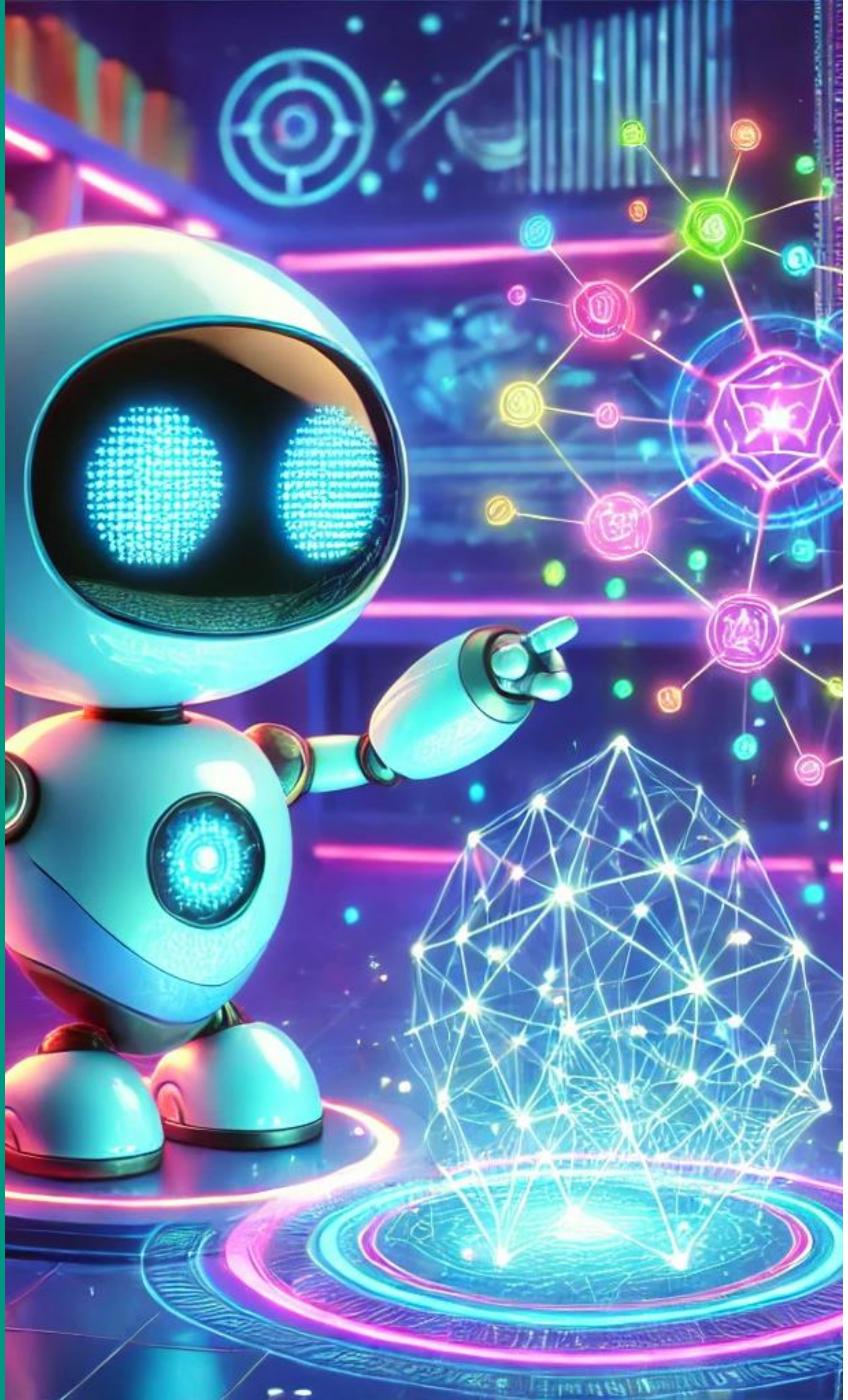
HubLink: Leveraging Language Models for Enhanced Scholarly Information Retrieval on Research Knowledge Graphs

by Marco Schneider



Master Thesis

supervised by Angelika Kaplan & Jan Keim

Motivation
Literature Research is Hard

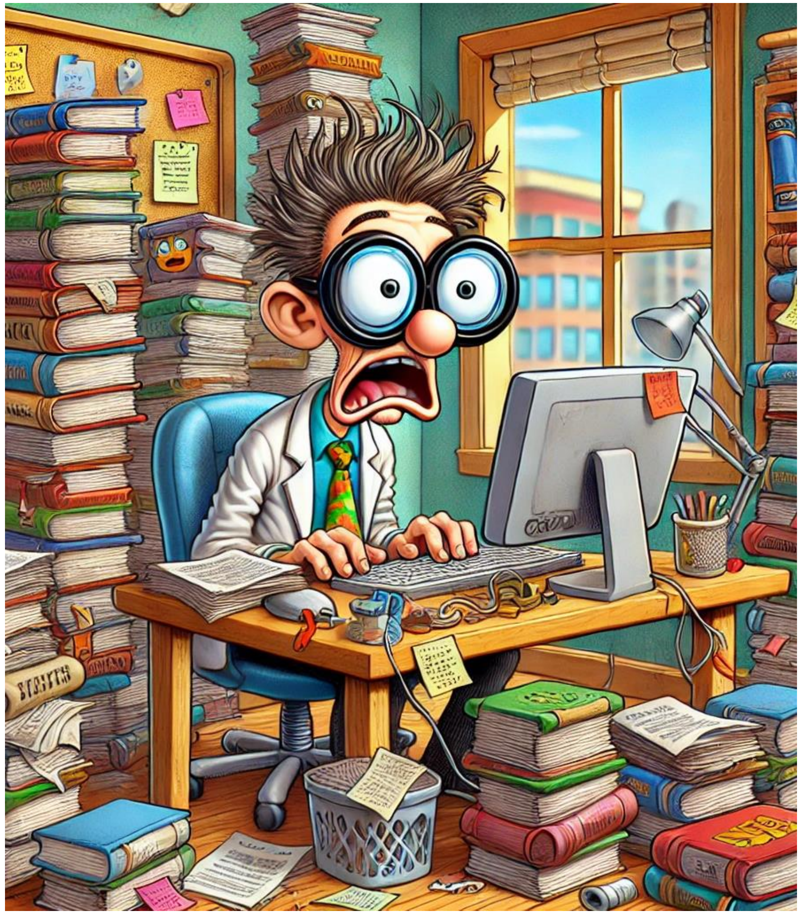
- Current practice to conduct literature search is a time-consuming and cognitively taxing process.

Two interesting concepts:

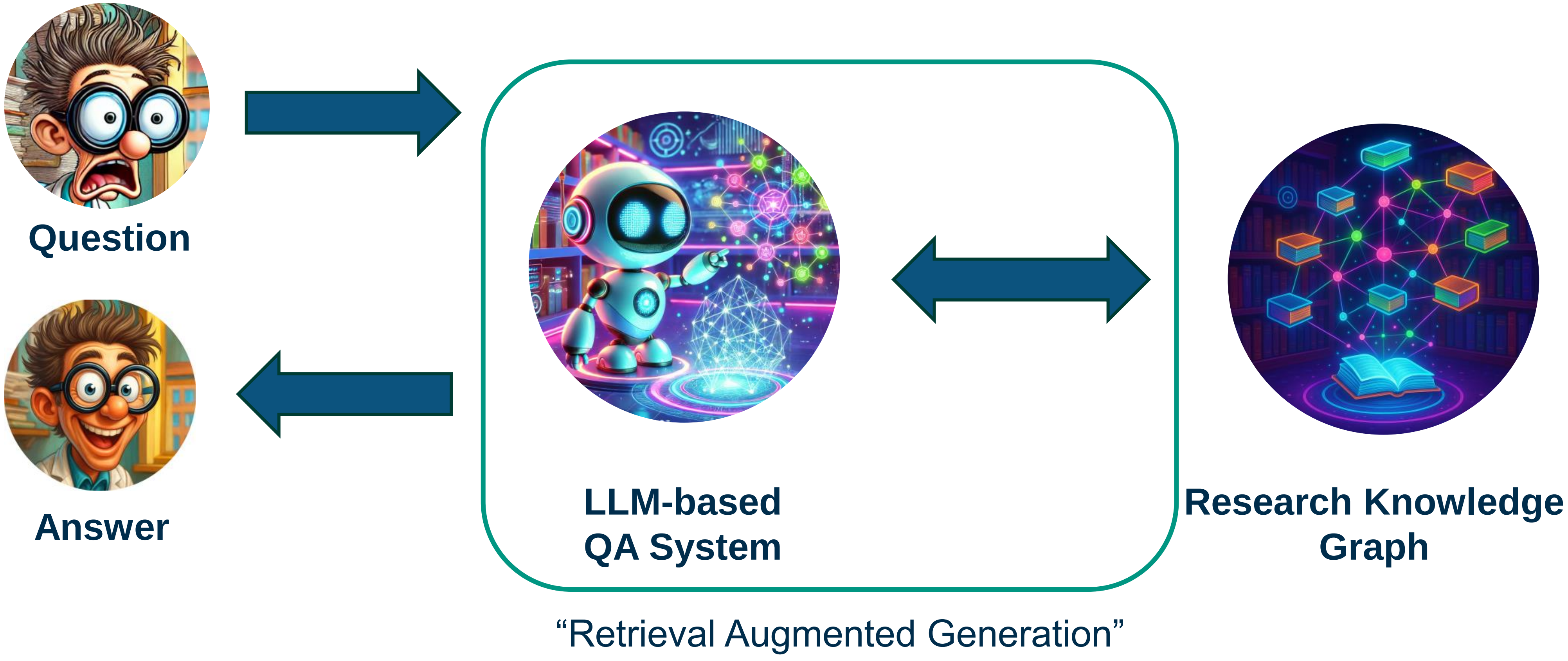
- **Research Knowledge Graphs (RKGs)**
  a. Store and interconnect scientific findings instead of embedding them in long texts.
  b. But are difficult for users to access.

- **Large Language Models (LLMs)**
  a. Demonstrate powerful capabilities in Natural Language Processing.
  b. But have difficulties with specialized domains and hallucinations.

Typical Academic Researcher



OpenAI. (2024). Cartoon-style image of a frustrated academic researcher. Created with DALL-E. Retrieved on July 4, 2024



Idea
Unifying LLMs with RKGs

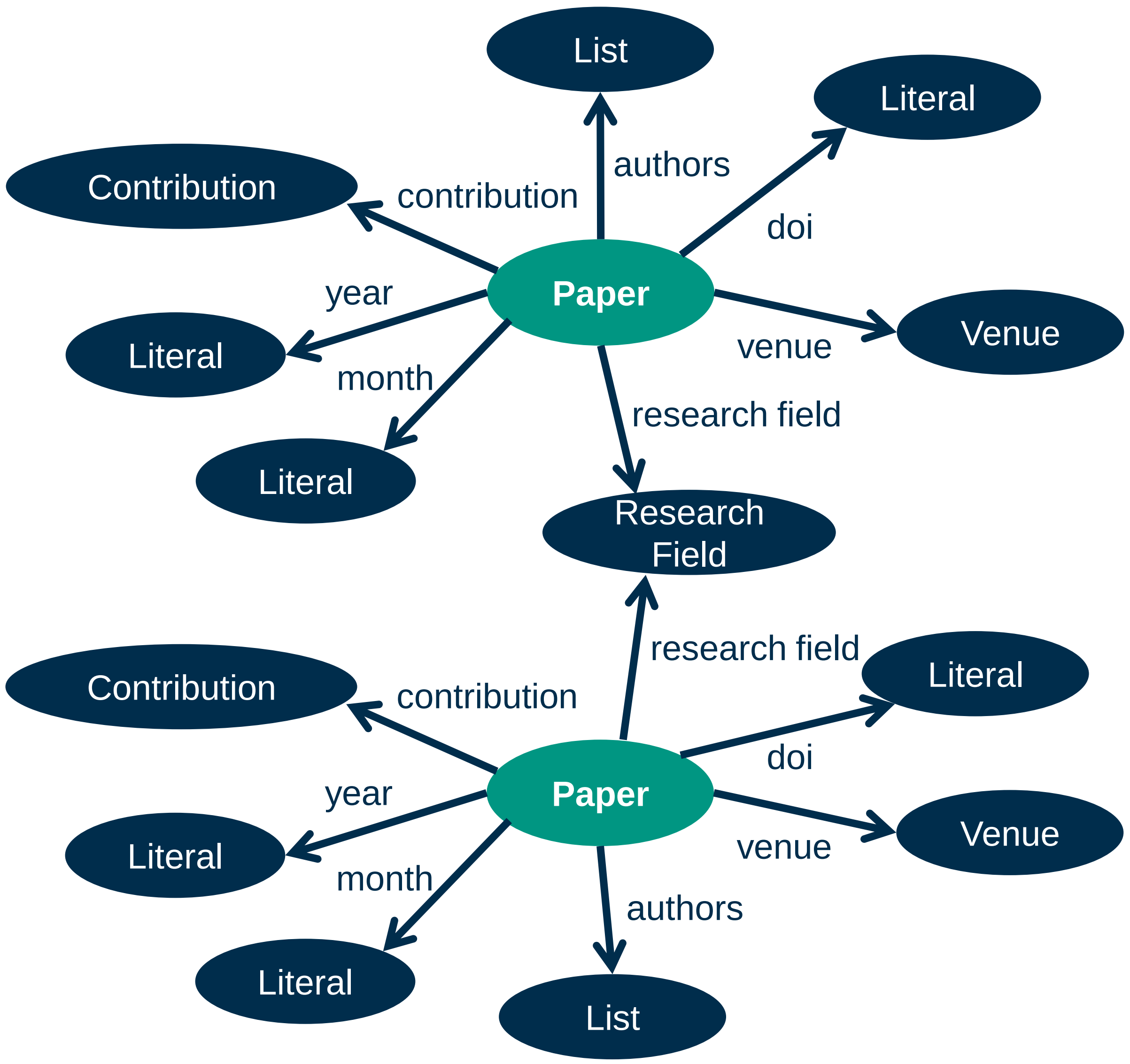
Idea: Combine both concepts in a Knowledge Graph Question Answering (KGQA) setting to enhance scholarly information retrieval.

Fundamentals
For this Presentation

- **Knowledge Graphs (KGs)** *[Banerjee2024, Pan2024]*
  a. Are networks that store structured data.
  b. Consists of triples in the RDF format: (subject, predicate, object).

- **Research Knowledge Graphs (RKGs)** *[Auer2018]*
  a. Are KGs that are focused on storing scholarly statements.
  b. For example, the Open Research Knowledge Graph (RKG).

- **Knowledge Graph Question Answering (KGQA)** *[Banerjee2024, Chakraborty2021, Pan2024, Yani2022]*
  a. Research area focused on empowering users to access information stored within a KG by formulating questions in natural language.



Visualization of the ORKG Graph

Related Work
Scholarly KGQA Approaches

- Current efforts for scholarly KGQA are primarily **Semantic Parsing (SP)** methods *[Jaradeh2020, Devlin2019, Banerjee2023, Raffel2023, Li2024, Taffa2023, Lehmann2024, Jiang2023]*.
- These approaches translate natural language into formal queries, such as SPARQL, by identifying relevant entities and relations in the RKG *[Zhang2023]*.

1. **Not Schema-Agnostic:** These approaches degrade when applied to larger and dynamic graphs where unseen schema components and entities are common *[Gu2022]*.
2. **Require Training-Data:** The approaches rely on task-specific training examples for efficient retrieval, limiting adaptability and scalability.

Related Work
Training-Free & Schema Agnostic KGQA Approaches

- Current related training-free and schema-agnostic KGQA approaches in the literature can be differentiated between:
  a. **Stepwise-Reasoning** for example FiDeLiS *[Sui2024]*.
  b. **Subgraph Construction** for example Mindmap *[Wen2024]*.
  c. **Vector-Based** for example DiFaR *[Baek2023]*.

- They all demonstrate **state-of-the-art performance** in their respective experiments.

- **However,** they have only been evaluated on open-domain graphs and not yet been tested in the scholarly domain **!**

- **And** they do not consider the source of knowledge during inference.

Summary
Problems and Contributions

I identified…

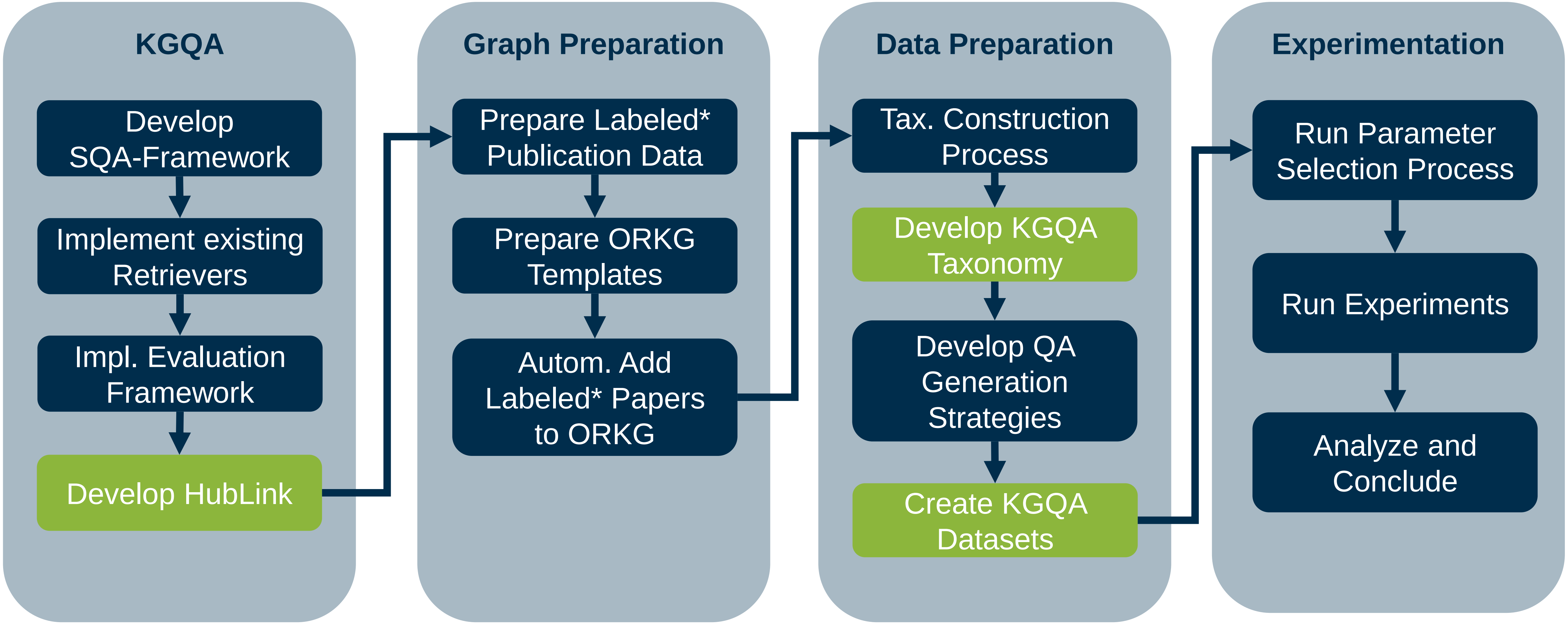
- … a **research gap** for applying alternative (non semantic parsing) and LLM-based strategies on the scholarly KGQA task.
- … that current KGQA approaches do not consider the **source of knowledge** during inference.

- … a lack of a **taxonomy** to classify the characteristics of questions posed to KGQA retrieval systems for scholarly tasks.

I contribute…

- … a novel **LLM-based KGQA** approach named "HubLink". *(schema-agnostic, no training data, source-aware inference)*

- … a **Question Taxonomy** for classifying questions targeting scholarly KGQA. *(to evaluate performance and capabilities of systems)*
- … new **KGQA Datasets** for the ORKG. *(to benchmark performance and robustness)*



Actions
The Research Workflow

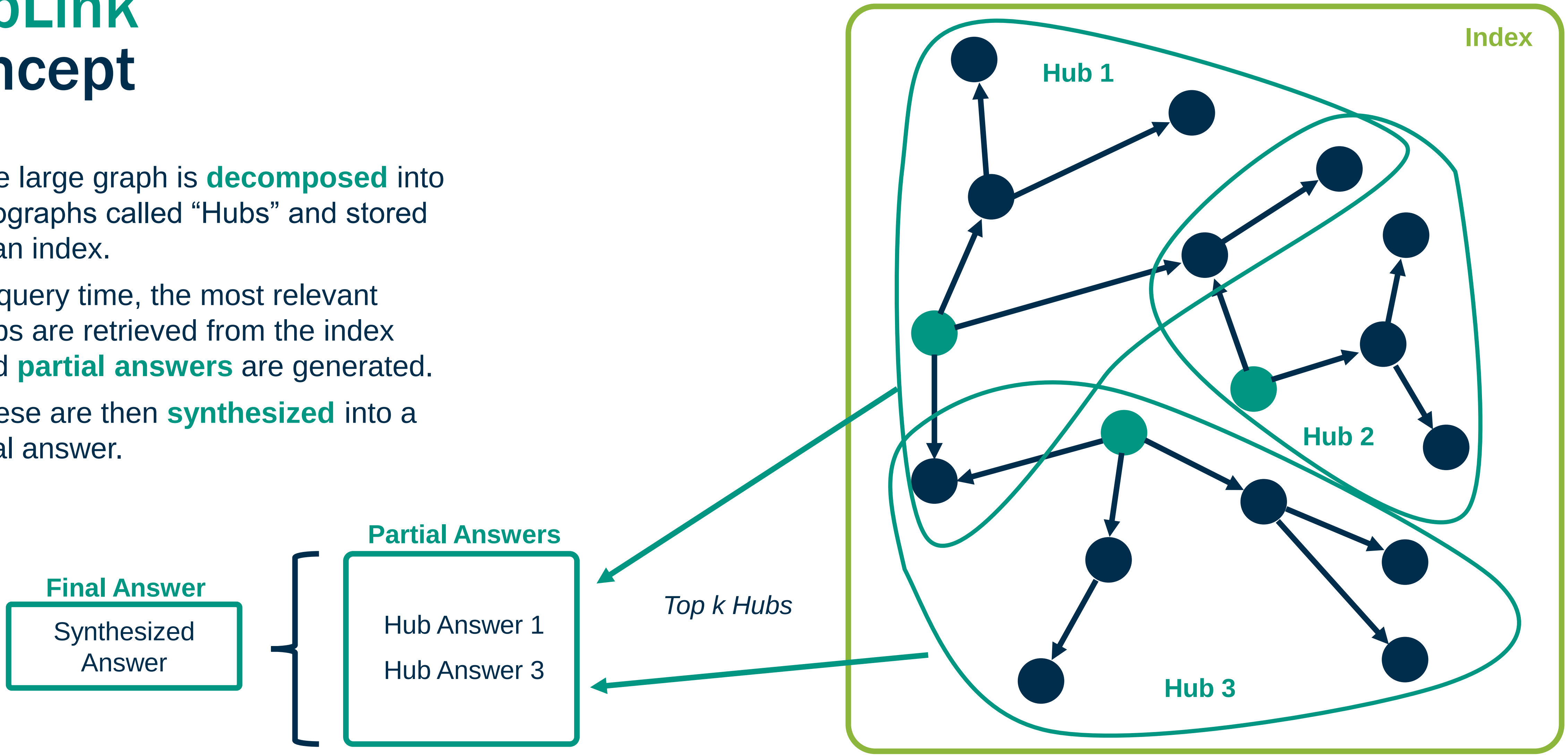
** Based on the work by [Konersmann2022]*

HubLink

KGQA Retrieval Approach

HubLink
Concept

- The large graph is **decomposed** into subgraphs called “Hubs” and stored in an index.
- At query time, the most relevant hubs are retrieved from the index and **partial answers** are generated.
- These are then **synthesized** into a final answer.

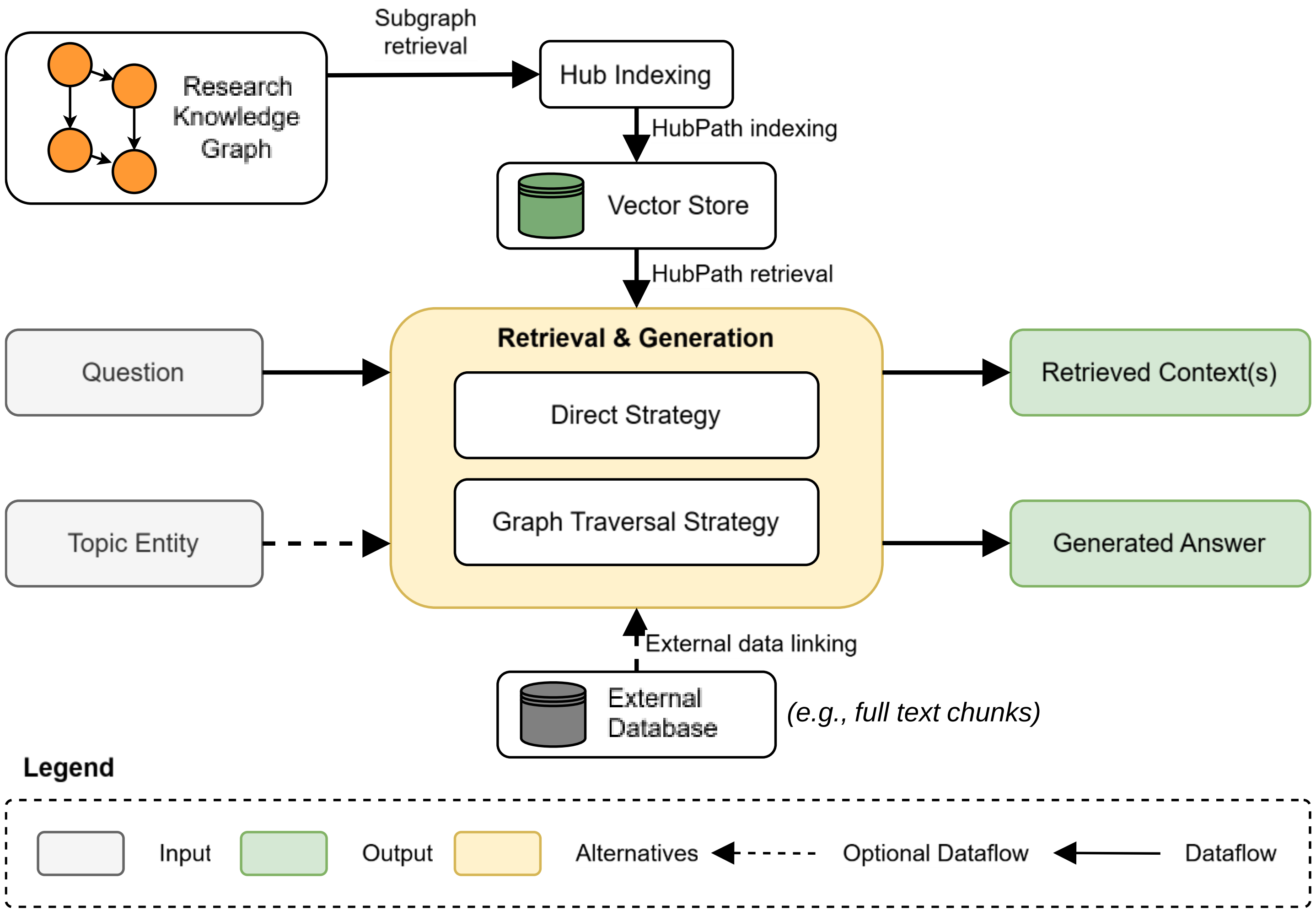


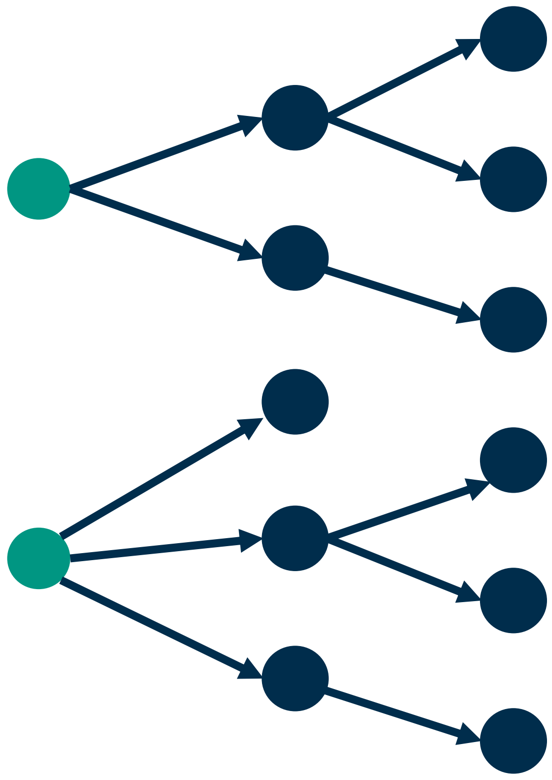
HubLink
Overview

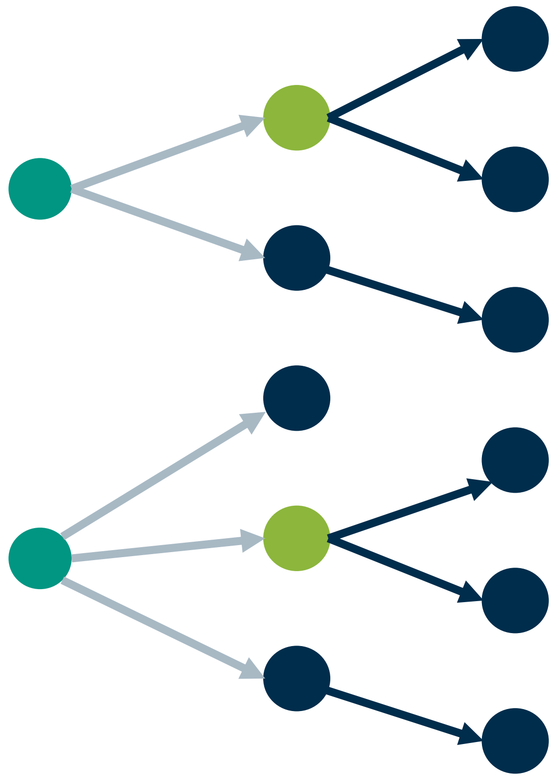
The new KGQA approach has three phases:

Three Phases

1. Indexing
2. Retrieval
3. Generation

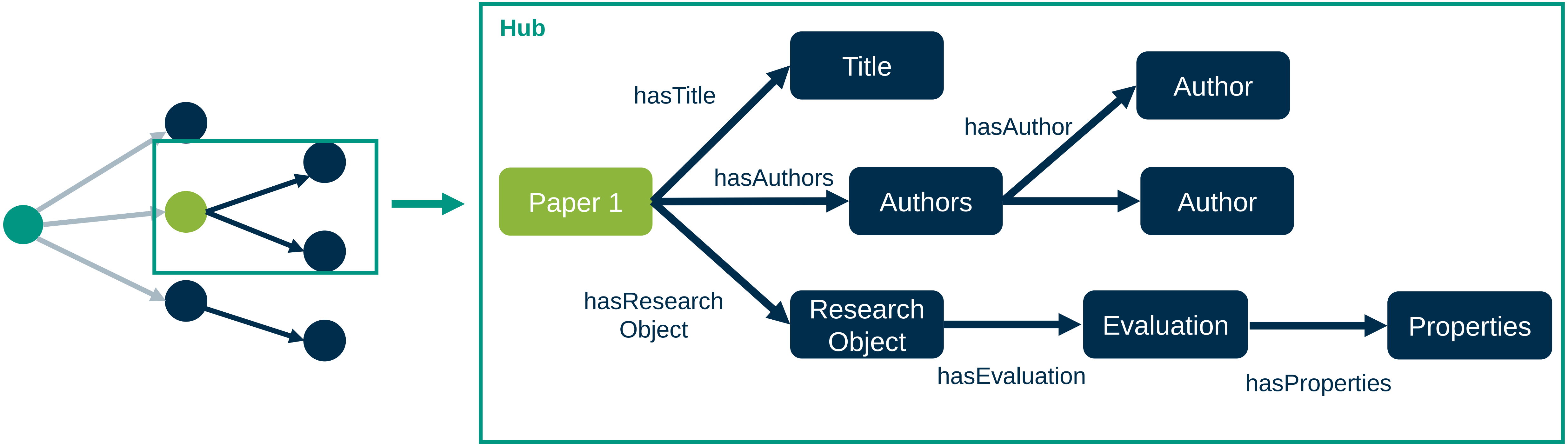






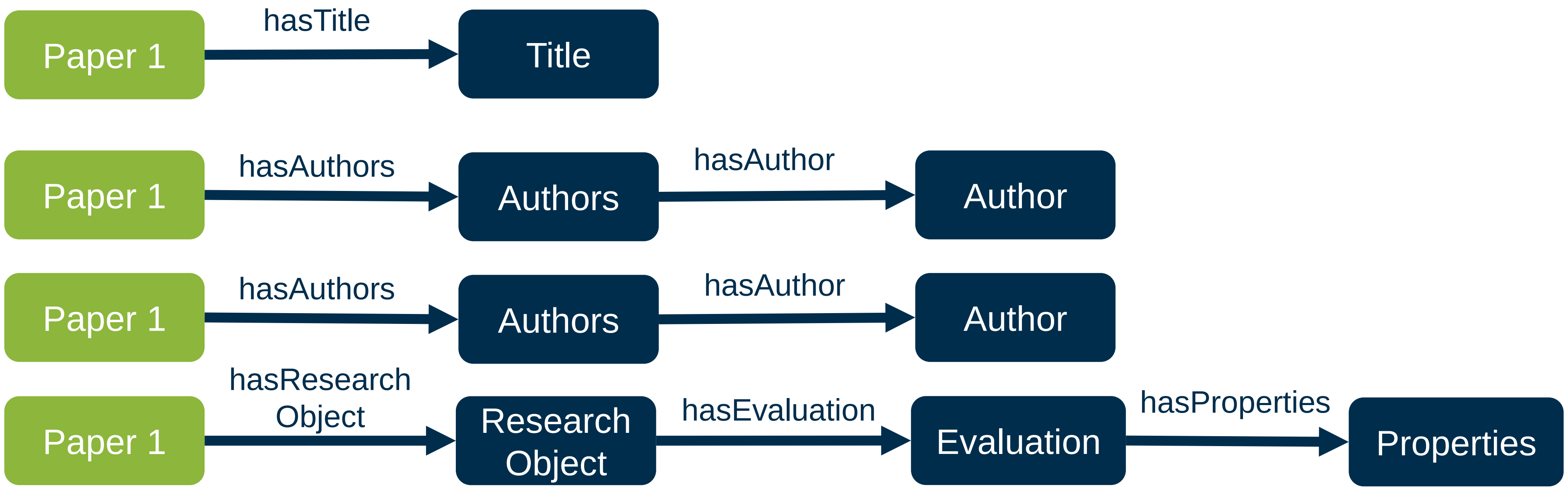
1. **Given:** Is a list of Start Points from which the indexing is started.
2. **Search:** The graph is traversed to find entities that are roots of Hubs.

HubLink
Indexing Process



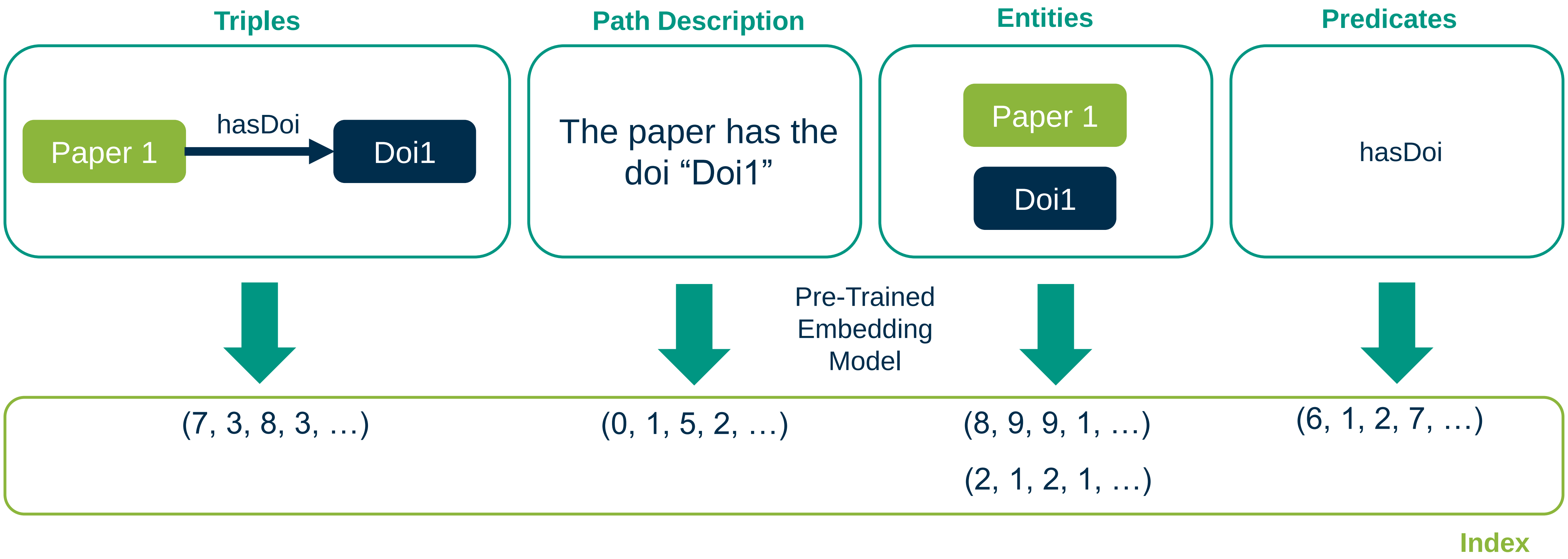
3. **Building Hubs:** Each Hub is then build starting from the root entity and subsequently stored in the vector store.

HubLink
Indexing Process



4. **Building Hubs:** The triple paths of the hub are extracted from the graph.

HubLink
Indexing Process



5. **Building Hubs:** The paths are converted to four different vectors at different content levels which are then stored in a vector index (including metadata for tracing).

HubLink
Indexing Process

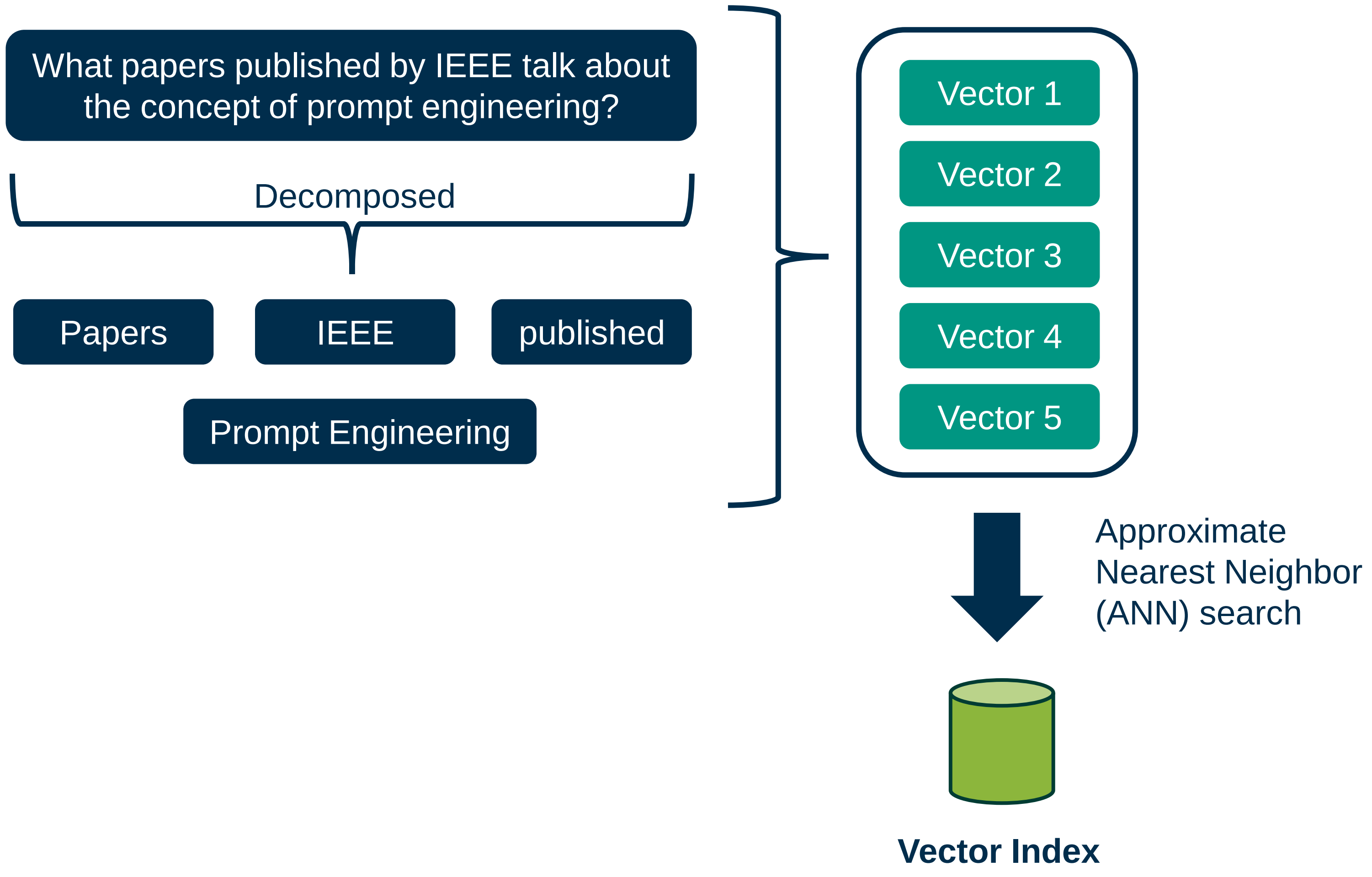
6. **Finding more Hubs:** The graph is further traversed to find and index more hubs.

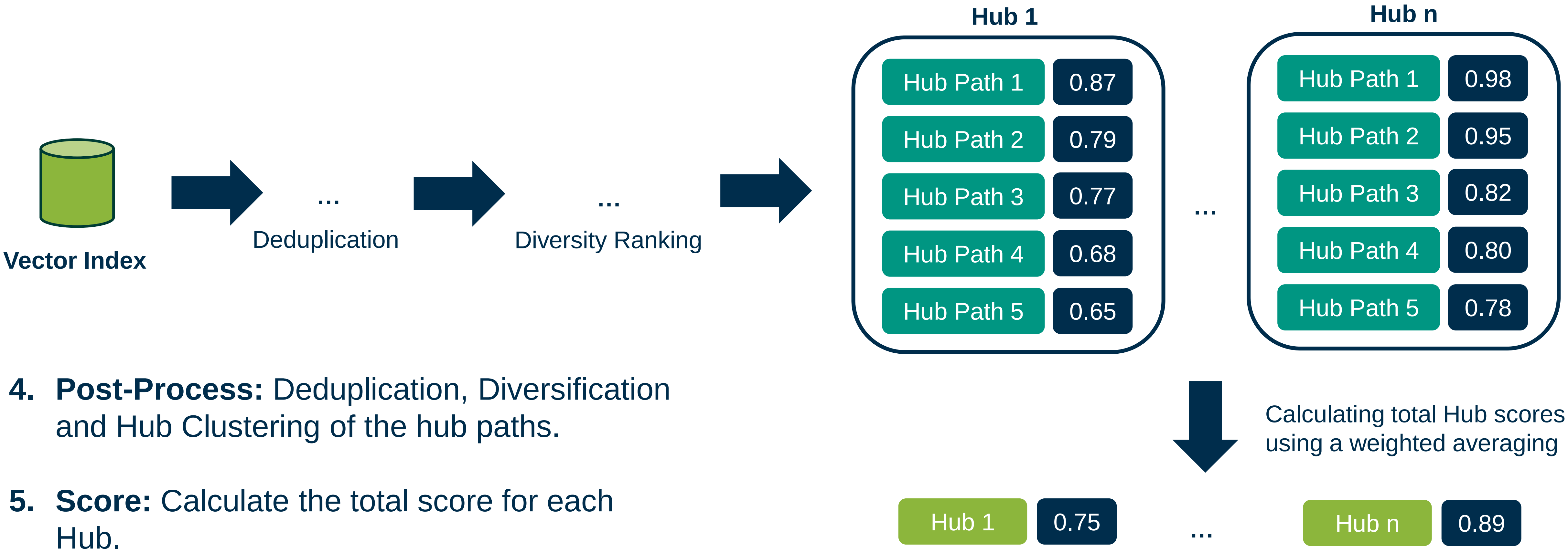
HubLink
Direct Retrieval Strategy

1. **Decompose:** Decompose the question into components.

2. **Embed:** Embed the question and the components into vectors.

3. **Retrieve:** Query the index to find relevant hub paths.

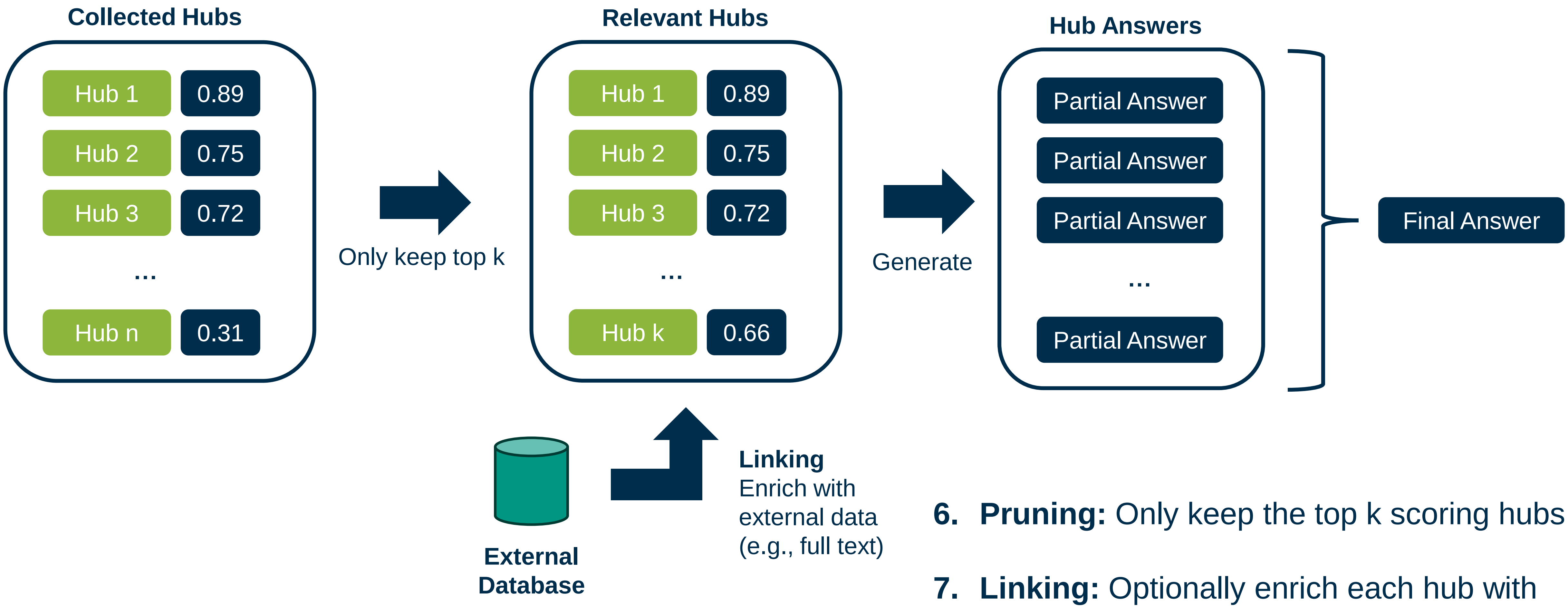




4. **Post-Process:** Deduplication, Diversification and Hub Clustering of the hub paths.

5. **Score:** Calculate the total score for each Hub.

HubLink
Direct Retrieval Strategy

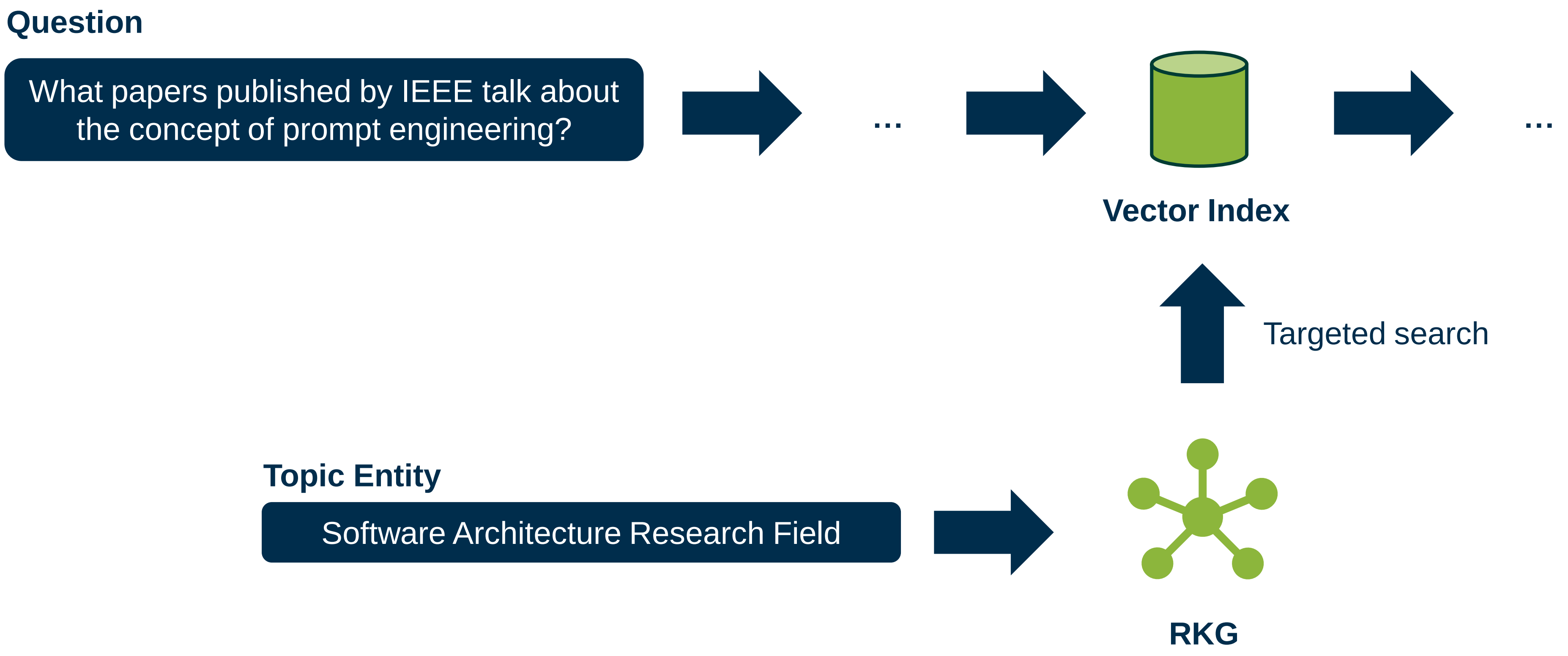


6. **Pruning:** Only keep the top k scoring hubs.

7. **Linking:** Optionally enrich each hub with external data.

8. **Answer Generation:** Generate for each hub a partial answer and synthesize to a final answer.

HubLink
Direct Retrieval Strategy



HubLink
Graph Traversal Strategy

- In Principle the same as the direct strategy but
  - … requires a **topic entity.**
  - … also **traverses** the graph to utilize a targeted search.

HubLink
General Limitations

- Requires the **maintenance** of an index, as only data that is indexed can be retrieved.
- Requires **well-defined criteria** for hub structures to index all relevant knowledge.
- **Inherits limitations** of embedding-based systems.
- Performance strongly **depends** on the LLM that is used.

Research Data

The data that was used to conduct the experimentation

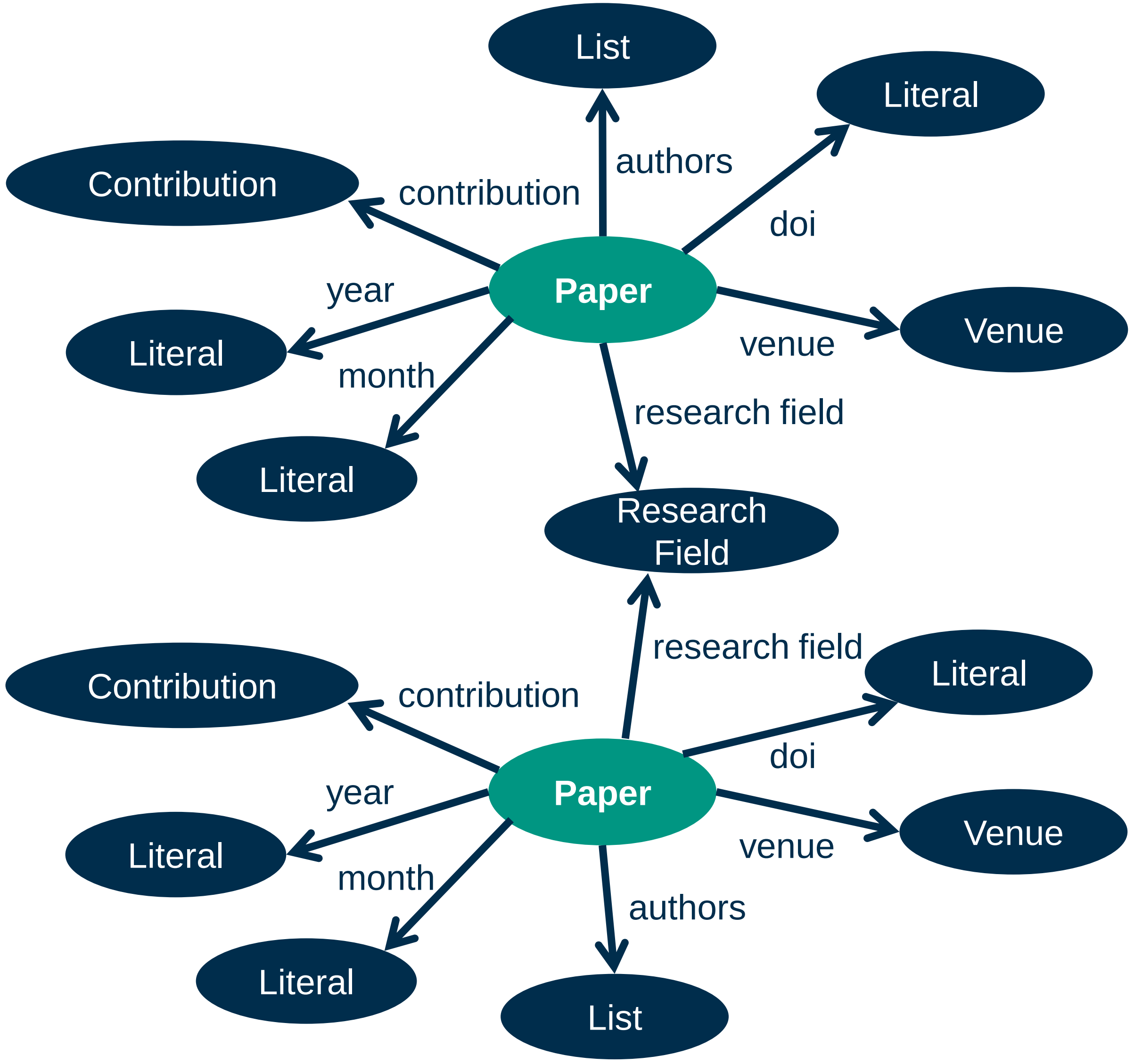
Research Data
Open Research Knowledge Graph

- Papers are added in a community effort by adding them based on templates.

Templates

- Are **predefined structures** that guide users in adding a new paper to the ORKG graph.

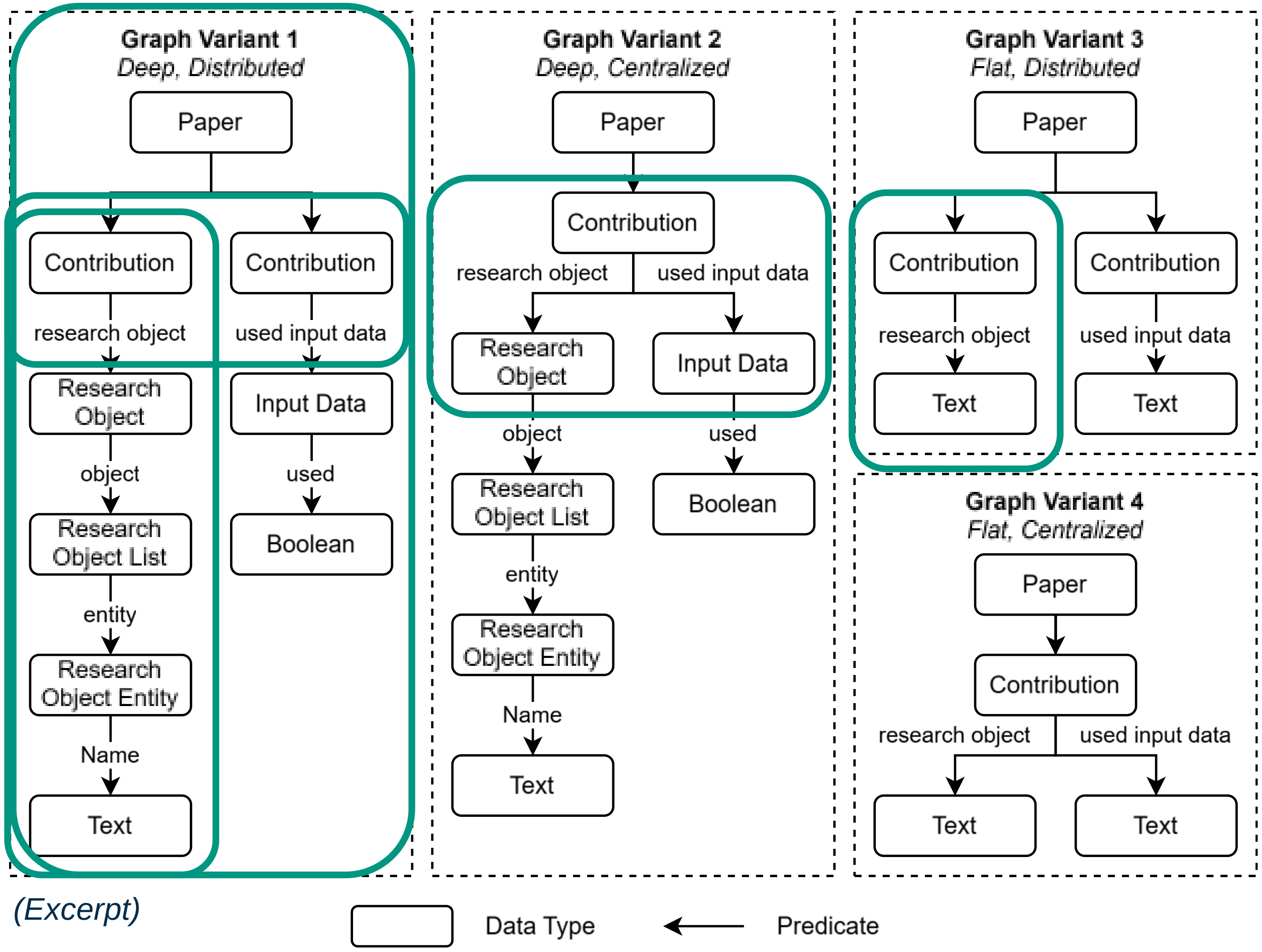
- Help **standardize** the representation of various types of research contributions.



Visualization of the ORKG Graph

Research Data
Template Variants

- I created **four** different templates based on a Software Architecture classification schema *[Konersmann2022].*
- Variations between **deep** paths and **distributed** paths.
- For experiments only involving one variant, I chose **GV1** as it is the model most realistically used in the real world.



Evaluation

Evaluating HubLink against state-of-the-art KGQA approaches on the ORKG

Evaluation
Goal-Question-Metric Plan

Goal 1: Relevance & robustness of retrieved contexts.

a. **Q1:** Improvement of overall retrieval performance.
b. **Q2:** Impact of retrieval operation.
c. **Q3:** Performance on six concrete literature search use cases.
d. **Q4:** Impact of type information in question.
e. **Q5:** Robustness to alternative graph schemas.
f. **Q6:** Runtime & token efficiency.
g. **Q7:** Environmental impact.

Metrics: *Precision, Recall, F1, Hits@k, EM@k, MRR@k, MAP@k*

Retrieval

Generation

Goal 2: Semantic and factual answer alignment to reference answers.

a. **Q8:** Improvement of semantic and factual alignment.

Metrics: *BLEU, ROUGE, Semantic Similarity, String Similarity, Bert-Score, Factual Correctness (LLM)*

Goal 3: Answer alignment with intent and content of the question.

a. **Q9:** Improvement of semantic intent.
b. **Q10:** Improvement of instruction following.

Metrics: *Answer Relevancy (LLM), Instruction Following (LLM)*

Goal 4: Answer alignment to retrieved contexts.

a. **Q11:** Faithfulness of generated answers.

Metrics: *Faithfulness (LLM)*

Evaluation
Tested HubLink Configurations

HubLink (T)

Parameter	Value
LLM	Gpt-o3-mini
Embedding	Text-embedding-3-large
Strategy	Graph Traversal
Hub Paths to Keep	10
Number of Hubs	30

HubLink (D)

Parameter	Value
LLM	Gpt-o3-mini
Embedding	Text-embedding-3-large
Strategy	Direct
Hub Paths to Keep	10
Number of Hubs	30

HubLink (F)

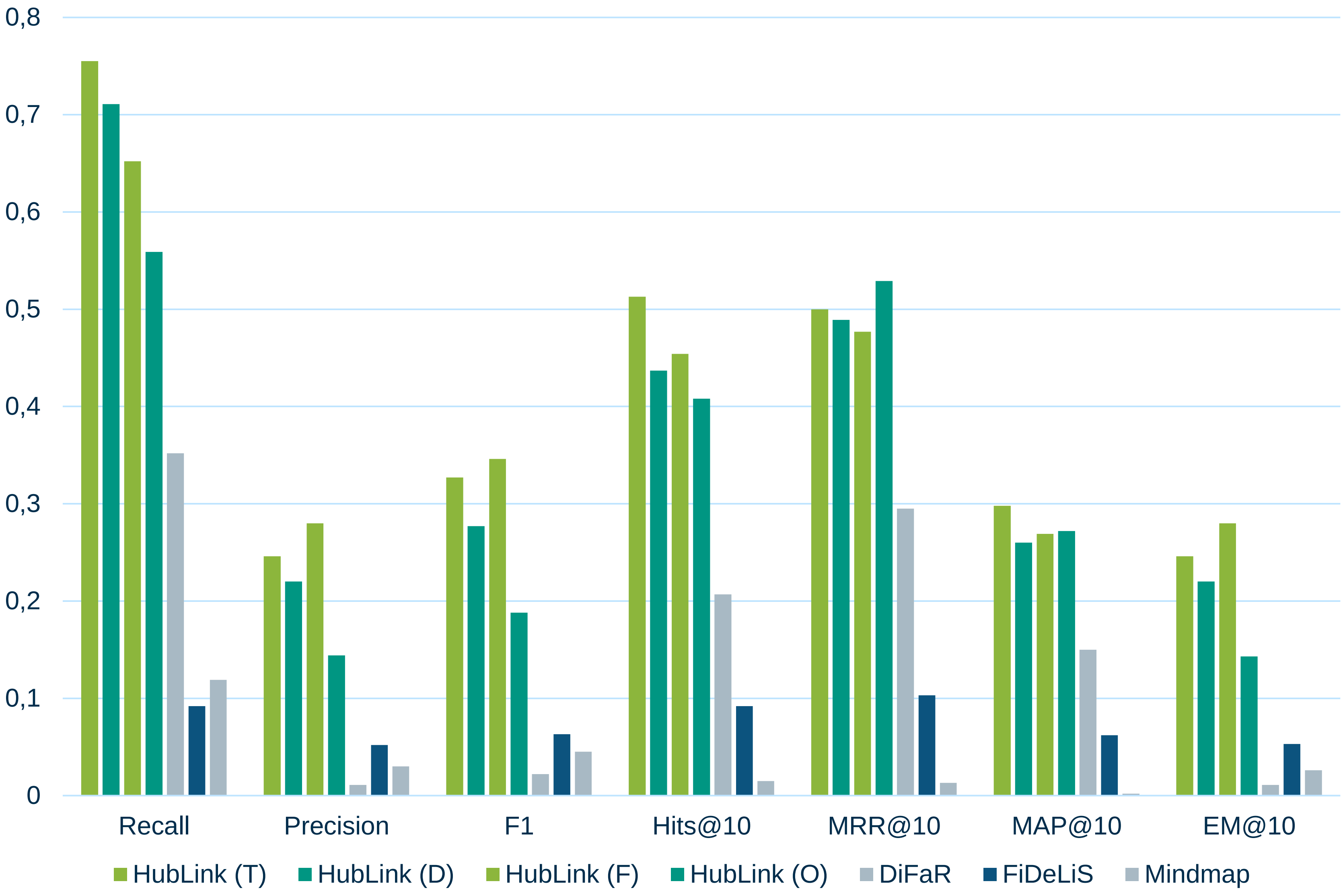
Parameter	Value
LLM	Gpt-o3-mini
Embedding	Text-embedding-3-large
Strategy	Direct
Hub Paths to Keep	10
Number of Hubs	10

HubLink (O)

Parameter	Value
LLM	Qwen2.5-14B
Embedding	mxbai-embed-large
Strategy	Graph Traversal
Hub Paths to Keep	10
Number of Hubs	30

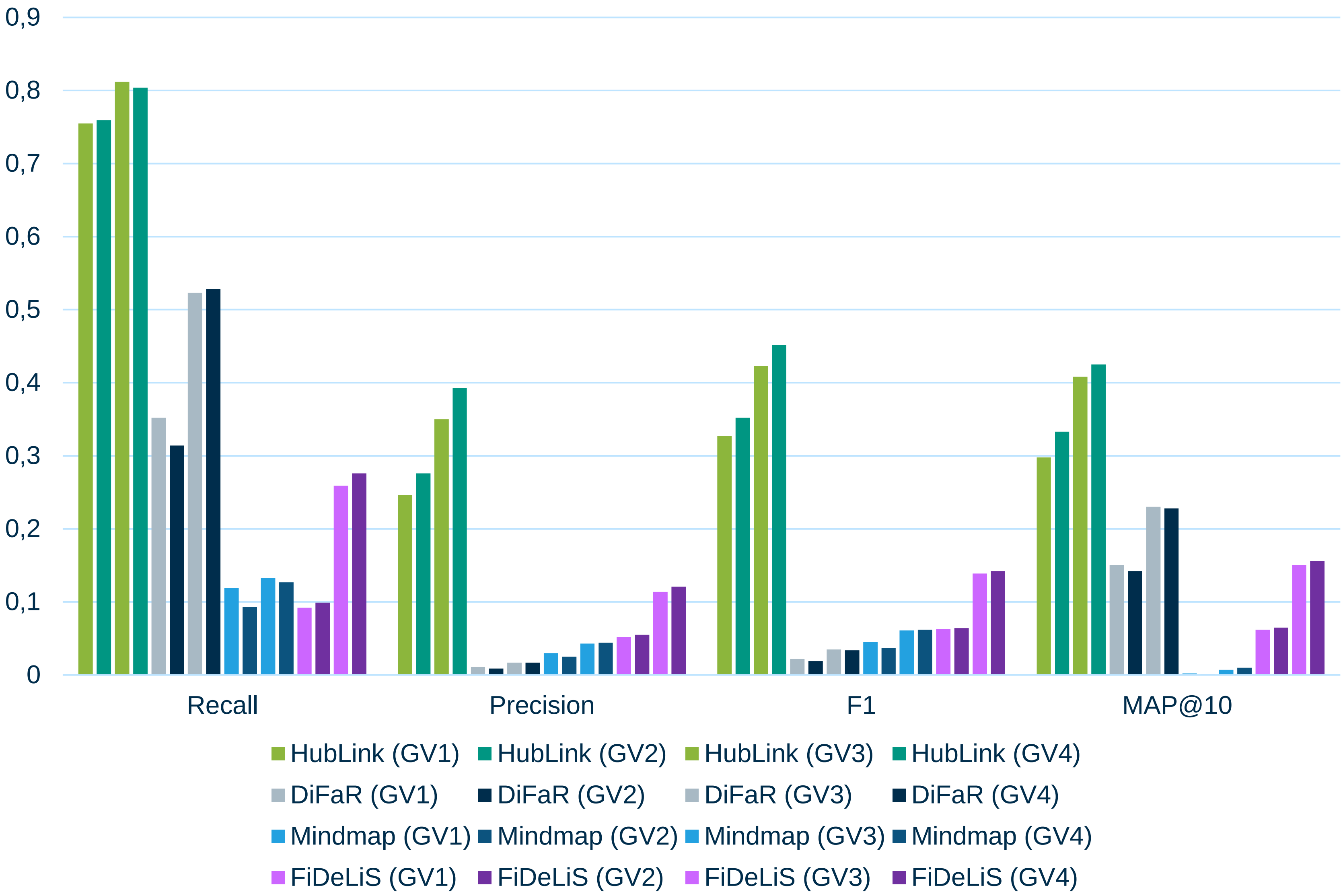
Evaluation
Overall Retrieval Performance

- Significantly improved performance compared to baselines
- Graph traversal strategy is superior
- Fewer hubs improve precision
- LLM models have a great influence
- Precision and ranking scores indicate limitations



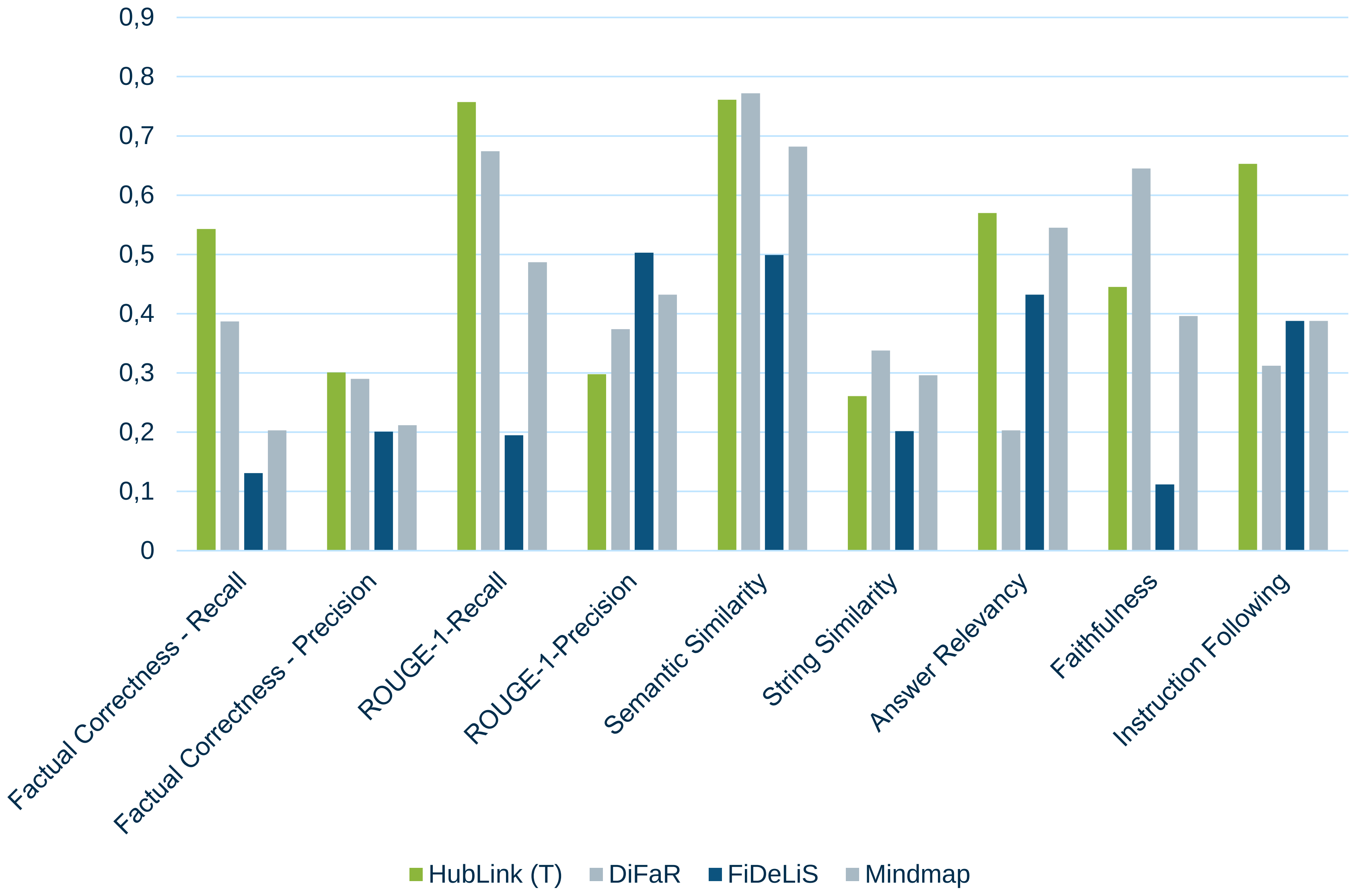
Evaluation
Graph Robustness

- Baselines perform significantly better on GV3 and GV4
- HubLink also shows a better performance but less pronounced
- Precision and ranking performance has the greatest increase
- HubLink can be applied without adaption to other graph variants which shows that it is schema-agnostic



Evaluation
Answer Generation

- HubLink tends to generate comprehensive answers compared to the reference answers
- Overall inclusion of facts seems to have limitations



Conclusion

HubLink to improve scholarly information retrieval in KGQA

Final Remarks

- **HubLink** has the potential to improve the current scholarly literature research workflow and provides a substantial advancement for schema-agnostic and training-free KGQA approaches.
- The **Taxonomy** can help in the development and evaluation of KGQA systems.
- The **KGQA Datasets** can be used to evaluate the performance of KGQA systems and assess their retrieval capabilities on the ORKG.

Future Work

- Addressing **limitations** of HubLink.
- Applying HubLink on other RKGs or KGs for **other domains.**
- Apply HubLink in a **document-based setting** (returns document chunks, uses graph as supporting structure) utilizing the linking feature of HubLink.
- **Extending Hubs** with additional data such as summaries or precomputed insights.

Conclusion

Problem

Literature research is a cumbersome process and current KGQA approaches have difficulties in real world settings.

Idea

Develop a new KGQA approach and taxonomy to foster research on how LLMs can be leveraged to enhance the quality and reliability of content retrieved from RKGs.

Benefit

1. A new approach to reduce barriers for accessing scholarly information and speed up research tasks.
2. A new taxonomy to support development and evaluation of scholarly KGQA systems.
3. Push forward research for training-free and schema-agnostic KGQA approaches in the scholarly domain.

Actions

1. Develop a new KGQA approach that is training-free, schema-agnostic and conducts source aware retrieval.
2. Develop a new Taxonomy to classify questions to understand capabilities of scholarly KGQA systems.
3. Create KGQA datasets on the ORKG to facilitate the evaluation of approaches.

Problem, Idea, Benefit, Actions (PIBA)

References

References

- [Banerjee2024] Banerjee, Debayan. „Knowledge Graph Question Answering With Generative Language Models“. *PQDT - Global*. Ph.D., Universitaet Hamburg (Germany), 2024. https://www.proquest.com/docview/3143985996/abstract/3893F0C5845D4D1DPQ/1.
- [Pan2024] Pan, Shirui, Yizhen Zheng, und Yixin Liu. „Integrating Graphs with Large Language Models: Methods and Prospects“. *IEEE Intelligent Systems* 39, Nr. 1 (January 2024): 64–68. https://doi.org/10.1109/MIS.2023.3332242.
- [Auer2018] Auer, Sören, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. „Towards a Knowledge Graph for Science“. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, 1–6. WIMS ’18. New York, NY, USA: Association for Computing Machinery, 2018. https://doi.org/10.1145/3227609.3227689.
- [Chakraborty2021] Chakraborty, Nilesh, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. „Introduction to Neural Network-Based Question Answering over Knowledge Graphs“. WIREs Data Mining and Knowledge Discovery 11, Nr. 3 (2021): e1389. https://doi.org/10.1002/widm.1389.
- [Yani2022] Yani, Mohammad, Adila Alfa Krisnadhi, and Indra Budi. „A better entity detection of question for knowledge graph question answering through extracting position-based patterns“. Journal of Big Data 9, Nr. 1 (17. June 2022): 80. https://doi.org/10.1186/s40537-022-00631-1.

References

- [Jaradeh2020] Jaradeh, Mohamad Yaser, Markus Stocker, and Sören Auer. „Question Answering on Scholarly Knowledge Graphs“. arXiv, 2. June 2020. https://doi.org/10.48550/arXiv.2006.01527.
- [Devlin2019] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, und Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), published by Jill Burstein, Christy Doran, und Thamar Solorio, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423.
- [Raffel2023] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” arXiv, September 19, 2023. https://doi.org/10.48550/arXiv.1910.10683.
- [Li2024] Li, Mingze, Haoran Yang, Zhaotai Liu, Mirza Mohtashim Alam, Ebrahim, Harald Sack, and Genet Asefa Gesese. “KGMistral: Towards Boosting the Performance of Large Language Models for Question Answering with Knowledge Graph Integration,” 2024. https://openreview.net/forum?id=JzL0qm3YA8.
- [Taffa2023] Taffa, Tilahun Abedissa, and Ricardo Usbeck. “Leveraging LLMs in Scholarly Knowledge Graph Question Answering,” 2023. https://doi.org/10.48550/ARXIV.2311.09841.

References

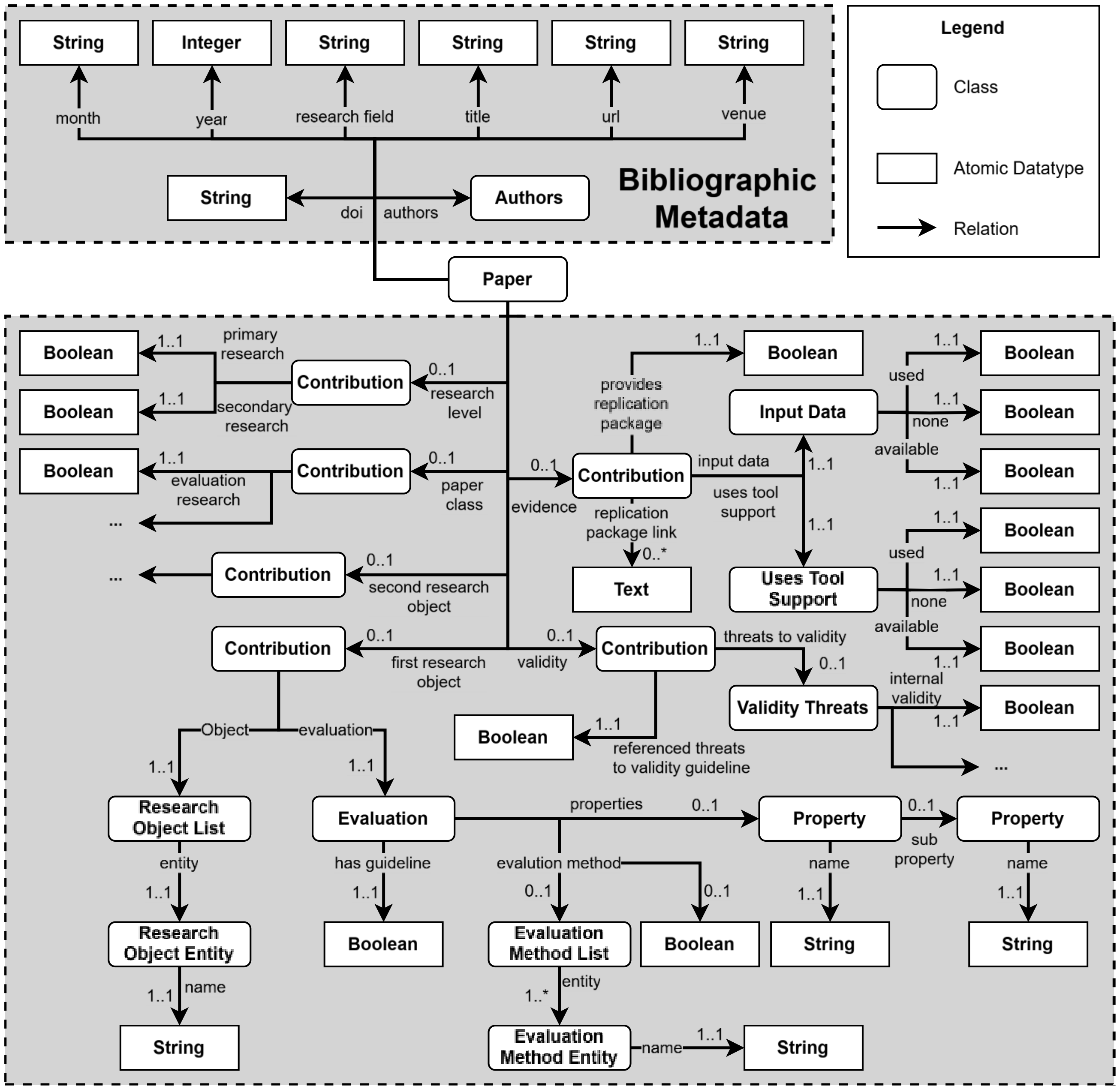
- [Lehmann2024] Lehmann, Jens, Antonello Meloni, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Antonio Salatino, and Sahar Vahdati. "Large Language Models for Scientific Question Answering: An Extensive Analysis of the SciQA Benchmark." In The Semantic Web, edited by Albert Meroño Peñuela, Anastasia Dimou, Raphaël Troncy, Olaf Hartig, Maribel Acosta, Mehwish Alam, Heiko Paulheim, and Pasquale Lisena, 199–217. Cham: Springer Nature Switzerland, 2024. https://doi.org/10.1007/978-3-031-60626-7_11.
- [Jiang2023] Jiang, Jinhao, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. "StructGPT: A General Framework for Large Language Model to Reason over Structured Data." arXiv, October 23, 2023. http://arxiv.org/abs/2305.09645.
- [Zhang2023] Zhang, Lingxi, Jing Zhang, Xirui Ke, Haoyang Li, Xinmei Huang, Zhonghui Shao, Shulin Cao, and Xin Lv. "A Survey on Complex Factual Question Answering." AI Open 4 (January 1, 2023): 1–12. https://doi.org/10.1016/j.aiopen.2022.12.003.
- [Gu2022] Gu, Yu, Vardaan Pahuja, Gong Cheng, and Yu Su. "Knowledge Base Question Answering: A Semantic Parsing Perspective." arXiv, October 24, 2022. https://doi.org/10.48550/arXiv.2209.04994.

References

- [Sui2024] Sui, Yuan, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. "FiDeLiS: Faithful Reasoning in Large Language Model for Knowledge Graph Question Answering." arXiv, October 10, 2024. https://doi.org/10.48550/arXiv.2405.13873.
- [Wen2024] Wen, Yilin, Zifeng Wang, and Jimeng Sun. "MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models." arXiv, March 2, 2024. http://arxiv.org/abs/2308.09729.
- [Baek2023] Baek, Jinheon, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. "Direct Fact Retrieval from Knowledge Graphs without Entity Linking." arXiv, May 21, 2023. https://doi.org/10.48550/arXiv.2305.12416.
- [Konersmann2022] Konersmann, Marco, Angelika Kaplan, Thomas Kuhn, Robert Heinrich, Anne Koziolek, Ralf Reussner, Jan Jurjens, et al. "Evaluation Methods and Replicability of Software Architecture Research Objects." In 2022 IEEE 19th International Conference on Software Architecture (ICSA), 157–68. Honolulu, HI, USA: IEEE, 2022. https://doi.org/10.1109/ICSA53651.2022.00023.

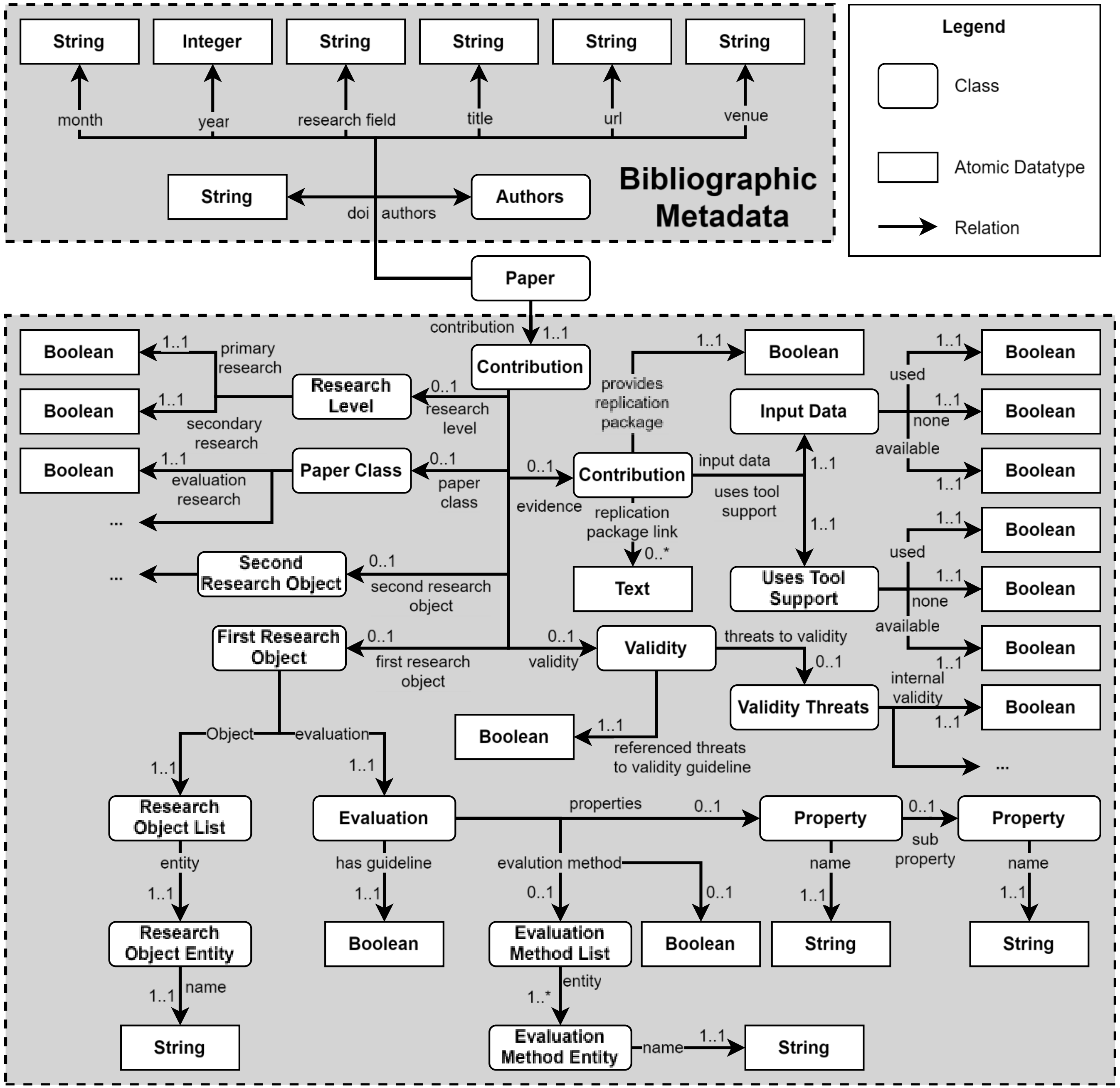
Appendix

Appendix
ORKG Templates



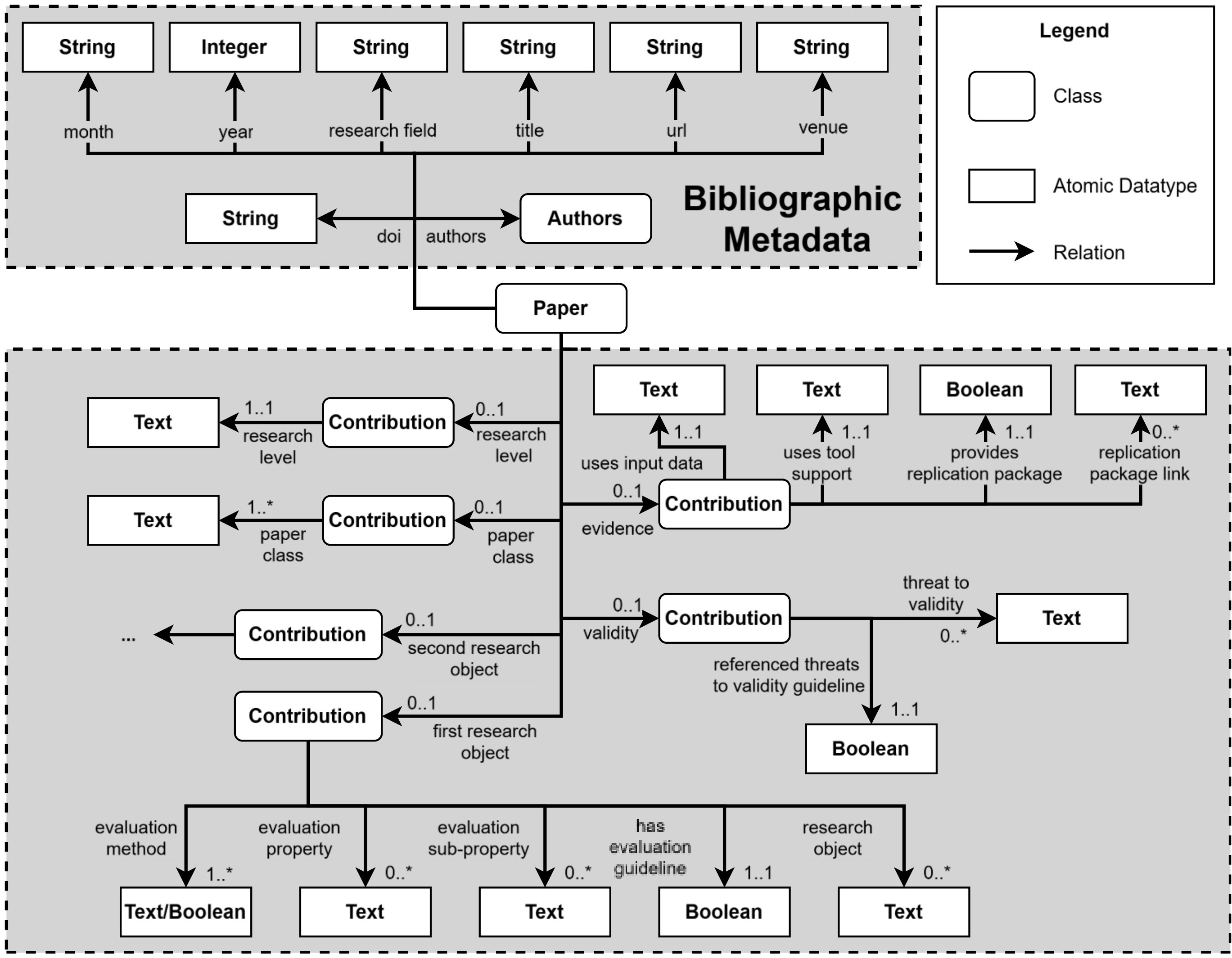
Template of Graph Variant (1)

Appendix
ORKG Templates



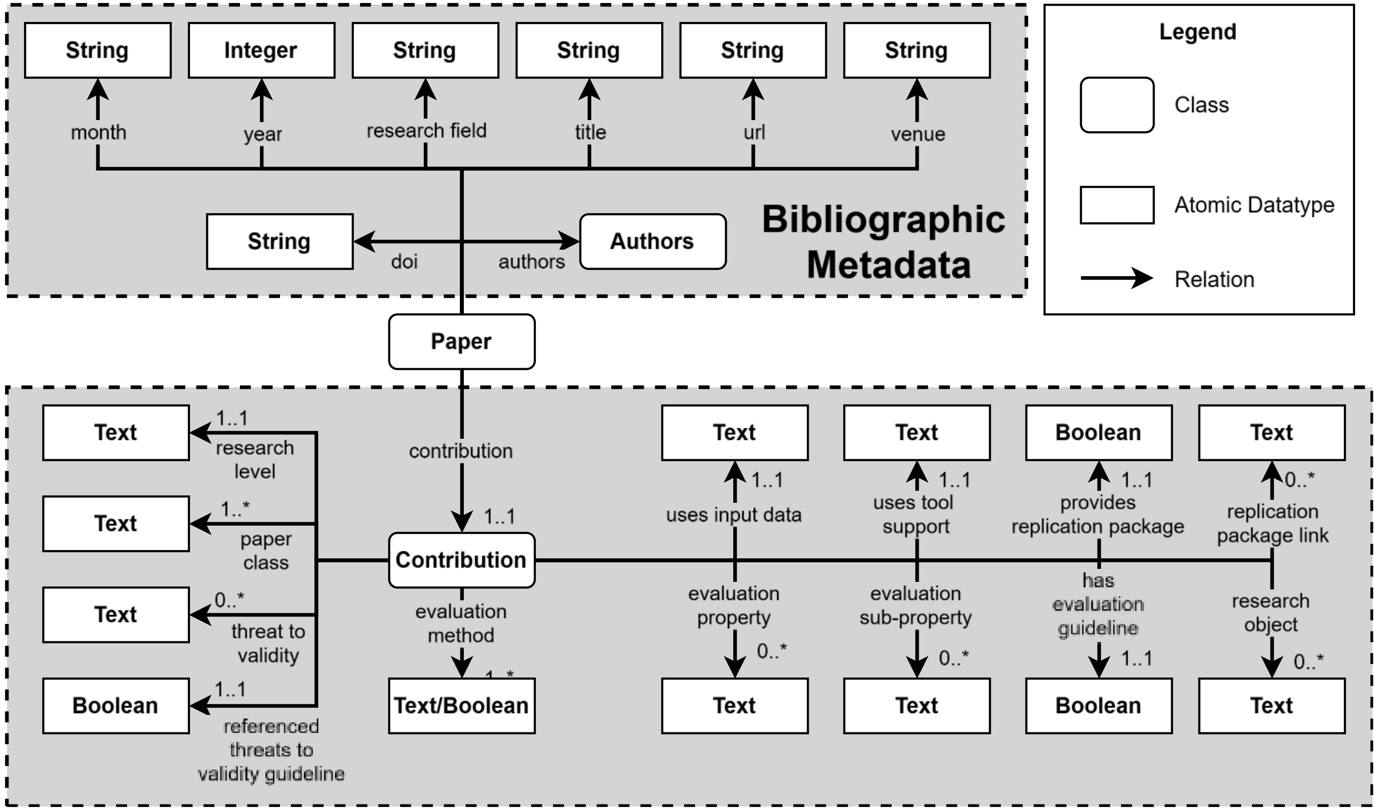
Template of Graph Variant (2)

Appendix
ORKG Templates



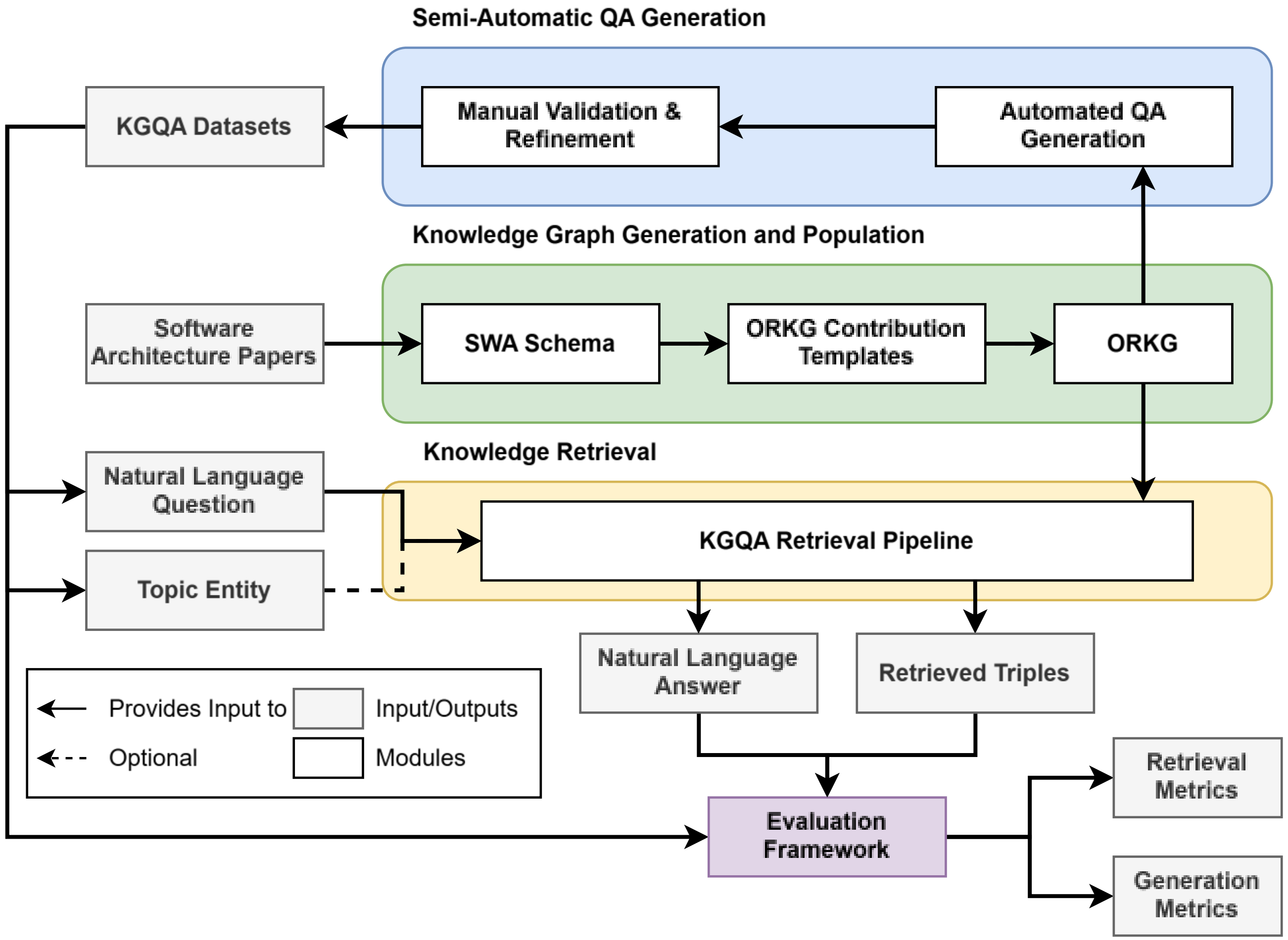
Template of Graph Variant (3)

Appendix
ORKG Templates

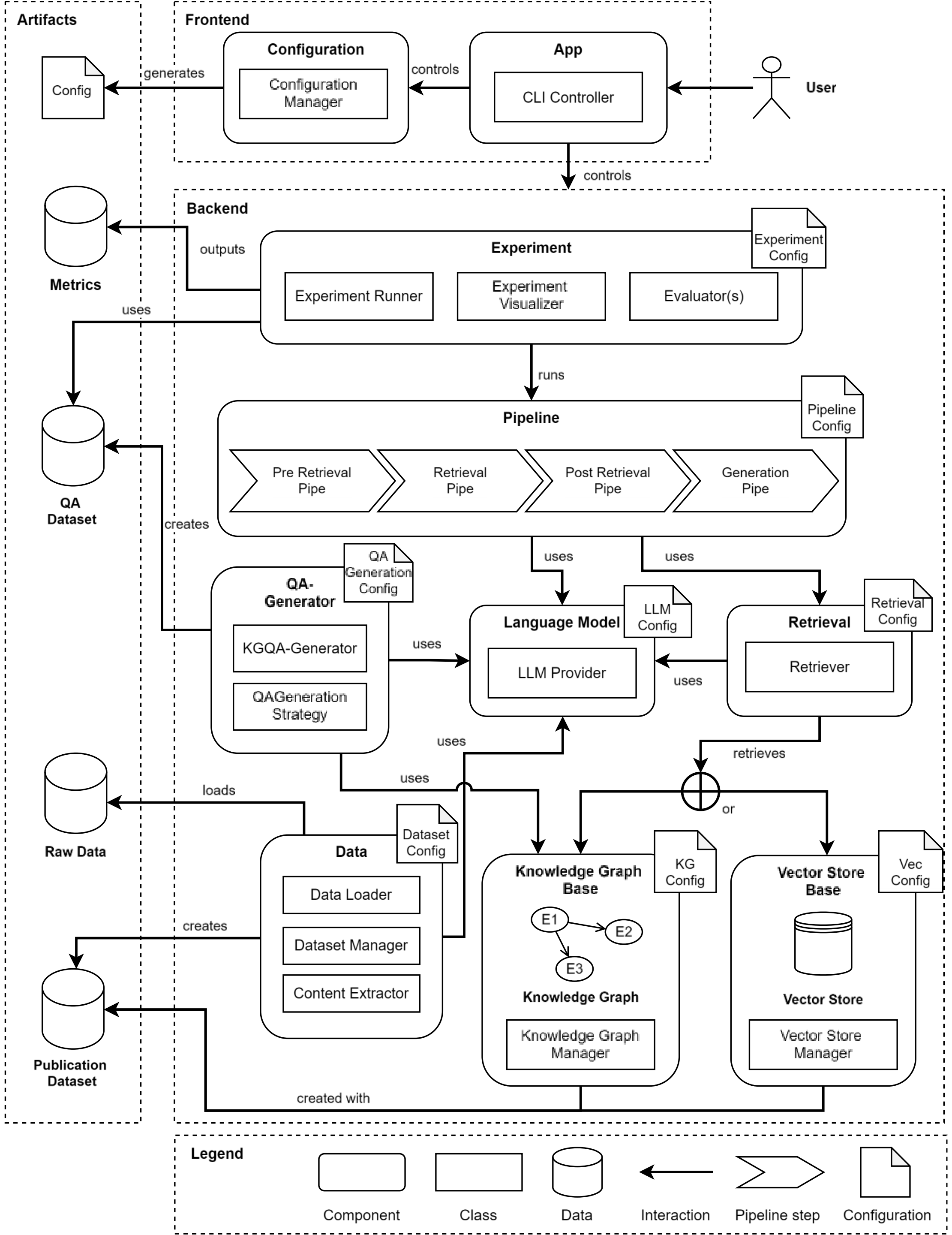


Template of Graph Variant (4)

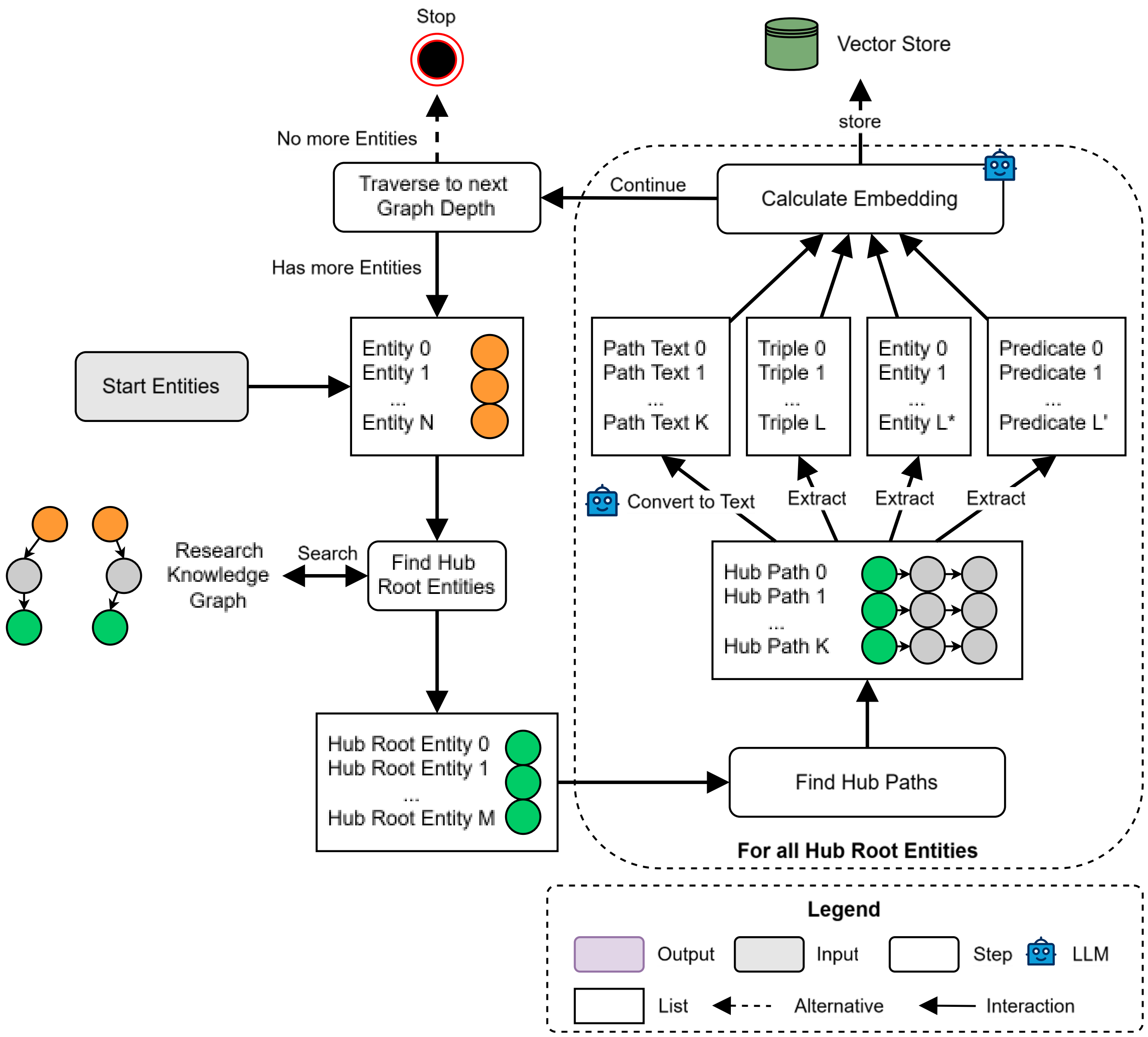
Appendix
Experimentation Workflow



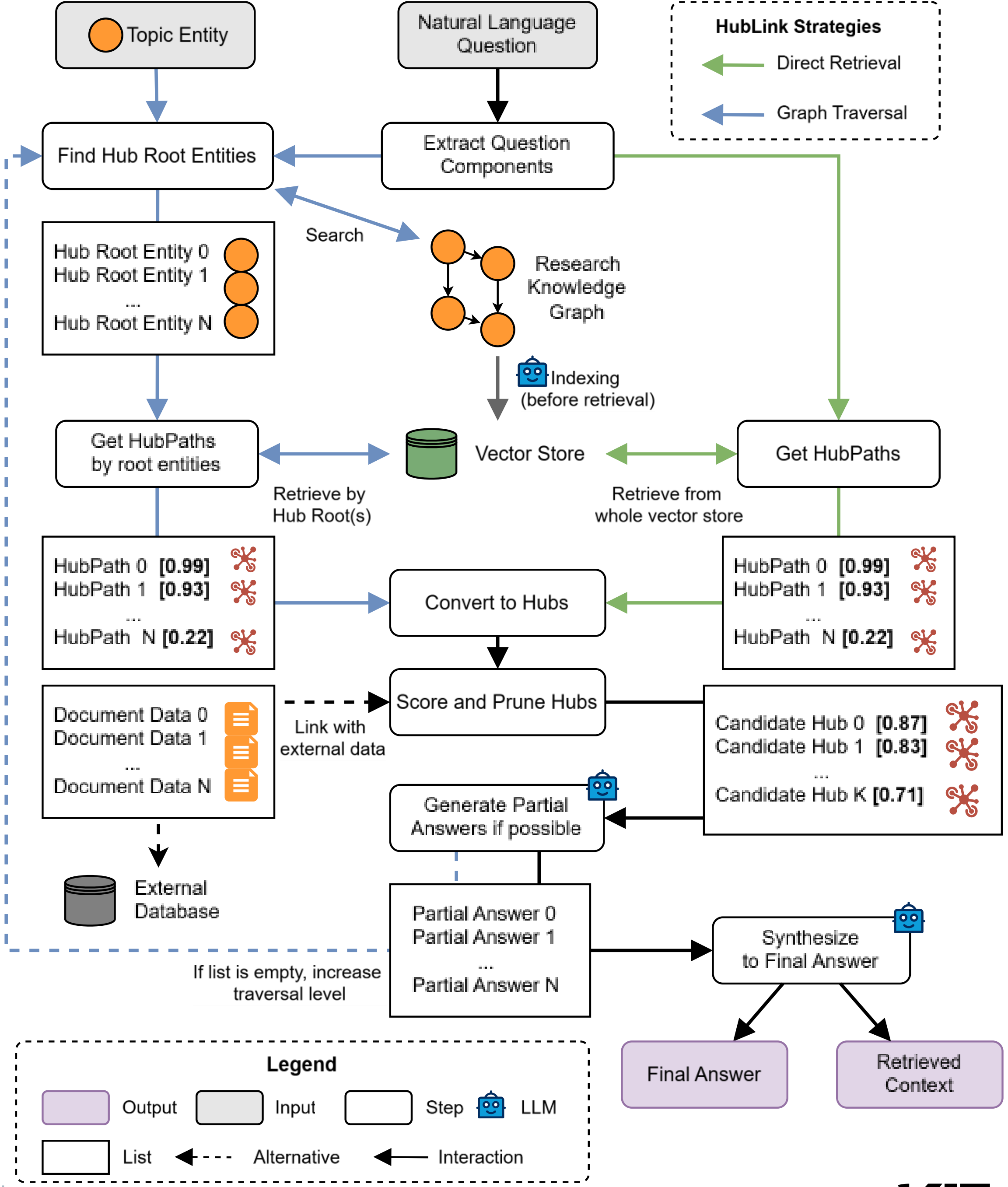
Appendix
SQA Framework



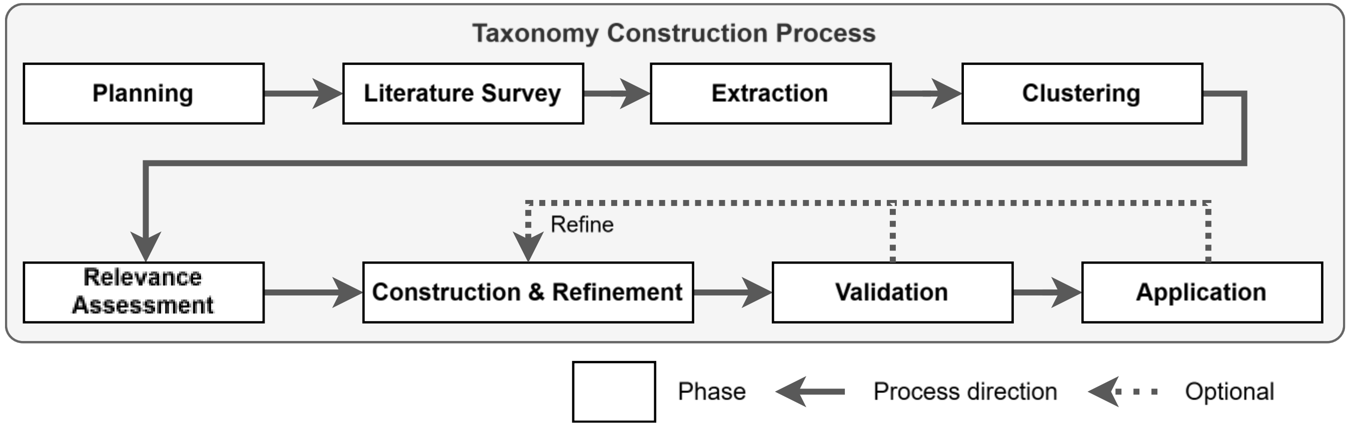
Appendix
HubLink Indexing



Appendix
HubLink Retrieval and Generation



Appendix
Taxonomy Construction Process



- **Planning**
  - planning.md
- **Literature Survey**
  - research_process.json
- **Extraction**
  - extraction.json
  - analysis_of_extraction.ipynb
- **Clustering**
  - cluster_increments.json
  - cluster_analysis.ipynb
- **Relevance Assessment**
  - relevance_assessment.md
- **Validation**
  - Metric Calculations
    - Increment 1
    - ...
  - validation.ipynb
- **Construction and Refinement**
  - taxonomy_increments.json
  - taxonomy_analysis.ipynb
- **Application**
  - calculate_significance.py
  - Taxonomy Applications
    - main_application.json
    - other_application.json
    - ...

Appendix
KGQA Retrieval Taxonomy

Graph Representation
- Single Fact
- Multi Fact

Answer Type
- Named Entity
- Description
- Temporal
- Quantitative
- Boolean
- Other Type

Answer Format
- Simple
- Enumerative
- Explanatory
- Other Format

Condition Type
- Named Entity
- Description
- Temporal
- Quantitative
- Other Type

Retrieval Operation
- Basic
- Relationship
- Negation
- Aggregation
- Counting
- Superlative
- Ranking
- Comparison

Intention Count
- Single Intention
- Multiple Intentions

Question Goal
- Information Lookup
- Reasoning
- Problem Solving
- Problematization
- Improvement
- Prediction
- Other Goal

Answer Credibility
- Subjective
- Objective
- Normative

Appendix
Retrieval Operation Performance

Retrieval Operation	Recall	Precision	F1	Hits@10	Map@10	MRR@10	EM@10
HubLink (T)							
basic	**0.917**	**0.382**	**0.480**	**0.917**	**0.445**	0.490	**0.389**
aggregation	0.810	0.209	0.285	0.497	0.225	0.347	0.240
counting	0.840	0.275	0.372	0.644	0.357	0.526	0.340
ranking	0.817	0.321	0.414	0.561	0.360	**0.576**	0.363
comparative	0.742	0.262	0.366	0.456	0.320	0.560	0.296
relationship	0.628	0.254	0.314	0.410	0.298	0.528	0.331
negation	0.584	0.072	0.122	0.244	0.125	0.419	0.144
superlative	0.656	0.129	0.193	0.319	0.207	0.540	0.237
HubLink (D)							
basic	**0.861**	0.217	0.276	**0.611**	0.297	0.332	0.228
aggregation	0.730	0.166	0.217	0.388	0.188	0.365	0.200
counting	0.723	0.218	0.293	0.481	0.287	0.410	0.253
ranking	0.659	0.221	0.278	0.428	0.278	0.494	0.269
comparative	0.701	0.314	**0.376**	0.444	0.287	0.537	0.339
relationship	0.689	**0.347**	**0.376**	0.456	**0.314**	0.627	**0.411**
negation	0.639	0.065	0.118	0.325	0.169	0.534	0.200
superlative	0.690	0.133	0.204	0.332	0.229	**0.635**	0.244

Appendix
Use Case Performance

Use Case	Recall	Precision	F1	Hits@10	Map@10	MRR@10	EM@10
HubLink (T)							
1	0.800	**0.507**	**0.575**	**0.767**	**0.557**	**0.644**	**0.552**
2	**0.848**	0.252	0.364	0.729	0.301	0.341	0.281
3	0.768	0.252	0.343	0.507	0.268	0.543	0.287
4	0.663	0.198	0.277	0.395	0.266	0.561	0.255
5	0.702	0.122	0.186	0.350	0.184	0.408	0.213
6	0.779	0.206	0.286	0.428	0.278	0.512	0.257
HubLink (D)							
1	**0.791**	**0.450**	**0.510**	**0.745**	**0.495**	0.556	**0.489**
2	0.715	0.073	0.127	0.410	0.155	0.281	0.111
3	0.675	0.265	0.332	0.444	0.234	0.459	0.298
4	0.543	0.195	0.225	0.302	0.184	0.444	0.234
5	0.756	0.144	0.191	0.317	0.183	0.481	0.236
6	0.790	0.213	0.295	0.463	0.341	**0.691**	0.284

Appendix
Retrieval Operation Performance

Use Case 1

Input: Metadata
Output: Metadata

Example:
- Find publications by a specific author
- Search for paper titles by keyword

Use Case 2

Input: Metadata
Output: Content

Example:
- Ask for the conclusions of a specific paper
- Ask what research problems an author has addressed

Appendix
Retrieval Operation Performance

Use Case 3

Input: Content
Output: Metadata

Example:
- Find publications that contain a specific evaluation method
- Search for papers treating a certain research problem

Use Case 4

Input: Content
Output: Content

Example:
- Find reference architectures proposed for a specific problem
- Search for definitions/explanations of a concept

Appendix
Retrieval Operation Performance

Use Case 5

Input: Metadata + Content
Output: Content

Example:
- Summary of conclusions in a specific time frame for a research problem
- Content answers constrained by both metadata and content

Use Case 6

Input: Metadata + Content
Output: Metadata

Example:
- Find publications that applied a method in a specific field
- Titles of papers matching both content and metadata constraints

Appendix
Retrieval Operation Performance

1. In which venue has the paper '[Metadata: paper title]' been published?
2. Which publications have been published by the author [Metadata: author name] in the year [publication year]?
3. Which threats to validity does the paper with the title '[Metadata: paper title]' discuss? Rank the threats to validity in descending alphabetical order.
4. How many times have [Content Data: research object name] been investigated in [Metadata: year] in comparison to [Metadata: year]?What evaluation methods have been used by the author [Metadata: author name] and were not applied with an evaluation guideline?
5. What is the distribution of papers that investigate the object [Content Data: research object name] between [Metadata: year] and [Metadata: year]?