Do Graph-based Approaches Outperform Vector-based Approaches in Retrieval Augmented Generation for Complex Question Answering? - A Study Using Wikipedia and the Mintaka Dataset

Philippe Saadé

03/06/2024, Master's Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Motivation

Research Questions

- Existing RAG Approaches
- Datasets
- Evaluation Technique
- Results

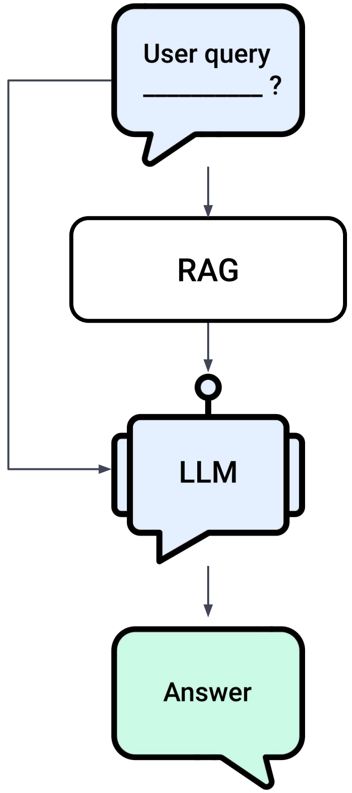
Insights

- Conclusion
- Limitations & Future Work

Motivation

Do Graph-based Approaches Outperform Vector-based Approaches in Retrieval Augmented Generation for Complex Question Answering?

- Compare different vector-based and graph-based RAG techniques to analyze their abilities in augmenting LLMs.
- Understand how vector and graph database structures influence the performance across various question complexities.
- This study evaluates RAG on the general domain of Wikipedia and Wikidata.



Outline

Motivation

Research Questions
- Existing RAG Approaches
- Datasets
- Evaluation Technique
- Results

Insights
- Conclusion
- Limitations & Future Work

Research Questions

1. **What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?**

2. How to align a vector database with a graph database to include the same information and be comparable in terms of retrieval performance?

3. How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?

4. How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?

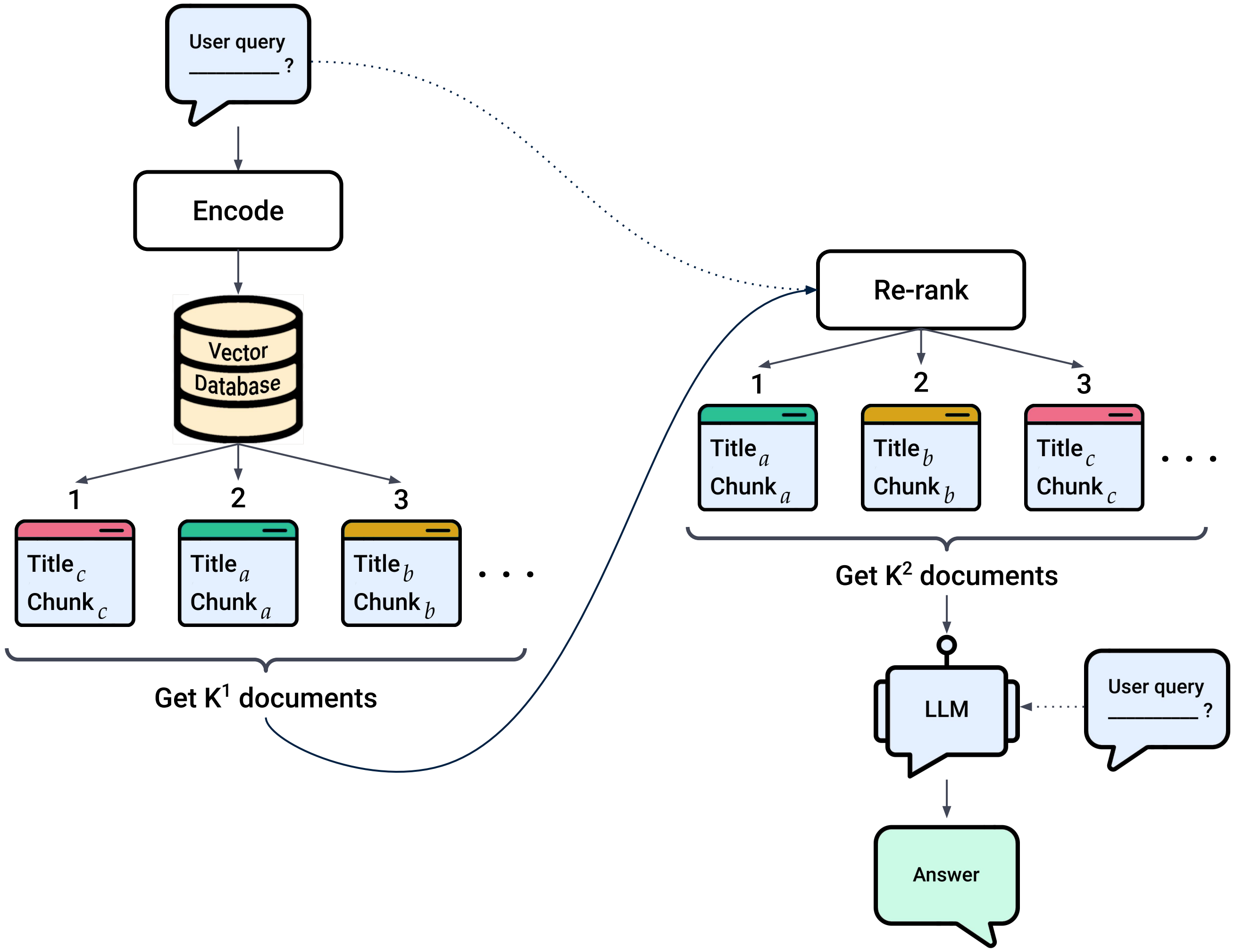
Vector-based RAG

Dense Passage Retrieval (DPR)

1. Encode the user query and retrieve the top K^1 similar documents from the Vector database.

2. Use a Re-ranker model to reorder the relevance of documents given the user query.

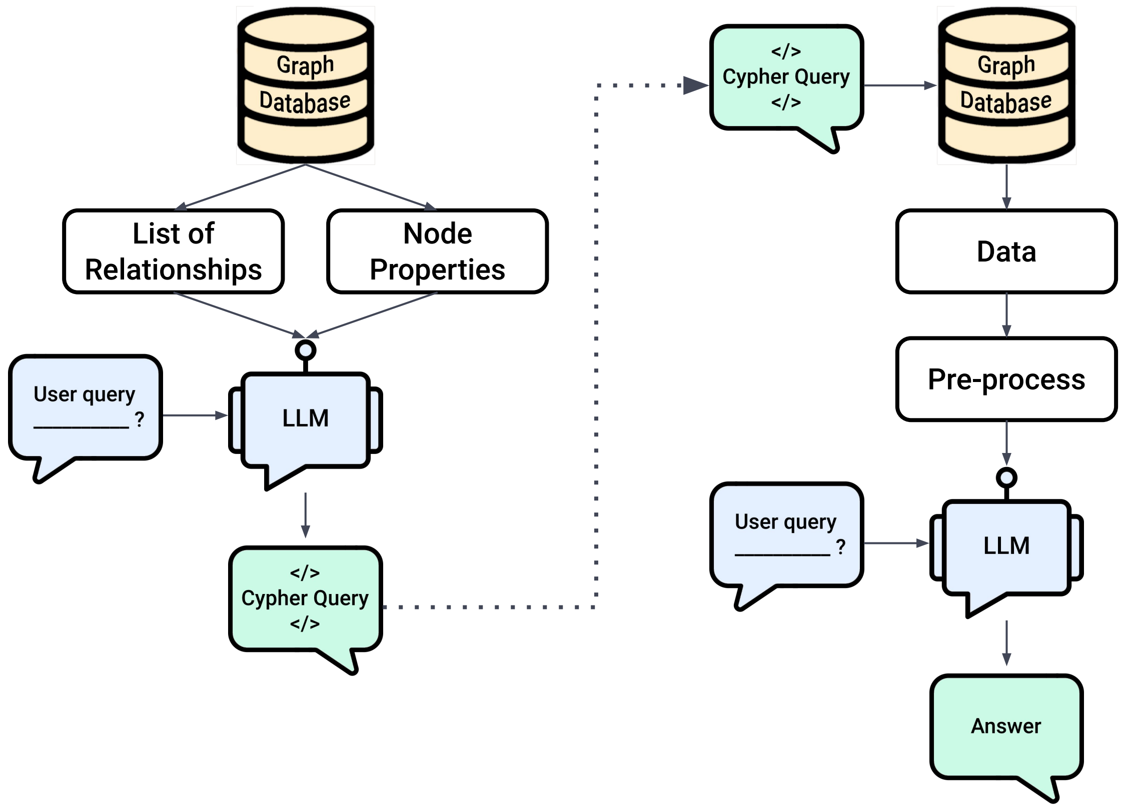
3. Provide the LLM with the top K^2 documents after re-ranking



Cypher Query Generation

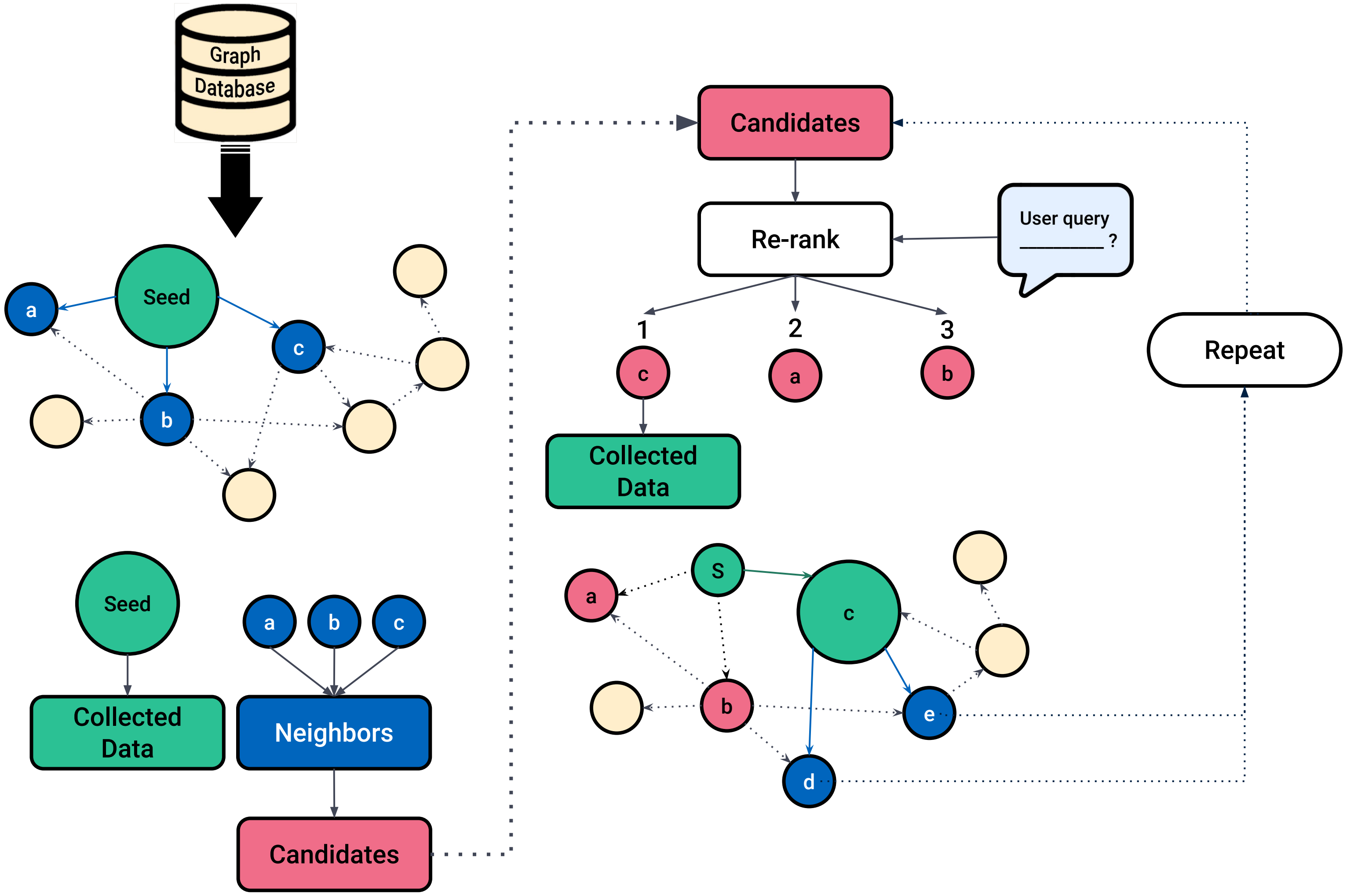
1. Given the Graph database structure and the user query, generate an Cypher query using an LLM.

2. Execute the Cypher query and provide the data to the LLM



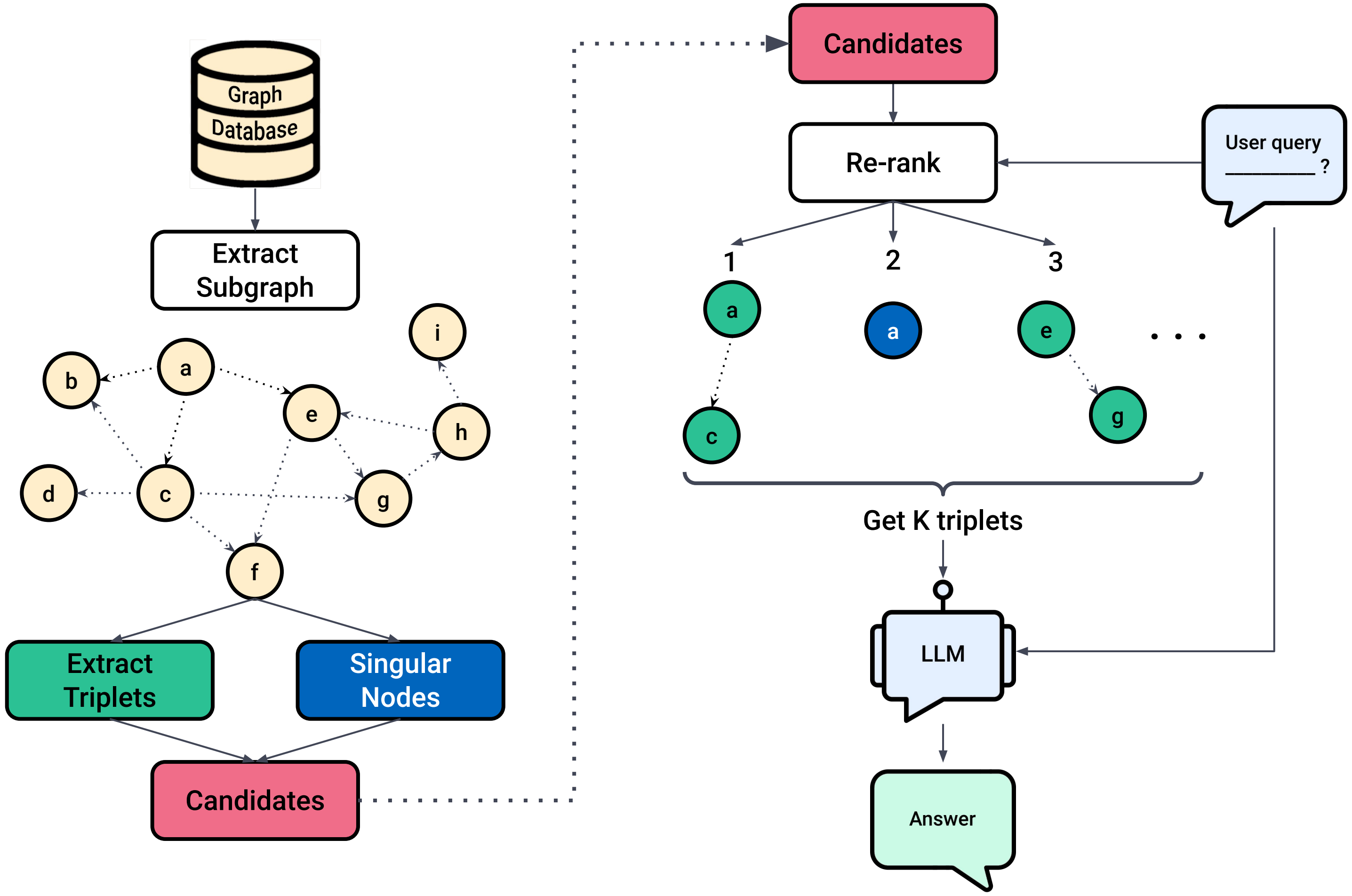
Graph Traversal

1. Start with a seed node that is relevant to the question.
2. Add the neighboring nodes to the list of candidates
3. At each iteration, rank the candidates and extract the most relevant one.
4. Add the top candidate to the collected data and its neighbors to the list of candidates.
5. Repeat.



Triplets Extraction

1. Extract a relevant Subgraph from the Graph database.
2. Extract all the triplets contained within the Subgraph along with the nodes.
3. Rank the content of the triplets and nodes
4. Augment the LLM with the top ranked candidates



Research Questions

1. What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?

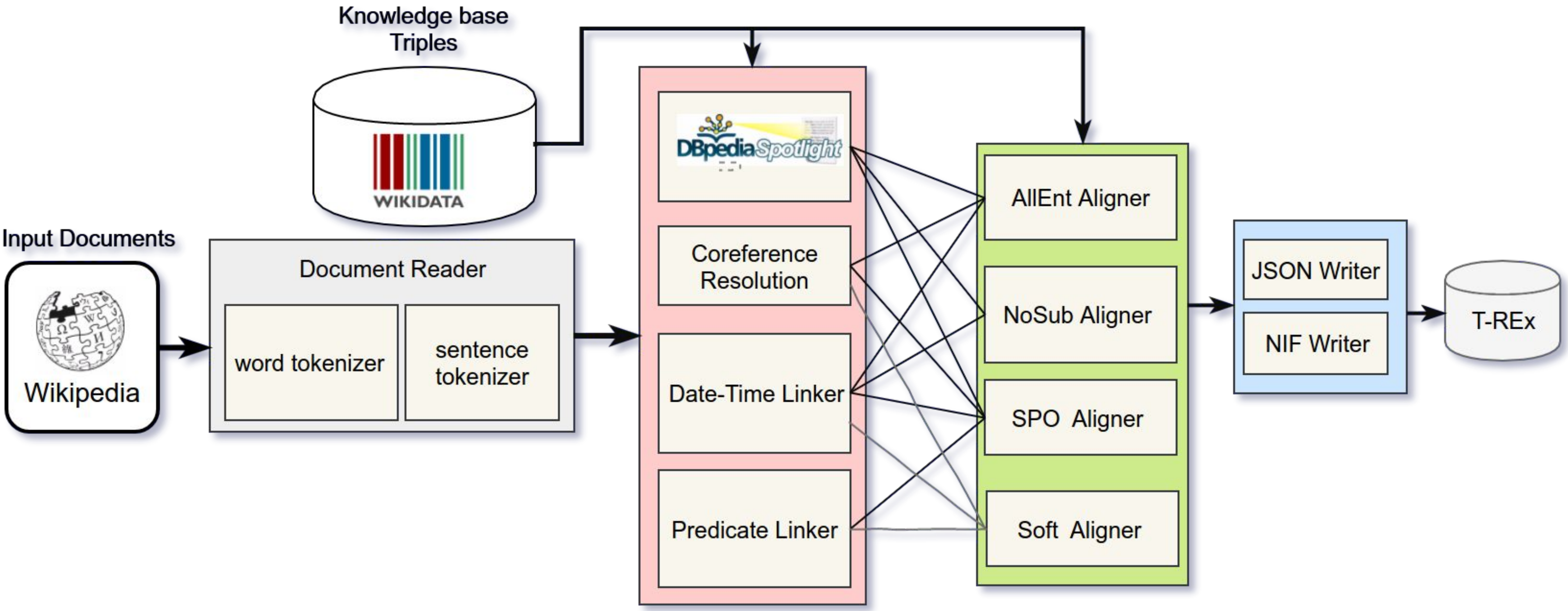
2. **How to align a vector database with a graph database to include the same information and be comparable in terms of retrieval performance?**

3. How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?

4. How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?

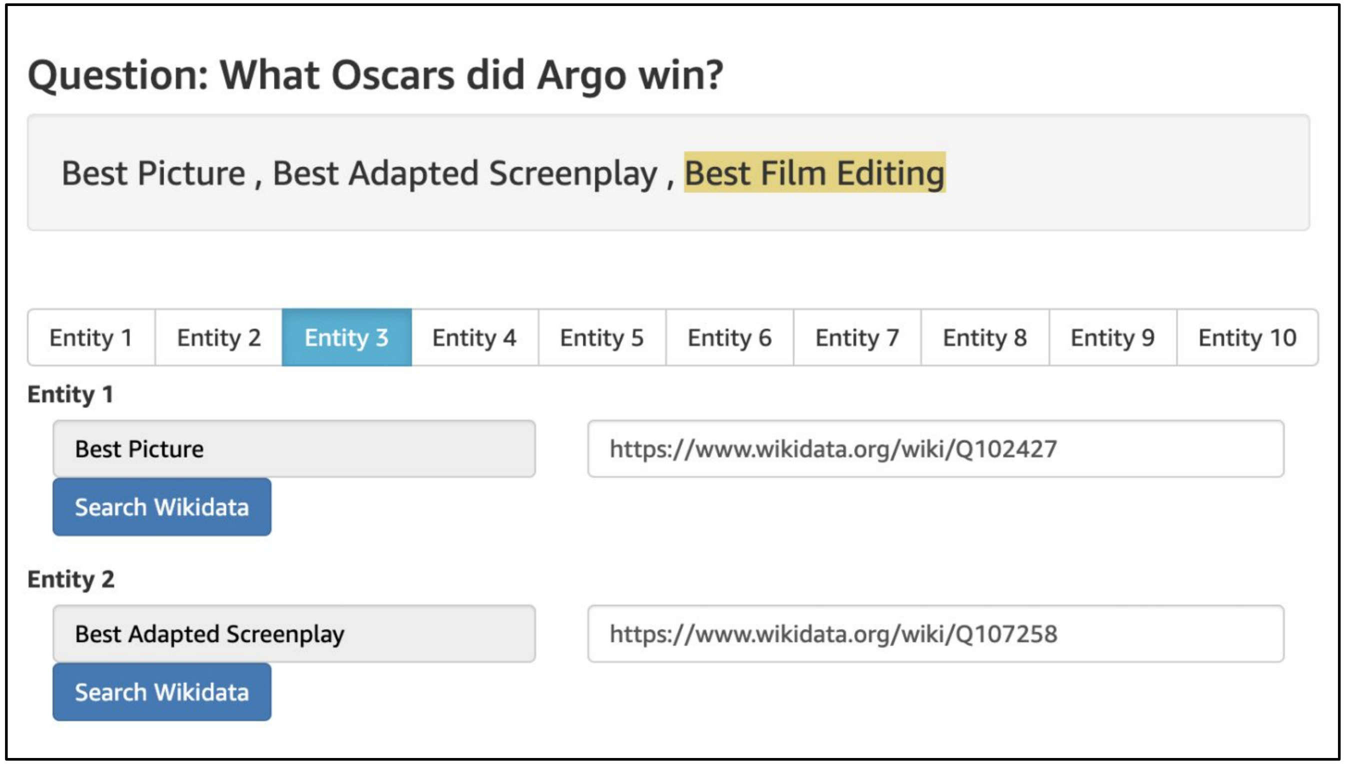
Setup

T-REx Dataset



Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., & Simperl, E. (2018, May). T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Mintaka Dataset



Benefits of using Mintaka:

- Answers are connected to Wikidata entities
- Questions categorized by type of answer or difficulty

Sen, P., Aji, A. F., & Saffari, A. (2022). Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613*.

Mintaka Dataset

Questions Complexities

Type	Description	Example
Yes/No	Answer is a Yes or No	Has Lady Gaga ever made a song with Ariana Grande?
Comparative	Compare 2 items by an attribute	Is Mont Blanc taller than Mount Rainier?
Generic	Simple questions	Where was Michael Phelps born?
Intersection	Requires multiple conditions	Which movie was directed by Denis Villeneuve and stars Timothee Chalamet?
Ordinal	Based on item's position in a list	Who was the last Ptolemaic ruler of Egypt?
Count	Answer requires counting	How many astronauts have been elected to Congress?
Difference	Contains a negation	Which Mario Kart game did Yoshi not appear in?
Superlative	Max or Min of given attribute	Who was the youngest tribute in the Hunger Games?
Multi-hop	Requires multiple steps to answer	Who was the quarterback of the team that won Super Bowl 50?

Sen, P., Aji, A. F., & Saffari, A. (2022). Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613*.

Research Questions

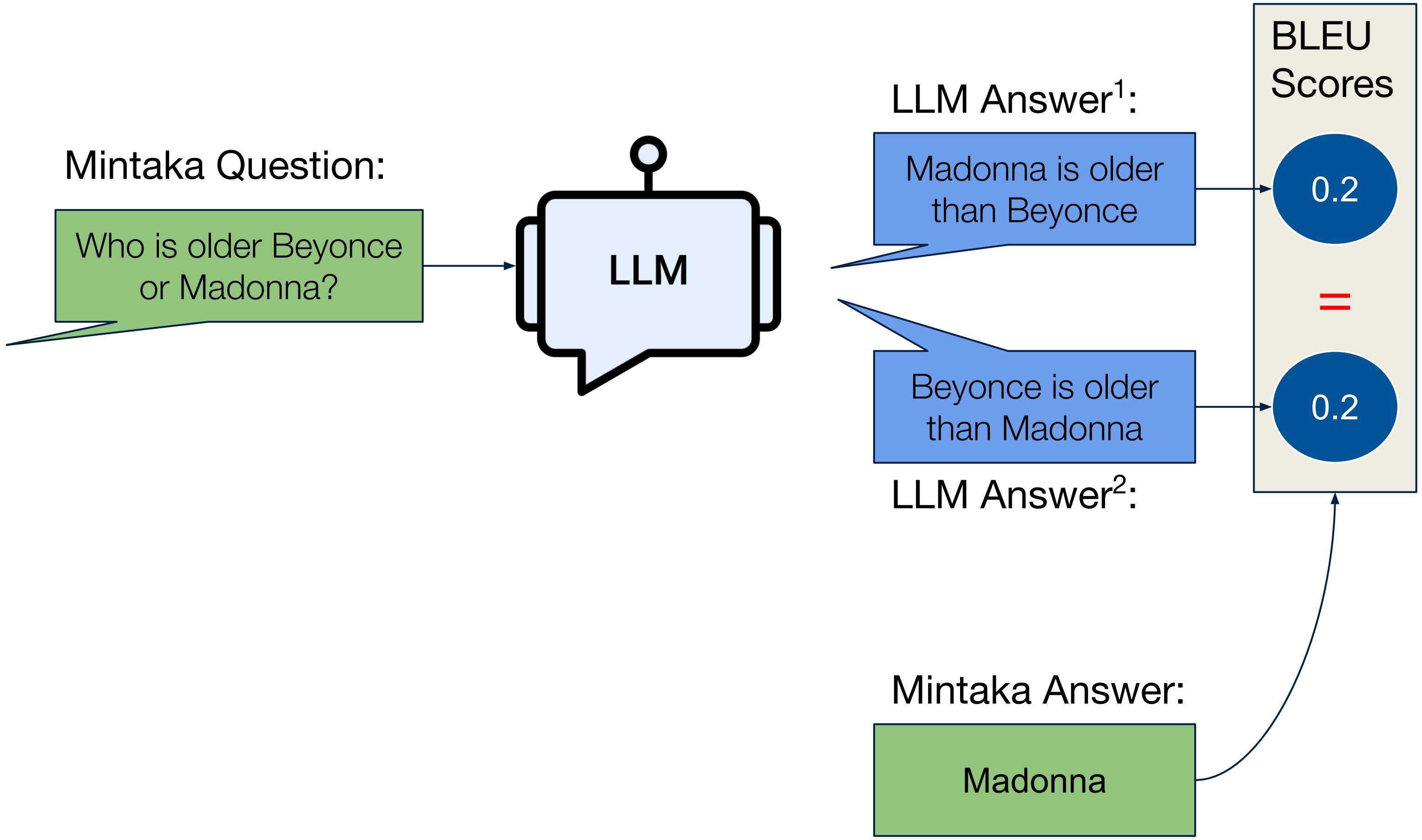
1. What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?

2. How to align a vector database with a graph database to include the same information and be comparable in terms of retrieval performance?

3. **How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?**

4. How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?

Evaluation Metric: BLEU, ROUGE & BERT Score



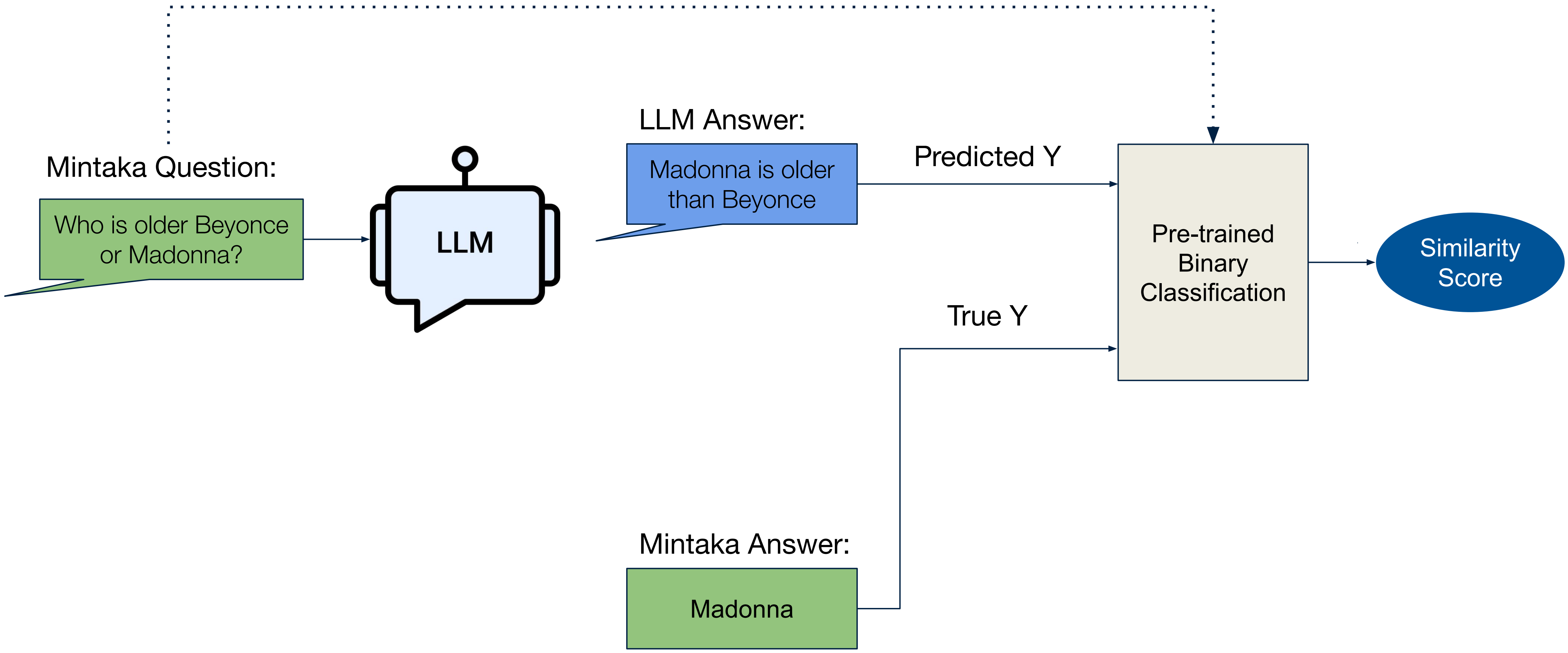
Limitations of BLEU & ROUGE:

- Doesn't work on comparative questions, numbers, dates…

Limitations of BERT Score:

- Designed for text generation & summarization. Doesn't work for short answers

Evaluation Metric: Classification



Research Questions

1. What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?

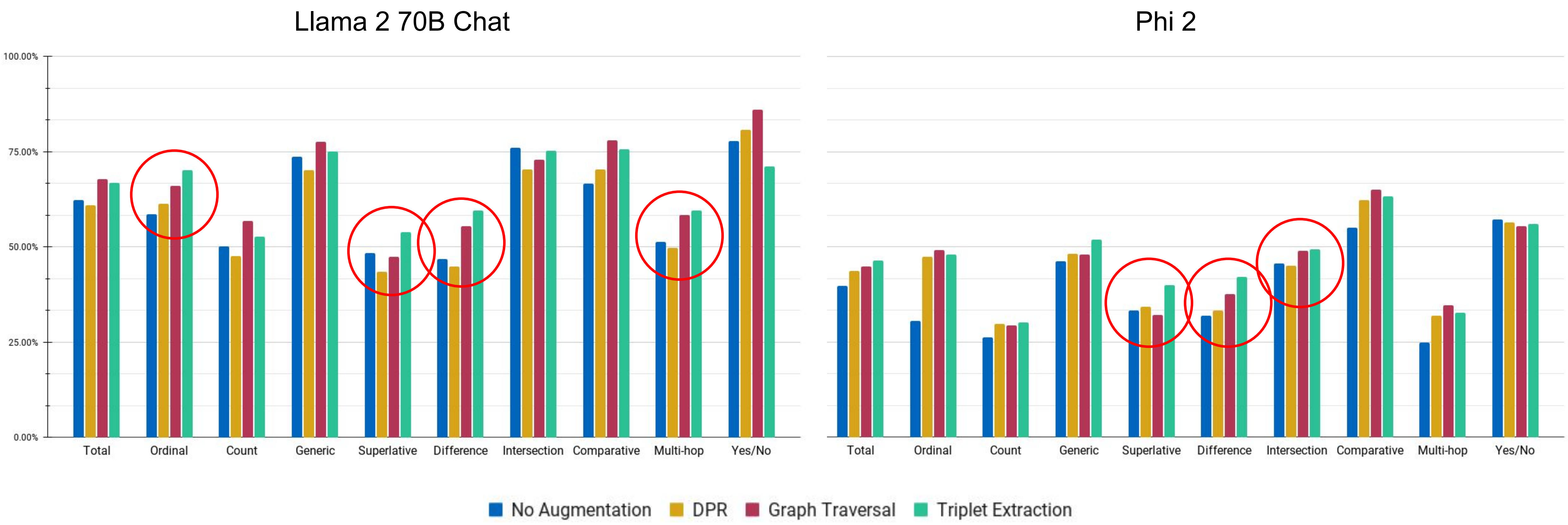
2. How to align a vector database with a graph database to include the same information and be comparable in terms of retrieval performance?

3. How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?

4. **How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?**

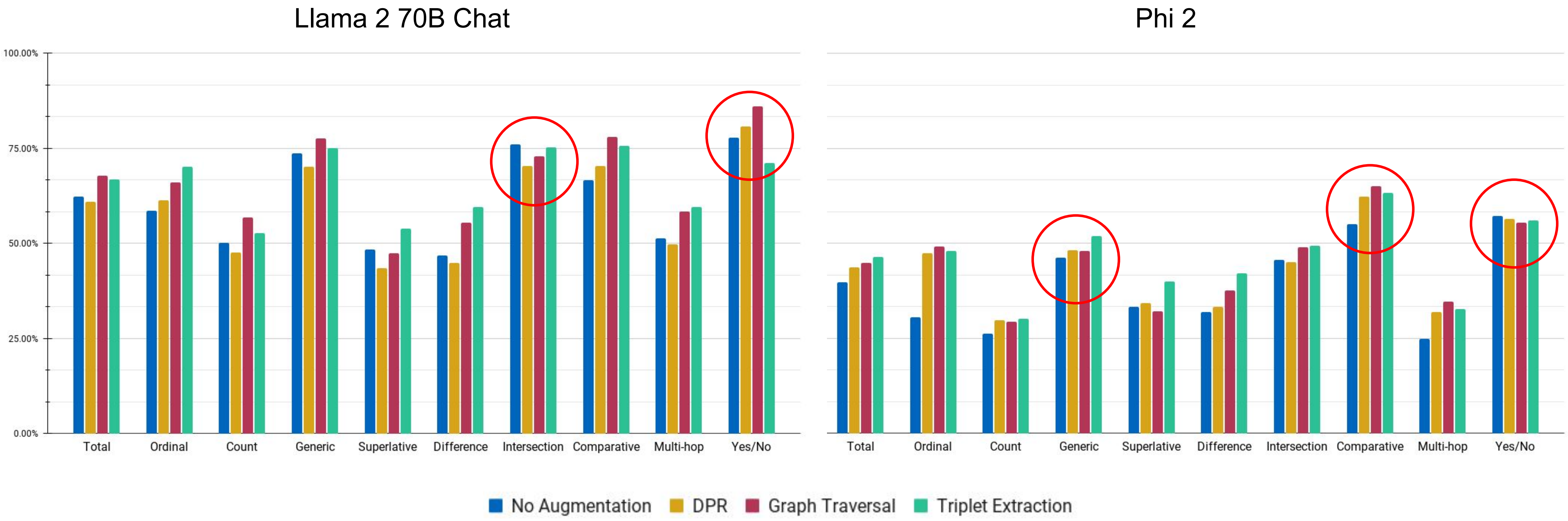
Results

per question complexity



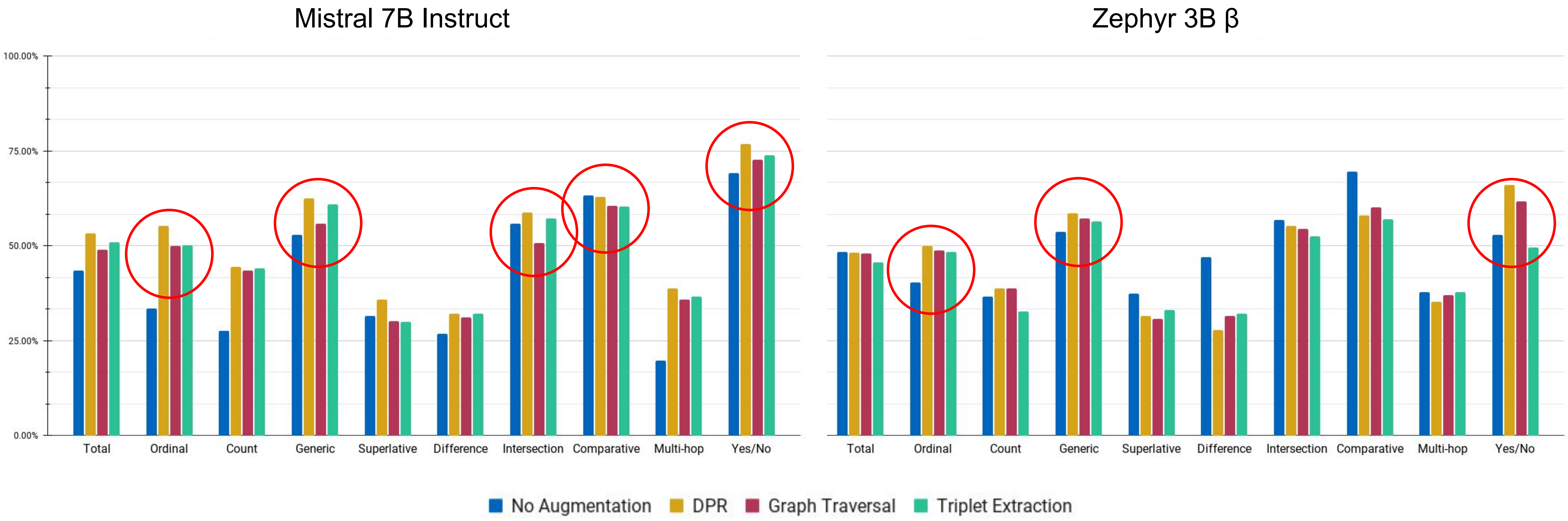
Results

per question complexity



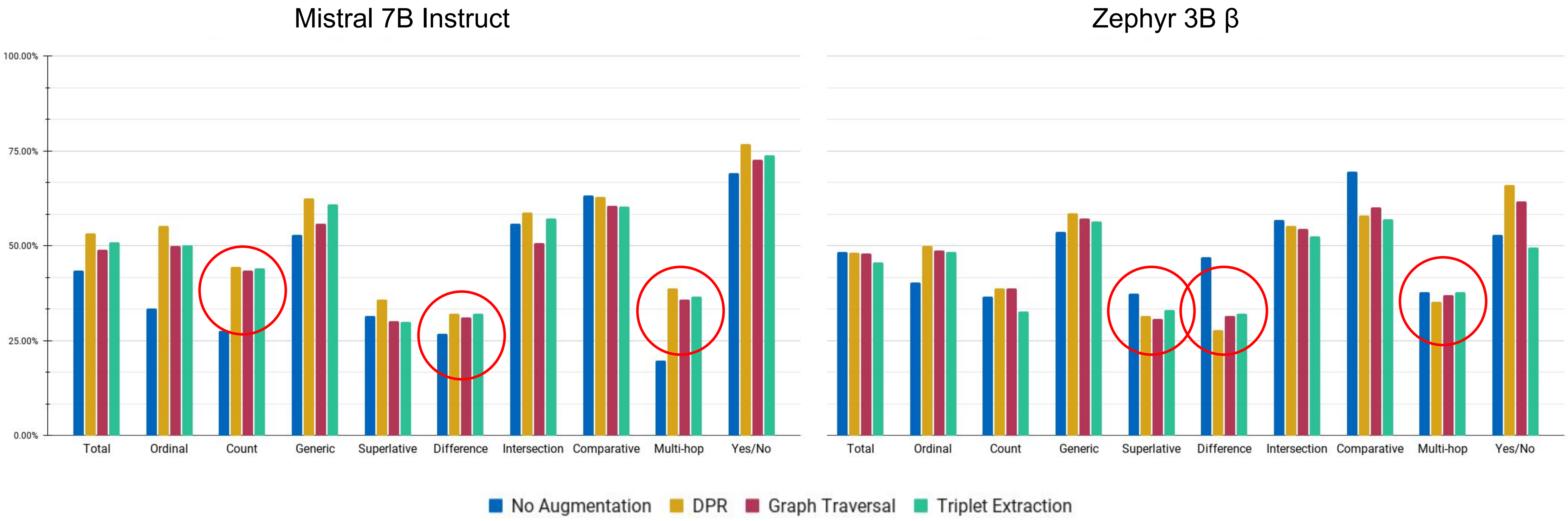
Results

per question complexity



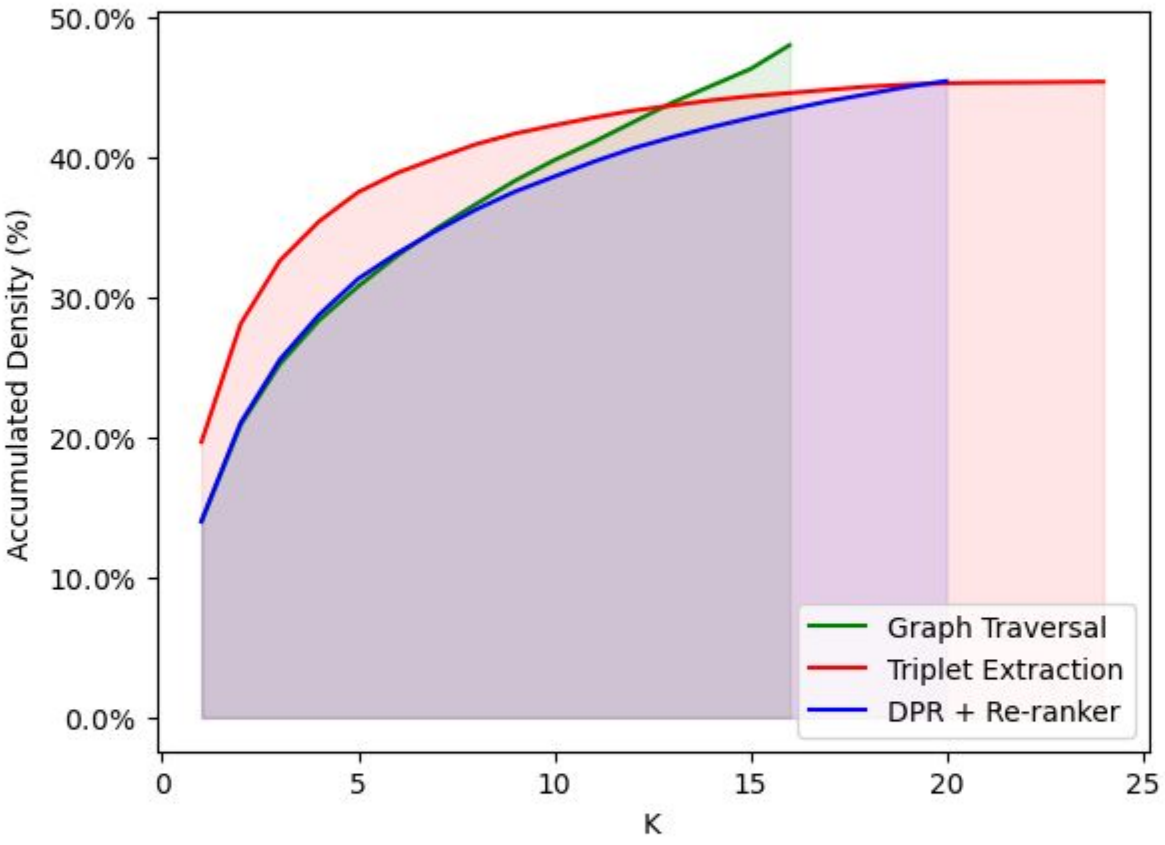
Results

per question complexity



Results

of retrieval only



Cumulative accuracy of the retrieval of each RAG technique

RAG	MRR@15	NDCG@15
Dense Passage Retrieval	0.41426	0.35626
Graph Traversal	0.40982	0.36820
Triplets Extraction	0.48982	0.41803

MRR and NDCG of each RAG technique

Outline

Motivation

Research Questions

- Existing RAG Approaches
- Datasets
- Evaluation Technique
- Results

Insights

- Conclusion
- Limitations & Future Work

Conclusion

1. Including edge information improves the accuracy of data ranking
2. Vector-based RAG excels at accurately retrieving an initial document, making it more suitable for simple questions.
3. Graph-based RAG is more effective at retrieving an accurate collection of documents over time, which is suitable for addressing more complex questions.
4. An LLM with higher understanding capabilities is necessary to fully benefit from the information provided by graph-based RAG techniques.

Limitations & Future Work

1. The re-ranker model used in all RAG techniques has a short context window.
   Use a generative LLM with a larger context window for ranking data and consider previous retrieval steps.

2. For Cypher generation, we were limited to less advanced LLMs.
   Use an LLM fine-tuned on Cypher queries and refine the generated queries by prompting with query results/errors

3. This study is limited to the general domain of Wikidata and Wikipedia.
   Analyse RAG systems with information not found in the training data of the LLMs.

Philippe Saadé

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

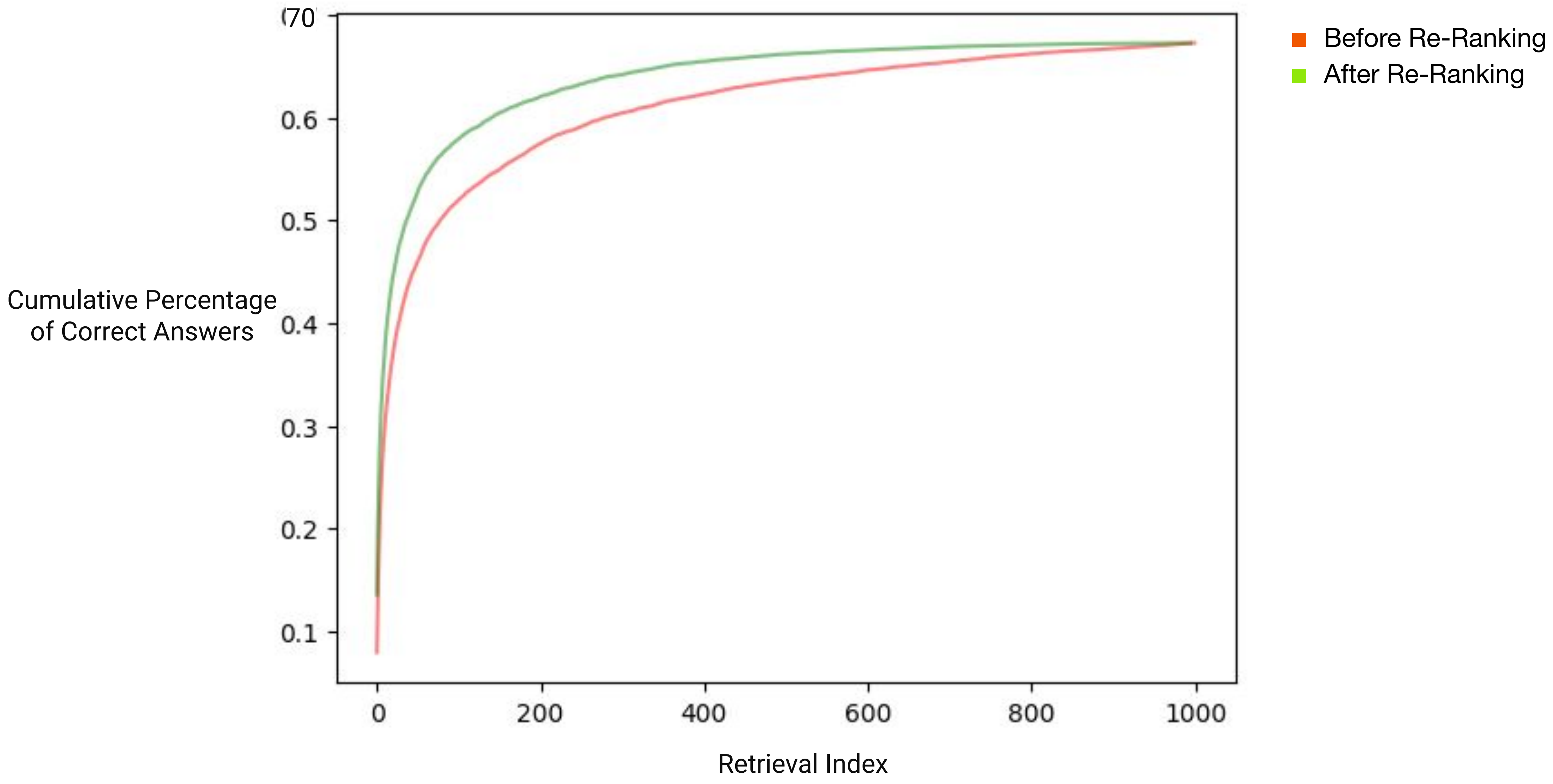
+49.89.289.

wwwmatthes.in.tum.de

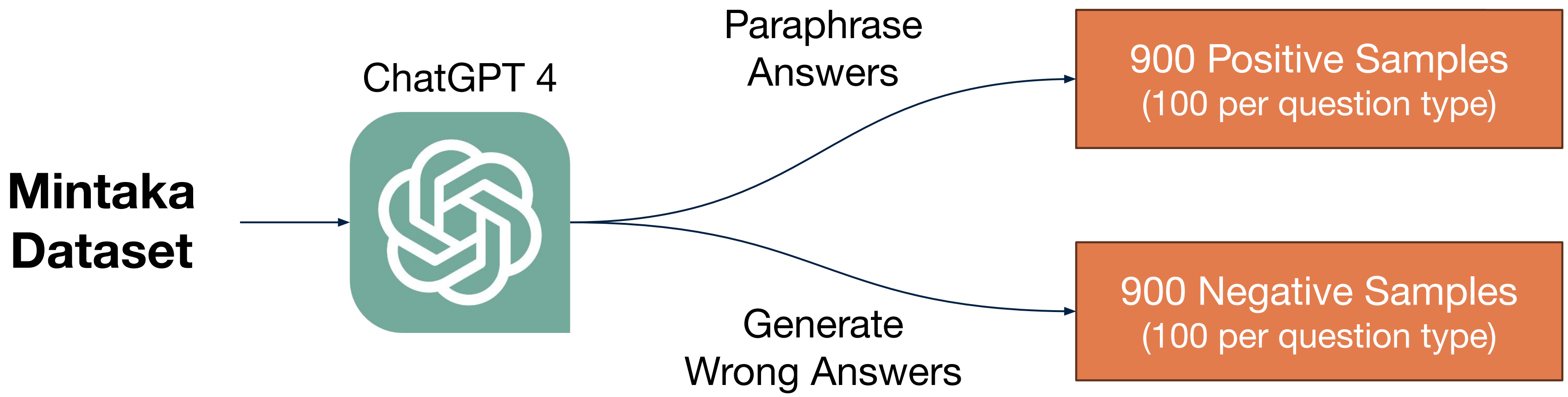
Results: Vector Database

Chunking	**Embedding Model**	MRR	MRR after re-ranking
Split by words	multi-qa-mpnet-base-dot-v1	0.09334	0.11414
Split by tokens	msmarco-distilbert-base-tas-b	0.12399	0.20302
Split by sentences using NLTK	multi-qa-mpnet-base-dot-v1	0.13746	0.21251
Split by sentences using Spacy	msmarco-distilbert-base-tas-b	0.12853	0.20990
Split by sentences using Spacy	multi-qa-mpnet-base-dot-v1	0.14817	0.21310

Results: Vector Database

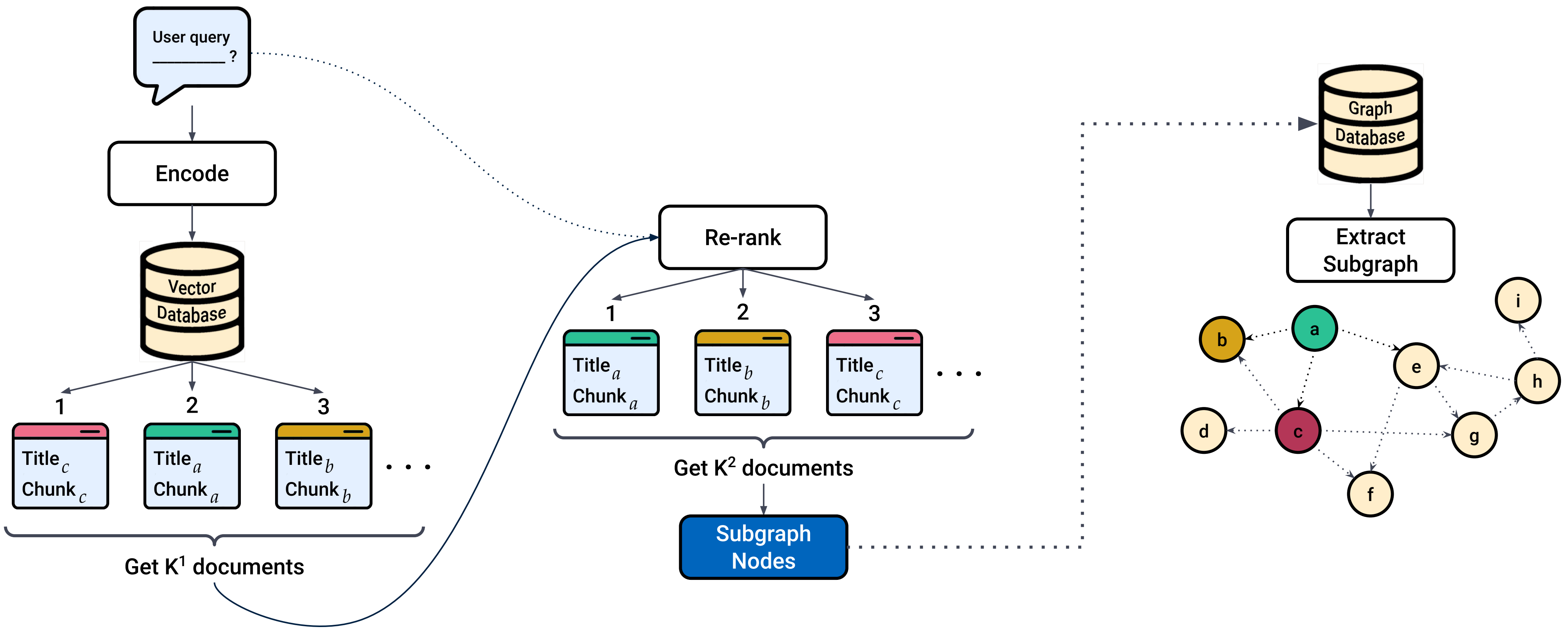


Evaluation Metric: Classification



Zero-Shot Classifier Model	Accuracy (Threshold 0.5)	Average Scores	Prediction Time
facebook/bart-large-mnli	95.5%	0.9467	0.1575 sec
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli	95.6%	0.9476	0.059 sec
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	97.9%	0.9744	0.18 sec

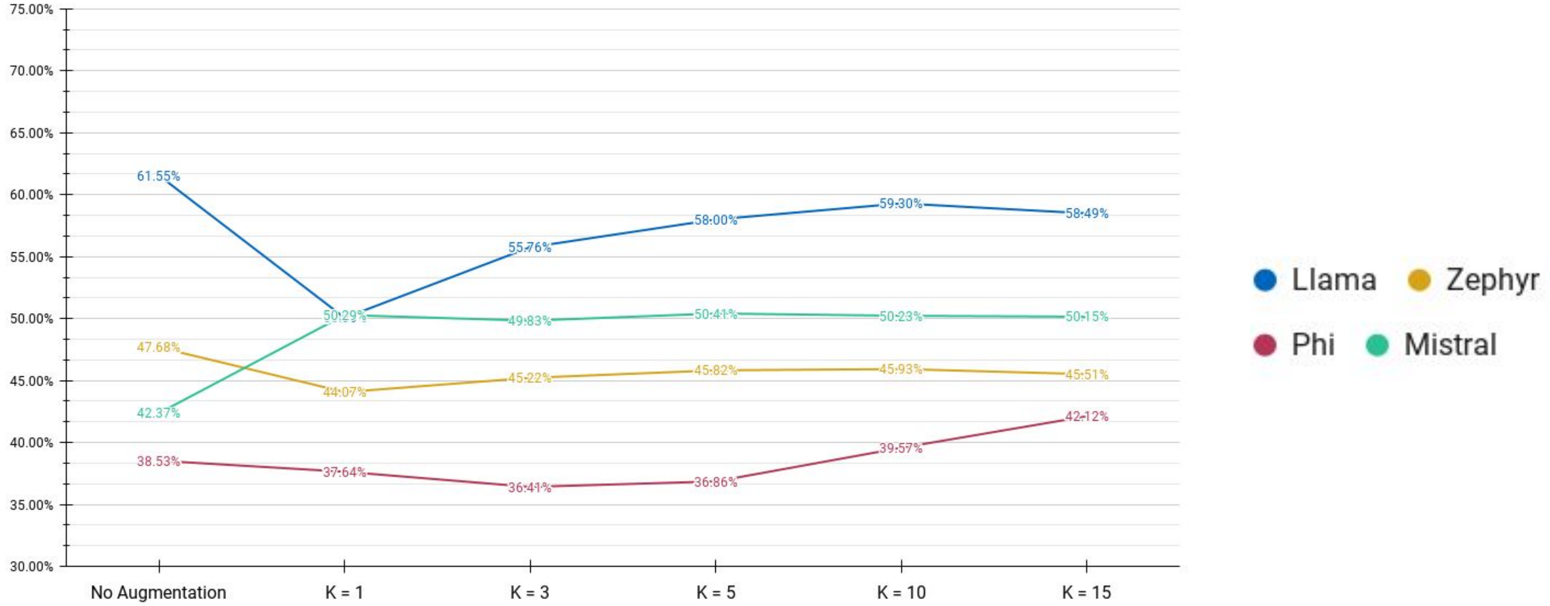
Subgraph Extraction



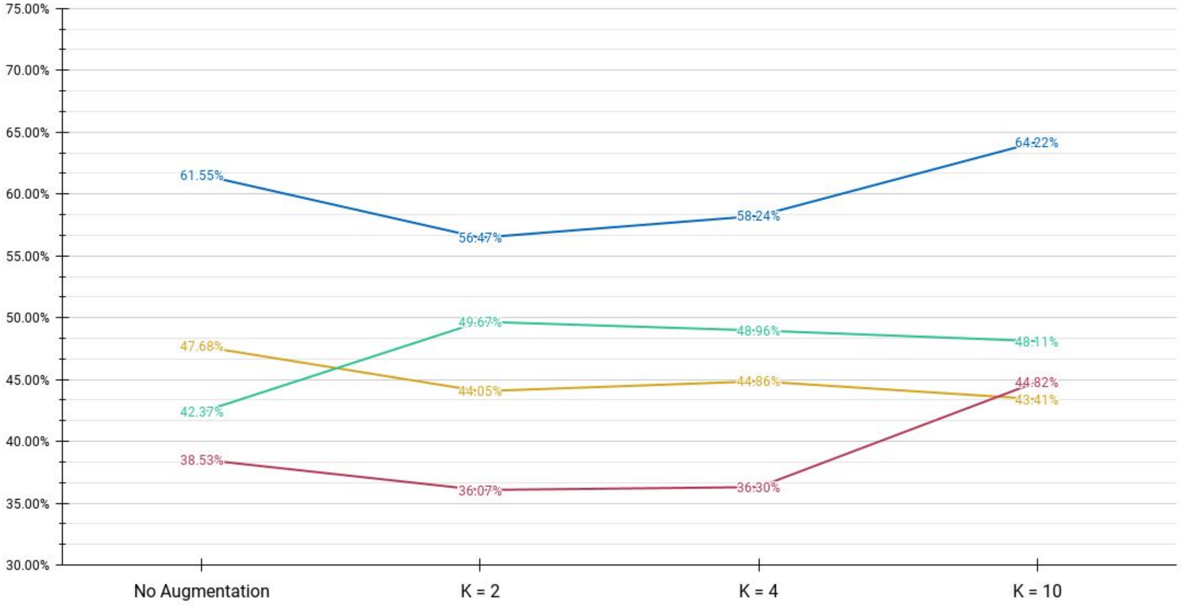
Results

per retrieval steps

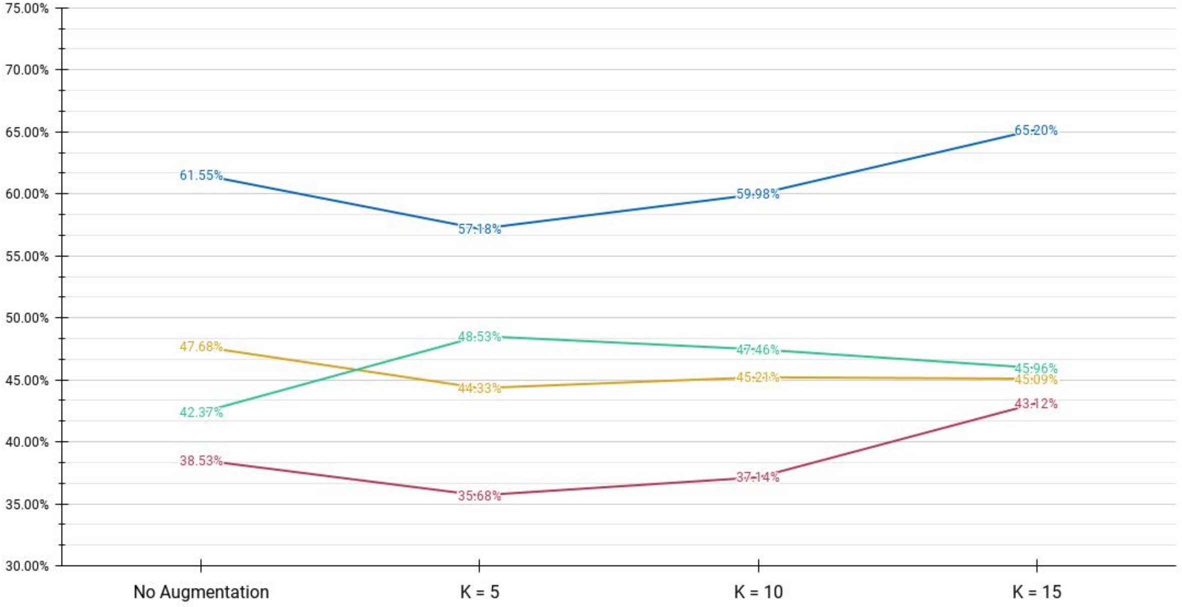
Dense Passage Retrieval Accuracy



Triplets Extraction Accuracy

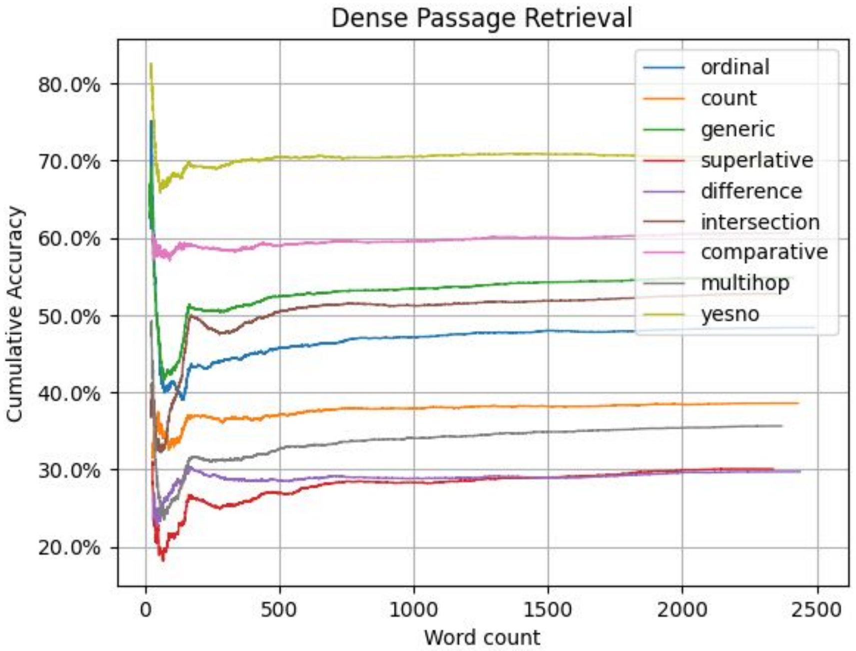


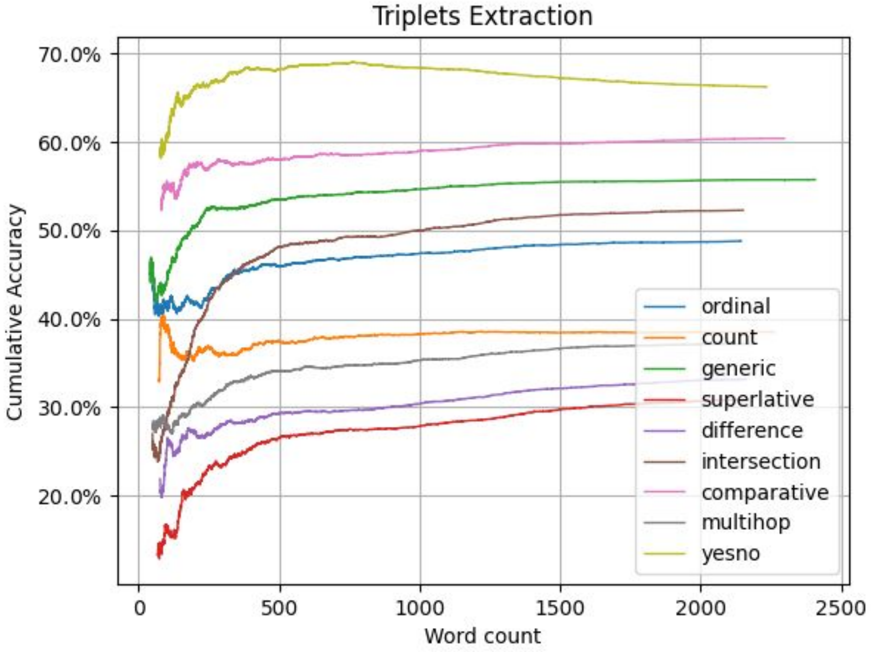
Graph Traversal Accuracy

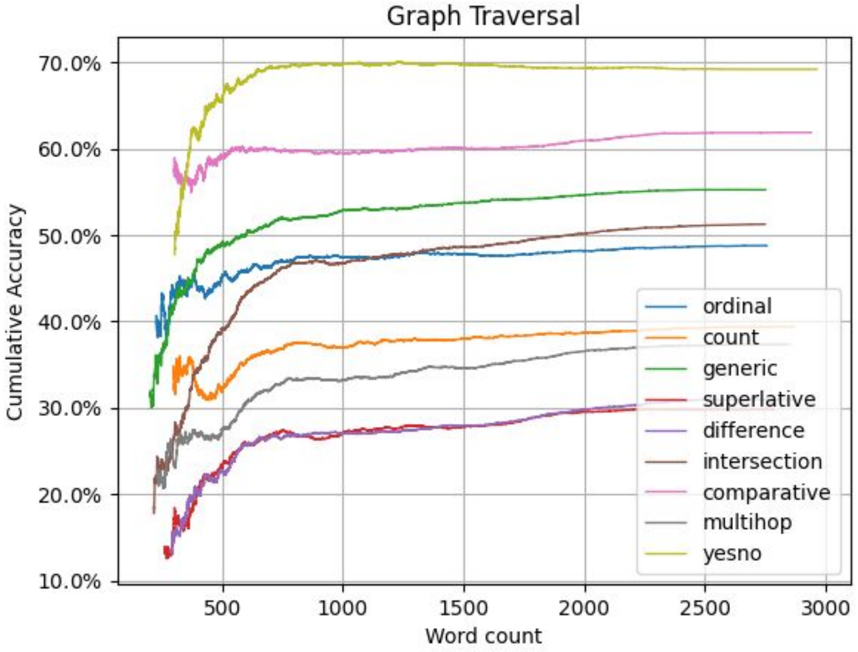


Results

per question type



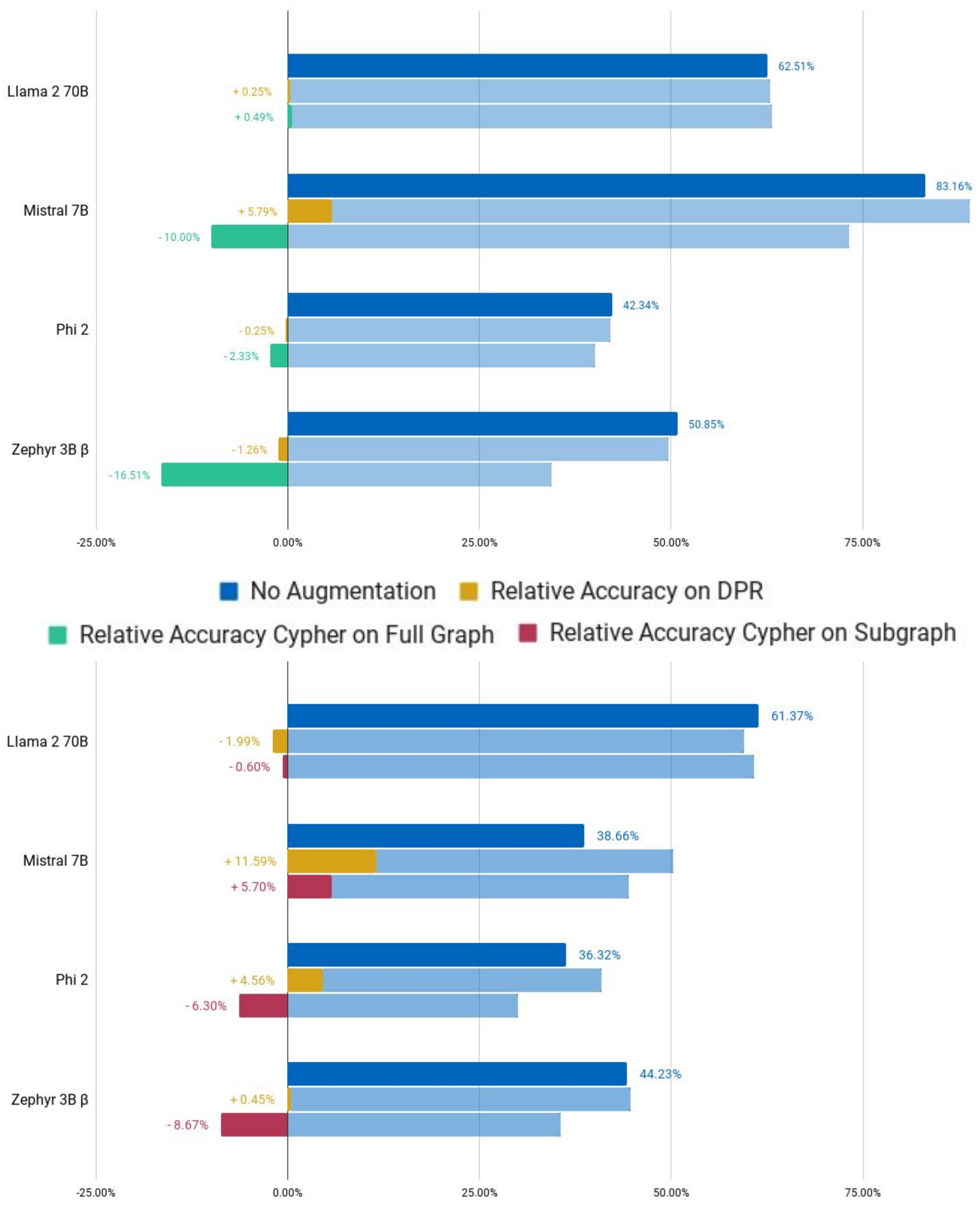




Results

of Cypher Generation

		Cypher Query on the Full Graph	
		Invalid Query	Valid Query
Cypher Query on a subgraph	Invalid Query	14740	1292
	Valid Query	3673	295



Cypher Query Generation on a Subgraph

1. Given the Graph database structure and the user query, generate an Cypher query using an LLM.

2. Execute the Cypher query and provide the data to the LLM