LLM-Enabled Scientific Knowledge Diffusion Analysis

Uttam Rao, Madhav Marathe

Biocomplexity Institute and Dept. of Computer Science
University of Virginia
ur6yr@virginia.edu, mvm7hz@virginia.edu

Abstract

Bibliometric and science-of-science studies have yielded valuable insights into co-authorship and citation networks, yet most analyses rely on static datasets and limited relation types. We introduce a multi-agent AI 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 and research agents 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.

Extended version (also on arXiv) —
https://www.uttamrao.com/papers/BITR202513.pdf

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 One such scholar arrives with expertise in a specialized subfield, bringing collaborators, unpublished ideas, and advanced methods.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.

Related Work

Early studies treated citation and collaboration data as networks, with Price (Price 1965) and Newman (Newman 2001) 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 (Fortunato et al. 2018). Recent work also highlights global fragmentation, where citation preferences restrict cross-border diffusion (Gates, Gao, and Mane 2025). Complementary strands have examined academic genealogy (advisor–advisee relations) (Arslan, Gunes, and Yuksel 2011), international mobility of researchers (Franzoni, Scellato, and Stephan 2012), and the growing linkage between patents and scholarly papers (Narin, Hamilton, and Olivastro 1997), supported by datasets such as PatCit. Large bibliographic graphs such as Microsoft Academic Graph, OpenAlex (Priem, Piwowar, and Orr 2022), and AMiner (Tang et al. 2008) 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., (Zhu et al. 2024)). 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 for extended discussion of related work.)

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 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 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

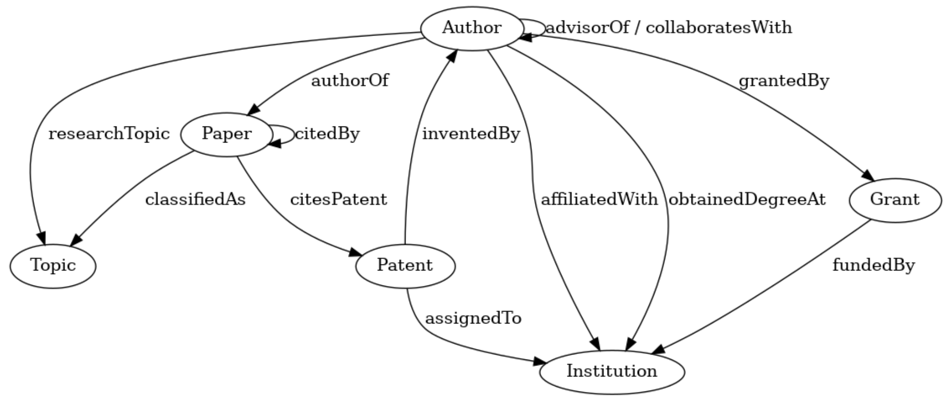

Figure 1: Graph schema representing entities and relationships in the scientific knowledge network. Though not explicitly shown in the figure above, all edges are temporal. See Supplementary Information section S2.

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).

System Architecture

Our system leverages `AutoGen` (Wu et al. 2024) 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).

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).

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.

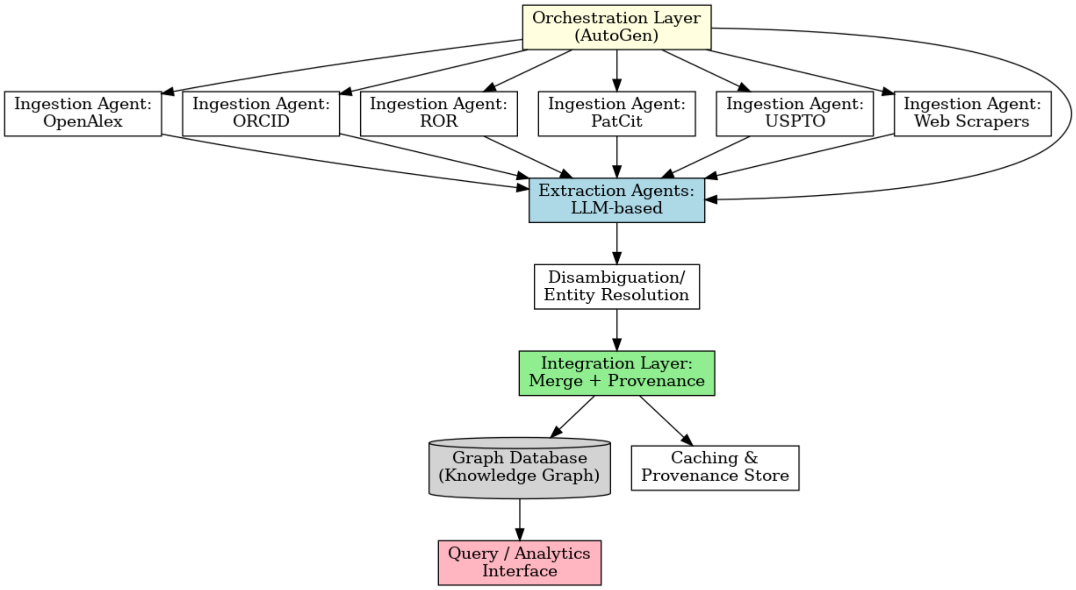

Figure 2: High Level System Architecture: orchestration layer coordinates specialized agents across ingestion, extraction, disambiguation, integration, and analysis.

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.

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.

Implementation Details

Most of the pipeline was executed utilizing an A100 GPU via a `vLLM` (Kwon et al. 2023) 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 (Touvron et al. 2023). 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 (Webber 2012) is used as the graph database.

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 includes further discussion of each data source and their uses.

Evaluation

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)

Dataset	Scope / Coverage	Key Properties (size, etc.)
OpenAlex (Priem, Piwowar, and Orr 2022)	Scholarly publications (papers, journals, conferences) with authors, affiliations, citations, topics.	∼240M works; ∼50M author entities; covers all disciplines (we focus on CS/AI subset); updated daily; open API access.
ORCID (Haak et al. 2012)	Researcher profiles (self-reported): education, employment, publications, etc.	15M+ registered researchers; rich metadata for career timelines; used for author disambiguation via unique IDs.
ROR (Research Organization Registry 2019)	Research organization registry (institutions, companies, etc.).	∼100k institutions worldwide; each with unique ID and metadata (names, aliases, location); helps unify affiliation names.
USPTO Patents (United States Patent and Trademark Office 2025)	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 (de Rassenfosse et al. 2020)	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 for further discussion.

- **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, and (c) Llama 3.1 8B.

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 (Zhao et al. 2023; Shen et al. 2023), 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 (Schuff et al. 2023), 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 includes further examples of our evaluation.

Results and Comparison

Table 2 summarizes the evaluation results (across the 20 queries for each type).

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, presented as averaged numbers (as exact values are less important than 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.

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.

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.

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.

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.

Acknowledgments

The authors would like to thank members of the Biocomplexity Institute and the National Security Data and Policy Institute. This work is partially supported by University of Virginia Strategic Investment Fund Awards: SIF176A. The research is also partly supported by the National Security Data & Policy Institute, Contracting Activity 2024-24070100001.

References

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

de Rassenfosse, G.; Verluise, C.; Cristelli, G.; Higham, K.; and Violon, L. 2020. PatCit: A Comprehensive Dataset of Patent Citations. Available at: https://doi.org/10.5281/zenodo.3710993. Accessed: 2025-08-18.

Fortunato, S.; Bergstrom, C. T.; Börner, K.; Evans, J. A.; Helbing, D.; Milojević, S.; Petersen, A. M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; Vespignani, A.; Waltman, L.; Wang, D.; and Barabási, A.-L. 2018. Science of science. *Science*, 359(6379): eaao0185.

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

Gates, A. J.; Gao, J.; and Mane, I. 2025. The increasing fragmentation of global science limits the diffusion of ideas. *arXiv preprint arXiv:2404.05861*.

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

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, 611–626.

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

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

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

Priem, J.; Piwowar, H.; and Orr, R. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833.

Research Organization Registry. 2019. ROR: Research Organization Registry. Community-led registry. Available at: https://ror.org. Accessed: 2025-08-18.

Schuff, H.; Vanderlyn, L.; Adel, H.; and Vu, N. T. 2023. How to Do Human Evaluation: A Brief Introduction to User Studies in NLP. *Natural Language Engineering*, 29(5): 1199–1222.

Shen, C.; Cheng, L.; Nguyen, X.-P.; You, Y.; and Bing, L. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. *arXiv preprint arXiv:2305.13091*.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 990–998.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

United States Patent and Trademark Office. 2025. USPTO Bulk Data Directory. Government database. Available at: https://data.uspto.gov/bulkdata/datasets. Accessed: 2025-08-18.

Webber, J. 2012. A programmatic introduction to Neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, 217–218. ACM.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E. E.; Jiang, L.; Zhang, X.; Zhang, S.; Awadallah, A.; White, R. W.; Burger, D.; and Wang, C. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *COLM 2024*.

Zhao, C.; Gella, S.; Kim, S.; Jin, D.; Hazarika, D.; Papangelis, A.; Hedayatnia, B.; Namazifar, M.; Liu, Y.; and Hakkani-Tur, D. 2023. "What do others think?": Task-Oriented Conversational Modeling with Subjective Knowledge. In *SIGDIAL 2023*, 309–323. Association for Computational Linguistics.

Zhu, Y.; Wang, X.; Chen, J.; Qiao, S.; Ou, Y.; Yao, Y.; Deng, S.; Chen, H.; and Zhang, N. 2024. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. *World Wide Web*, 27.