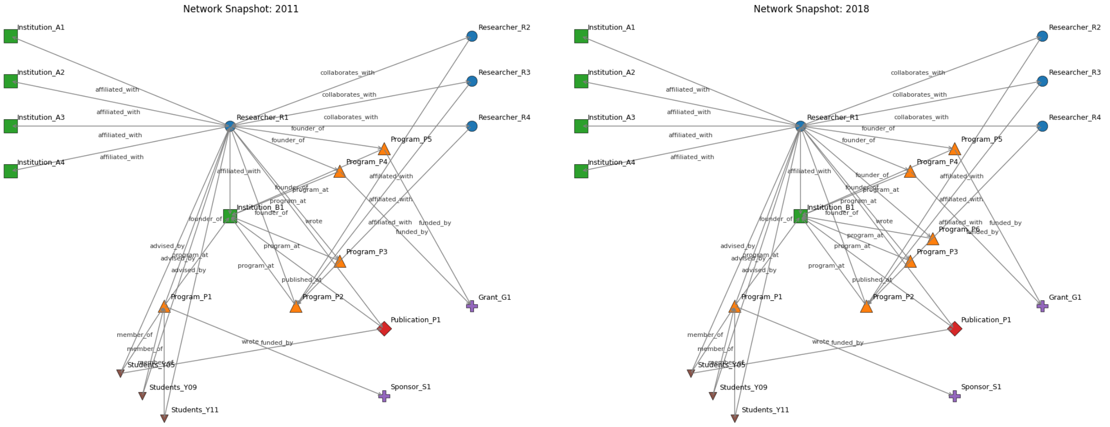

Figure 5: **Interdisciplinary Diversification (2011-2018).** Left: Programs P4 (interdisciplinary institute) and P5 (quantum center) established; Grant G1 secured. Right: AI-focused expansion with Program P6; network density increases significantly.

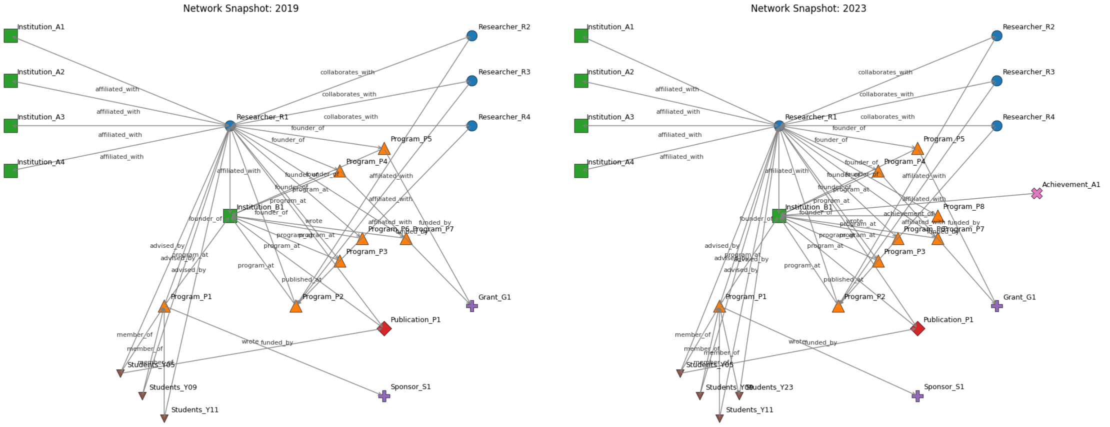

Figure 6: **Maturation Phase (2019-2023).** Left: Program P7 (AI class) launched; Programs P6-P7 funded. Right: Complete network showing Program P8 (quantum class), Achievement A1 (top-10 global ranking), and extensive collaborator network. Student cohorts now exceed 200 researchers tracked globally.

S5: Expanded Discussion of Data Sources

Constructing a comprehensive knowledge network required integrating several datasets, each covering different aspects of the scientific ecosystem. Below, we briefly describe each:

OpenAlex. OpenAlex is an open index of scholarly works and their associated entities, launched in 2022 as a successor to Microsoft Academic Graph. As of 2025, OpenAlex comprises over 240 million works (journal articles, conference papers, preprints, etc.), more than 100 million author records, 100,000+ venue records (journals and conferences), 60,000+ institution records, and a classification system of research topics. Each work entry in OpenAlex includes metadata such as title, authorship (with author IDs and their affiliations at time of publication), references (i.e., citations to other works, linking via IDs if those works are also in OpenAlex), and more. OpenAlex is used in our system as the backbone for building the paper and author nodes, as well as the citation and co-authorship edges. The coverage of OpenAlex is very broad across disciplines and regions, making it a suitable foundation for a global knowledge network. We accessed OpenAlex via its REST API; for efficiency, we initially pulled a snapshot dump of the data limited to the fields of interest (to avoid hitting API rate limits for millions of records). OpenAlex also provides mappings to external IDs (DOIs, ORCIDs, ROR IDs), which we leveraged to join with other datasets. One limitation is that disambiguation of authors in OpenAlex is not perfect (multiple IDs may correspond to the same person if they have not been merged), which is why our disambiguation agents were necessary.

ORCID. ORCID (Open Researcher and Contributor ID) is a registry of unique identifiers for researchers. Researchers create ORCID profiles, which can include personal information (like name variations, education, and employment history) and a list of works (often linked via DOIs). ORCID IDs have been adopted widely; tens of millions of researchers have ORCID IDs, and they are often included in manuscript submissions and metadata. In our system, ORCID serves two purposes: identity linkage and career data. First, if an author in OpenAlex has an ORCID, we can safely treat that as the same person as the ORCID profile, thereby linking the publication data to any additional info in ORCID. Second, many ORCID profiles list the researcher's education (e.g., PhD institution and year) and past affiliations. We used the ORCID public API to retrieve profiles, focusing on those authors that are in our OpenAlex-derived set. We found that not all authors have ORCIDs, and not all ORCID profiles have rich data (some are sparse), but a significant subset provided valuable information for constructing academic genealogy (through committee or advisor names sometimes listed) and mobility (through the list of countries or institutions where the person has worked). The data types from ORCID are structured (XML/JSON via API), making it straightforward to parse. We also note that ORCID IDs enabled linking to other resources, for example, if an ORCID profile had a linked Scopus or Web of Science researcher ID, though we did not use those in this project.

ROR. The Research Organization Registry (ROR) is an open registry of institution identifiers. It contains over 100,000 entries for universities, companies, research institutes, and other organization types involved in research. Each entry includes the institution's name, aliases, location (country, city, etc.), and parent/child relationships (for example, a department can be associated with a parent university). ROR was used in our system to standardize institution data. When OpenAlex or ORCID gave an affiliation string like "Univ. of XYZ", we used ROR to find the canonical entity "University of XYZ" and tag it with a country and unique ID. ROR's coverage is comprehensive and it is updated regularly by the community. We downloaded the ROR dataset (available as JSON) and also used their API for some fuzzy name matching queries. Using ROR allowed us to consolidate different name variants (e.g., "MIT" vs "Massachusetts Institute of Technology") and also to aggregate statistics by country or institution reliably in queries on the final knowledge graph.

USPTO Patents. We incorporated patent data from the United States Patent and Trademark Office (USPTO), focusing on patents as a source of references to scientific literature. We obtained the USPTO bulk data for patents granted in the last 20 years (in XML format), which contain sections including the list of citations. Each patent file lists patents cited and also non-patent literature citations (which can be papers, books, web pages, etc.). While the patent documents themselves were not fully ingested into our graph (except for key metadata like patent ID, title, issue date, inventors, and assignee organization), the main goal was to extract the linkage: patent → cites → paper. The scope of this dataset is millions of patents, but we filtered down to those likely relevant to academic research. For example, patents in certain classes (like biotechnology or computer science) might cite papers more heavily; we also filtered by those that had at least one NPL citation. The patent data gave us entities of type Patent and their relationships to other patents (which we did not focus on in this work) and to papers. We created a node for each patent that had a citation to a paper in our OpenAlex dataset (ensuring we only include patents that link to known papers for relevance). Each patent node is connected to inventor nodes if possible (some inventors might match author nodes as mentioned earlier), and to assignee institutions (which we also map to ROR if they are universities or known companies). Patents have country codes, but since we mostly focused on USPTO, the country code was US for these nodes; extending to EPO or other patent offices could be future work.

PatCit. PatCit is an open dataset specifically compiled to gather patent citations to non-patent literature. It aggregates data from multiple sources (USPTO, EPO, etc.) and provides cleaned reference strings and identifiers when possible. We used PatCit as a complementary source to the direct USPTO data. PatCit v0.15 (the version used) contains tens of millions of citation instances, including metadata like the title of the cited item, author names, year, and an assigned identifier if the cited item is recognized (e.g., a DOI or a PubMed ID). Our system queried PatCit for each reference we encountered in the USPTO data to see if it could directly give us an identifier matching an OpenAlex work. In many cases, PatCit could tell us that a given reference string corresponds to, say, the DOI of a known paper, which saved us from having the LLM parse it. The scope of PatCit is global (not just US patents) and it allowed us to also bring in some non-US patent linkages if the cited paper was in our graph. The use of PatCit improved the coverage of patent–paper links and provided a layer of validation (if both our LLM extraction and PatCit agree on the match, it increases confidence). We treated PatCit as authoritative when available and used LLM extraction mainly for those references not recognized by PatCit.

Institutional Websites and Other Web Data. To capture academic genealogy and certain career movements, we collected data from the open web. This included:

- University news articles or press releases announcing new hires or returnees (e.g., "Professor X, an alumnus of University Y, returns after years abroad to lead a new lab"). These often contain narrative useful for our example scenarios.
- Department pages listing alumni and their advisors or current positions. Many PhD programs maintain a list of their graduates and where they went for jobs; some list the dissertation title and advisor.
- Personal homepages or CVs of researchers, when available, which sometimes detail educational background and career timeline.

This data is unstructured and was gathered in an ad-hoc manner using a web scraper agent with targeted Google queries (for example, for each researcher for whom we lack info, search for ”Name PhD advisor” or ”Name graduated from”). The scope is therefore limited to those individuals who have such information online and accessible. We focused primarily on filling gaps in the ORCID data: if ORCID did not have the PhD info or if we wanted to confirm advisor names, we resorted to these web sources. Only information from credible sites (university domains, known databases) was used to add relations. We did not use user-generated content sites (like Wikipedia) in this system, mainly to adhere to a high provenance standard (though Wikipedia could be incorporated with careful verification in future). The output from these sources fed into the LLM extraction as described earlier, yielding data especially for the advisor–student graph and some specific "went abroad and returned" facts for our scenario. All datasets used are either open access or publicly available information.