Deep Semantic Linking of Scientific Knowledge: An Agentic AI Framework for Knowledge Graph Construction

Sandra Schaftner[1 [0009−0008−3235−3042]] and
Martin Gaedke[1 [0000−0002−6729−2912]⋆]

Chemnitz University of Technology, Chemnitz, Germany
{sandra.schaftner,martin.gaedke}@informatik.tu-chemnitz.de

Abstract. The foundation of scientific research is the comprehensive analysis of existing literature. However, the exponential growth of published research leaves scientists increasingly overwhelmed, making it difficult to maintain a complete overview of the state-of-the-art or to discover hidden synergies between studies. The root of this problem lies in the traditional format of scholarly communication: crucial knowledge about applied methods, datasets, and metrics remains locked in semantically unlinked documents. While this format is optimal for human reading, it is highly inefficient for machine processing. Even modern AI research assistants frequently fail to provide complete, verifiable, and hallucination-free answers to complex research queries. To enable true machine-assisted exploration and verification, scholarly literature must be transformed from isolated documents into deeply interlinked, machine-readable structures. To achieve this, we introduce an agentic AI framework that automatically extracts key research entities, seamlessly interlinking the literature by mapping them to standard knowledge bases via unified URIs.

Keywords: Scientific Knowledge Graphs · Agentic AI · Large Language Models

1 Introduction

In daily digital life, graph-based representations are the invisible backbone of modern AI applications. Whether for Amazon product recommendations, Spotify playlists, or Google's semantic search, they rely on representing information as a machine-readable web of interlinked concepts.

In academia, this structured data foundation is missing. Scholarly communication relies on isolated documents, leaving valuable research knowledge – such as applied methods, metrics, or datasets – "frozen" in unstructured text. Relying on current AI tools for research causes three fundamental challenges: (1) a lack of verifiability, as the opaque nature of Large Language Models (LLMs) prevents

⋆ PhD supervisor of the first author.

reliable traceability to original sources and makes them prone to hallucinations; (2) a lack of completeness, where complex aggregation queries (e.g., assessing method frequencies across thousands of papers) fail; and (3) a lack of interlinking, preventing the discovery of cross-disciplinary or -institutional synergies [17].

Bioinformatics demonstrates how structured knowledge representation overcomes these issues. The UniProt database exemplifies a highly precise Knowledge Graph (KG): protein knowledge is deep-semantically structured into triples, every fact links to its original paper (provenance), and entities seamlessly connect to external databases [15]. While other academic disciplines lag behind this ideal, replicating UniProt's manual expert curation is unfeasible given their sheer literature volume [15]. Automation is the only logical path forward, yet relying on uncontrolled LLMs fails to deliver the required scientific precision [7].

This dilemma leads to our central research question: *How can research knowledge, isolated in unstructured text documents, be scalably transformed into a machine-readable, semantically interoperable, and verifiable representation?*

To answer this, we conceptualize ALLMaC-SKG – Agentic LLM-augmented Construction of Scientific Knowledge Graphs (SKGs). This scalable framework establishes a verifiable data foundation required for advanced scientific discovery through (1) deep semantic mapping to standardized ontologies, (2) agentic workflow automation, and (3) federated exploration and interlinking.

2 Problem Statement and Motivation

Advanced scientific discovery is fundamentally hindered by a lack of deep semantic interlinking within published research. Persistent reliance on unstructured documents traps valuable research within isolated boundaries, severely limiting machine-actionability, resulting in poor input quality for modern LLMs, and impeding both automated extraction and manual exploration.

Limitations of Current AI Assistants. To address literature research challenges, current AI-powered assistants (e.g., Elicit, Consensus, SciSpace and Scite) leverage LLMs to synthesize answers. While they rely on underlying graphs, these are largely restricted to metadata – such as the Semantic Scholar Academic Graph (S2AG) – combined with stochastic text retrieval and vector embeddings [9]. While useful for basic discovery, these approaches inherently fail at precise quantitative aggregations or federated cross-repository retrieval. Current systems cannot reliably answer aggregation queries like: "How many papers in the last three years applied a Whisper model to the Common Voice dataset?" or "What unexploited research intersections exist between TU Chemnitz, the University of Udine, and the University of Girona?" To reliably answer such complex queries, AI requires structured graphs of explicit semantic facts, not just similarity-based vector embeddings.

Industry and Policy Needs for Structured Knowledge. Recognizing the limitations of purely text-based and shallow-structured approaches, both industry and

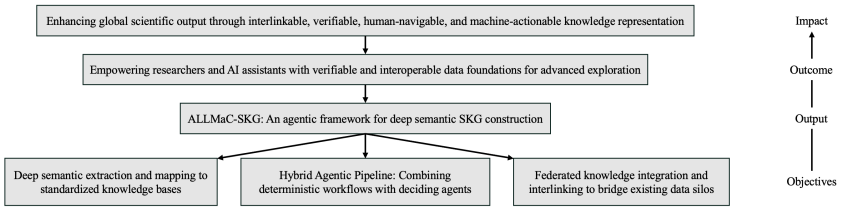

Fig. 1. The ALLMaC-SKG objectives tree, illustrating the path to overarching impact.

academia strongly push toward KGs for knowledge representation. In the enterprise sector, Gartner highlights that meeting high accuracy thresholds[1] makes it "crucial to bridge the gap between data and AI by using knowledge graphs"[2]. Parallel to industry, scientific initiatives – such as the Research Data Alliance (RDA), the European Open Science Cloud (EOSC), the Australian Research Data Commons (ARDC), and Germany's NFDI – aggressively pursue federated SKGs [2, 14]. To achieve true FAIR compliance, these infrastructures rely on RDF and Linked Data principles. The overarching goal is to semantically interlink distributed knowledge via reusable standard ontologies and knowledge bases. Consequently, evaluating these systems requires strict adherence to FAIR principles, particularly semantic Interoperability (I) and Reusability (R). To effectively combat AI hallucinations, frameworks must further prioritize structural accuracy, semantic consistency, and provenance to the source document.

3 Related Work

The transition toward machine-actionable scholarly communication has driven the development of various SKGs [17]. However, achieving both deep semantic representation and high scalability remains an unsolved challenge. While LLMs are increasingly utilized, one of the latest paradigms of deploying them – Agentic AI – remains largely unexplored in the realm of KG construction (KGC).

Current Landscape of SKGs. Most existing SKGs model only shallow metadata (e.g., authorship, venues, citation networks), lacking detailed methodological content [17]. Apart from highly specialized domain KGs like UniProt [15], literature reports only two SKGs deeply model scientific claims: the Open Research Knowledge Graph (ORKG) [1] and the Computer Science Knowledge Graph (CS-KG) [6]. While UniProt excellently demonstrates deep semantics, scaling manually curated, niche-specific micro-graphs to every scientific domain is resource-wise unfeasible. However, existing broad-scope attempts face severe limitations. ORKG's reliance on manual crowdsourcing fails to scale and struggles with consistency [11]. CS-KG is criticized as sparse and incomplete [4].

[1] Gartner Report. https://www.gartner.com/en/documents/7444326

[2] Gartner Press Release. https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-data-and-analytics-summit-2025-orlando-day-3-highlights

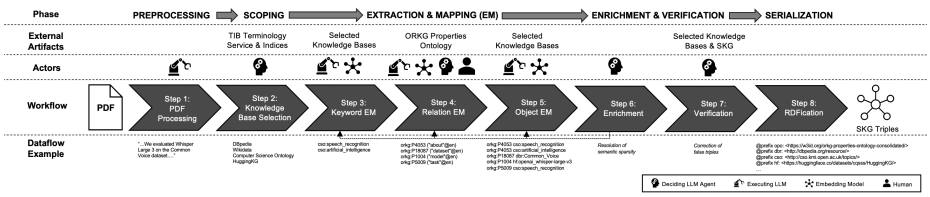

Fig. 2. The ALLMaC-SKG pipeline: Deterministic workflow with deciding agents.

Automated Construction and Agentic Workflows. To overcome manual bottlenecks, recent KGC pipelines increasingly leverage LLMs [3, 17], yet rarely map entities to standardized vocabularies. Recently, KGC research has seen a surge in multi-agent frameworks [10]. However, many simply chain static LLM prompts together and label them as "agents," lacking the autonomous, dynamic plan-act-reflect loops defining true agentic systems [5]. One exception is Kaplunovich et al. [8], who apply a LangGraph KGC framework to the movie domain.

The Research Gap. There is currently no overarching automated framework for the deep semantic construction of SKGs from text. By fully exploiting advancements in agentic AI, this work proposes a concept and an architecture that entirely bridges the gap between high scalability through automation and deep semantic precision.

4 Objectives and Contributions

To fundamentally improve how researchers and AI systems interact with published literature, we introduce the ALLMaC-SKG framework. To realize the impact detailed in Figure 1, this work is driven by three core objectives that constitute our main contributions:

Deep Semantic Mapping. Moving beyond superficial metadata, ALLMaC-SKG captures fine-grained entities like methods, datasets, and specific research properties. To answer complex multi-hop queries reliably, heterogeneous textual mentions must be resolved. The methodology robustly maps both extracted entities and properties to URIs via mapping agents utilizing vector search. Rather than creating redundant hierarchies, extractions are strictly linked to standard knowledge bases and ontologies like the ORKG Properties Ontology (OPO) [12, 13], ensuring FAIR compliance and exploiting existing structural features.

Agentic Workflow Automation. ALLMaC-SKG utilizes a hybrid agentic pipeline, building on a deterministic workflow to guarantee consistent, sequential extraction routing, which is highly resource-efficient. According to LangGraph principles[3], workflows follow predetermined code paths, whereas true agents operate in

[3] LangChain AI: Workflows and Agents. https://docs.langchain.com/oss/python/langgraph/workflows-agents

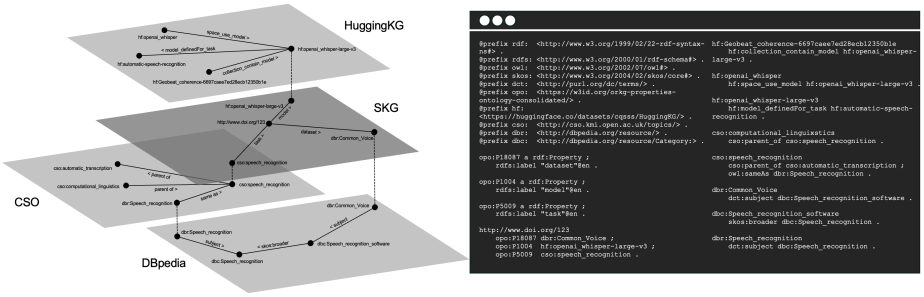

Fig. 3. Visualization of semantic snapping in ALLMaC-SKG. Dashed vertical lines represent the alignment of identical entities across the central SKG and external KGs.

dynamic feedback loops, directing their own actions. Such agents are employed at critical junctures requiring complex reasoning: (1) determining required domain-specific knowledge bases via search tools; (2) evaluating new relations for the OPO via Human-in-the-Loop review; (3) assessing semantic depth and deciding where iterative refinement is needed; and (4) conducting critical, tool-assisted verification before finalizing a paper. As seen in Figure 2, other LLMs are utilized merely as processing tools within the deterministic path.

Federated Exploration and Semantic Snapping. To enable boundary-spanning knowledge discovery, isolated research must be placed into a global context. Mirroring bioinformatics initiatives [16], the methodology dynamically links publications with datasets, domain knowledge bases and global encyclopedic graphs like DBpedia. This envisions a meta-registry (similar to EOSC [14]) enabling "Semantic Snapping". Architecturally, ALLMaC-SKG can dock onto existing metadata foundations like S2AG [9], enriching them with deep semantic triples. Because all artifacts share global URIs, users can modularly combine KGs (e.g., institutional KGs + DBpedia + domain KGs), which semantically "snap" together at matching URI intersection nodes (cf. Figure 3), dynamically interconnecting isolated knowledge spaces to uncover hidden research intersections.

5 Conclusion and Outlook

Constructing robust, semantically rich web-based knowledge architectures from unstructured text is a critical Web Engineering challenge. Even if future LLMs achieve near-perfect zero-shot extraction, the rigorous mapping of text to standard knowledge bases via deterministic agentic workflows remains a highly significant, FAIR-compliant blueprint transcending proprietary black-box products.

Toward this goal, initial efforts focused on foundational property extraction, resulting in the LOPE (LLM-driven Ontology-based Property Extraction) method and the OPO consolidation [12, 13]. Future work will implement and rigorously evaluate the remaining ALLMaC-SKG modules against SOTA ap-

proaches. Finally, deploying the framework within the EU-funded Across Alliance[4] will assess its scalability, interoperability, and practical utility. This real-world application aims to overcome knowledge fragmentation, providing AI assistants and researchers with the data foundation demanded by modern research.

Acknowledgments. This work is supported by the European Union's Erasmus+ Programme under grant agreement No 101177485, project Across (European University for Cross-Border Knowledge Sharing), and by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Auer, S., et al.: Improving access to scientific literature with knowledge graphs. Bibl. Forsch. Prax. **44**(3), 516–529 (2020). https://doi.org/10.1515/bfp-2020-2042
2. Bernard, L., et al.: Base4nfdi – basic services for nfdi (2023)
3. Bian, H.: LLM-empowered knowledge graph construction: A survey (2025)
4. Borrego, A., et al.: Completing scientific facts in knowledge graphs of research concepts. IEEE Access **10**, 125867–125880 (2022)
5. Chand, B., et al.: Synergistic ai agents: Integrating knowledge graphs and large language models for scholarly communication. Open Conference Proc. **8** (2026)
6. Dessì, D., et al.: SCICERO: A deep learning and nlp approach for generating scientific knowledge graphs. Knowl.-Based Syst. **258**, 109945 (2022)
7. Huang, L., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**(2) (2025)
8. Kaplunovich, A.: Langgraph-orchestrated LLM agents for scalable movie knowledge graphs and question answering. In: Proc. ICAIR 2025 (2025)
9. Kinney, R., et al.: The semantic scholar open data platform (2025)
10. Lu, Y., et al.: Karma: Leveraging multi-agent llms for automated knowledge graph enrichment (2026), arXiv:2502.06472
11. Nechakhin, V., et al.: Evaluating LLMs for structured science summarization in the ORKG. Information **15**(6) (2024). https://doi.org/10.3390/info15060328
12. Schaftner, S., Gaedke, M.: The lope method: Improving consistent property extraction for scientific knowledge graphs using llms. In: WWW '26 Companion (2026)
13. Schaftner, S., Gaedke, M.: Orkg properties ontology consolidated: LLM-driven refinement of crowdsourced knowledge for machine-actionability. In: WWW '26 Companion (2026)
14. Schirrwagen, J., et al.: Data sources and persistent identifiers in the open science research graph of openaire. Int. J. Digital Curation **15**, 5 (2020)
15. The UniProt Consortium: Uniprot: A worldwide hub of protein knowledge. Nucleic Acids Res. **47**(D1), D506–D515 (2018). https://doi.org/10.1093/nar/gky1049
16. Waagmeester, A., et al.: Wikidata as a knowledge graph for the life sciences. eLife **9**, e52614 (2020). https://doi.org/10.7554/eLife.52614
17. Zloch, M., et al.: Research knowledge graphs: The shifting paradigm of scholarly information representation. In: The Semantic Web. pp. 140–154 (2025)

[4] Across Alliance. https://www.across-alliance.eu/