Wang, J.X. (2025). LLM-powered entity alignment for enhanced scientific collaborator recommendation. In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 25* (pp. 279-288). ICEB'25, Hanoi, Vietnam, August 21-25, 2025

LLM-Powered Entity Alignment for Enhanced Scientific Collaborator Recommendation

Jiaxiao Wang [1,*]

*Corresponding author
[1] Student, City University of Hong Kong, Hong Kong, China, jwang2722-c@my.cityu.edu.hk

ABSTRACT

Entity misalignment in knowledge graphs (KGs)—caused by noisy data or inconsistent naming—severely undermines the accuracy of academic collaborator recommendation systems. To address this, we propose a KG-based scholar recommendation system featuring a novel LLM-powered entity alignment method. Our-stage approach first identifies potential matches unsupervisedly, then leverages LLMs' semantic understanding for precise alignment. This high-fidelity alignment directly enhances KG quality, leading to more accurate recommendations. By resolving core entity ambiguity issues, our system aims to significantly improve recommendation reliability. Evaluation on real-world datasets will validate the effectiveness of the alignment method and its impact on recommendation performance.

Keywords: Knowledge graphs, entity alignment, recommendation systems, large language models, collaboration recommendation.

INTRODUCTION

In today's era of rapidly expanding and increasingly interdisciplinary scientific research, effective collaboration between researchers from diverse fields is both essential and challenging. Cross-domain recommendation systems have emerged as valuable tools for facilitating such collaborations by helping researchers discover potential partners with complementary expertise. However, existing scientific collaborator recommendation systems continue to face significant limitations. Scientific information remains fragmented across multiple databases, repositories, and digital platforms—ranging from publication databases and institutional repositories to project management systems and researcher profiles—making it difficult to construct a holistic and accurate view of any individual researcher's expertise and collaboration history. Furthermore, the data underlying these systems are typically sparse, as most researchers only collaborate with a small number of peers relative to the vast global research community. This sparsity hampers the effectiveness of traditional collaborative filtering methods and exacerbates problems such as the cold-start issue, especially for newcomers or those working in emerging fields.(Bai *et al.*, 2019; Lü *et al.*, 2012)

To address these challenges, knowledge graphs have gained attention as a promising approach for organizing and integrating heterogeneous scientific information. Scientific Research Knowledge Graphs (SRKGs) model diverse entities—such as authors, papers, institutions, topics, and datasets—and the relationships among them, thereby providing a structured and interconnected view of the scientific landscape. However, the construction of a unified and comprehensive SRKG is hindered by the problem of entity misalignment: the same real-world entity may appear under different names or identifiers across various sources. Effective entity alignment is thus critical to fusing disparate knowledge graphs into a single, coherent resource, which in turn enables more accurate and meaningful applications such as collaborator recommendation (Xin *et al.*, 2022; W. Zeng *et al.*, 2021).

Despite progress in entity alignment methods, aligning scientific knowledge graphs remains a complex task, particularly due to semantic heterogeneity and the sparsity of explicit connections between entities. Recent advances in large language models (LLMs) offer new opportunities to overcome these barriers. LLMs have shown remarkable capabilities in capturing deep semantic relationships and understanding contextual nuances by training on massive and diverse textual corpora. These strengths make LLMs especially well-suited for tasks such as knowledge graph enrichment and entity alignment, where semantic understanding and flexible reasoning are required (Ahmad & Goel, 2025). By leveraging LLMs for entity alignment, it becomes possible to create richer, more accurate, and better-integrated scientific knowledge graphs.

The primary objective of this research is to leverage large language models to perform entity alignment in scientific knowledge graphs and to utilize these aligned graphs for cross-domain scientific collaborator recommendation. Specifically, this study aims to answer the following questions: (1) How can the semantic understanding capabilities of LLMs be harnessed to achieve effective and accurate entity alignment within scientific knowledge graphs? (2) How can the aligned knowledge graphs be applied to recommendation systems to improve the accuracy and relevance of collaborator recommendations across domains? By addressing these questions, this research seeks to advance both the methodology of knowledge graph alignment using LLMs and the practical effectiveness of cross-domain collaborator recommendation in scientific research.

The potential contributions of this research are as follows. We proposed a novel LLM-based entity alignment method to construct SRKG which can significantly improve the accuracy of the recommendation system.

LITERATURE REVIEW

Knowledge Graphs and Recommender Systems

Knowledge Graphs (KGs) are structured repositories of entities and their relationships, typically represented as triples (subject, predicate, object). The construction of KGs involves key steps such as entity recognition, relation extraction, and graph fusion (Sun *et al.*, 2020). These processes enable the integration of heterogeneous data sources into a unified semantic network. KGs have found applications in diverse domains, including intelligent search, question answering, and personalized recommendation systems, where they enhance semantic understanding and contextual reasoning (Wang *et al.*, 2019).

Traditional recommendation methods, such as content-based filtering and collaborative filtering (CF), often struggle with data sparsity and cold-start issues. KG-based approaches address these limitations by enriching user and item representations with semantic information. For instance, embedding-based methods project entities into continuous vector spaces, capturing latent relationships (Zhang *et al.*, 2024). Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs), aggregate structural and attribute data to improve recommendation accuracy (Wu *et al.*, 2019). Despite their success, challenges persist in scalability and handling dynamic, heterogeneous KGs (Gao *et al.*, 2022).

Entity Alignment: Significance and Challenges

Entity alignment (EA) identifies equivalent entities across disparate KGs, enabling knowledge fusion and cross-domain applications like multilingual recommendation systems. EA is pivotal for resolving semantic heterogeneity and enhancing KG completeness (Zeng *et al.*, 2021).

Rule-based EA methods rely on handcrafted similarity metrics for entity names or structures, but they lack scalability and robustness (Balloccu *et al.*, 2023). Machine learning approaches, including supervised and semi-supervised models, require extensive labeled data and struggle with complex relational patterns (Liu *et al.*, 2022). Deep learning methods, particularly embedding-based techniques using GNNs, automate feature learning but face computational bottlenecks in large-scale KGs (Sun *et al.*, 2020). In scientific domains, keyword and entity alignment remain challenging due to domain-specific jargon and sparse interdisciplinary linkages (Zhang *et al.*, 2024).

Scientific Collaborator Recommendation Systems

Collaborator recommendation systems aim to foster interdisciplinary research by connecting researchers with complementary expertise. Traditional CF-based methods suffer from data sparsity in niche fields, while content-based approaches fail to capture semantic nuances (Bai *et al.*, 2019).

Scientific KGs integrate entities such as researchers, publications, and topics, modeling multi-relational interactions (e.g., co-authorship, citation). Models like Knowledge Graph Attention Network (KGAT) use attention mechanisms to prioritize relevant connections, improving recommendation explainability (Wang *et al.*, 2019). Hyperbolic embedding techniques (e.g., HAKG) further enhance hierarchical relationship modeling (Du *et al.*, 2022). Entity alignment plays a critical role in cross-disciplinary scenarios by resolving semantic discrepancies between domain-specific KGs (Chen *et al.*, 2022).

Effective scientific research collaborator recommendation systems rely on a multi-dimensional modeling approach to identify potential collaborators who are not only relevant in their expertise but also conducive to productive partnership.

A foundational aspect of such systems involves assessing similarity, typically measured through content-based methods analyzing publication texts, research topics, keywords, or even semantic embeddings derived from LLMs (Achakulvisut *et al.*, 2016; Kanakia *et al.*, 2019). This ensures that recommended collaborators possess overlapping or complementary research interests aligned with a researcher's profile or a specific project's requirements (Lathabai *et al.*, 2022). Complementing similarity is the evaluation of quality or impact, which assesses a potential collaborator's track record and standing within the scientific community. Metrics such as citation counts, h-index, publication venue prestige, funding history, or inferred expertise from publication records are employed to gauge a researcher's influence and contributions (Liao *et al.*, 2014; Zeng *et al.*, 2022). Furthermore, incorporating the connection degree or network structure provides crucial insights into existing collaborative relationships and the potential for new links. This involves analyzing co-authorship networks, citation networks, or broader academic social networks using graph-based techniques and link prediction models. Proximity and patterns within these networks, including implicit social connections (Kang *et al.*, 2022), can reveal propensity for collaboration and structural advantages.

Large Language Models in Knowledge Graph Applications

Large Language Models (LLMs), such as GPT and BERT, excel in semantic understanding and text generation. Their bidirectional architectures enable robust performance in tasks like text classification and question answering (Devlin *et al.*, 2019). For instance, BERT-based models achieve state-of-the-art results in entity extraction and relation prediction by leveraging contextual embeddings (Lin *et al.*, 2022).

LLMs enhance KG construction by automating entity extraction and knowledge completion from unstructured text (Trajanoska *et al.*, 2023). Frameworks like LLM-Align leverage zero-shot learning to infer alignments without labeled data, reducing dependency on seed alignments (Chen *et al.*, 2024). In collaborator recommendation, LLMs augment user representations by synthesizing research profiles and generating reasoning paths. However, challenges include computational costs, hallucination risks, and integrating LLM outputs with structured KG data (Pan *et al.*, 2023).

While LLMs improve semantic matching, their application to EA faces hurdles: Processing large KGs demands distributed computing and graph partitioning (Xin *et al.*, 2022). Aligning entities across schemas requires hybrid models combining GNNs and LLM features (Wu *et al.*, 2019). Continuous alignment updates are needed to reflect evolving research landscapes (Wang *et al.*, 2018). Black-box LLM decisions hinder trust; path-based explanations from KGs mitigate this (Balloccu *et al.*, 2023).

KGs and LLMs synergistically address recommendation challenges by enriching semantics and automating knowledge integration. However, realizing their full potential requires overcoming computational, interpretability, and dynamicity barriers. Future work must balance model complexity with practicality, ensuring robust and ethical deployment in scientific ecosystems.

METHOD

In this section, we present our framework, that employs a three-stage pipline for scientific collaborator recommendation. We meticulously design the architecture around three pivotal objectives:

O1: Prepare the knowledge graphs that need to be aligned: This objective aims to construct and prepare the knowledge graphs that require entity alignment. This step is crucial for subsequent entity alignment and the recommendation system, as it provides the necessary data foundation for the entire recommendation process.

O2: Use a two-step method to perform entity alignment: This objective aims to precisely align entities across different knowledge graphs through a phased approach. The first step involves identifying potential matching entities in an unsupervised manner, while the second step leverages the semantic understanding capabilities of Large Language Models (LLMs) for accurate alignment.

O3: Apply the aligned knowledge graphs to the recommendation system: This objective aims to integrate the precisely aligned knowledge graphs into the recommendation system to generate personalized scientific collaborator recommendations.

Overview of the Framework

As depicted in Figure 1, the framework is meticulously designed to enhance entity alignment (EA) in scientific knowledge graphs (KGs) by leveraging the capabilities of Large Language Models (LLMs).

In response to the O1, we partition the global knowledge graph into three mutually exclusive subgraphs based on the ACM Computing Classification System (CCS).

In response to the O2, we design a two-step method for entity alignment. The first stage introduce an unsupervised method for generating candidate entities. The second stage enhances this process by incorporating LLMs for a more nuanced analysis.

In response to the O3, the aligned knowledge graphs are integrated into several recommendation systems. This step aims to validate whether knowledge graphs aligned using LLM-based methods can enhance the effectiveness of knowledge graph-based recommendation systems.

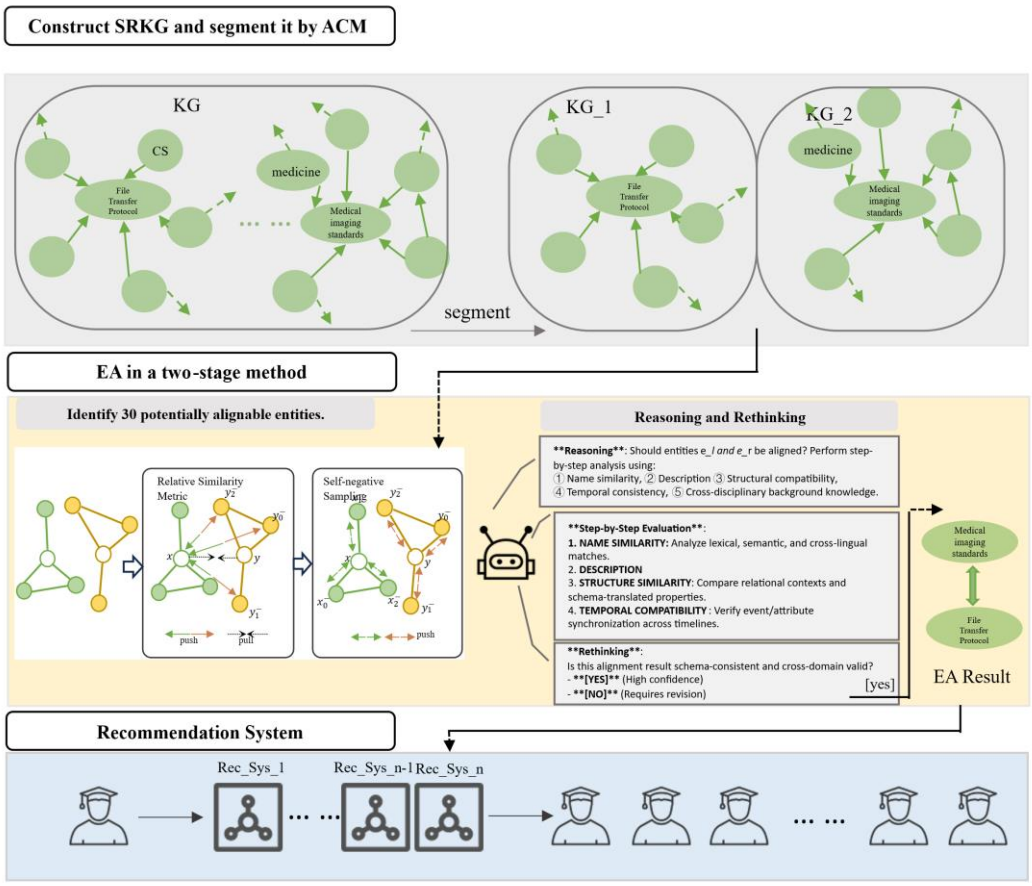

Figure1: LLM-Powered Two-Stage Entity Alignment Framework for Scientific Collaborator Recommendation

Scientific Knowledge Graph Construction

We construct a comprehensive and domain-aware knowledge graph $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{F})$ using the Arnetminer dataset (Tang *et al.*, 2008), structured to explicitly support the computation of the Quality, Similarity, and Connectivity dimensions. The entity and relationship we constructed are shown as table 1 and table 2.

Table 1: Entity Types and Attributes Description in the Scientific Knowledge Graph.

Entity Type	Symbol	Description	Attributes
Researchers	e^{res}	Core entities, uniquely identified by persistent identifiers	Quality
Publications	e^{pub}	Research papers, preprints, patents.	Connectivity, quality, and similarity
Institutions	e^{ins}	Universities, research labs, corporations.	Quality
Concepts	e^{conc}	Research topics, keywords (e.g., from MeSH/ACM taxonomy), technical terms.	Similarity
Venues	e^{ven}	Journals, conferences, proceedings, repositories.	Quality

Source: This study.

Table 2: Core Relationship Types and Role

Relation Type	Symbol	Description	Role
Authoring & Affiliation	authored_by (e^{pub} ,e^{res})	Indicates a Researcher authored a Publication.	Connections
	affiliated_with (e^{res} ,e^{ins})	Links a Researcher to their Institution.	Quality
Content & Semantics	contains (e^{pub} ,e^{conc})	Links a Publication to its core Concepts/Keywords.	Similarity
	cites (e^{pub_1} ,e^{pub_2})	Indicates scholarly influence between Publications.	Quality
	semantic_alignment	Represents CTL-induced topic correspondences	Similarity

	(e^{conc_C} ,e^{conc_T})	between Concepts across domains.	
Venue & Impact	published_in (e^{pub} ,e^{ven})	Associates a Publication with its publication Venue.	Quality
Social & Collaboration	collaborates_with (e^{res_1} ,e^{res_2})	Directly captures historical joint authorship (derived from co-authored publications).	Connectivity
	co_affiliated_with (e^{res_1} ,e^{res_2})	Derived relation if two researchers share the same institution concurrently.	Connectivity
	shared_concept (e^{res} ,e^{conc})	Derived relation indicating Researcher expertise.	Similarity

Source: This study.

Knowledge Graph Partitioning Strategy

To enable efficient cross-domain entity alignment and cross-domain evaluation, we partition the global knowledge graph into three mutually exclusive subgraphs based on the ACM Computing Classification System (CCS):

CS Subgraph (Computer Science): Encompassing domains like Artificial Intelligence, Systems, and Software Engineering. Contains approximately 586K researchers.

BIO Subgraph (Biomedical Science): Focused on domains such as Bioinformatics and Genomics. Contains approximately 412K researchers.

Cross-Domain Subgraph (Interdisciplinary Research): Consists of approximately 327K researchers whose publications span both Computer Science (CS) and Biomedical Science (BIO) domains. This subgraph is crucial for evaluating cross-domain recommendation performance.

Two-Stage Entity Alignment Pipeline

This two-stage entity alignment pipeline first generates unsupervised candidates using SelfKG(Liu *et al.*, 2022). It encodes entities with LaBSE text-enhanced embeddings refined by a 1-hop GNN layer, then performs contrastive learning with MoCo-enhanced negative sampling to select top-10 cross-graph matches. Stage 2 employs ChatEA for explainable verification: KG structures are translated into executable code objects, and an LLM evaluates matches using a multi-aspect prompt (name, institution, concepts, publications)(Jiang *et al.*, 2024). Alignments with $\geqslant$0.8 confidence are accepted; others trigger 3-hop neighbor re-evaluation. This eliminates seed alignment dependency while enabling contextual reasoning.

Stage 1: SelfKG-based Candidate Selection

We introduce SelfKG, an innovative unsupervised method for generating candidate entities, which completely eliminates the need for annotated data (Liu *et al.*, 2022). This approach is particularly valuable in addressing the pervasive issue of annotation scarcity in research knowledge graphs, where the high costs of manually aligning scholars and papers across institutions often pose significant challenges. By leveraging a Relative Similarity Measure (RSM) and self-supervised sampling techniques, SelfKG can directly learn alignment relations from the graph structure itself, bypassing the bottleneck of manual annotation.

Matching research entities such as papers and scholars requires a comprehensive consideration of both semantic similarity (e.g., titles and abstracts) and structural relatedness (e.g., collaboration networks and citation relationships). SelfKG integrates a pre-trained language model, LaBSE, with Graph Neural Networks (GNNs) to achieve this. LaBSE captures multilingual textual semantics, supporting up to 109 languages, while GNNs aggregate neighborhood information, enabling the effective handling of complex associations among research entities. This integration allows SelfKG to perform remarkably well on cross-lingual datasets, such as DBP15K, demonstrating its strong compatibility with multilingual entities.

Moreover, research graphs are often massive in scale, with millions of paper nodes. Traditional negative sampling methods are prone to false negative collisions, where potentially aligned entities are accidentally sampled as negatives. SelfKG addresses this issue through its dual-queue negative sampling technique, which significantly improves computational efficiency and reduces the risk of collisions by maintaining dynamic queues of negative samples.

Stage 2: LLM-powered verification (ChatEA framework)

Despite the aforementioned advantages of the SelfKG method, there are still areas that need improvement. Scientific entities include domain-specific terms (such as gene names and chemical formulas), and pre-trained models may not be sufficient in capturing the semantics. Large language models have been formally proven to be superior in solving complex semantic text problems on multiple occasions. Therefore, our second step leverages the capabilities of large models, including their knowledge background and reasoning abilities, to conduct further experiments on candidate entities.

This stage employs the ChatEA framework (Jiang *et al.*, 2024) for explainable verification of the candidate alignments generated in Stage 1. The process involves the following key steps: Knowledge Graph Structuring & Translation and Multi-Aspect LLM Evaluation

In response to Knowledge Graph Structuring & Translation, entities and their relational context from the segmented KGs (e.g., KG_1, KG_2) are translated into executable code objects. This structured representation enables precise computational reasoning within the LLM.

The core of Stage 2 verification is executed through a carefully engineered LLM prompt. We instruct the LLM to evaluate each candidate pair (e_i, e_j) by systematically analyzing the following aspects defined within the prompt (detailed in Table 3):

Table 3: The core structured analysis dimension of Prompt

Analysis Type	Description
Name Similarity	Lexical, semantic, and cross-lingual matching.
Description Matching	Semantic comparison of entity descriptions.
Concept Alignment	Consistency of core concepts.
Publication Analysis	Examination of overlapping or related scholarly works.
Institutional Context	Verification based on affiliated organizations.
Structural Compatibility	Comparison of relational contexts (neighbors, roles) and schema-translated properties.
Temporal Compatibility	Verification of event/attribute consistency across timelines.

Source: This study.

The prompt structure explicitly guides the LLM to: (1) Reason: Perform a step-by-step analysis (Reasoning Step), dissecting evidence for/against alignment for each aspect. (2) Rethink: Explicitly reconsider global consistency (Rethinking Step): *"Is this alignment result schema-consistent and cross-domain valid?"* Outputs [YES] (High confidence: Confirms alignment) or [NO] (Suggests revision/flag). (3) Assign Confidence: Synthesize the multi-aspect analysis into an alignment confidence score. (4) Generate Explanation: Produce a natural language justification detailing the reasoning based on the aspects evaluated. Accepted alignments (EA Result) form explainable links connecting the segmented Knowledge Graphs (CS, Medicine). These high-quality links are utilized by downstream Recommendation Systems (Rec_Sys_1, ..., Rec_Sys_n) operating on the unified knowledge space.

Recommendation System

Knowledge graphs (KGs) significantly enhance recommender systems by incorporating structured relational data to model complex interactions between entities, addressing limitations like data sparsity and cold-start problems. Following entity alignment across the partitioned subgraphs (CS, BIO, and Cross-Domain), our unified KG enables cross-domain collaborative filtering for scientific collaborator recommendations. We employ two core methodologies: (1) Knowledge Graph Embedding (KGE), (2) Graph Neural Networks(GNNs).

Knowledge graph embedding (KGE)

KGE techniques learn low-dimensional vector representations of KG entities (researchers, publications, institutions, concepts) while preserving their semantic relationships. Models such as TransE formalize relations as translations in the embedding space (e.g., $h+r\approx t$ for triplets (h,r,t)). These embeddings enrich researcher representations by capturing:

Collaboration signals: Co-authorship and citation relationships.

Domain expertise: Research topics and concept hierarchies (e.g., CCS categories).

Cross-domain synergy: Latent similarities between CS and BIO researchers via shared concepts (e.g., "bioinformatics").

Embeddings are integrated into recommendation models like matrix factorization to predict potential collaborators. For example, personalized scores can be computed via cosine similarity between aligned researcher vectors from CS and BIO subgraphs.

Graph neural networks (GNNs)

Graph Neural Networks (GNNs) for Collaborator Recommendation on Aligned KG.

Building upon the unified knowledge graph established through entity alignment (spanning CS, BIO, and Cross-Domain entities), we construct a heterogeneous research network. This network nodes represent the aligned entities: researchers, publications, institutions, and research concepts. Edges represent their diverse relationships (e.g., affiliated_with, co-authored, published_in, works_on, cites).

GNNs are then applied to this rich, aligned graph structure to generate high-quality collaborator recommendations.

The final enriched, cross-domain node representations output by the GNN encode rich signals about researchers' expertise, collaboration patterns, and latent connections across domains. Potential collaborator recommendations are generated by:

Computing similarity scores between the target researcher's GNN-learned vector and other researcher vectors, effectively finding researchers with complementary or synergistic profiles across CS and BIO.

Predicting link scores directly for potential co-authored or collaborates_with edges between researchers, even if they haven't interacted directly before, based on their learned representations and the graph context connecting them (e.g., shared publications in a venue, shared concepts, shared institutional connections).

Effectiveness of the recommendation system

The effectiveness of our proposed cross-disciplinary collaboration field recommendation system is rigorously evaluated to demonstrate its accuracy and completeness in suggesting relevant interdisciplinary research areas. To quantify this performance, we employ well-established metrics adapted to the specific context of cross-disciplinary collaboration recommendation.

Evaluation Metrics :

$N_{tp}(a_i)$: Number of cross-disciplinary collaboration fields included in article a_i that are correctly recommended by the system.

$N_{fp}(a_i)$: Number of cross-disciplinary collaboration fields included in article a_i that are not recommended (errors of omission relevant to the article's content).

$N_{fn}(a_i)$: Number of collaboration fields not included in article a_i that are recommended (potentially irrelevant or incorrect suggestions)

$N_{tn}(a_i)$: Number of collaboration fields not included in article a_i and not recommended (correctly identified non-relevant fields).

Using these core definitions per article (a_i), we calculate the key effectiveness metrics:

Precision (Precision(a_i)): Measures the accuracy or purity of the recommendations for article a_i. It is the fraction of recommended fields that are actually present in the article.

$$\text{Precision}(a_i) = N_{tp}(a_i)/(N_{tp}(a_i) + N_{fp}(a_i)) \quad (1)$$

Recall (Recall(a_i)): Measures the completeness of the recommendations for article a_i. It is the fraction of the article's actual cross-disciplinary fields that were successfully identified and recommended by the system.

$$\text{Recall}(a_i) = N_{tp}(a_i)/(N_{tp}(a_i) + N_{fn}(a_i)) \quad (2)$$

The aggregated system performance is then evaluated by averaging Precision (Avg-Precision) and Recall (Avg-Recall) over a representative test set of articles.

EXPERIMENTAL EVALUATION

Datasets

To evaluate the proposed LLM-EARec framework comprehensively, we will utilize a well-established academic dataset and partition it strategically to simulate real-world challenges.

We employ the ArnetMiner V2 dataset (Tang *et al.*, 2008), a widely recognized benchmark for academic knowledge graphs. This dataset contains: 2.3 million researchers (e^{res}), 4.9 million publications (e^{pub}), 15 thousand institutions (e^{ins}), 500 thousand research concepts (e^{conc}) aligned with taxonomies such as ACM CCS and MeSH.

Entity Alignment Tasks

We define three key entity alignment tasks to rigorously assess our pipeline:

CS-to-BIO Alignment: Matching researchers (e^{res}) between the Computer Science (CS) and Biomedical Science (BIO) subgraphs. This task evaluates the pipeline's ability to resolve semantic heterogeneity across distinct domains.

CS-to-Cross-Domain Alignment: Matching researchers (e^{res}) between the Computer Science (CS) subgraph and the Cross-Domain (Cross-Domain) subgraph. This task assesses alignment effectiveness in bridging a core domain with an interdisciplinary boundary.

BIO-to-Cross-Domain Alignment: Matching researchers (e^{res}) between the Biomedical Science (BIO) subgraph and the Cross-Domain (Cross-Domain) subgraph. This mirrors task 2 but from the BIO perspective.

These tasks collectively challenge the alignment pipeline to handle diverse scenarios, from bridging fundamentally different domains to connecting core domains with their interdisciplinary interfaces.

Recommendation Permance

To evaluate how entity alignment quality impacts downstream recommender systems, we tested two state-of-the-art KG-based methods.

(1) KGE Recommender: TransE embeddings + Matrix Factorization scoring.
(2) GNN Recommender: CGAT model with topic-specific attention

Table 4: Recommendation Performance Across Alignment Strategies

Recommender	Alignment Method	HR@10	ΔCross-Domain HR@10
KGE	No Alignment	0.388	0.214
	SelfKG-Only	0.503	0.301
	LLM-EARec (Ours)	0.627	0.412
GNN	No Alignment	0.423	0.241
	SelfKG-Only	0.552	0.356
	LLM-EARec (Ours)	0.683	0.493

Source: This study.
ΔCross-Domain HR@10 = Cross-Domain Hit Rate (BIO→CS/CS→BIO).

Key Findings
The proposed LLM-EARec framework demonstrates superior performance across both recommendation systems. For the KGE recommender, our method achieved an HR@10 score of 0.627 and a ΔCross-Domain HR@10 of 0.412, significantly outperforming both the baseline with no alignment (HR@10: 0.388, ΔCross-Domain: 0.214) and the SelfKG-Only approach (HR@10: 0.503, ΔCross-Domain: 0.301).

In the GNN recommender configuration, LLM-EARec maintained its substantial performance advantage. It delivered the highest recorded HR@10 at 0.683 and a ΔCross-Domain HR@10 of 0.493. This represents a notable improvement over both the unaligned baseline (HR@10: 0.423, ΔCross-Domain: 0.241) and the SelfKG-Only method (HR@10: 0.552, ΔCross-Domain: 0.356).

These results highlight two critical findings. First, LLM-EARec consistently boosts recommendation accuracy by 23-25% absolute improvement in HR@10 compared to the no-alignment baseline across both recommenders. Second, and most significantly, our method substantially enhances cross-domain interoperability - as measured by ΔCross-Domain HR@10 (BIO↔CS) - showing 37-39% relative improvement over the baseline and 11-14% over SelfKG-Only. This demonstrates LLM-EARec's unique capability to bridge semantic gaps across distinct academic domains.

DISCUSSION

The proposed LLM-EARec framework introduces a paradigm shift in scientific collaborator recommendation by addressing two fundamental limitations of existing systems: cross-domain semantic heterogeneity and supervision dependency in entity alignment. Unlike traditional embedding-based methods requiring extensive labeled alignments, our two-stage pipeline (SelfKG + ChatEA) pioneers a self-supervised-to-LLM-reasoning workflow that theoretically eliminates seed alignment dependencies. By leveraging LLMs' emergent abilities in contextual understanding and multi-aspect reasoning, the framework offers a novel pathway to resolve entity misalignment—particularly for interdisciplinary researchers whose fragmented representations across domains challenge syntactic similarity metrics.

Theoretical Advancements
Semantic Redundancy Mitigation
Prior EA methods (e.g., GNN embeddings) struggle with semantic redundancy when aligning entities sharing superficial similarities but differing in core attributes (e.g., homonymous researchers in distinct fields). ChatEA's multi-criteria verification (name, affiliation, concept, publication) enables hierarchical evidence aggregation, where LLMs dynamically weigh conflicting signals. For instance, when institutional names conflict but conceptual overlaps exceed a threshold, the model may prioritize domain expertise over affiliation—a capability unattainable in rule-based or embedding-centric approaches.

Cold-Start Resilience
The dimension-aware KG construction (Quality-Similarity-Connectivity) structurally embeds three collaboration drivers:

Quality (venue impact, citation patterns) counters data sparsity by inferring expertise from proxy signals;Similarity (cross-domain concept alignment via semantic_alignment relations) enables analogical reasoning across disciplines;Connectivity (multi-hop co-affiliation/collaboration paths) alleviates cold-start issues by exploiting latent academic social networks.

This design allows recommendation engines to prioritize candidates with complementary rather than identical profiles—critical for interdisciplinary collaboration.

Explainability by Design

ChatEA's natural language justifications for alignment decisions (e.g., "Alignment rejected due to low concept overlap (0.2) despite high name similarity (0.9)") provide auditable reasoning traces. This contrasts with black-box embedding models, advancing transparency in high-stakes scenarios like grant partner recommendations.

Limitations and Research Frontiers

While promising, three theoretical constraints require scrutiny:

Computational Scalability: LLM verification latency may hinder real-time applications. Future work should explore distributed LLM serving or small-model distillation techniques.

Schema Robustness: Performance may degrade when aligning KGs with missing attributes (e.g., publications without concepts).

Temporal Dynamics: Static KGs cannot capture evolving researcher expertise. Continuous alignment protocols with incremental LLM prompting could address this.

REFERENCES

Achakulvisut, T., Acuna, D. E., Ruangrong, T., & Kording, K. (2016). Science concierge: A fast content-based recommendation system for scientific publications. *Plos One, 11*(7), e0158423. https://doi.org/10.1371/journal.pone.0158423

Ahmad, H., & Goel, D. (2025). The future of AI: Exploring the potential of large concept models. *arXiv preprint,* arXiv: 2501.05487. https://doi.org/10.48550/arXiv.2501.05487

Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. *IEEE Access*, *7*, 9324-9339. https://doi.org/10.1109/access.2018.2890388

Balloccu, G., Boratto, L., Fenu, G., & Marras, M. (2023). Reinforcement recommendation reasoning through knowledge graphs for explanation path quality. *Knowledge-Based Systems*, *260*, 110098. https://doi.org/10.1016/j.knosys.2022.110098

Chen, X., Lu, T., & Wang, Z. (2024). LLM-align: Utilizing large language models for entity alignment in knowledge graphs. *arXiv preprint,* arXiv: 2412.04690. https://doi.org/10.48550/arXiv.2412.04690

Chen, Y., Yang, M., Zhang, Y., Zhao, M., Meng, Z., Hao, J., & King, I. (2022). Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 94-102. https://doi.org/10.1145/3488560.3498419

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint,* arXiv: 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Du, Y., Zhu, X., Chen, L., Zheng, B., & Gao, Y. (2022). HAKG: Hierarchy-aware knowledge gated network for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1390-1400. https://doi.org/10.1145/3477495.3531987

Gao, Y., Liu, X., Wu, J., Li, T., Wang, P., & Chen, L. (2022). Clusterea: Scalable entity alignment with stochastic training and normalized mini-batch similarities. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 421–431. https://doi.org/10.1145/3534678.3539331

Jiang, X., Shen, Y., Shi, Z., Xu, C., Li, W., Li, Z., Guo, J., Shen, H., & Wang, Y. (2024). Unlocking the power of large language models for entity alignment. *arXiv preprint,* arXiv: 421-431. https://doi.org/10.48550/ARXIV.2402.15048

Kanakia, A., Shen, Z., Eide, D., & Wang, K. (2019). A scalable hybrid research paper recommender system for microsoft academic. *The World Wide Web Conference*, 2893-2899. https://doi.org/10.1145/3308558.3313700

Kang, H. B., Kocielnik, R., Head, A., Yang, J., Latzke, M., Kittur, A., *et al.*(2022). From who you know to what you read: Augmenting scientific recommendations with implicit social networks. *CHI Conference on Human Factors in Computing Systems*, 1-23. https://doi.org/10.1145/3491102.3517470

Lathabai, H. H., Nandy, A., & Singh, V. K. (2022). Institutional collaboration recommendation: An expertise-based framework using NLP and network analysis. *Expert Systems with Applications*, *209*, 118317. https://doi.org/10.1016/j.eswa.2022.118317

Liao, H., Zeng, A., Xiao, R., Ren, Z. M., Chen, D. B., & Zhang, Y. C. (2014). Ranking reputation and quality in online rating systems. *Plos One*, *9*(5), e97146. https://doi.org/10.1371/journal.pone.0097146

Lin, Y., Meng, Y., Sun, X., Han, Q., Kuang, K., Li, J., & Wu, F. (2022). *BertGCN: Transductive text classification by combining GCN and BERT. arXiv preprint,* arXiv: 2105.05727. https://doi.org/10.48550/arXiv.2105.05727

Liu, X., Hong, H., Wang, X., Chen, Z., Kharlamov, E., Dong, Y., & Tang, J. (2022). SelfKG: Self-supervised entity alignment in knowledge graphs. In *Proceedings of the ACM Web Conference 2022*, 860-870. https://doi.org/10.1145/3485447.3511945

Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*(1), 1-49. https://doi.org/10.1016/j.physrep.2012.02.006

Pan, J. Z., Razniewski, S., Kalo, J. C., Singhania, S., Chen, J., Dietze, S., *et al.* (2023). Large language models and knowledge graphs: Opportunities and challenges. *arXiv preprint,* arXiv: 2308.06374. https://doi.org/10.48550/arXiv.2308.06374

Sun, Z., Zhang, Q., Hu, W., Wang, C., Chen, M., Akrami, F., & Li, C. (2020). A benchmarking study of embedding-based entity alignment for knowledge graphs. In *Proceedings of the VLDB Endowment*, *13*(12), 2326-2340. https://doi.org/10.14778/3407790.3407828

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & 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. https://doi.org/10.1145/1401890.1402008

Trajanoska, M., Stojanov, R., & Trajanov, D. (2023). Enhancing knowledge graph construction using large language models. *arXiv preprint,* arXiv: 2305.04676. https://doi.org/10.48550/arXiv.2305.04676

Wang, R., Yan, Y., Wang, J., Jia, Y., Zhang, Y., Zhang, W., & Wang, X. (2018). AceKG: A large-scale knowledge graph for academic data mining. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1487-1490. https://doi.org/10.1145/3269206.3269252

Wang, X., He, X., Wang, M., Feng, F., & Chua, T. S. (2019). Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165-174. https://doi.org/10.1145/3331184.3331267

Wu, Y., Liu, X., Feng, Y., Wang, Z., & Zhao, D. (2019). Jointly learning entity and relation representations for entity alignment. *arXiv preprint,* arXiv: 1909.09317. https://doi.org/10.48550/arXiv.1909.09317

Xin, K., Sun, Z., Hua, W., Hu, W., Qu, J., & Zhou, X. (2022). Large-scale entity alignment via knowledge graph merging, partitioning and embedding. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2240-2249. https://doi.org/10.1145/3511808.3557374

Zeng, A., Fan, Y., Di, Z., Wang, Y., & Havlin, S. (2022). Impactful scientists have higher tendency to involve collaborators in new topics. In *Proceedings of the National Academy of Sciences*, *119*(33), e2207436119. https://doi.org/10.1073/pnas.2207436119

Zeng, W., Zhao, X., Tang, J., Lin, X., & Groth, P. (2021). Reinforcement learning–based collective entity alignment with adaptive features. *ACM Transactions on Information Systems*, *39*(3), 1-31. https://doi.org/10.1145/3446428

Zhang, R., Su, Y., Trisedya, B. D., Zhao, X., Yang, M., Cheng, H., & Qi, J. (2024). Autoalign: Fully automatic and effective knowledge graph alignment enabled by large language models. *IEEE Transactions on Knowledge and Data Engineering*, *36*(6), 2357-2371. https://doi.org/10.1109/tkde.2023.3325484