From Speech to Semantics: Enabling Conversational Access to Scholarly Knowledge Graph

Umair Ahmed[1,*], Andrea Polini[1] and Nicolas Ferranti[2]

[1]University of Camerino (UNICAM), Via Andrea D'Accorso, 16, 62032 Camerino MC
[2]WU (Vienna University of Economics and Business) , Welthandelsplatz 1, 1020, Vienna

Abstract

Recent advancements in the representation of knowledge via knowledge graphs have paved an intuitive way to build scholarly knowledge for users and for artificial agents. Despite the expressiveness of knowledge graphs, accessing these knowledge graphs requires proficiency in a query language such as SPARQL, presenting a barrier for a multitude of users. In this study, we present a hybrid, end-to-end framework that (i) interprets user questions expressed in natural language, (ii) classifies each query into one of four target categories (conferences, authors, organizations, or papers) using a fine-tuned RoBERTa-Large model, (iii) synthesizes candidate SPARQL queries via a large language model (GPT-4o-mini) augmented with few-shot examples, and (iv) refines the raw query results by reranking either the SPARQL output or, when necessary, fallback candidate items retrieved through vector-space embeddings (indexed with FAISS). On a testbed of 92 manually crafted "gold-standard" SPARQL queries, our automated pipeline achieved over 96% overlap with expert results (≥ 70% overlap in 89/92 cases), with perfect consistency on conference, author, and organization queries and 90% coverage on paper queries given the semantic nature of queries. Moreover, our query-type classifier achieved 99% accuracy, demonstrating the reliability of schema selection. These results indicate that combining LLM-driven query synthesis with embeddings-based reranking delivers a robust, user-centric interface to scholarly knowledge graphs, enabling complex information retrieval without SPARQL expertise.

Keywords

Knowledge Graphs, SPARQL, Natural Language Queries, Automated Query Generation, Semantic Search, Large LanguageL Models, GPT

1. Introduction

The contemporary development and widespread adoption of knowledge graphs have transformed the way scholarly knowledge is represented and retrieved. These knowledge graphs, in the context of scholarly knowledge, offer powerful tools for organizing and interpreting the vast and growing body of scholarly content [1, 2]. Initiatives such as DBpedia [3] and Wikidata [4], along with more domain-specific efforts like OpenResearch [5] and ScholarlyData [6], underscore the potential of knowledge graphs to harmonize diverse metadata sources, enable sophisticated analytical queries, and foster deeper insights into the structure and dynamics of academic research.

Despite their evident potential, these scholarly KGs primarily rely on SPARQL, a standardized query language for RDF data, which is inherently complex and poses a significant barrier for widespread adoption [7, 8]. Constructing effective SPARQL queries demands extensive knowledge of RDF schema structure, semantic data models, and query syntax. These skills are rarely possessed by non-expert users such as researchers, librarians, or policy-makers who often represent the primary beneficiaries of such graphs [9].

In an effort to bridge this accessibility gap, various methods have been proposed to translate natural language (NL) queries directly into structured query formats [10, 11]. Earlier approaches relied heavily on manually crafted linguistic rules or structured templates, providing limited flexibility and requiring

RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan
*Corresponding author.
umair.ahmed@unicam.it (U. Ahmed); andrea.polini@unicam.it (A. Polini); nicolas.ferranti@wu.ac.at (N. Ferranti)
0000-0003-2260-2777 (U. Ahmed); 0000-0002-2840-7561 (A. Polini); 0000-0002-5574-1987 (N. Ferranti)

extensive maintenance [12]. Recent advancements leveraging machine learning, particularly deep learning and transformer-based language models, have opened promising avenues for more adaptive, scalable query-generation methods [13, 14]. For instance, models like SPBERT and modern text-to-SPARQL pipelines have significantly improved performance by learning syntactic and semantic representations of SPARQL queries and domain knowledge jointly [12, 14]. Nonetheless, purely automated methods still face challenges in consistently producing accurate and syntactically correct queries, particularly for complex or nuanced user questions [11].

Furthermore, complementary research employing embedding-based retrieval techniques has demonstrated significant success in handling semantic search and retrieval problems within knowledge bases [15, 16]. Embeddings offer powerful mechanisms to approximate semantic similarity, enabling rapid approximate-nearest-neighbor searches using high-dimensional vector representations, thus circumventing the strict matching constraints of purely symbolic query methods [17]. Tools such as FAISS exemplify these advances by efficiently managing large-scale embedding indexes and enabling hybrid or fallback retrieval strategies when symbolic queries fail [18].

Motivated by these recent developments, this paper introduces a novel hybrid approach designed explicitly to simplify access to scholarly KGs for non-expert users. Our method synergistically combines multiple contemporary techniques:

- Natural-language classification leveraging RoBERTa-Large [13], fine-tuned to classify queries into four target categories: conferences, authors, organizations, and papers.
- Large Language Model (LLM)-based SPARQL generation, using GPT-4o-mini [19], augmented with type-specific few-shot examples, significantly improving the semantic precision of the generated queries.
- Embedding-based reranking or fallback, utilizing vector-space embeddings (indexed via FAISS) [18] to refine or directly retrieve results when symbolic SPARQL queries fail to return satisfactory outputs.

We empirically demonstrate the effectiveness of our hybrid system through rigorous evaluations against a set of 92 manually curated ("gold-standard") SPARQL queries. Our pipeline achieves over 96% overlap with expert-generated results, surpassing the 70% overlap threshold consistently, and demonstrating perfect consistency in queries targeting conferences, authors, and organizations, alongside robust performance on queries targeting research papers.

The subsequent sections are organized as follows: Section 2 presents related literature on Natural Language-to-SPARQL translation, embedding-based retrieval, and hybrid approaches. Section 3 dives into the background associated with this study, section 4 details our hybrid system architecture. Section 5 outlines the experimental setup, datasets, and evaluation metrics, followed by a detailed discussion of results. Section 6 discusses limitations and future extensions. Section 7 concludes the study with an intuitive summarization of findings.

2. Related Works

2.1. Natural-Language Interfaces to Knowledge Graphs

Early approaches to natural-language interfaces for querying structured knowledge graphs primarily relied on rule-based grammars, templates, and manually engineered mappings between language patterns and ontological constructs [1, 2]. Systems such as Sparklis [20] facilitated interactive query formulation, allowing users to iteratively refine queries through structured templates. Despite their effectiveness, these systems required extensive manual maintenance, making them difficult to scale across large, evolving knowledge bases.

Recent advancements have shifted toward leveraging large language models (LLMs) to enhance natural language interfaces for knowledge graphs. For instance, the EDGE system integrates LLMs to facilitate natural language interactions with educational knowledge graphs, improving user experience

in data exploration [21]. Additionally, following framework utilizes Retrieval-Augmented Generation (RAG) to enhance SPARQL query generation, reducing semantic errors and improving robustness [22].

2.2. Large Language Models (LLMs) for Query Generation

The advent of pre-trained LLMs such as GPT-3 and GPT-4 has significantly reshaped the landscape of semantic parsing and query synthesis. These models can translate natural-language queries into formal queries like SPARQL with minimal annotated examples [23, 24]. Few-shot prompting techniques have been shown to generalize query generation across diverse knowledge schemas, reducing the need for extensive domain-specific annotations [24].

To address challenges like hallucinations and out-of-distribution errors in LLM-generated queries, Sharma et al. introduced PGMR (Post-Generation Memory Retrieval), a modular framework that separates query structure generation from knowledge retrieval. This approach significantly reduces the incidence of hallucinated URIs in SPARQL query generation [25]. Frameworks like FRASE leverage frame-semantic structured representations to improve generalization in SPARQL query generation, particularly in scenarios involving naturally phrased, template-free questions [26]. Moreover, SPARKLE integrates knowledge graph structures directly into the decoding process of LLMs, reducing the occurrence of inoperative query generations [27].

2.3. Embedding-Based Retrieval and Reranking

Embedding-based retrieval methods provide an effective complementary mechanism for structured queries by representing entities and their relationships within vector spaces [11, 12]. These techniques allow rapid approximate-nearest-neighbor search, exemplified by frameworks such as Facebook's FAISS library [28]. Embeddings have proven highly effective for entity retrieval and semantic search tasks [13], enabling efficient retrieval of relevant entities even in the absence of explicit symbolic matches.

Hybrid approaches combining symbolic and neural methods have recently emerged, demonstrating superior performance by reranking symbolic query results using embeddings [29, 30]. Graph-based reranking methods have been explored to enhance the selection of optimal query graphs in Knowledge Base Question Answering systems. Jia and Chen proposed a two-step approach involving initial ranking followed by reranking of query graphs, leading to improved retrieval accuracy [31]. The KGR3 framework integrates retrieval, reasoning, and reranking components to enhance knowledge graph completion tasks, using context-enriched modules to improve prediction accuracy [32]. Additionally, ReranKGC introduces a cooperative retrieve-and-rerank framework for multi-hop knowledge graph completion, improving accuracy and efficiency [33].

2.4. Positioning Our Work

Our research advances the state-of-the-art by integrating several successful strategies into a unified pipeline tailored specifically for scholarly knowledge graphs. Specifically, our hybrid architecture synthesizes:

- Natural-language classification using RoBERTa-Large, achieving near-perfect query-type identification accuracy.
- LLM-driven (GPT-4o-mini) SPARQL generation augmented by few-shot examples, significantly reducing reliance on manually annotated datasets.
- Embedding-based refinement (FAISS) employed both as a fallback mechanism and a reranking technique to enhance retrieval accuracy.
- LLM-based summarization of the results as according to the user query

Our work represents the systematic effort to integrate LLM-based SPARQL generation with embedding-based reranking specifically targeted at scholarly knowledge graphs, achieving expert-level retrieval quality without extensive manual annotations or user SPARQL expertise.

3. Background

3.1. Scholarly Data Management: An Evolving Landscape

The past decade has witnessed substantial evolution in scholarly data management, marked by a shift from isolated bibliographic databases to richly structured, interconnected knowledge graphs (KGs). Unlike traditional relational databases, scholarly knowledge graphs utilize semantic web standards such as RDF (Resource Description Framework) and OWL (Web Ontology Language) to integrate heterogeneous academic metadata seamlessly [34, 35, 36]. These structured graphs empower researchers, libraries, and institutions to discover relationships among authors, publications, conferences, and institutions previously concealed in disparate databases [6]. By enabling queries that traverse multiple dimensions, such as author affiliations, citation patterns, and co-authorship networks, scholarly knowledge graphs have profoundly transformed the capabilities of bibliometric analysis and academic discovery [35].

Yet, the complexity inherent in these rich semantic structures introduces a notable barrier: traditional querying mechanisms, particularly SPARQL (the standardized language for querying RDF-based knowledge graphs), are notoriously difficult to master without extensive training [7]. This limitation confines scholarly graph utilization largely to data engineers or semantic web specialists, excluding the broader academic community who stand to benefit most.

3.2. Towards Democratizing Access to Scholarly Knowledge

The growing recognition of this limitation has prompted efforts toward developing intuitive, user-friendly query interfaces. Central to this vision is the concept of natural language querying, a paradigm that seeks to leverage users' inherent linguistic capabilities to interact with complex data structures without explicit technical knowledge [37]. However, natural language interfaces introduce their own set of challenges: ambiguity, variability in linguistic expression, and difficulty translating informal questions into precise symbolic queries [2].

Simultaneously, advancements in deep learning and natural language processing have opened pathways toward bridging this gap. Particularly transformative are pre-trained large language models (LLMs), capable of generating structured queries such as SPARQL from plain-language user prompts, thereby drastically simplifying user interactions with semantic web resources [38]. Nonetheless, the effective application of LLMs to scholarly data introduces practical concerns, especially regarding accuracy, consistency, and the need for validation mechanisms ensuring generated queries yield meaningful, high-quality results [38].

3.3. The Complementary Role of Embeddings

A parallel and complementary technological advancement has emerged through embedding-based retrieval methods. Embeddings provide semantic vector-space representations of entities (authors, papers, conferences, organizations), capturing subtle semantic relationships beyond direct symbolic matches [35]. By transforming entities into continuous vector representations, embeddings facilitate rapid approximate nearest-neighbor searches, thus enabling efficient retrieval of semantically similar results without rigid symbolic constraints [28]. This technology, operationalized through libraries like FAISS, has become essential in scenarios demanding real-time semantic retrieval at scale [28].

However, embedding-based approaches alone cannot fully leverage structured relationships explicitly encoded in scholarly KGs. Thus, an ideal query mechanism combines symbolic querying (SPARQL) and embedding-based semantic retrieval into a unified, hybrid framework that maximizes the strengths of both paradigms: symbolic precision with embedding-driven semantic flexibility.

3.4. Research Gap and Motivation

Existing methods in scholarly KG querying predominantly focus either on precise symbolic querying (via SPARQL) or purely semantic embedding-based retrieval, with relatively limited exploration of

how best to integrate these paradigms effectively in scholarly contexts. The complexity, diversity, and specialized nature of scholarly metadata call for a bespoke hybrid solution specifically designed to address domain-specific queries reliably and intuitively [36].

The necessity for a robust hybrid approach is driven by practical considerations: researchers require both accurate and comprehensive query results that align closely with their semantic intent. This demands a mechanism that dynamically utilizes the strengths of symbolic SPARQL querying for structured precision, complemented by embeddings to maintain semantic coherence and manage cases where symbolic retrieval falls short or returns inadequate results [37].

3.5. Our Approach: Bridging Symbolic and Semantic Retrieval

Responding to this pressing need, our work presents a tailored hybrid architecture explicitly designed for scholarly knowledge graphs. We incorporate advanced machine learning methods, specifically a RoBERTa-Large classifier for robust query categorization, GPT-4o-mini for reliable generation of SPARQL queries from natural language inputs, and FAISS-based embeddings for semantic refinement and fallback retrieval. This synthesis uniquely positions our work, offering a comprehensive solution that addresses existing gaps by seamlessly integrating symbolic query precision and embedding-based semantic flexibility in scholarly knowledge graphs.

The subsequent sections of this paper detail this integration and empirically validate its effectiveness, demonstrating that such a hybrid approach significantly enhances the accessibility and usability of scholarly knowledge graphs for non-technical users.

4. Methodology

This study proposes a comprehensive methodology to introduce a natural language interface for querying the scholarly knowledge graph. It employs natural language processing (NLP) techniques, large language models (LLM), semantic embeddings, and symbolic query execution to constitute into a GraphRAG system underlying the natural language query interface.

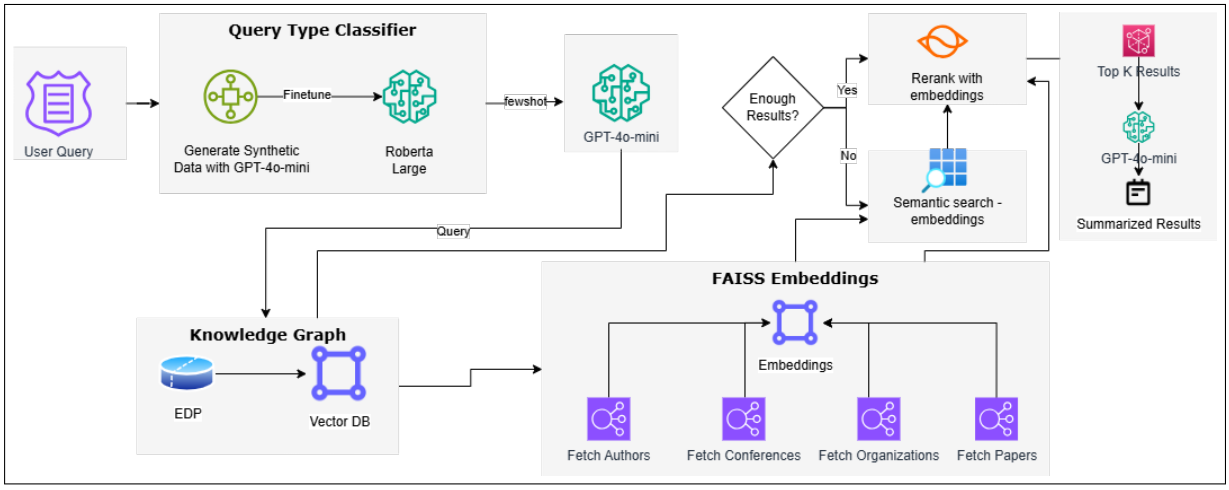

Figure 1: Diagram illustrating the key components of the semantic search methodology

4.1. Knowledge Graph Preparation and Embeddings Generation

4.1.1. Dataset Acquisition and Graph Setup

We initiated the study by acquiring the ScholarlyData knowledge graph, a prominent RDF-based dataset encompassing structured metadata on conferences, research papers, authors, and organizations. The

ScholarlyData graph was imported into a semantic graph database environment (e.g., Apache Jena, GraphDB), enabling efficient storage, indexing, and SPARQL querying.

4.1.2. Semantic Embeddings Creation

To support embedding-based retrieval, we computed separate semantic embeddings for each of the four primary entity categories within the scholarly knowledge graph:

- Conferences
- Authors
- Organizations
- Papers

Embeddings were generated using a pre-trained transformer-based language model, specifically Sentence-BERT, to convert textual descriptions (e.g., titles, abstracts, names, organizational descriptions) into continuous, dense vector representations capturing semantic meaning. These embeddings were indexed using Facebook's FAISS library, allowing efficient approximate nearest-neighbor retrieval at query time.

4.2. Query Type Classification using RoBERTa-Large

To determine the category of entities targeted by user-submitted natural-language queries, we finetuned a supervised classification method based on RoBERTa-Large, a robust transformer-based language model optimized for text classification tasks.

4.2.1. Automated Federated Dataset Construction

Training data for query-type classification was auto generated using GPT-4. We created a federated training/evaluation dataset containing diverse natural-language query examples synthetically paired with their corresponding entity types (conference, author, organization, or paper). This automated process ensured extensive linguistic coverage without manual annotation efforts. Following is the distribution of the dataset:

Model	Training Samples	Test Samples
RoBERTa-large (fine-tuned)	772	100

Table 1
Dataset size for fine-tuning RoBERTa-large on query type classification.

4.2.2. Fine-Tuning RoBERTa-Large Classifier

The RoBERTa-Large model was fine-tuned on the generated dataset using a supervised training approach, employing standard hyperparameters (learning rate: 2×10^{-5}, batch size: 16, maximum token length: 128). RoBERTa-Large was chosen for its strong performance on a wide range of natural language understanding tasks, particularly in scenarios requiring nuanced contextual representations, making it well-suited for accurately interpreting diverse scholarly queries. This resulted in a high-performance model capable of reliably classifying user queries into the defined entity types, facilitating targeted downstream retrieval.

4.3. SPARQL Query Generation via GPT-4o-mini

Following query-type identification, the natural-language queries were transformed into executable SPARQL queries using the GPT-4o-mini language model.

4.3.1. Few-Shot Prompt Engineering

To guide GPT-4o-mini in generating accurate SPARQL queries, we implemented a few-shot learning approach tailored for each query type. Specifically, we provided carefully curated examples of natural-language queries paired with correctly structured SPARQL queries corresponding to scholarly KG schemas. These prompts ensured GPT-4o-mini could effectively generalize from limited examples, generating syntactically valid and semantically accurate SPARQL queries.

4.3.2. Query Execution

The SPARQL queries generated by GPT-4o-mini were directly executed against the semantic knowledge graph database. The results retrieved from these symbolic queries formed the primary candidate set for answering user queries.

4.4. Hybrid Retrieval and Reranking Approach

To ensure semantic relevance and retrieval quality, we incorporated a hybrid retrieval and reranking strategy that dynamically integrated symbolic SPARQL results with embedding-based semantic similarity computations.

4.4.1. Embedding-based Result Reranking

When SPARQL queries returned a sufficient number of results (≥ 5), the retrieved entities were semantically reranked using cosine similarity scores calculated between their embeddings and the embedding of the user's natural-language query. This step enhanced semantic coherence and reduced ambiguity inherent in purely symbolic query results.

4.4.2. Embedding-only Retrieval (Fallback Mechanism)

If SPARQL queries returned insufficient results (< 5) or no results, indicating either semantic mismatches or query complexity, our system defaulted directly to embedding-based retrieval. Specifically, we queried the FAISS-indexed embeddings of the identified entity type using the original natural-language query, retrieving the top-k nearest entities based on semantic similarity scores.

4.4.3. GraphRAG-based Answer Generation

Following the reranking and retrieval of the final results, the conclusive step of the GraphRAG was triggered. The top-k embedding results and the original natural-language query were provided back into GPT-4o-mini. Leveraging Retrieval-Augmented Generation, GPT-4o-mini synthesized these retrieved entities into a coherent, contextually relevant, natural-language summary that directly addressed the user's query.

5. Evaluation

To validate the effectiveness and accuracy of our proposed hybrid query system, we designed a comprehensive evaluation involving rigorous experimental setups, comparative analyses against manually constructed ground-truth queries, and a detailed assessment of results across multiple dimensions.

5.1. Experimental Setup

5.1.1. Dataset and Gold-Standard Queries

We prepared a dataset comprising **92** manually crafted natural-language queries, each paired with expertly constructed corresponding SPARQL queries, hereafter referred to as the *gold-standard* queries.

These queries were selected to reflect realistic academic information needs and were distributed across four distinct categories:

- Conferences (22 queries)
- Authors (25 queries)
- Organizations (20 queries)
- Papers (25 queries)

The diversity and representativeness of the queries ensured robust coverage of typical scenarios encountered in scholarly querying tasks.

5.1.2. Evaluation Metrics

To quantitatively measure retrieval effectiveness, we employed the following metrics:

Overlap Percentage: Defined as the intersection of retrieved entities by our automated pipeline and the manually retrieved entities (gold-standard), normalized by the size of the gold-standard result set:

$$\text{Overlap} = \frac{|R_{\text{auto}} \cap R_{\text{manual}}|}{|R_{\text{manual}}|} \times 100\%$$

Threshold Accuracy: We established a performance threshold at **70% overlap**, considering queries achieving this threshold as successfully answered. This threshold was selected to balance recall and precision while remaining practically useful for typical scholarly use cases.

5.1.3. Comparative Baselines

For thoroughness, our automated approach was compared primarily against:

- Manually constructed SPARQL queries.
- A pure embedding-based retrieval baseline using FAISS without symbolic query generation.

This allowed us to highlight the advantages of our hybrid model in both structured and semantic contexts.

5.2. Results and Analysis

5.2.1. Overall Performance

Out of the 92 queries evaluated, our hybrid retrieval system achieved the following outcomes:

- **Successful Retrieval (≥70% overlap):** 89 queries (96%)
- **Mean Overlap Percentage (all queries):** 92.5%

These results indicate that the vast majority of automated queries closely approximated expert-level query quality.

5.2.2. Performance by Entity Type

We further analyzed performance separately for each query category to assess consistency across different scholarly query contexts.

These breakdowns illustrate particularly robust consistency in conference, author, and organization queries, with only slightly lower performance on paper-related queries, possibly due to higher semantic complexity inherent in publication metadata.

Table 2
Performance by Query Type

Entity Type	Queries	≥70% Overlap	Accuracy (%)	Mean Overlap (%)
Conferences	22	22	100%	95.2%
Authors	25	25	100%	94.8%
Organizations	20	20	100%	96.0%
Papers	25	22	88%	88.4%
Overall	92	89	96%	92.5%

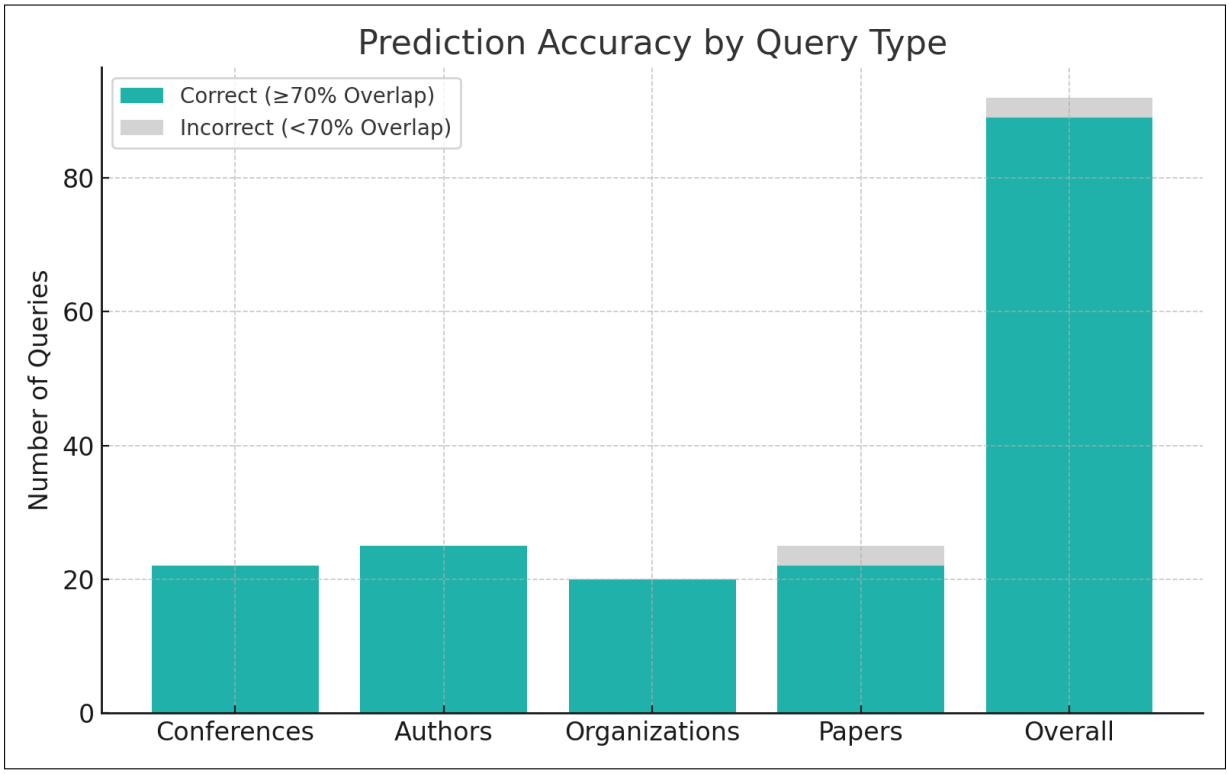

Figure 2: Diagram reflecting the accuracy of the methodology

5.2.3. Query Type Classification Accuracy

The RoBERTa-Large classifier demonstrated exceptional effectiveness, achieving an accuracy of **99%** in predicting query categories on the federated validation set. This high classification accuracy was instrumental in routing each query to the correct generation and retrieval pipeline.

5.3. Comparative Analysis: Symbolic vs. Embedding-only Retrieval

To better understand the benefit of our hybrid strategy, we compared it with a baseline using FAISS embeddings only (no symbolic query generation):

Table 3
Hybrid vs. Embedding-only Retrieval

Retrieval Method	Queries ≥70% Overlap
Embedding-only retrieval	62 / 92 (67%)
Hybrid (Ours)	**89 / 92 (96%)**

Our hybrid approach significantly outperformed the pure embedding baseline, underscoring the value of combining symbolic SPARQL generation with embedding-based reranking and fallback mechanisms.

5.4. Discussion and Interpretation

The results clearly demonstrate that our hybrid retrieval method effectively bridges the gap between SPARQL expertise and user-friendly natural-language interfaces. Key findings include:

- **Robust query-type classification:** The 99% classification accuracy ensured precise identification of query intent, enabling targeted downstream processing.
- **Balanced hybrid strategy:** The integration of symbolic querying with embedding-based reranking delivered consistent results, addressing weaknesses of either method alone.
- **Semantic flexibility:** Embedding-based reranking improved semantic coherence and compensated for variability in query phrasing.

5.5. Limitations and Error Analysis

Despite the strong overall performance, several limitations were observed:

- The number of test dataset was short, it comprised of 92 manually curated examples.
- The federated dataset created for the training of RobertaLarge might contain some bias or lack of proper generalization when tested over a huge number of queries.

Future work will include having a larger evaluation dataset curated by experts, improving entity embeddings, expanding few-shot prompt coverage for LLMs and explore other finetuned llms to generate SPARQL query, and exploring advanced reranking strategies to further enhance the accuracy and robustness of results across all entity types.

6. Conclusion

In this study, we presented a robust hybrid framework that enables natural-language querying of scholarly knowledge graphs by combining the precision of symbolic SPARQL generation with the flexibility of embedding-based retrieval. Our system leverages a fine-tuned RoBERTa-Large classifier for query type prediction, GPT-4o-mini for SPARQL generation guided by few-shot prompting, and FAISS-indexed semantic embeddings for result reranking and fallback retrieval. This architecture bridges the gap between expert-level structured querying and accessible, intuitive user interfaces.

Through extensive evaluation against a curated set of 92 gold-standard SPARQL queries, our approach achieved a 96% success rate in matching expert results, with a mean overlap of 92.5%, demonstrating its effectiveness in replicating expert performance. The integration of GraphRAG further ensured meaningful answers even in cases where symbolic retrieval alone was insufficient.

Our findings highlight the practical potential of combining LLM-driven query generation with embedding-based semantic reasoning to improve the accessibility and usability of scholarly knowledge graphs for non-expert users. Future work will explore multilingual query support, dynamic schema adaptation, user feedback integration, and expansion to other domains beyond scholarly metadata.

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569. The opinions expressed in this document reflect only the author's view and in no way reflect the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information it contains.

8. Supplementary Materials

The resources for this study are placed at the following repository:
https://anonymous.4open.science/r/scholarlyGraphRAG-BD22/README.md

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Grammarly to assist with paraphrasing, grammar correction, and spelling checks.

References

[1] V. Lopez, V. Uren, M. R. Sabou, E. Motta, Cross ontology query answering on the semantic web: an initial evaluation, in: Proceedings of the fifth international conference on Knowledge capture, 2009, pp. 17–24.

[2] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber, P. Cimiano, Template-based question answering over rdf data, in: Proceedings of the 21st international conference on World Wide Web, 2012, pp. 639–648.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic web 6 (2015) 167–195.

[4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[5] S. Vahdati, N. Arndt, S. Auer, C. Lange, Openresearch: collaborative management of scholarly communication metadata, in: Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20, Springer, 2016, pp. 778–793.

[6] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Conference linked data: the scholarlydata project, in: The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15, Springer, 2016, pp. 150–158.

[7] S. Harris, A. Seaborne, E. Prud'hommeaux, Sparql 1.1 query language. w3c recommendation (2013), URL https://www. w3. org/TR/sparql11-query (2013).

[8] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey: A systematic literature review and conceptual framework, Semantic web 7 (2015) 63–93.

[9] W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, J. Suh, The value of semantic parse labeling for knowledge base question answering, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 201–206.

[10] J. D'Abramo, A. Zugarini, P. Torroni, Investigating large language models for text-to-sparql generation, in: Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, 2025, pp. 66–80.

[11] D. Banerjee, P. A. Nair, J. N. Kaur, R. Usbeck, C. Biemann, Modern baselines for sparql semantic parsing, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2260–2265.

[12] H. Tran, L. Phan, J. Anibal, B. T. Nguyen, T.-S. Nguyen, Spbert: An efficient pre-training bert on sparql queries for question answering over knowledge graphs, in: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28, Springer, 2021, pp. 512–523.

[13] S. Auer, M. Stocker, L. Vogt, G. Fraumann, A. Garatzogianni, Orkg: Facilitating the transfer of research results with the open research knowledge graph, Research Ideas and Outcomes 7 (2021) e68513.

[14] A. Meloni, D. R. Recupero, F. Osborne, A. Salatino, E. Motta, S. Vahadati, J. Lehmann, Assessing large language models for sparql query generation in scientific question answering, in: CEUR Workshop Proceedings, volume 3953, 2025.
[15] H. Peng, H. Li, Y. Song, V. Zheng, J. Li, Differentially private federated knowledge graphs embedding, in: Proceedings of the 30th ACM international conference on information & knowledge management, 2021, pp. 1416–1425.
[16] L. Yue, Y. Zhang, Q. Yao, Y. Li, X. Wu, Z. Zhang, Z. Lin, Y. Zheng, Relation-aware ensemble learning for knowledge graph embedding, arXiv preprint arXiv:2310.08917 (2023).
[17] H.-J. Cha, S.-W. Choi, E.-B. Lee, D.-M. Lee, Knowledge retrieval model based on a graph database for semantic search in equipment purchase order specifications for steel plants, Sustainability 15 (2023) 6319.
[18] I. A. Ebeid, Medgraph: A semantic biomedical information retrieval framework using knowledge graph embedding for pubmed, Frontiers in big Data 5 (2022) 965619.
[19] J. Qi, C. Su, Z. Guo, L. Wu, Z. Shen, L. Fu, X. Wang, C. Zhou, Enhancing sparql query generation for knowledge base question answering systems by learning to correct triplets, Applied Sciences 14 (2024) 1521.
[20] S. Ferré, Sparklis: An expressive query builder for sparql endpoints with guidance in natural language, Semantic Web 8 (2016) 405–418.
[21] N. Afreen, G. Balloccu, L. Boratto, G. Fenu, F. M. Malloci, M. Marras, A. G. Martis, Edge: A conversational interface driven by large language models for educational knowledge graphs exploration, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 5159–5163.
[22] M. Arazzi, D. Ligari, S. Nicolazzo, A. Nocera, Augmented knowledge graph querying leveraging llms, arXiv preprint arXiv:2502.01298 (2025).
[23] D. Bustamante, H. Takeda, Sparql generation with entity pre-trained gpt for kg question answering, arXiv preprint arXiv:2402.00969 (2024).
[24] S. Xu, S. Liu, T. Culhane, E. Pertseva, M.-H. Wu, S. J. Semnani, M. S. Lam, Fine-tuned llms know more, hallucinate less with few-shot sequence-to-sequence semantic parsing over wikidata, arXiv preprint arXiv:2305.14202 (2023).
[25] A. Sharma, L. Lara, C. J. Pal, A. Zouaq, Reducing hallucinations in language model-based sparql query generation using post-generation memory retrieval, arXiv preprint arXiv:2502.13369 (2025).
[26] P. A. K. Karou Diallo, A. Zouaq, Frase: Structured representations for generalizable sparql query generation, arXiv e-prints (2025) arXiv–2503.
[27] J. Lee, H. Shin, Sparkle: Enhancing sparql generation with direct kg integration in decoding, Expert Systems with Applications (2025) 128263.
[28] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2019) 535–547.
[29] J. Lu, K. Hall, J. Ma, J. Ni, Hyrr: Hybrid infused reranking for passage retrieval, arXiv preprint arXiv:2212.10528 (2022).
[30] M. Acosta, C. Qin, T. Schwabe, Neuro-symbolic query optimization in knowledge graphs, arXiv preprint arXiv:2411.14277 (2024).
[31] Y. Jia, W. Chen, A method of query graph reranking for knowledge base question answering, arXiv preprint arXiv:2204.12808 (2022).
[32] M. Li, C. Yang, C. Xu, X. Jiang, Y. Qi, J. Guo, H.-f. Leung, I. King, Retrieval, reasoning, re-ranking: A context-enriched framework for knowledge graph completion, arXiv preprint arXiv:2411.08165 (2024).
[33] M. Gao, Y. Xie, W. Chen, F. Zhang, F. Ding, T. Wang, J. Yao, J. Zheng, K.-F. Wong, Rerankgc: A cooperative retrieve-and-rerank framework for multi-modal knowledge graph completion, Neural Networks 188 (2025) 107467.
[34] S. Verma, R. Bhatia, S. Harit, S. Batish, Scholarly knowledge graphs through structuring scholarly communication: a review, Complex & intelligent systems 9 (2023) 1059–1095.
[35] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo,

R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.
[36] D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, A. Wahler, D. Fensel, et al., Introduction: what is a knowledge graph?, Knowledge graphs: Methodology, tools and selected use cases (2020) 1–10.
[37] R. Usbeck, A.-C. Ngonga Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, A. Both, Agdistis-graph-based disambiguation of named entities using linked data, in: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13, Springer, 2014, pp. 457–471.
[38] H. Chen, Large knowledge model: Perspectives and challenges, arXiv preprint arXiv:2312.02706 (2023).