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

22/01/2024, Master’s Thesis Intermediate 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

Introduction

Research Questions

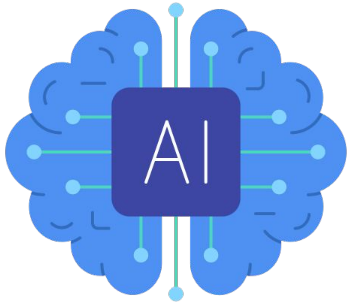
- Vector Database vs Graph Database
- Existing Approaches
- Evaluation Dataset
- Evaluation Technique

Progress

- Current Results
- Next Steps

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

What happened in the news last night?



I apologize, but my knowledge is based on information available up to September 2021.

Benefits of LLMs:

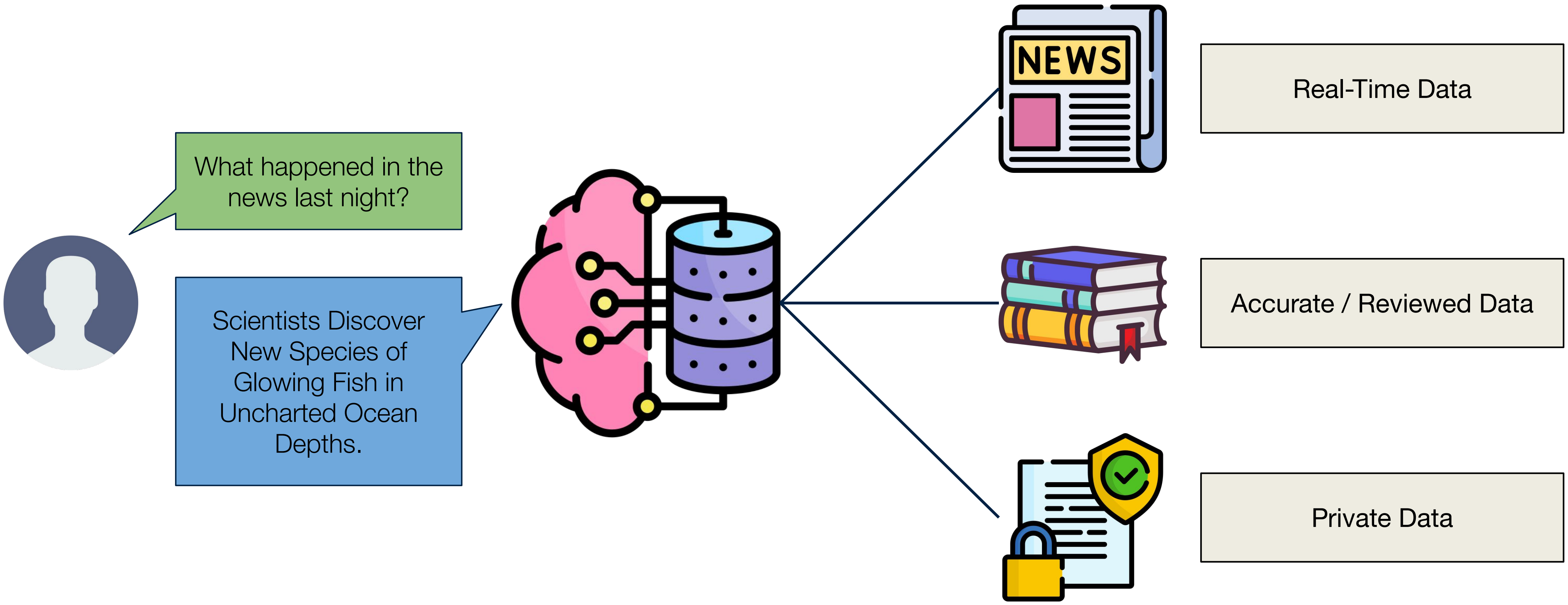
- Enable more natural and context-aware interactions in applications
- assist in various research fields in NLP by serving as pre-trained models for downstream tasks

Limitations of LLMs:

- Introducing new information in the current structure requires further training. It's difficult and not efficient.
- Limited control over the accuracy of the information that is provided by the model

Introduction

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



Outline

Introduction

Research Questions
- Vector Database vs Graph Database
- Existing Approaches
- Evaluation Dataset
- Evaluation Technique

Progress
- Current Results
- Next Steps

Research Questions

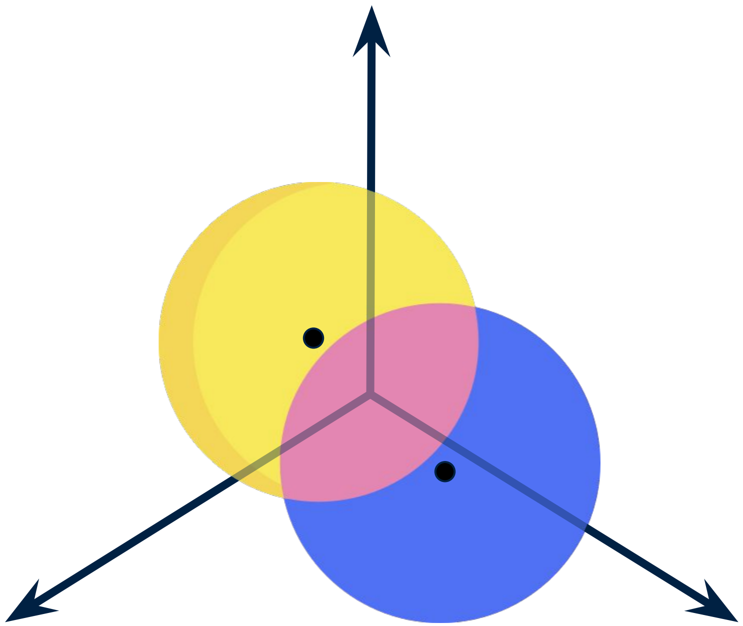
1. **How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?**

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

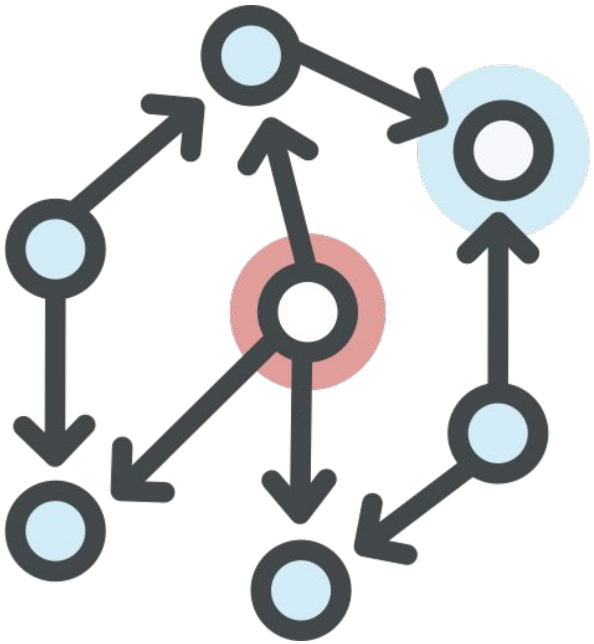
3. What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?

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

Vector Database vs Graph Database

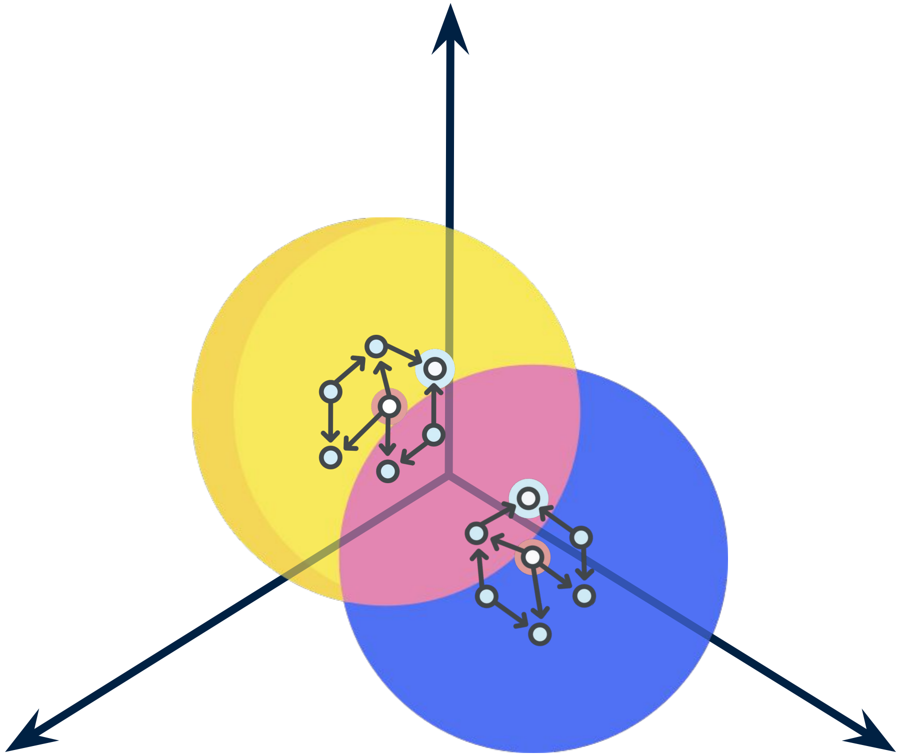


Hypothesis: Better for simple questions that require a general idea of a topic



Hypothesis: Better for more complex questions that include rules and conditions

vs Combination of both Databases



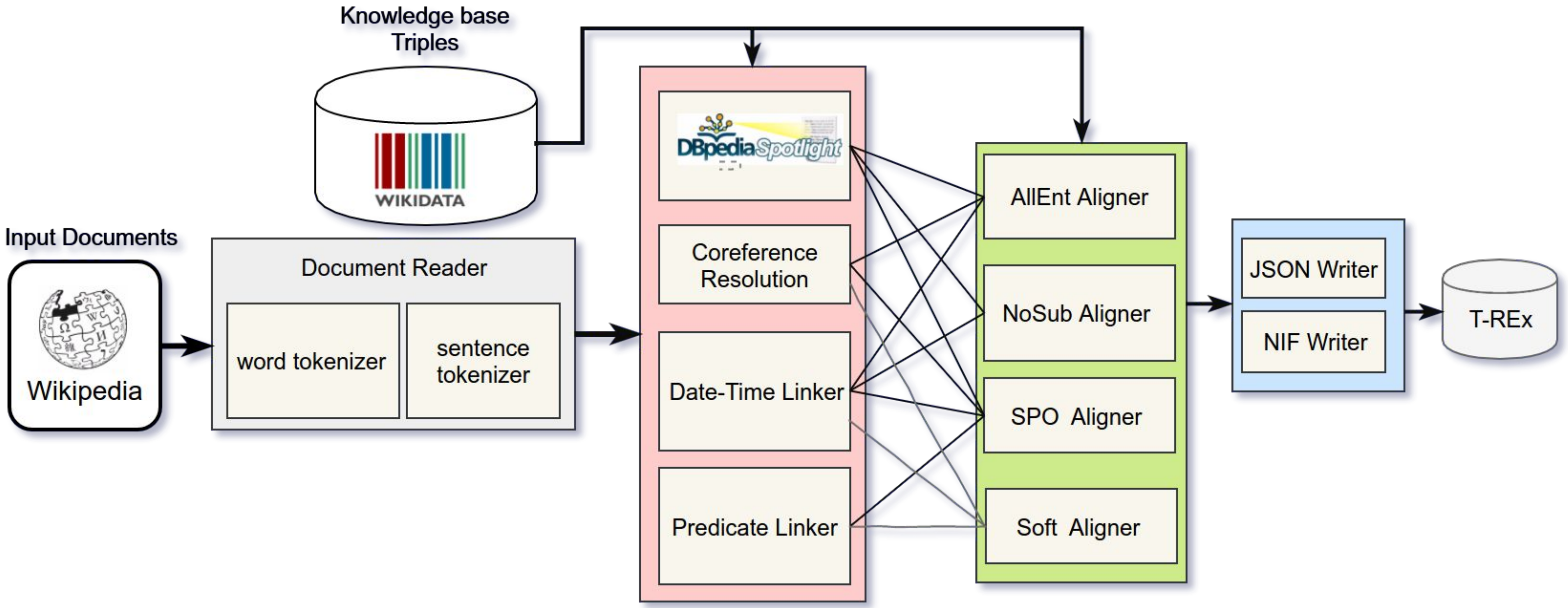
Hypothesis: Better performance overall, good compromise

Research Questions

1. How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?
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. What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?
4. How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?

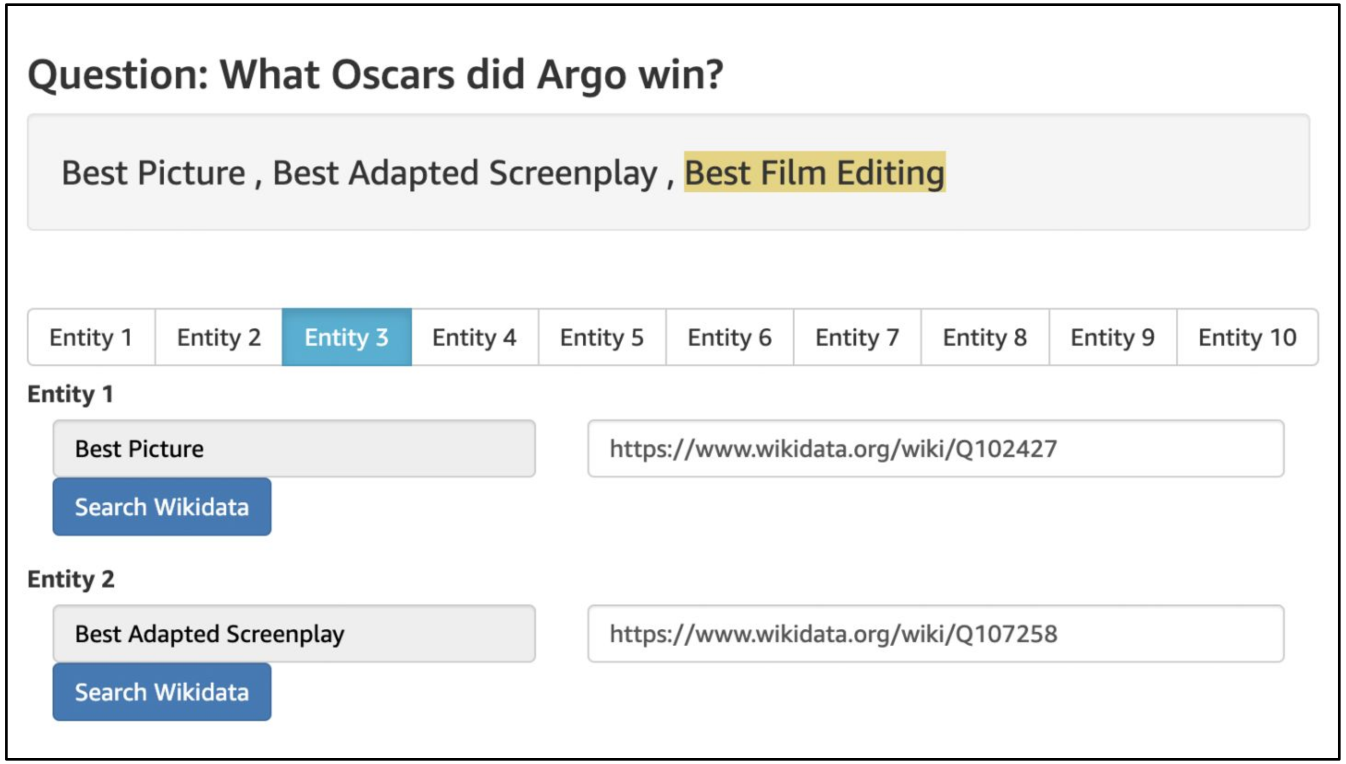
Setup

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

Types of questions:

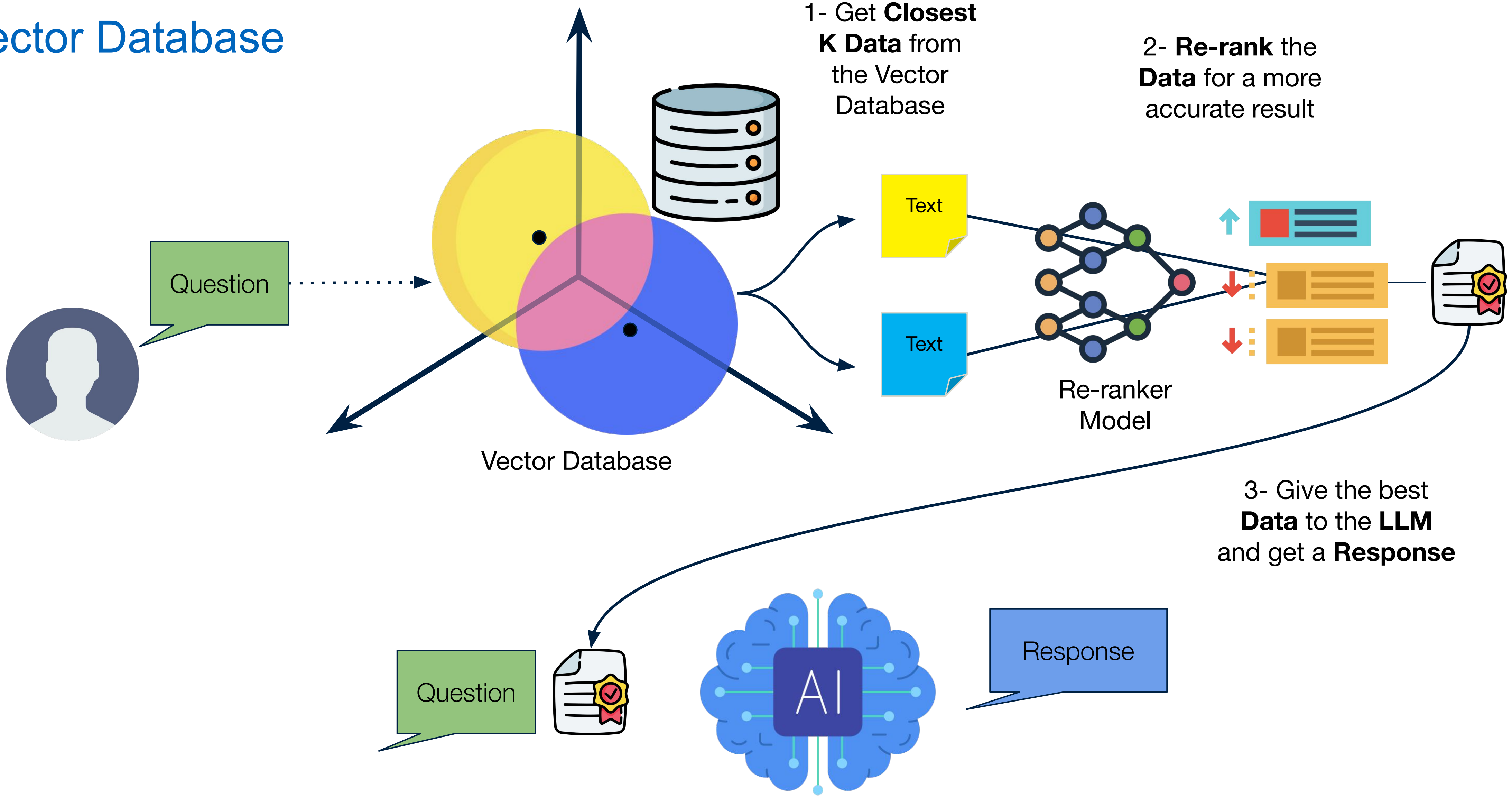
Type	Description	Example
Generic	Simple questions	Where was Michael Phelps born?
Yes/No	Answer is a Yes or No	Has Lady Gaga ever made a song with Ariana Grande?
Count	Answer requires counting	How many astronauts have been elected to Congress?
Superlative	Max or Min of given attribute	Who was the youngest tribute in the Hunger Games?
Comparative	Compare 2 items by an attribute	Is Mont Blanc taller than Mount Rainier?
Ordinal	Based on item's position in a list	Who was the last Ptolemaic ruler of Egypt?
Difference	Contains a negation	Which Mario Kart game did Yoshi not appear in?
Intersection	Requires multiple conditions	Which movie was directed by Denis Villeneuve and stars Timothee Chalamet?
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. How do vector databases and graph databases differ in their performance when augmenting LLMs in question answering tasks?
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. **What are existing retrieval approaches for retrieval augmented generation using vector databases and graph databases?**
4. How can the quality of question-answering performance be systematically evaluated across different Large Language Model-based Retrieval Augmented Generation systems?

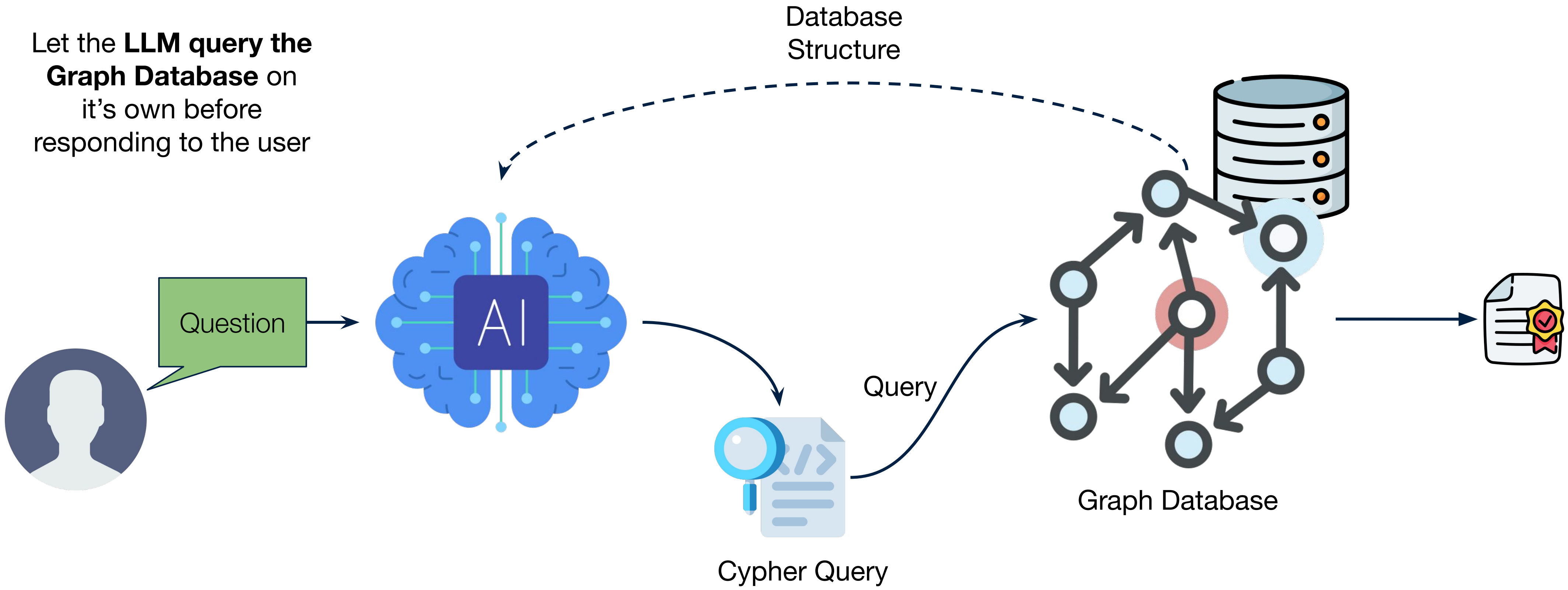
Vector Database



Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). [Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.]()

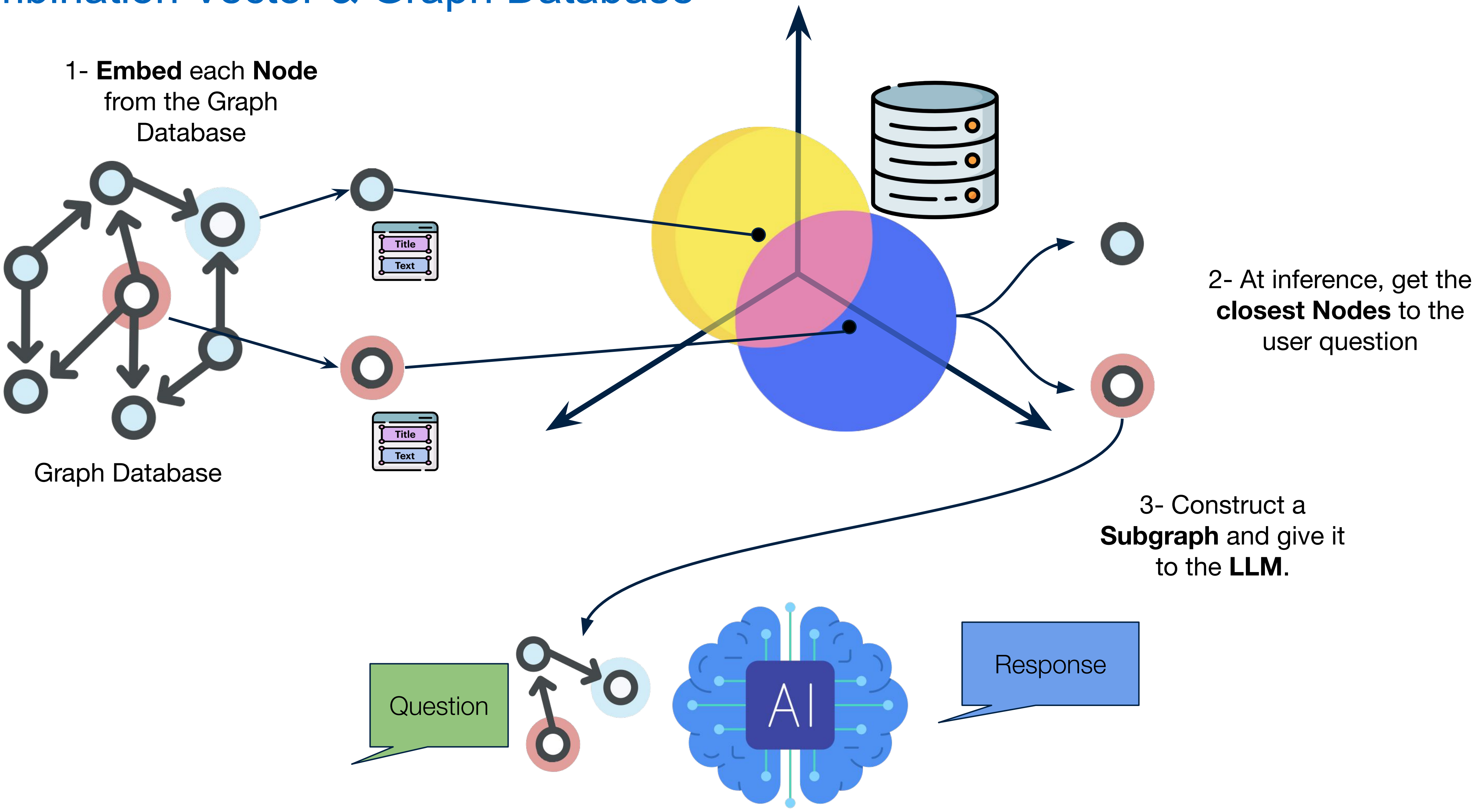
Graph Database

Let the **LLM query the Graph Database** on it's own before responding to the user



https://medium.com/neo4j/enhanced-qa-integrating-unstructured-and-graph-knowledge-using-neo4j-and-langchain-6abf6fc24c27

Combination Vector & Graph Database



https://medium.com/@nebulagraph/graph-rag-the-new-llm-stack-with-knowledge-graphs-e1e902c504ed

Research Questions

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

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

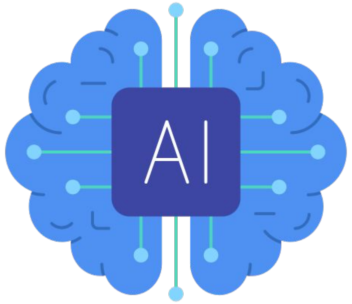
3. How to align a vector database with a graph database to include the same information and be comparable in terms of retrieval performance?

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

Evaluation Metric: Exact Match

Mintaka Question:

When was Michael Phelps born?



LLM Answer:

Michael Phelps is born in June 30, 1985

$\neq$

Mintaka Answer:

30/06/1985

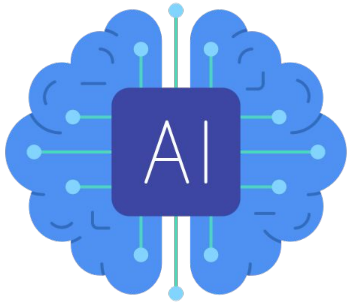
Limitations of Exact Match:

- Difficulty to control the LLM answer
- Many cases for each question type

Evaluation Metric: BLEU, ROUGE & BERT Score

Mintaka Question:

Who is older Beyonce or Madonna?



LLM Answer 1:

Madonna is older than Beyonce

BLEU 1

=

BLEU 2

Beyonce is older than Madonna

LLM Answer 2:

Mintaka Answer:

Madonna

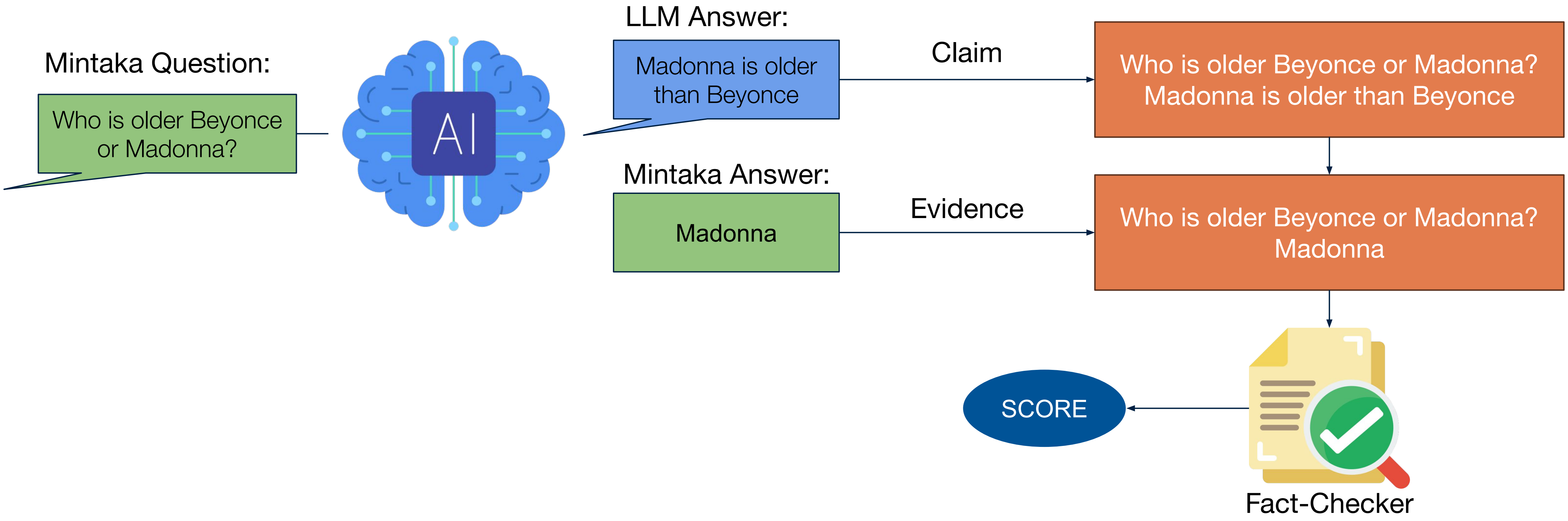
Limitations of BLEU & ROUGE:

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: Fact-Checking



Outline

Introduction

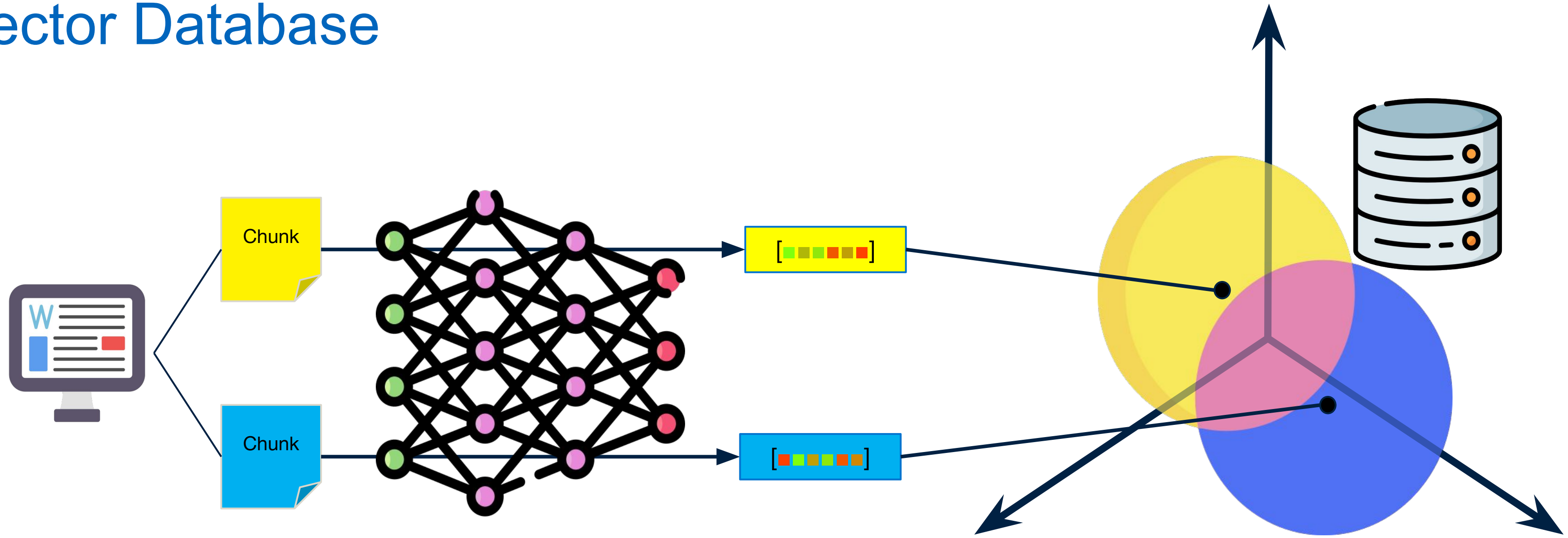
Research Questions

- Vector Database vs Graph Database
- Existing Approaches
- Evaluation Dataset
- Evaluation Technique

Progress

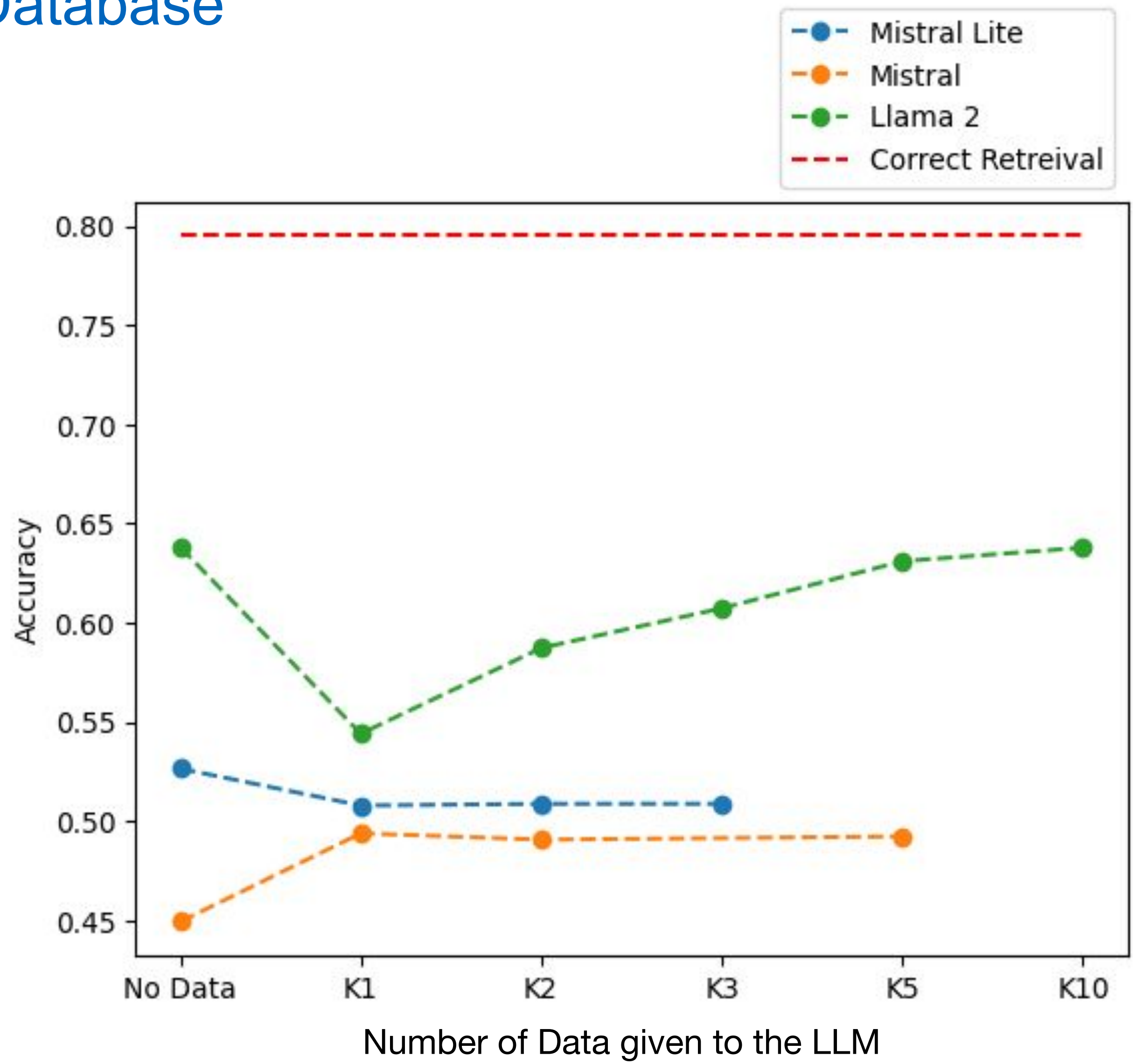
- Current Results
- Next Steps

Results: Vector Database

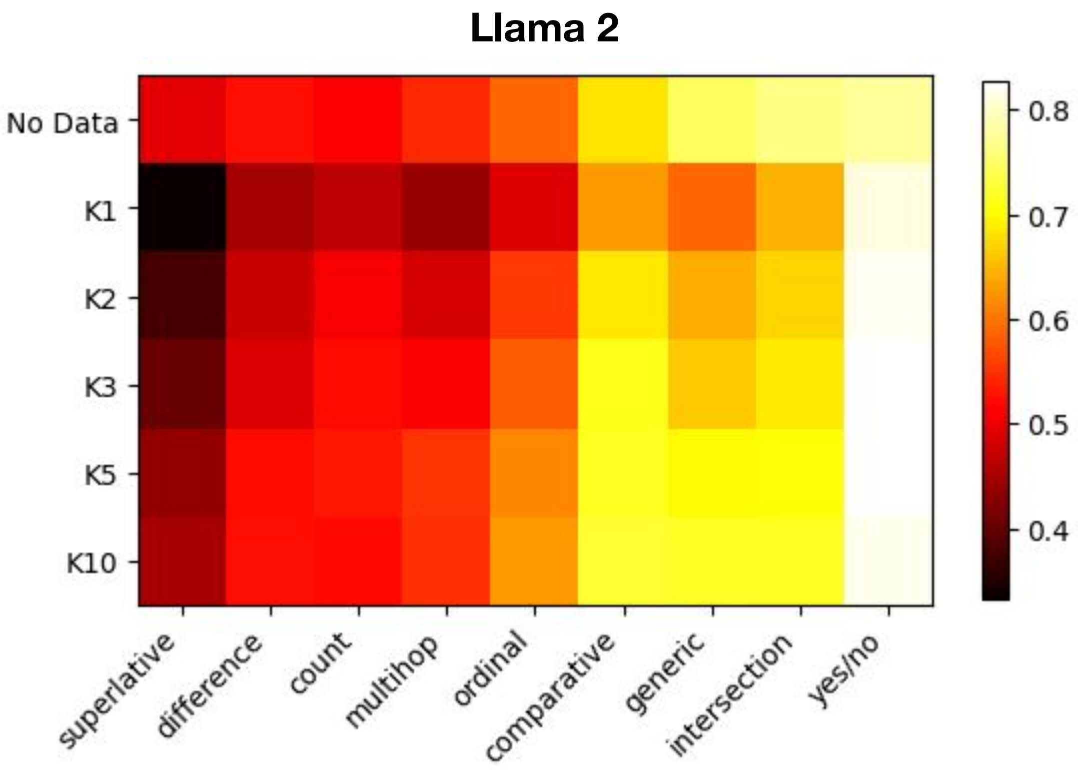


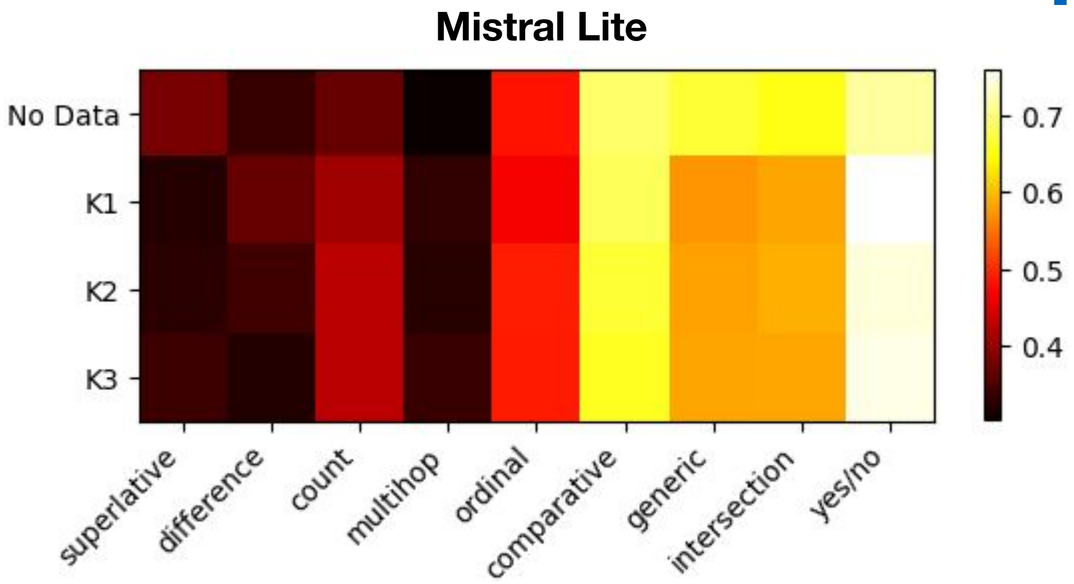
Chunking	**Embedding Model**	MRR	MRR after re-ranking
Split by words	multi-qa-mpnet-base-dot-v1	0.09334	0.11414
Split by tokens	msmarco-distilbert-base-tas-b	0.12399	0.20302
Split by sentences using NLTK	multi-qa-mpnet-base-dot-v1	0.13746	0.21251
Split by sentences using Spacy	msmarco-distilbert-base-tas-b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

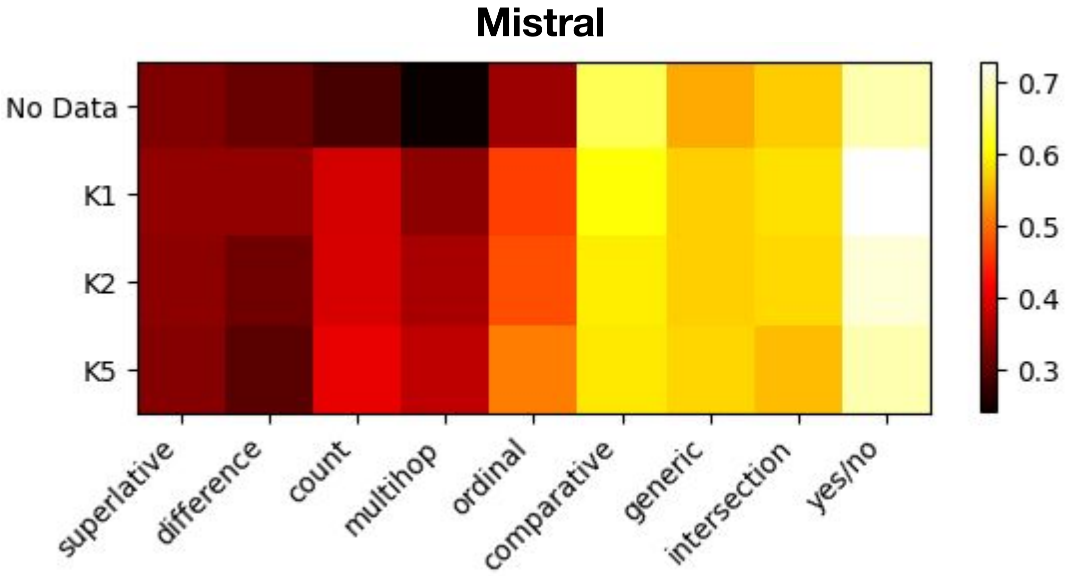
Results: Vector Database



Results: Vector Database







Next Steps

1. Implement and test **Graph Database** techniques
2. Implement and test advanced techniques with **Combined Databases**
3. Improve previous techniques if fitting
4. Analyse the results and write the thesis

References

Baek, J., Aji, A. F., Lehmann, J., & Hwang, S. J. (2023). Direct Fact Retrieval from Knowledge Graphs without Entity Linking. arXiv preprint arXiv:2305.12416.

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).

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906.

Oguz, B., Chen, X., Karpukhin, V., Peshterliev, S., Okhonko, D., Schlichtkrull, M., ... & Yih, S. (2020). Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. arXiv preprint arXiv:2012.14610.

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.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., ... & Yih, W. T. (2023). Replug: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.

Yu, W. (2022, July). Retrieval-augmented generation across heterogeneous knowledge. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop (pp. 52-58).

Philippe Saadé

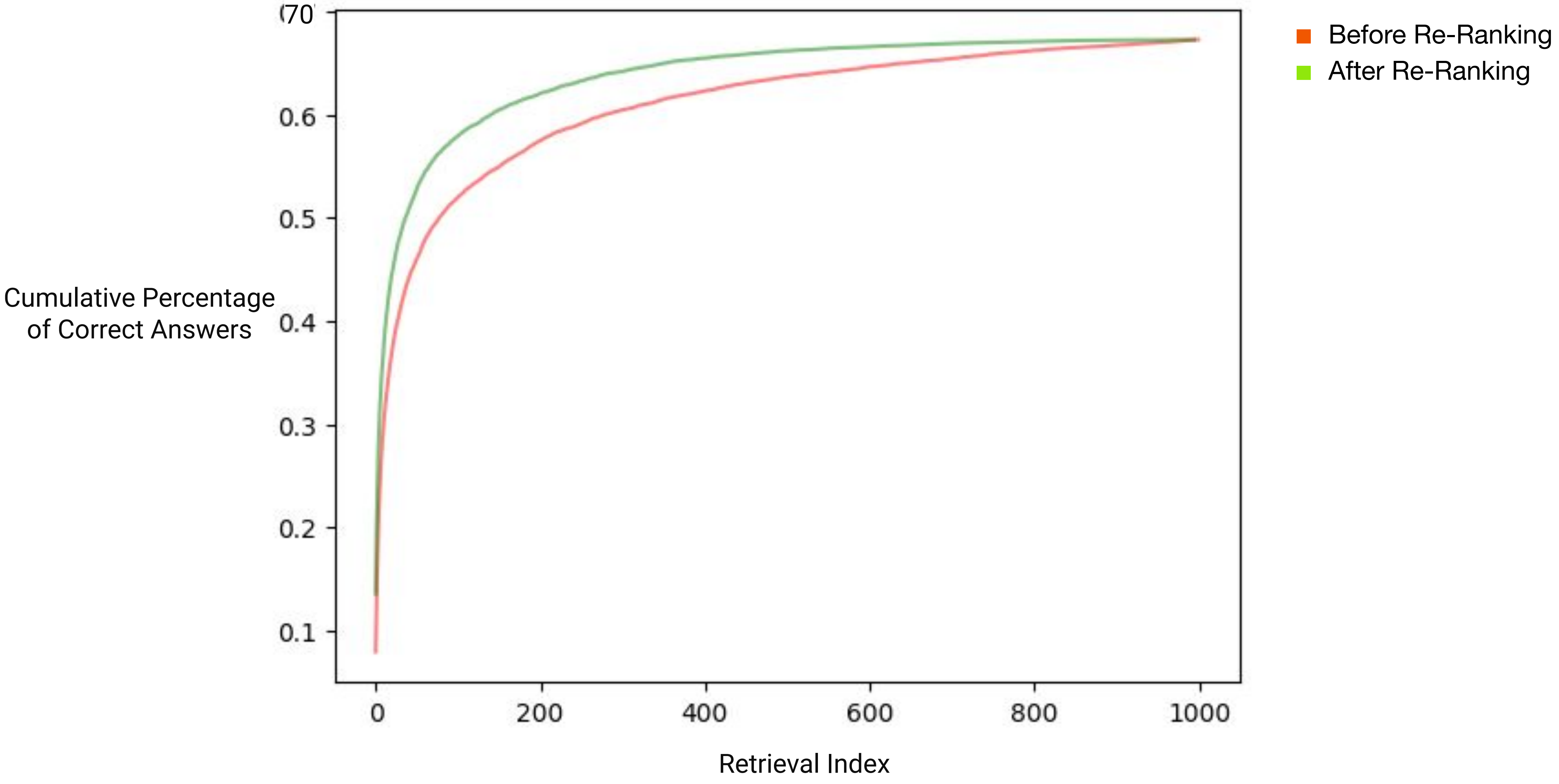
Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

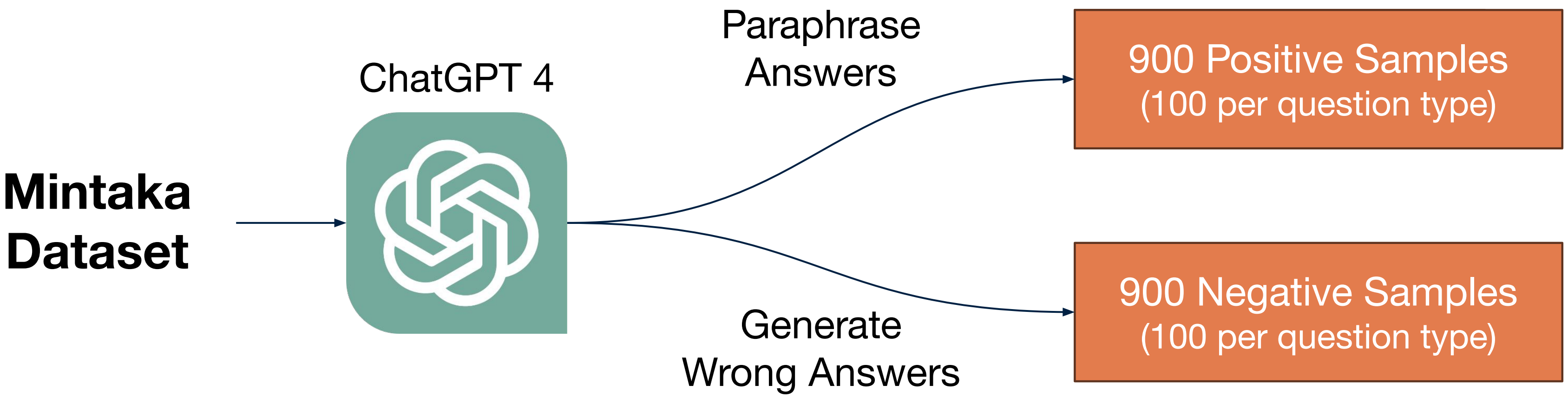
+49.89.289.

wwwmatthes.in.tum.de

Results: Vector Database

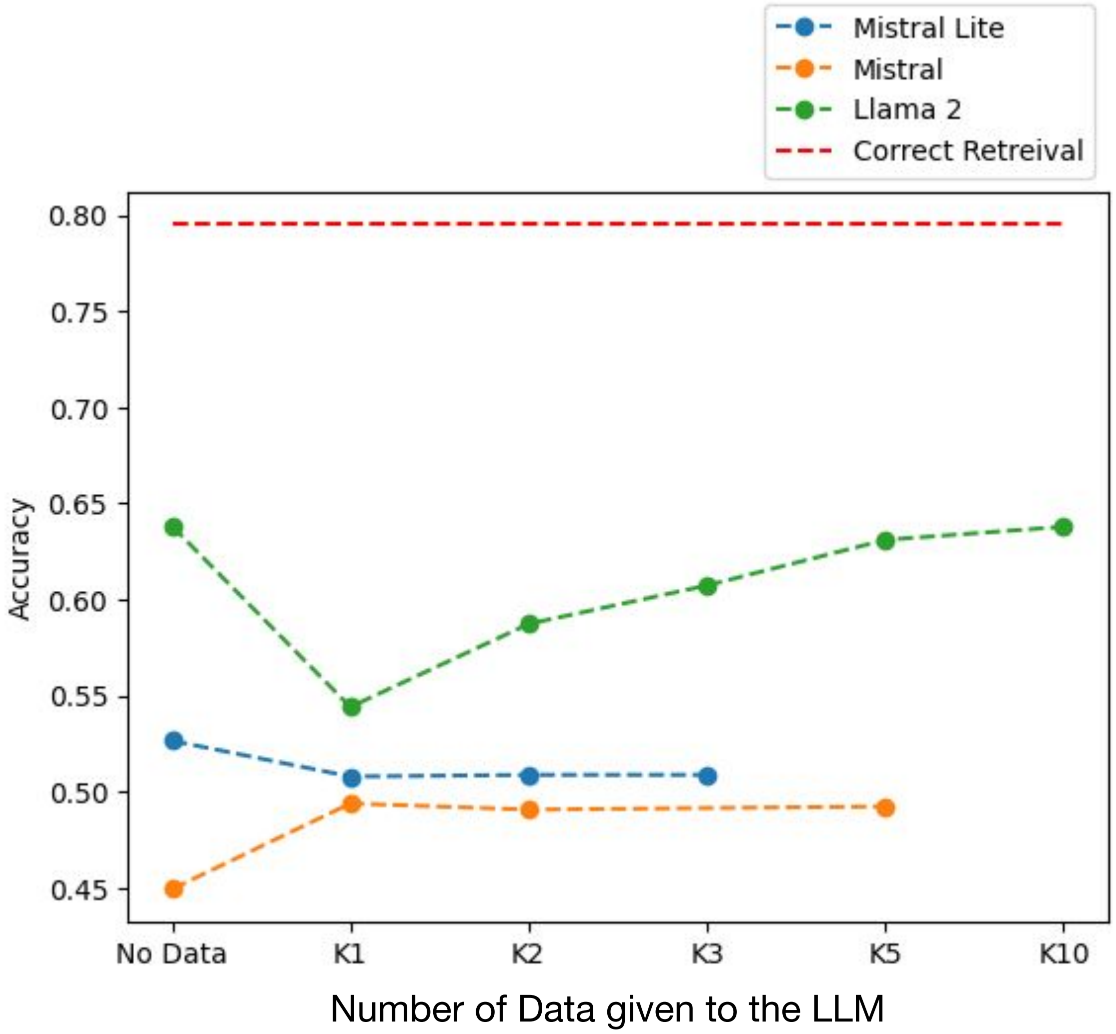


Evaluation Metric: Fact-Checking

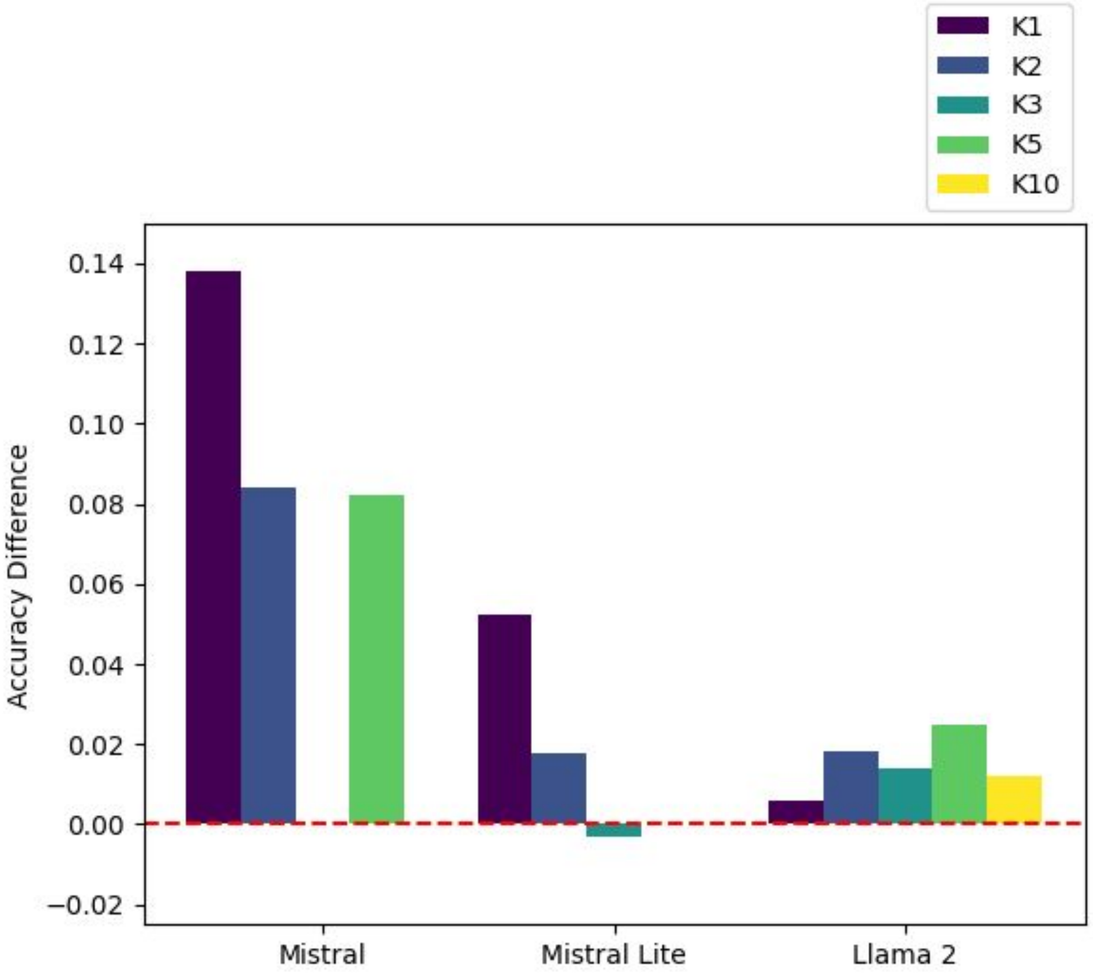


Fact-Checking Model	**Accuracy (Threshold 0.5)**	**Average Scores**	**Prediction Time**
facebook/bart-large-mnli	95.5%	0.9467	0.1575 sec
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli	95.6%	0.9476	0.059 sec
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	97.9%	0.9744	0.18 sec

Results: Vector Database

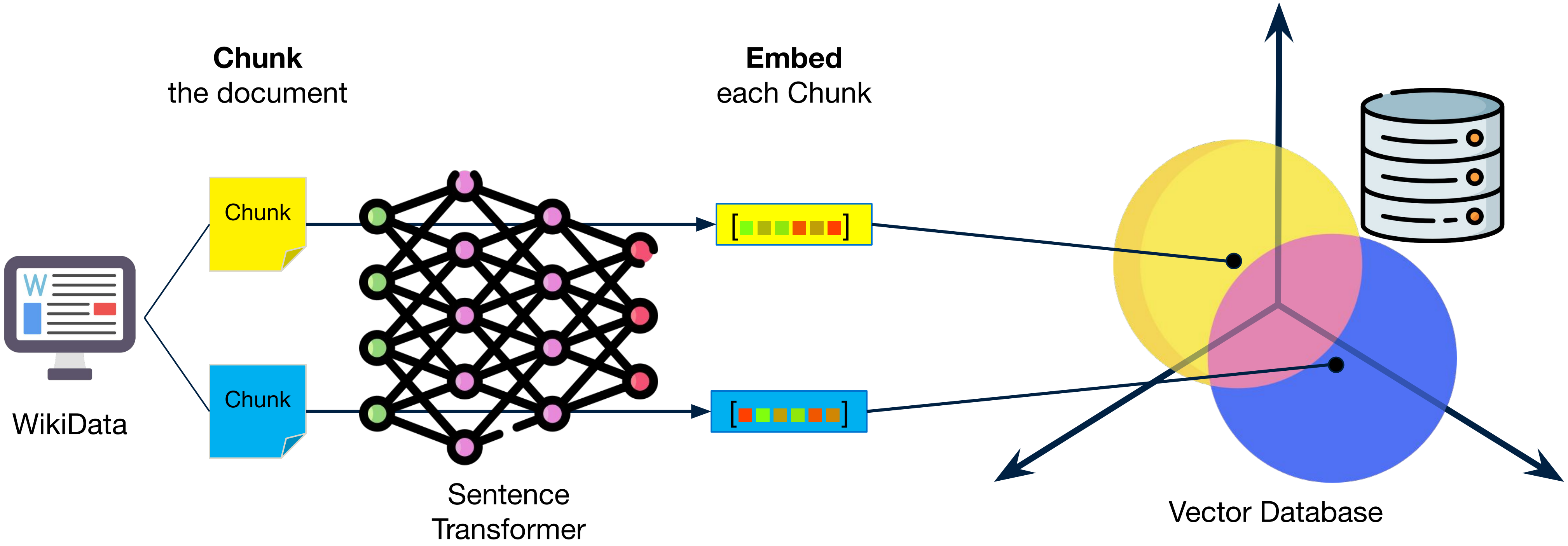


Overall Accuracy



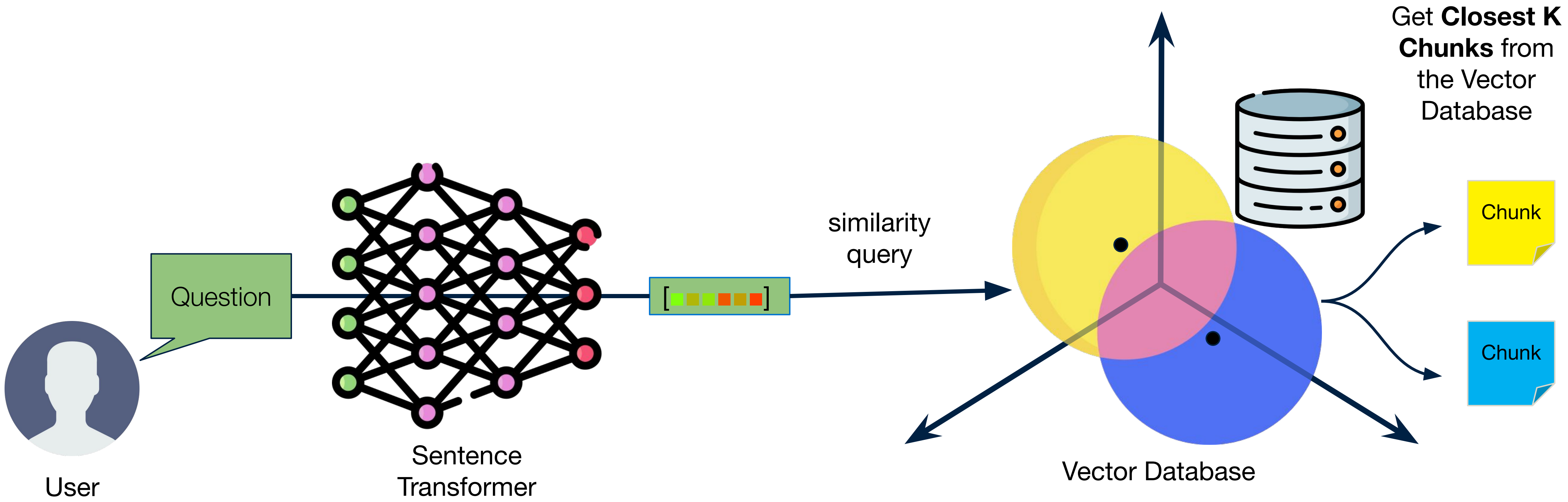
Accuracy improvement
where retriever got the correct data

Vector Database: Setup



Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

Vector Database: Inference

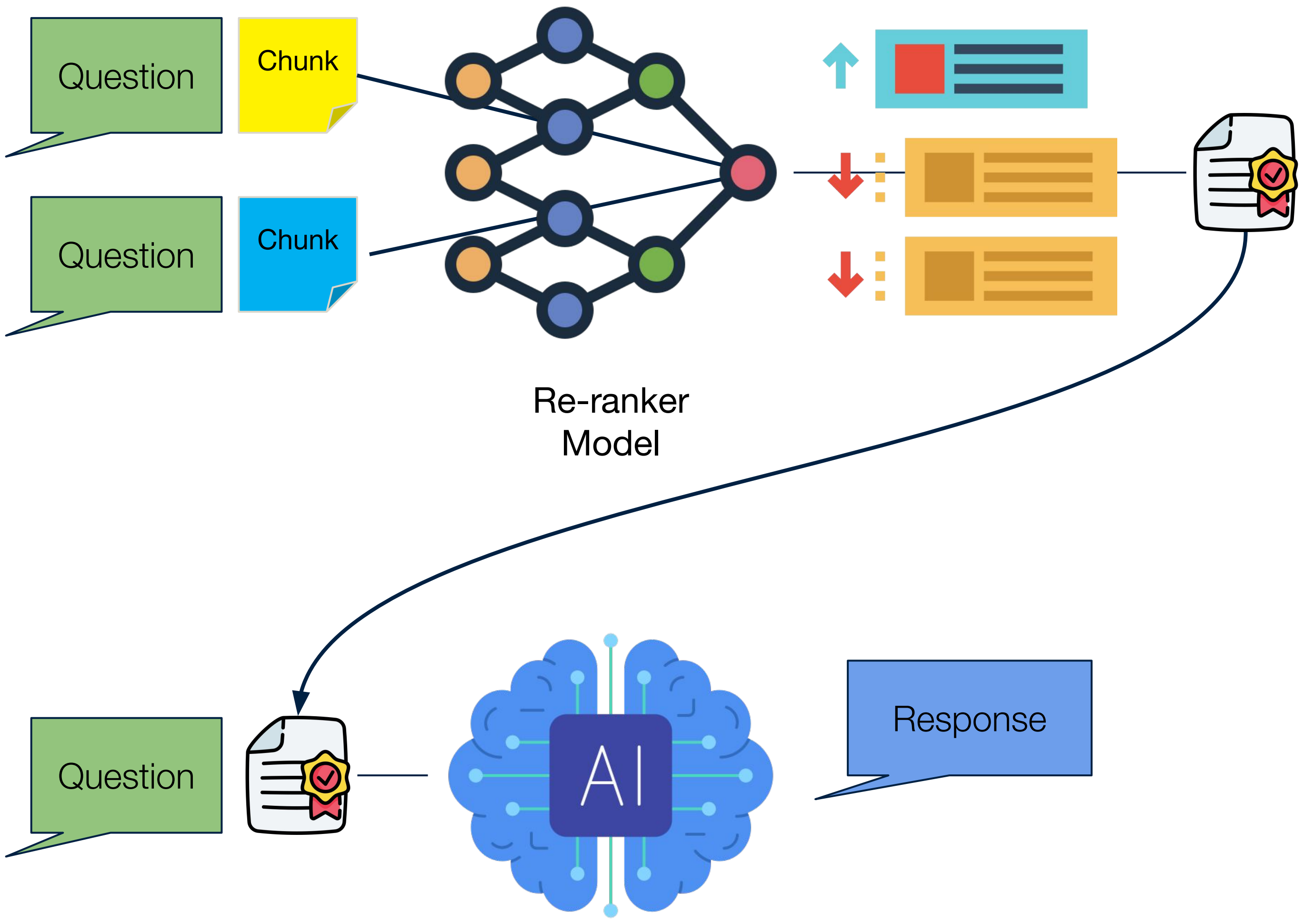


Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

Vector Database: Inference

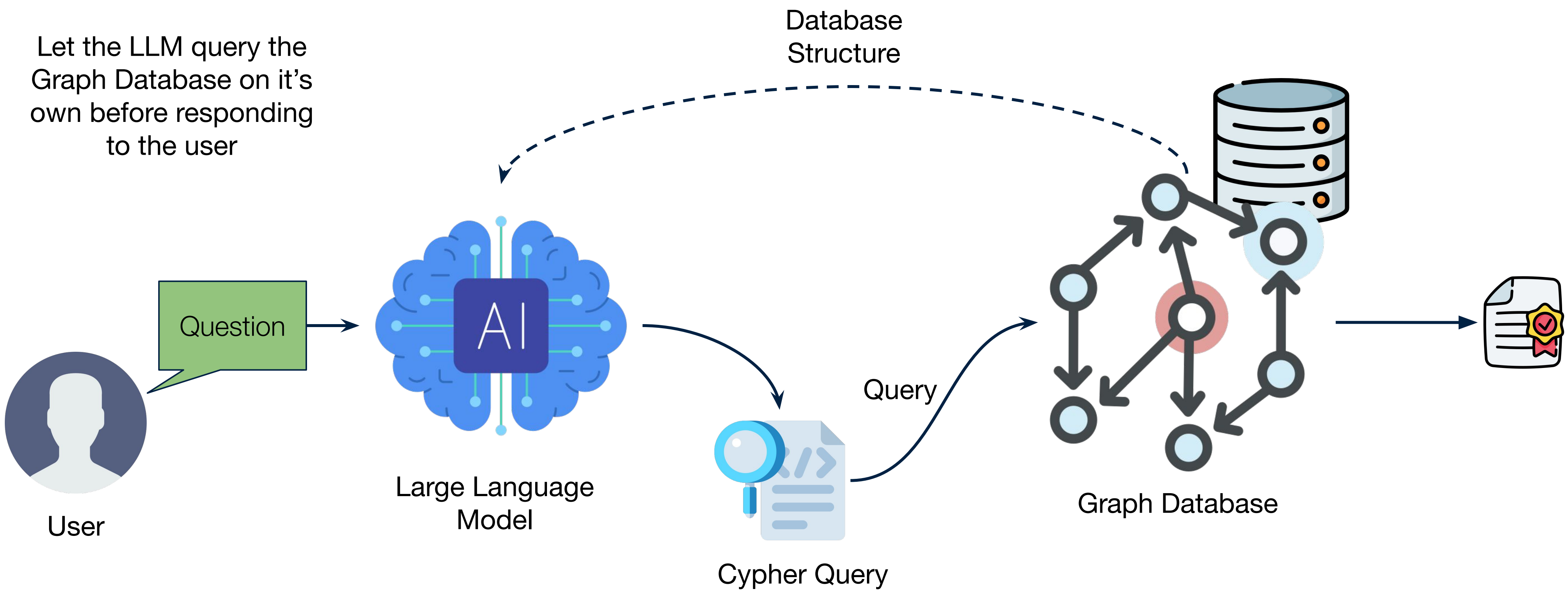
1- Re-rank the chunks for a more accurate result. Use the chunk with the highest rank.

2- Give the best chunk to the LLM model and get response.



Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

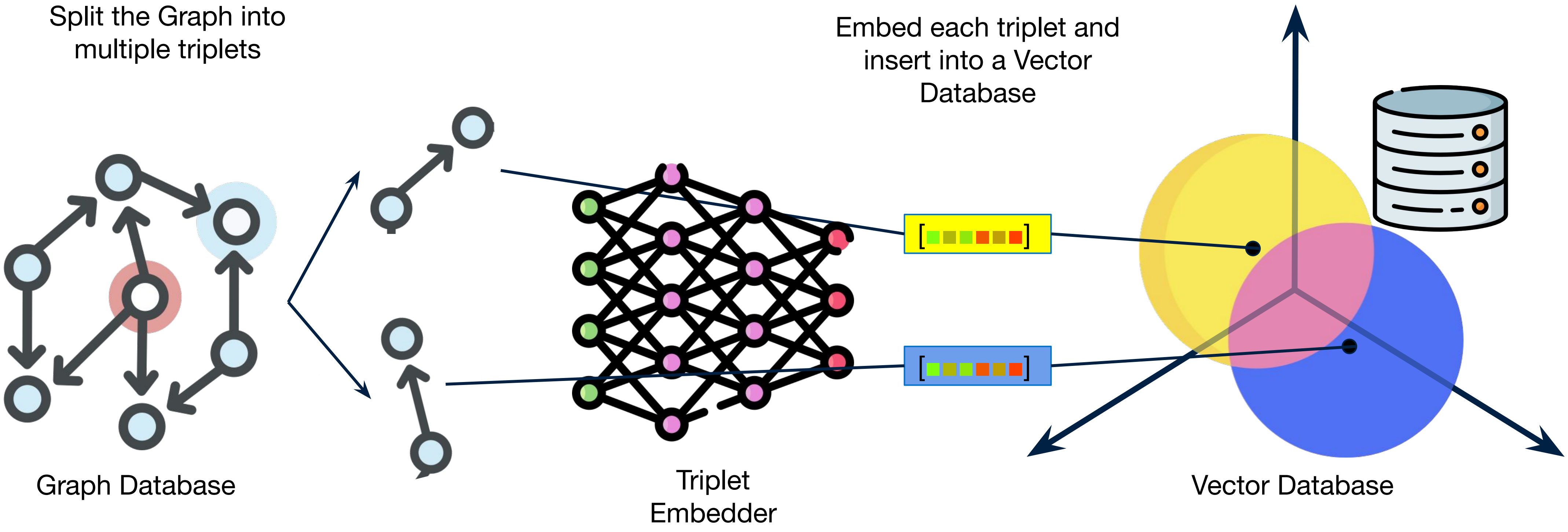
Graph Database: Inference



https://medium.com/neo4j/enhanced-qa-integrating-unstructured-and-graph-knowledge-using-neo4j-and-langchain-6abf6fc24c27

Combination Vector & Graph Database

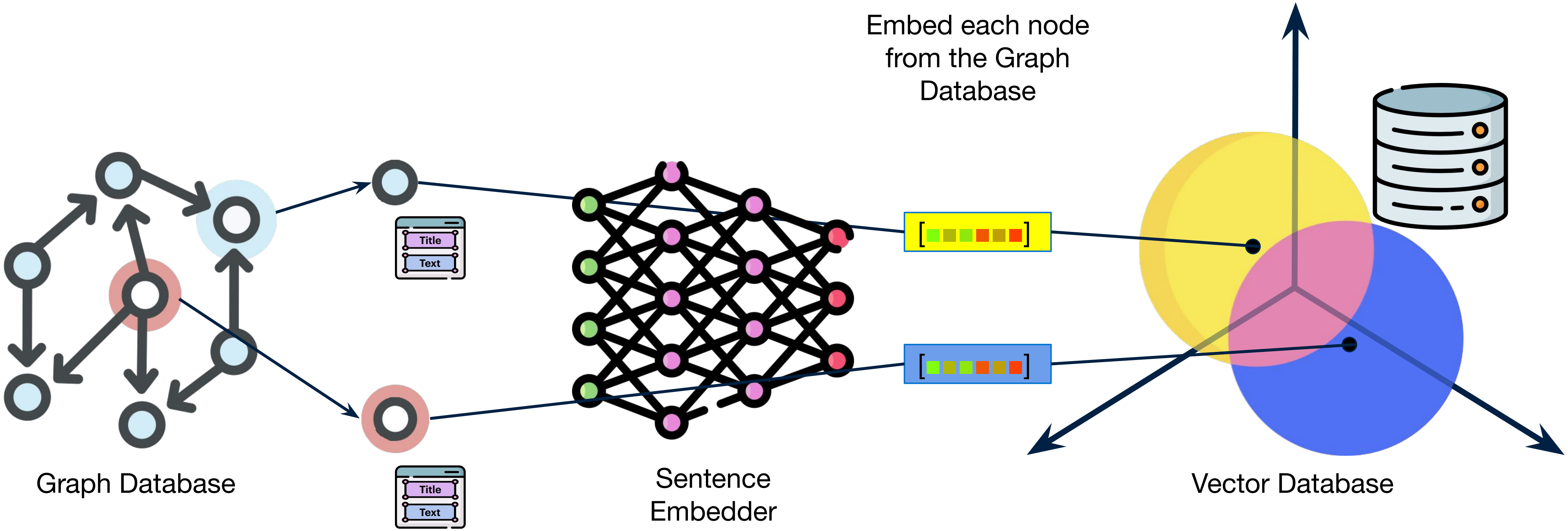
Method 1: Setup



Baek, J., Aji, A. F., Lehmann, J., & Hwang, S. J. (2023). [Direct Fact Retrieval from Knowledge Graphs without Entity Linking. arXiv preprint arXiv:2305.12416.]()

Combination Vector & Graph Database

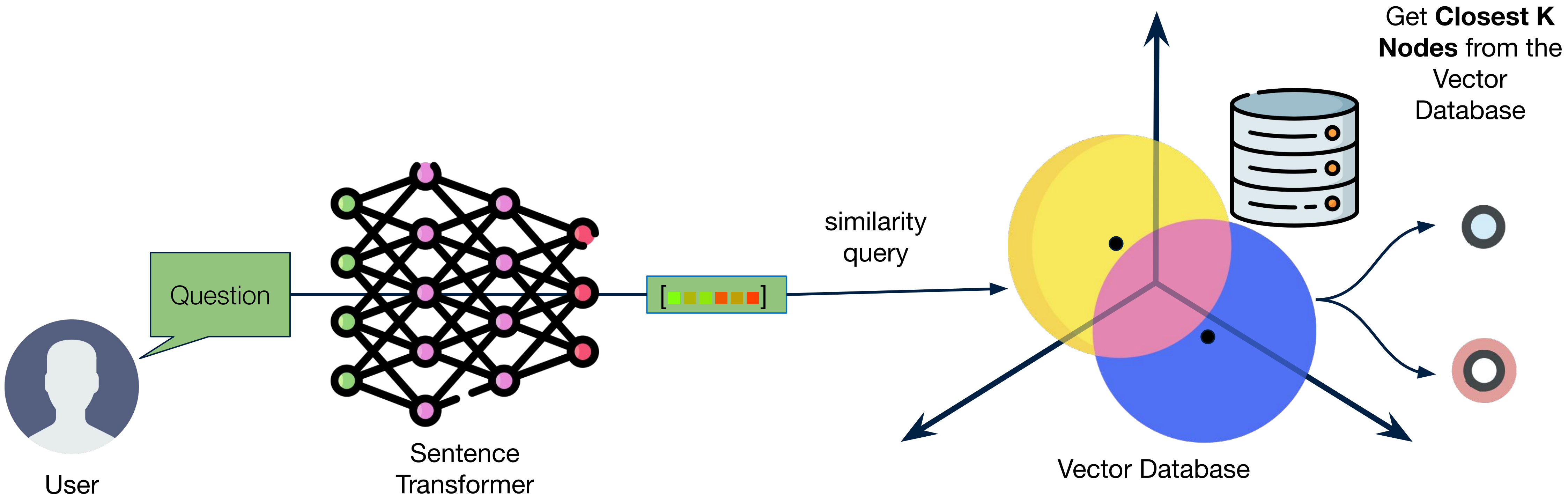
Method 2: Setup



https://medium.com/@nebulagraph/graph-rag-the-new-llm-stack-with-knowledge-graphs-e1e902c504ed

Combination Vector & Graph Database

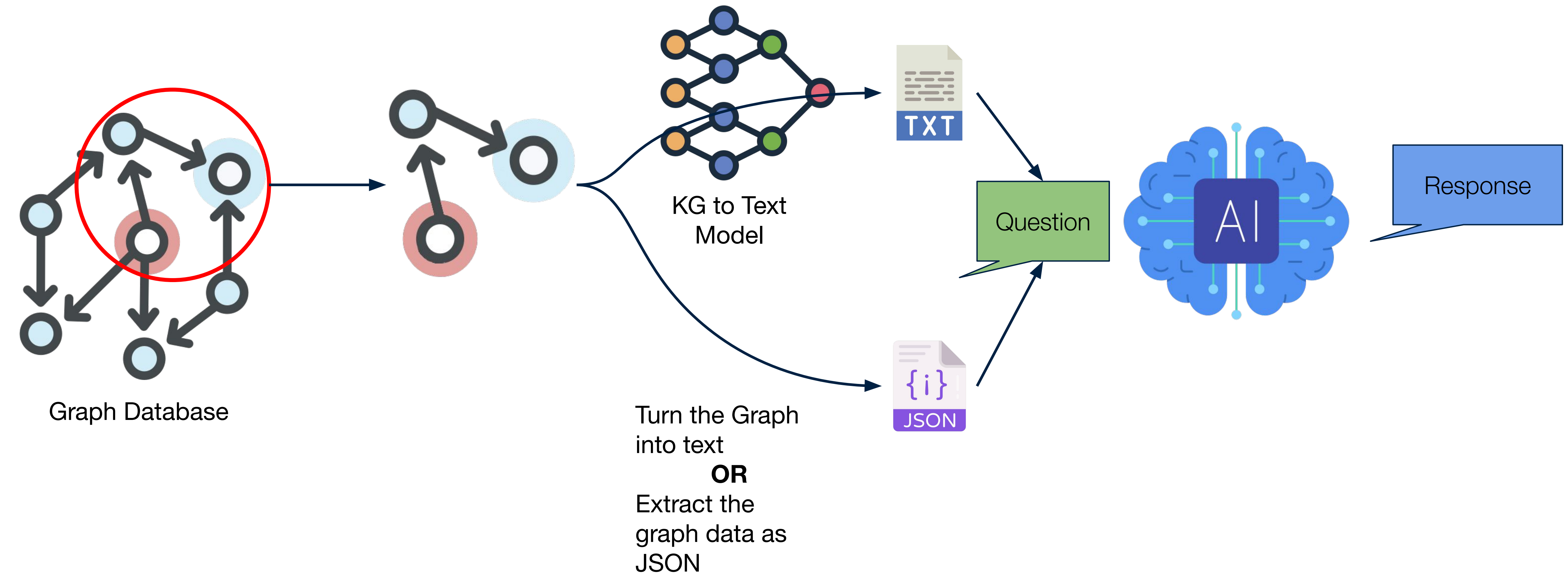
Method 2: Inference



https://medium.com/@nebulagraph/graph-rag-the-new-llm-stack-with-knowledge-graphs-e1e902c504ed

Combination Vector & Graph Database

Method 2: Inference



https://medium.com/@nebulagraph/graph-rag-the-new-llm-stack-with-knowledge-graphs-e1e902c504ed