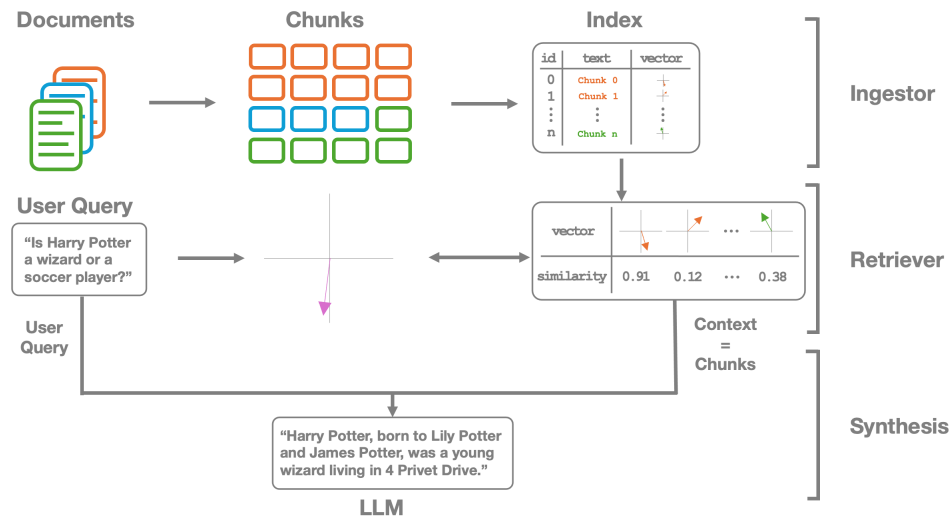


RAG for Source Text Summarization

Midterm Presentation 1

Guangyao Quan
Berke Göçmen
Luca Mattes Wiehe

IN2106: NLP Lab
Supervisor: Miriam Anschütz
10 June 2024



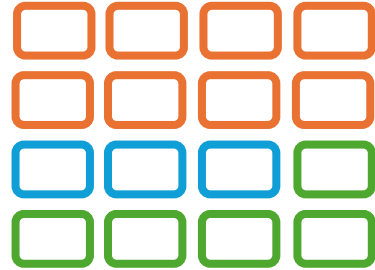
Part 1: Introduction

Traditional RAG Pipeline

Documents



Chunks



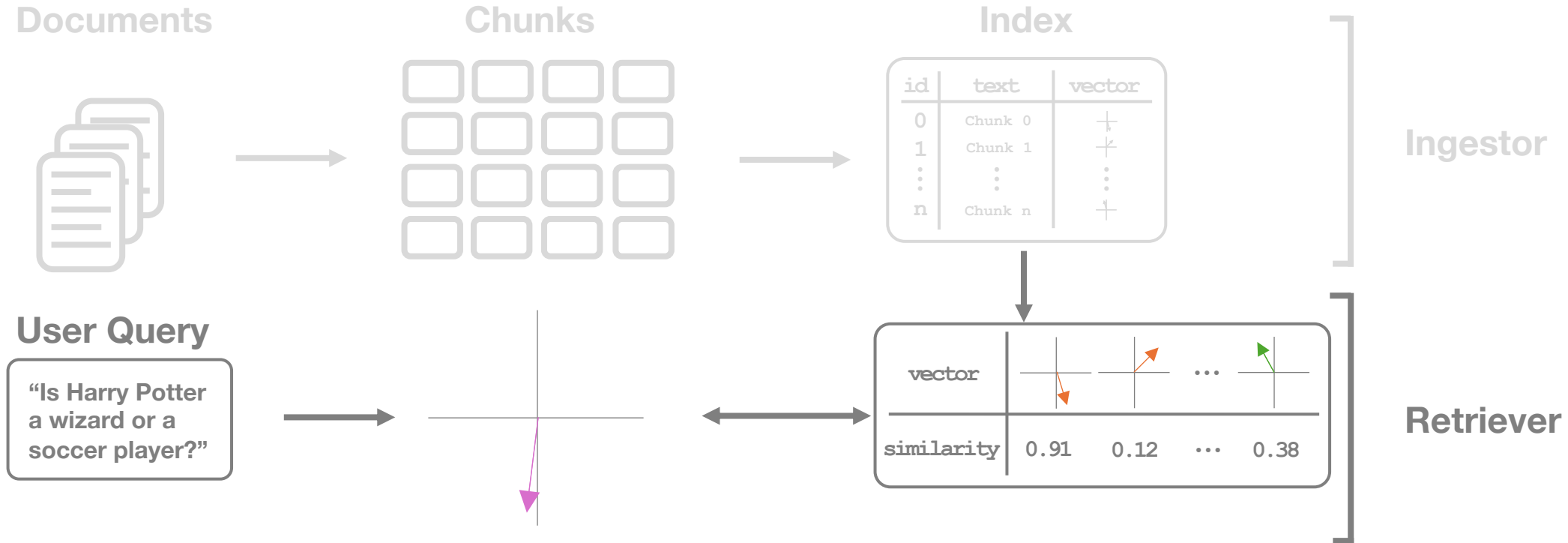
Index

id	text	vector
0	Chunk 0	+
1	Chunk 1	+
⋮	⋮	⋮
n	Chunk n	+

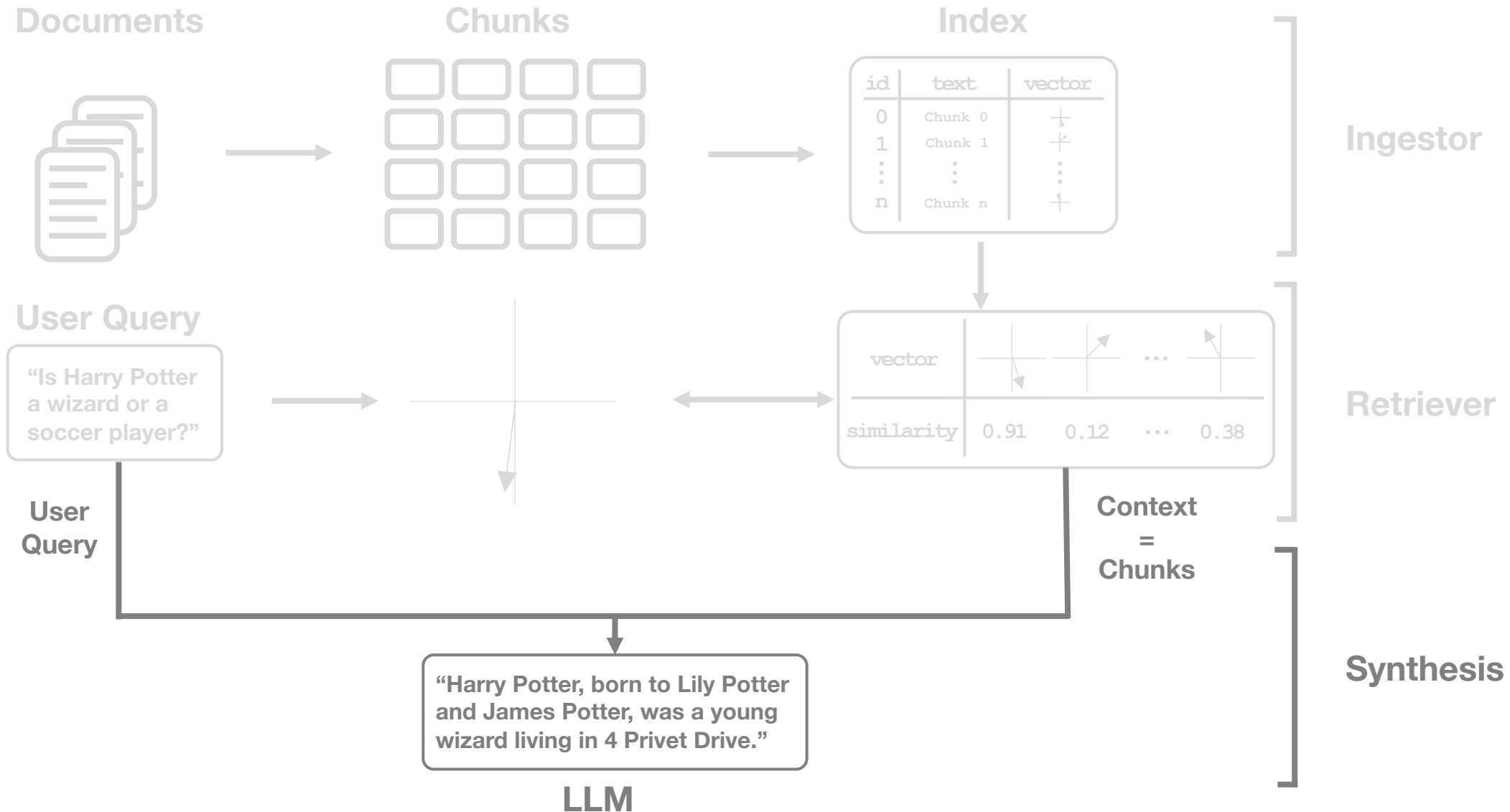


Ingestor

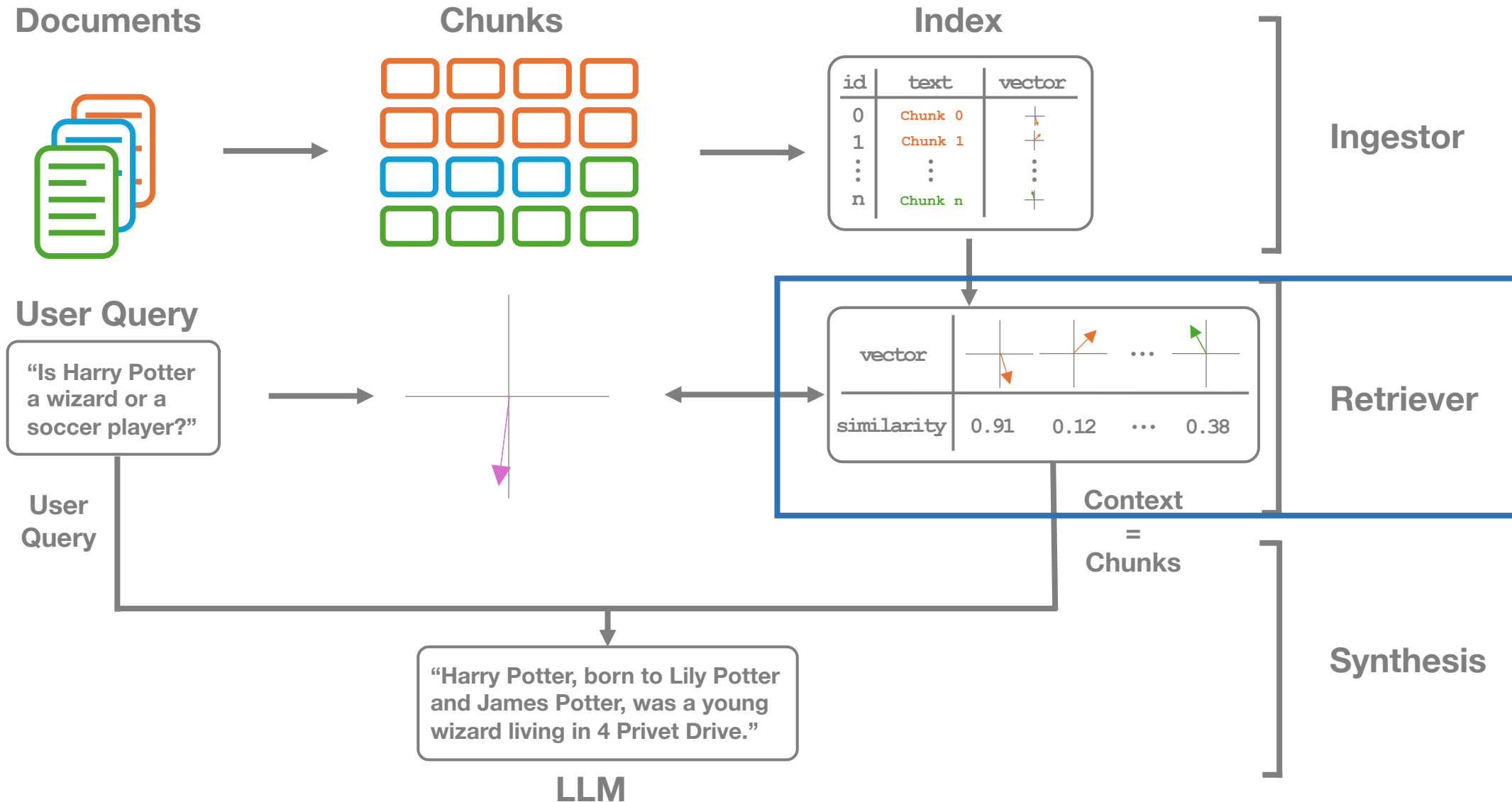
Traditional RAG Pipeline



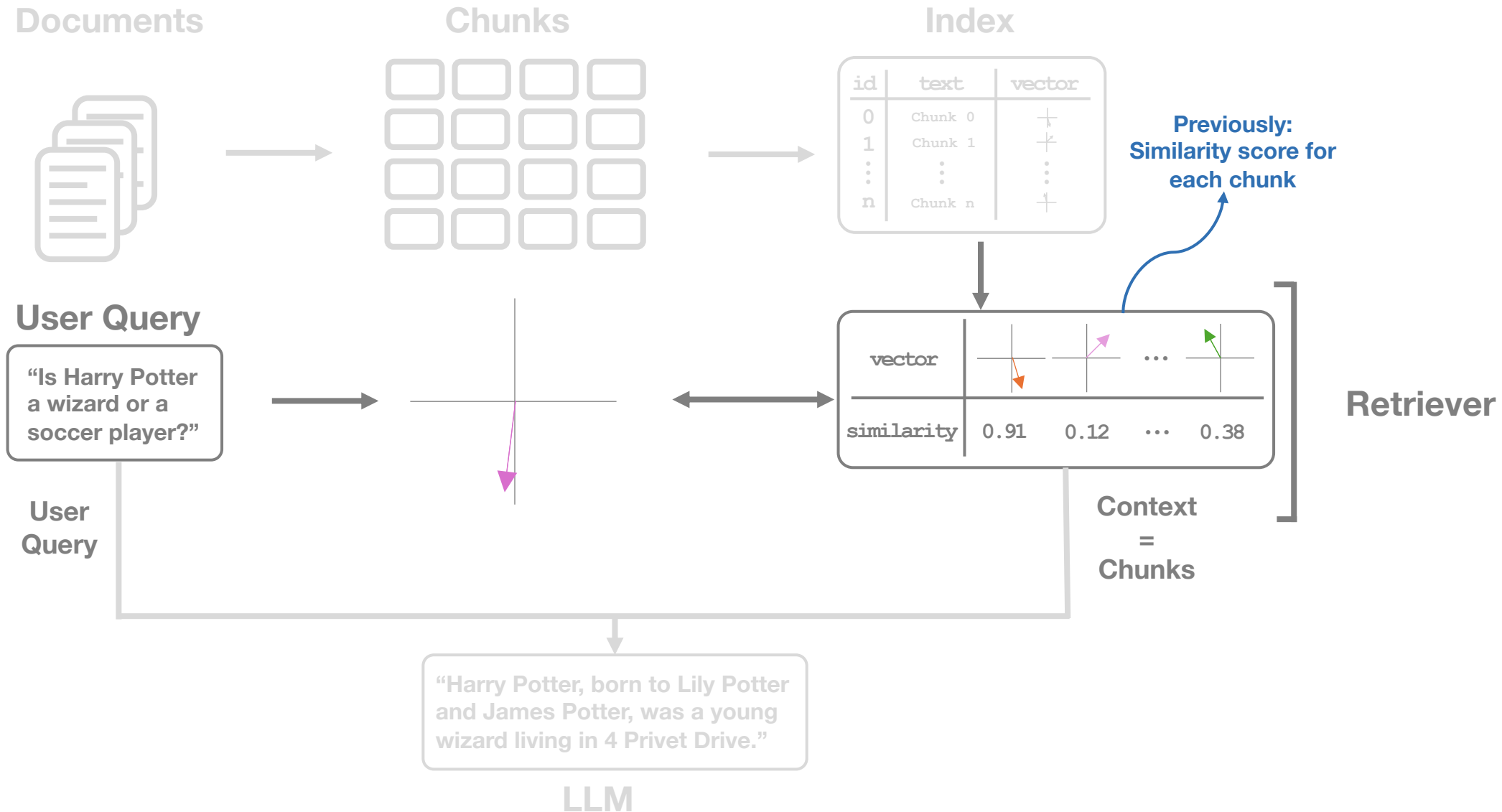
Traditional RAG Pipeline



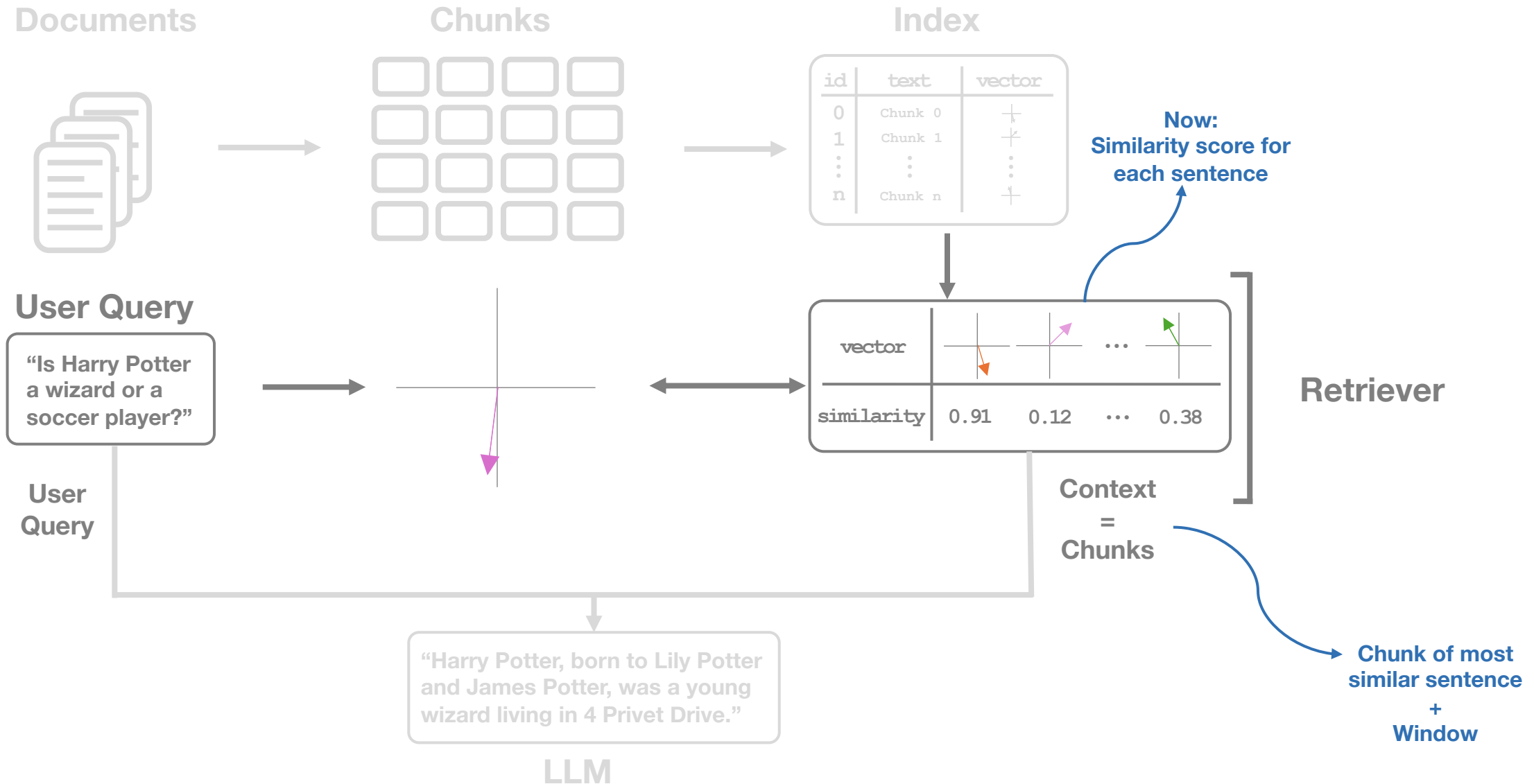
Sentence Window Retrieval



Sentence Window Retrieval



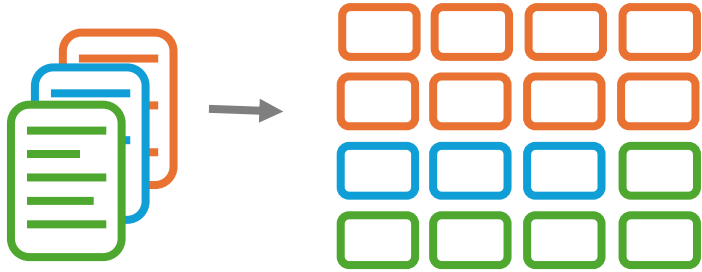
Sentence Window Retrieval



Auto-Merging Retrieval

Documents

Chunks

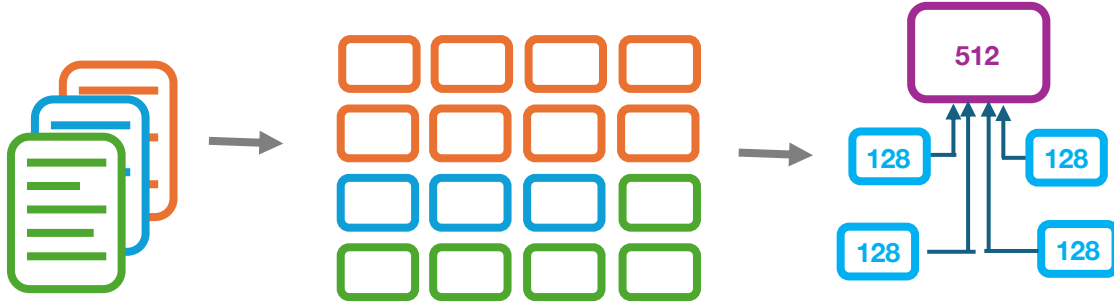


Auto-Merging Retrieval

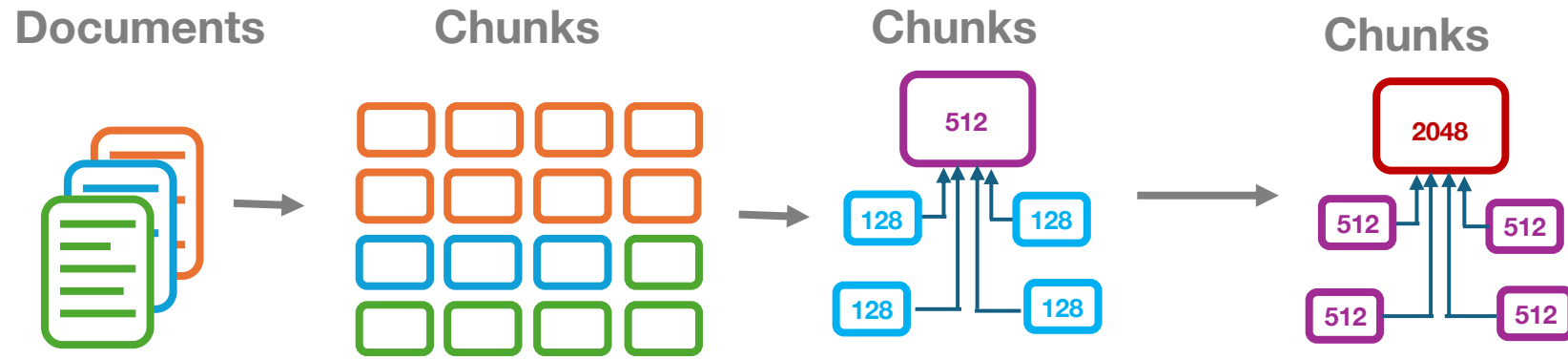
Documents

Chunks

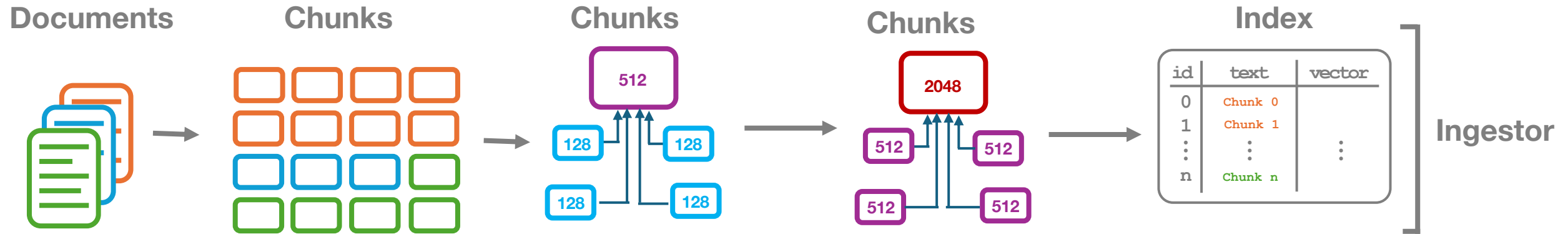
Chunks



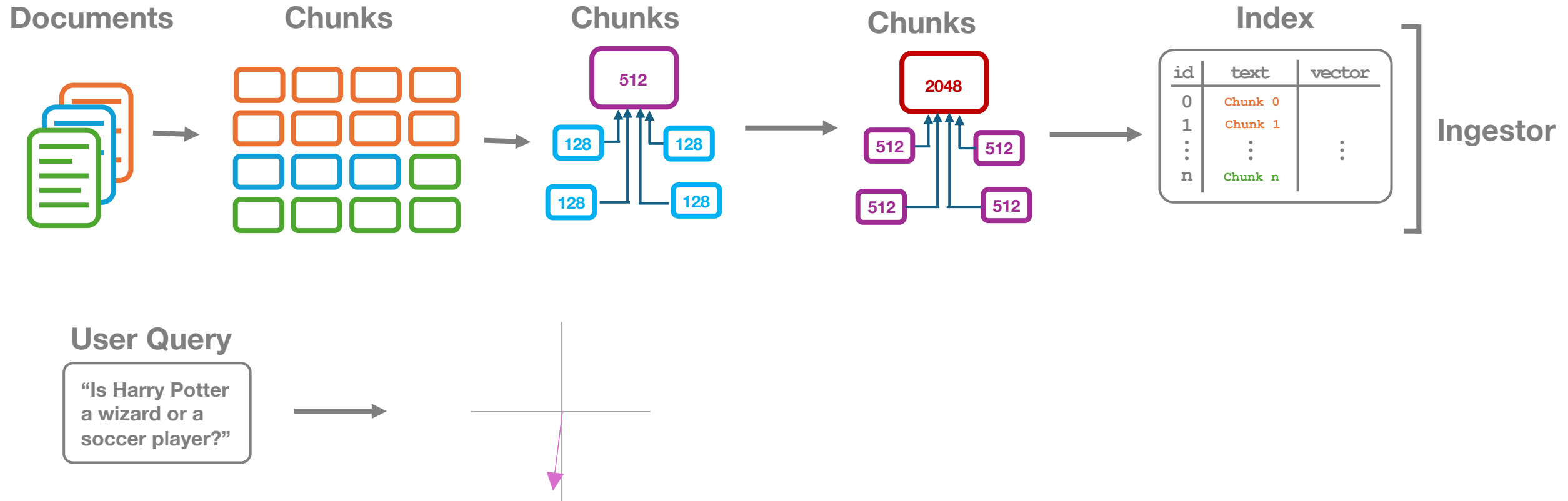
Auto-Merging Retrieval



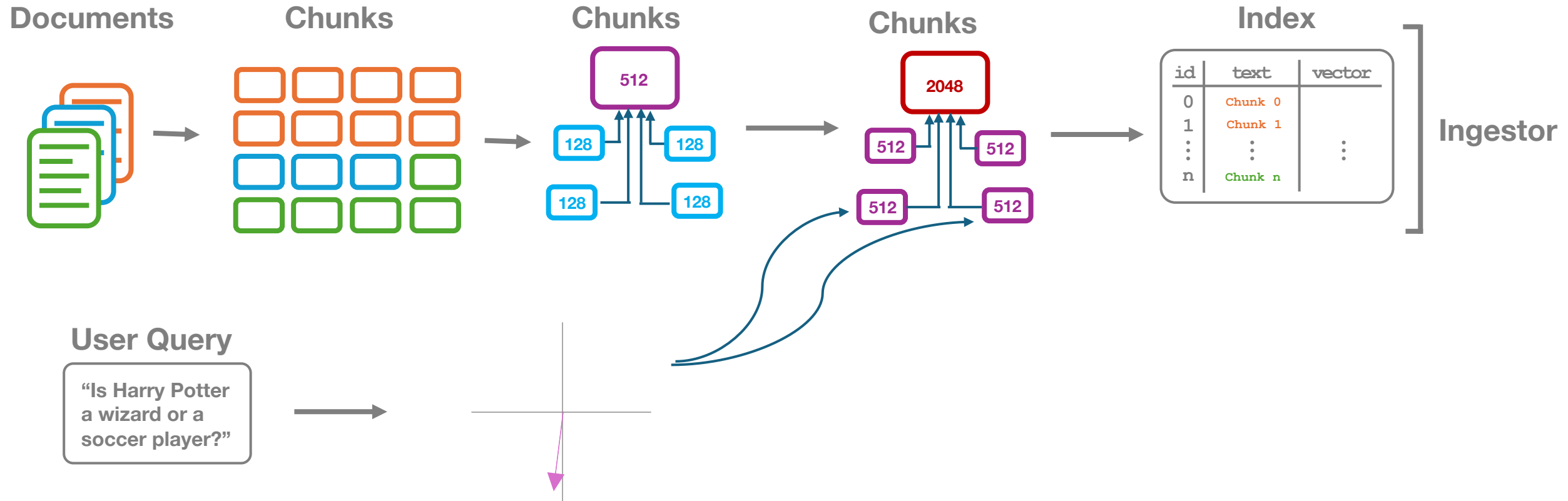
Auto-Merging Retrieval



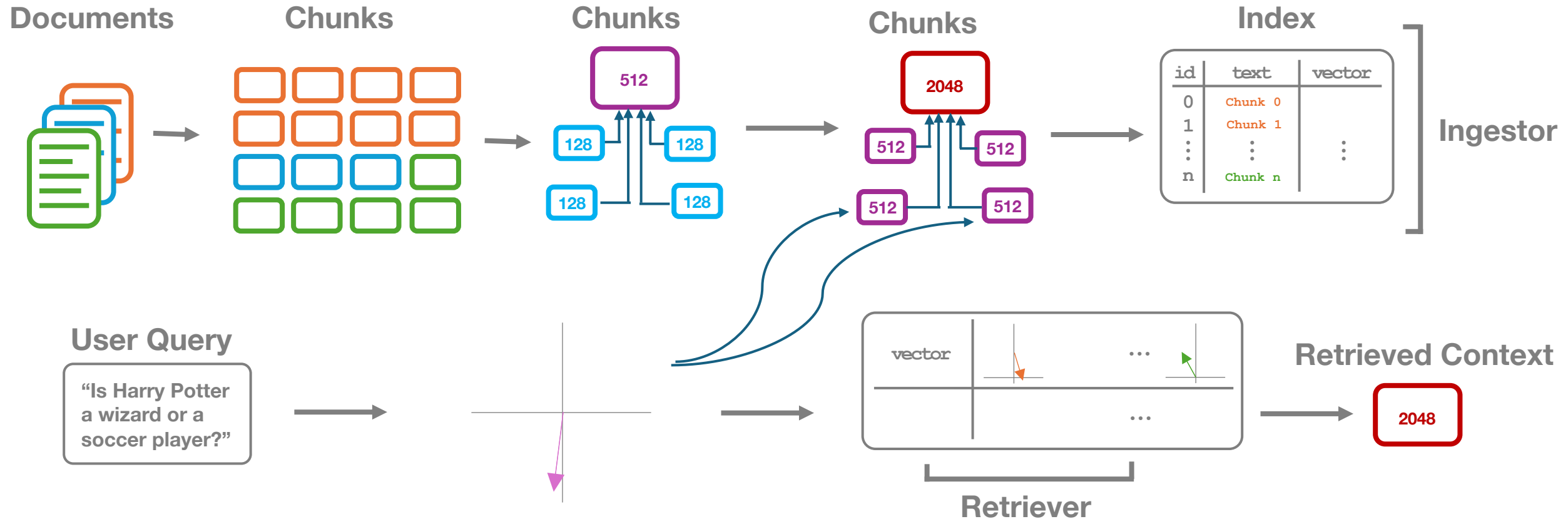
Auto-Merging Retrieval



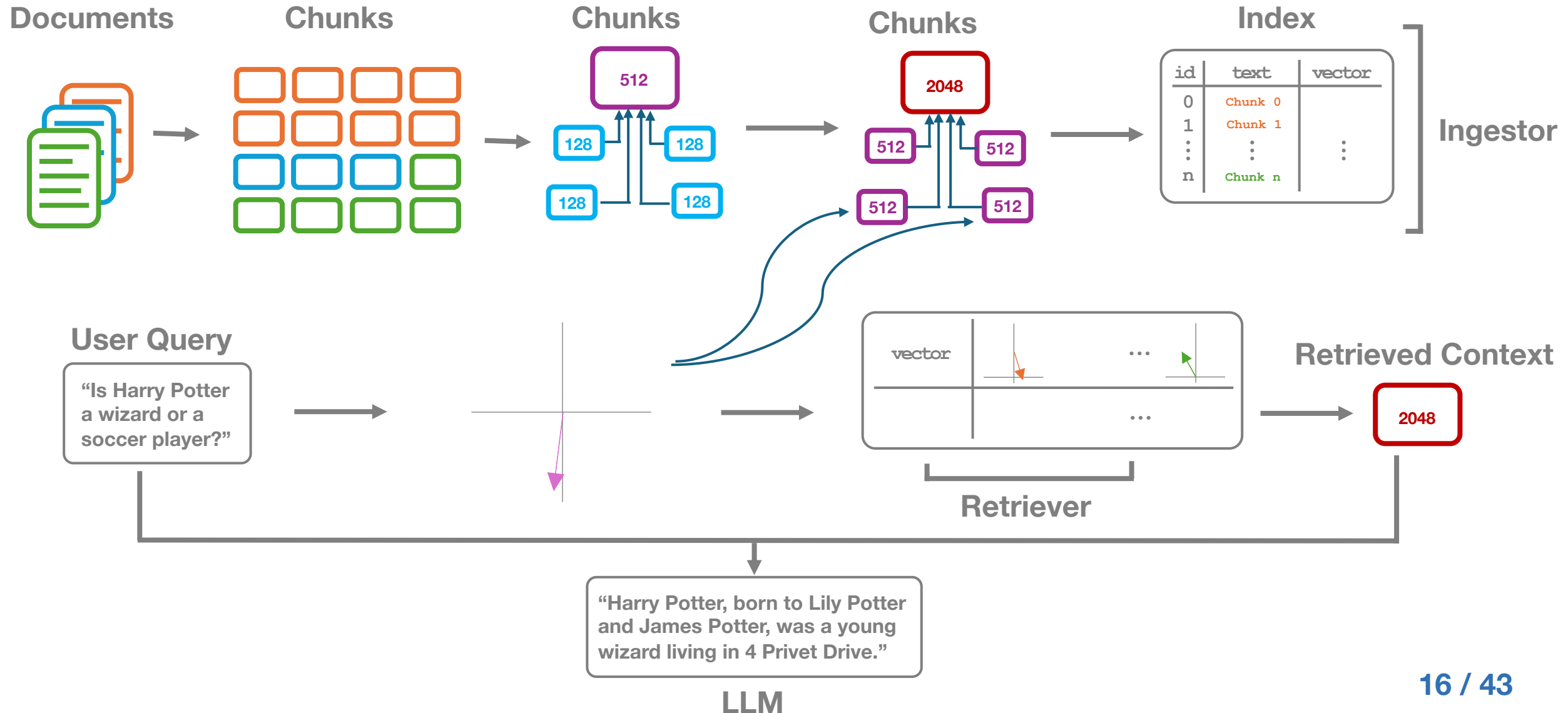
Auto-Merging Retrieval



Auto-Merging Retrieval

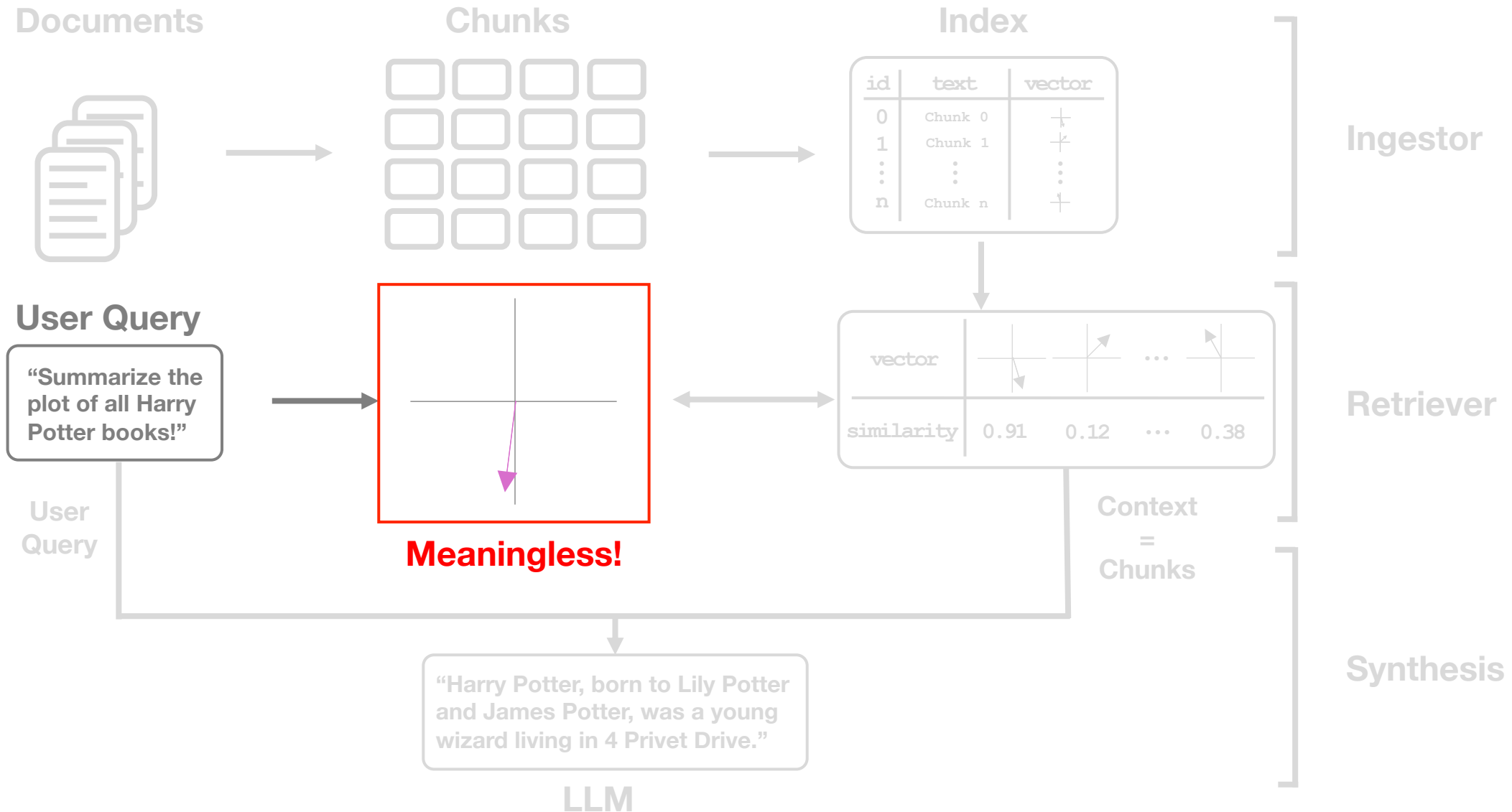


Auto-Merging Retrieval



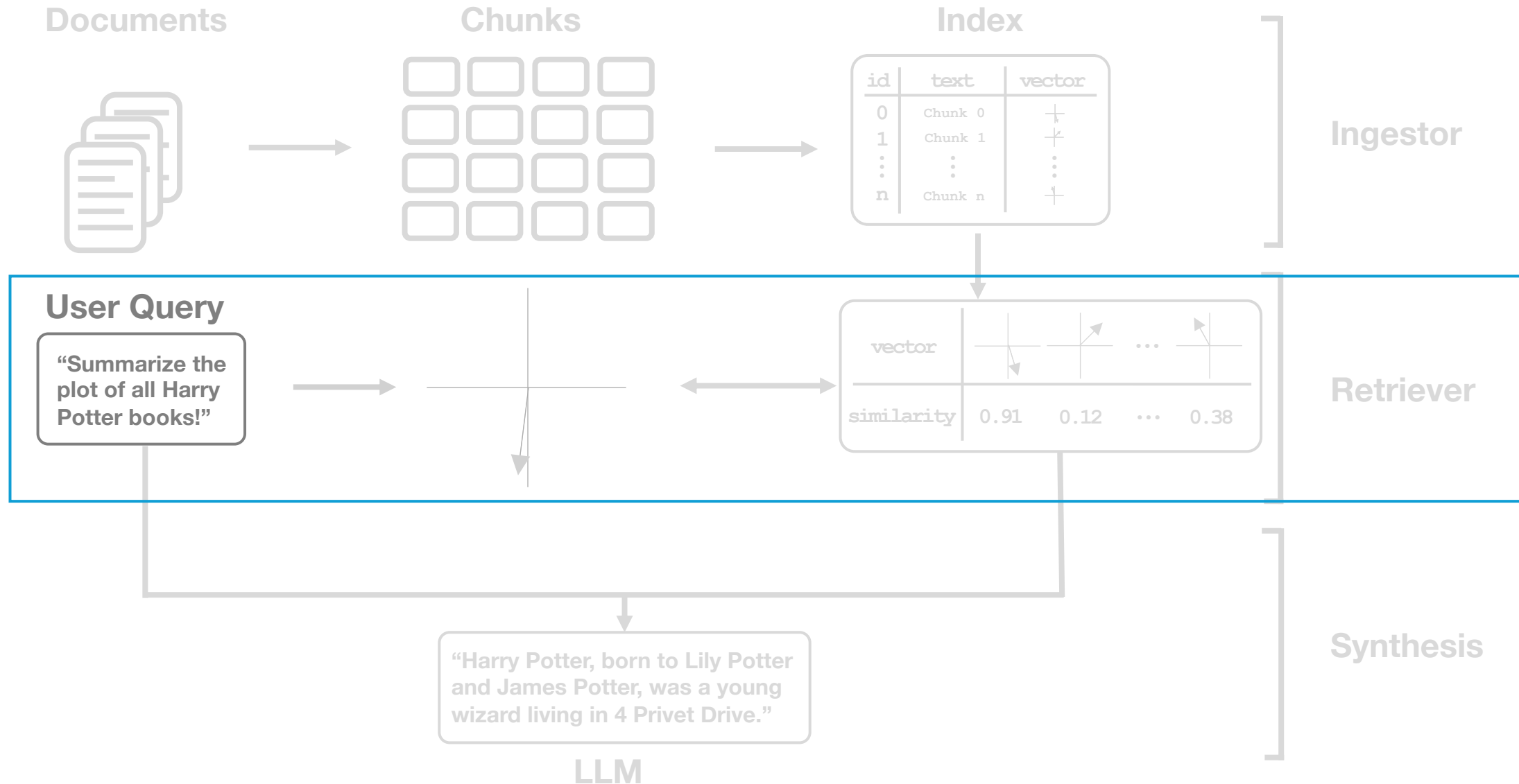
Part 2: Motivation

Motivation






Part 3: Methods

Topic Retrieval

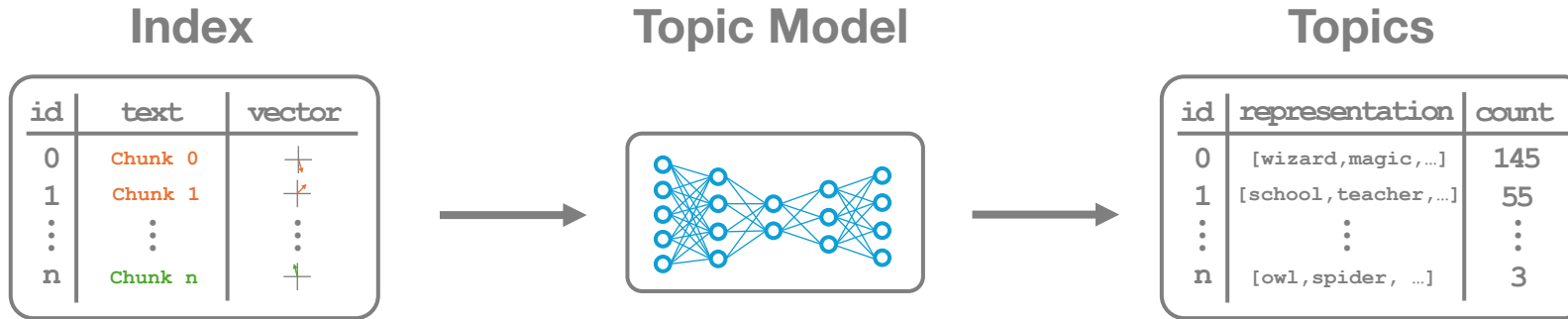


Topic Retrieval

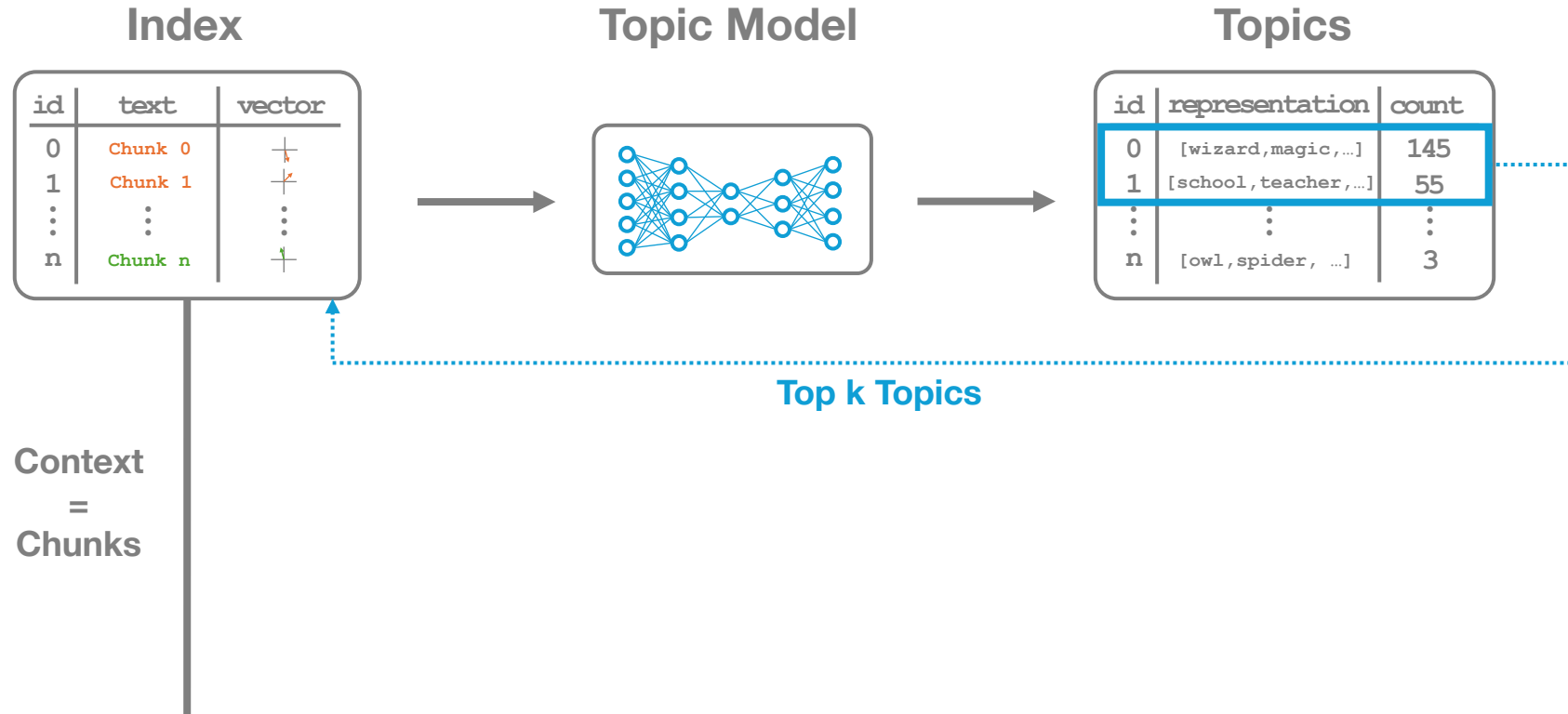
Index

id	text	vector
0	Chunk 0	
1	Chunk 1	
⋮	⋮	⋮
n	Chunk n	

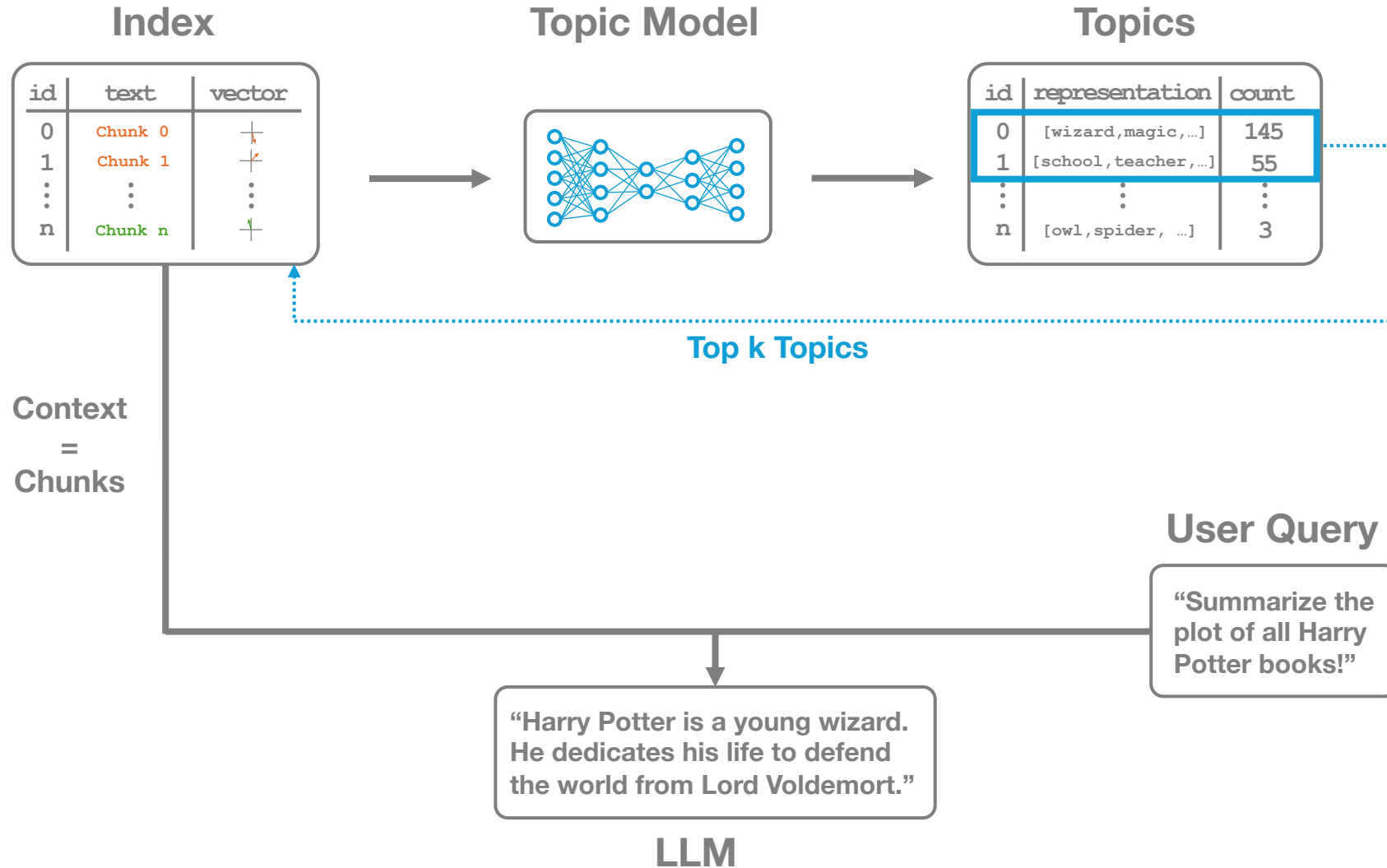
Topic Retrieval



Topic Retrieval



Topic Retrieval



Extractive Summarization



User Query

“Summarize the
following article :
....

Extractive Summarization

User Query

“Summarize the
following article :
....

Embedding Model



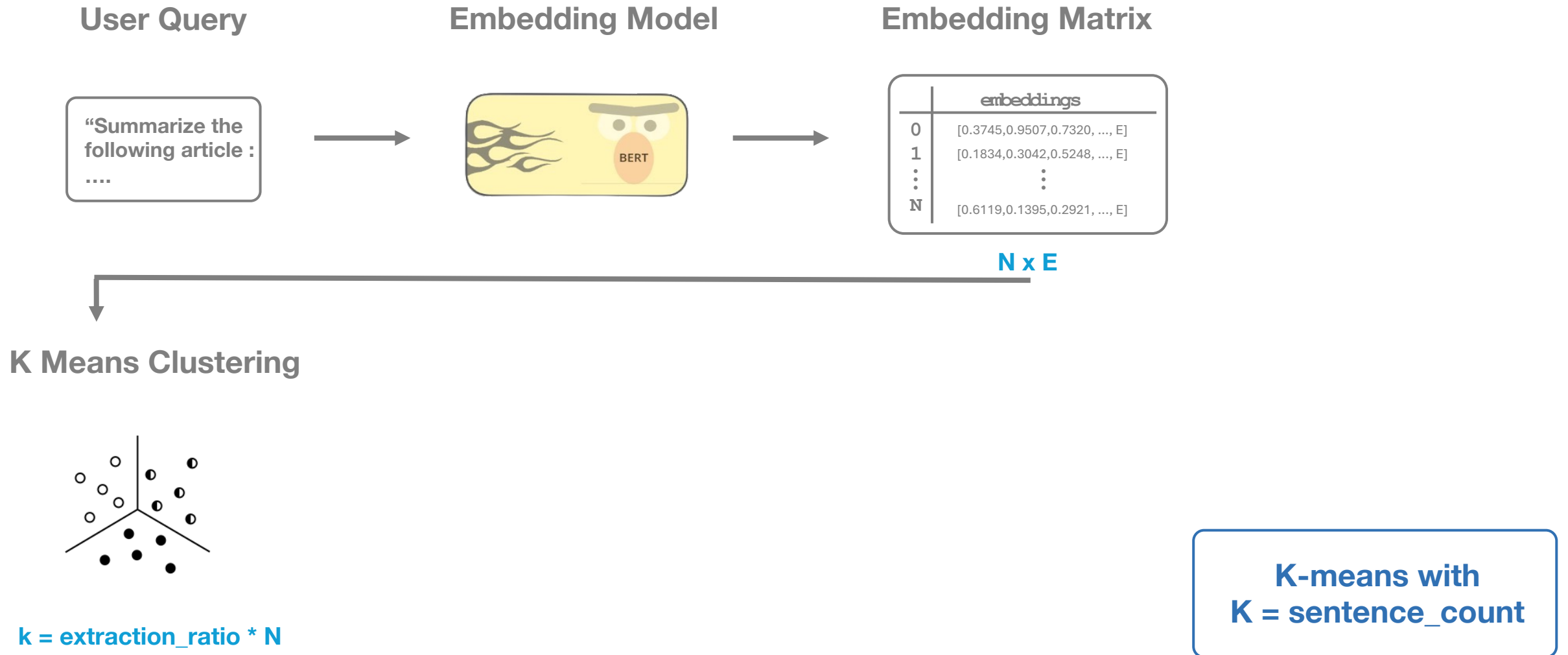
Embedding Matrix

	embeddings
0	[0.3745,0.9507,0.7320, ..., E]
1	[0.1834,0.3042,0.5248, ..., E]
⋮	⋮
N	[0.6119,0.1395,0.2921, ..., E]

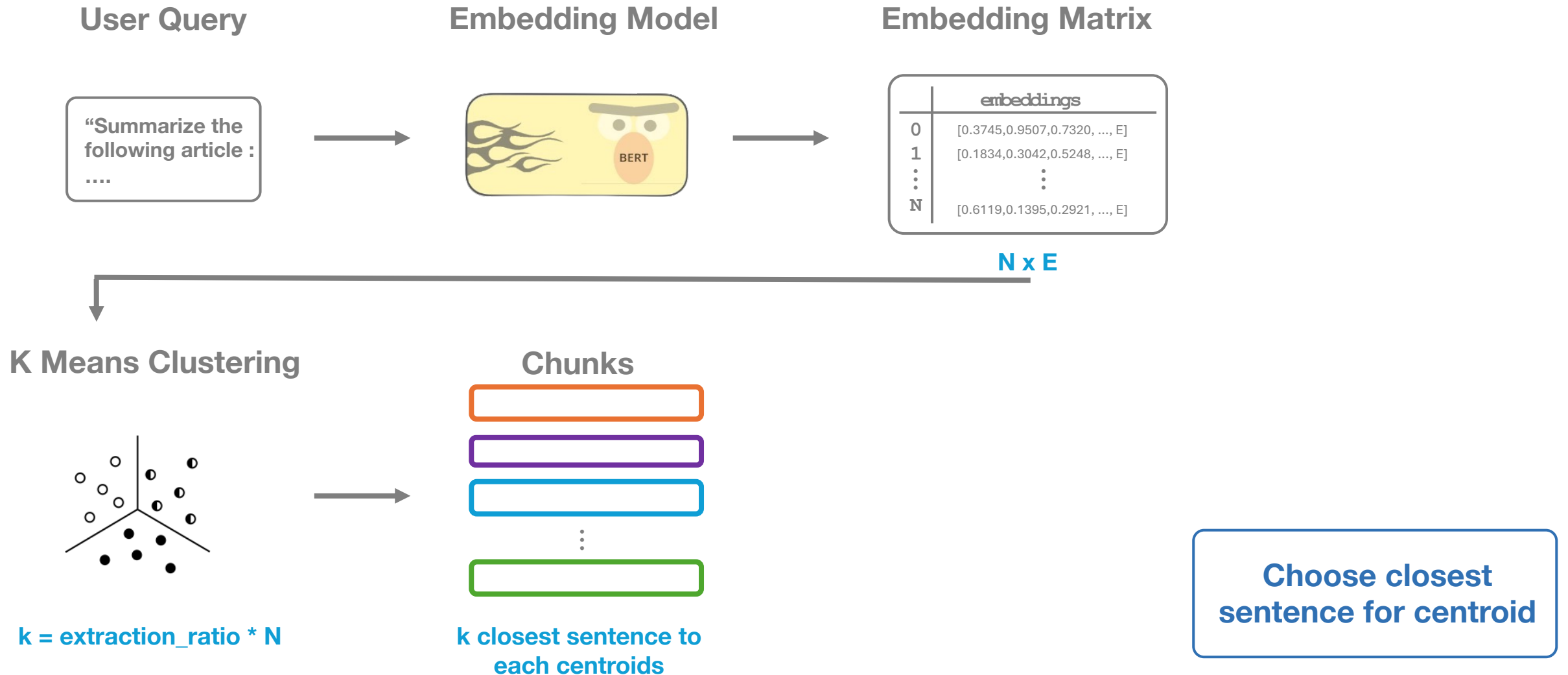
$N \times E$

Sentence Embeddings
instead of Index

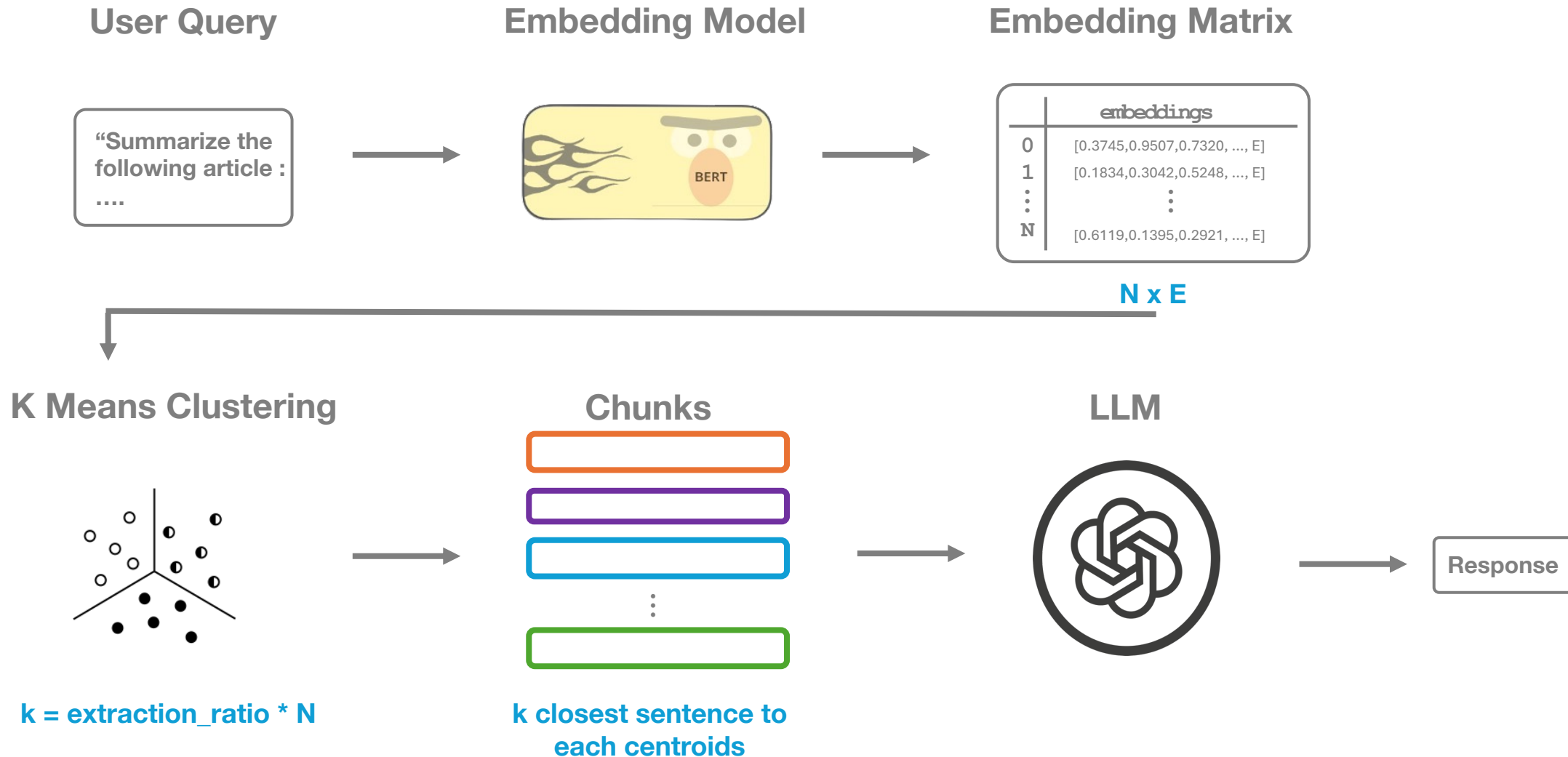
Extractive Summarization



Extractive Summarization

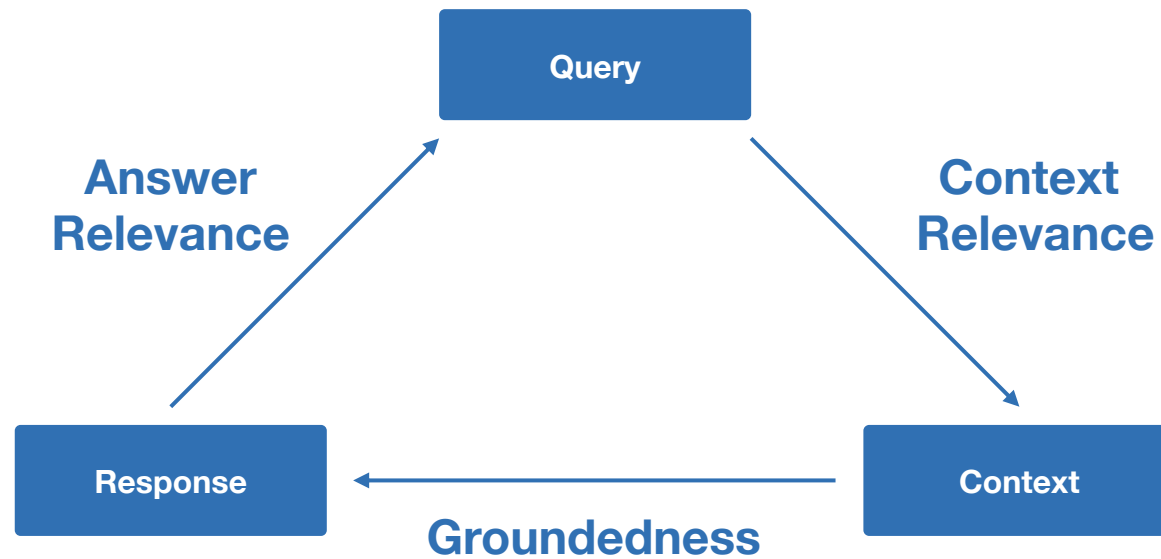


Extractive Summarization



Part 4: Measuring Performance

RAG Triad



Confidence whether LLM app is free from hallucination



Context Relevance: retrieval quality



Groundedness: evidence support for independent claims



Answer Relevance: relevance of the final response to the user input

Summarization Metrics

Bi-Lingual Evaluation Study

Overlap of **n-grams** between reference summary and generated summary. Penalty for differing summary **lengths**.

BLEU (2002)

Recall-Oriented Understudy for Gisting Evaluation

Recall between **n-grams**, **word sequences**, and **word pairs**

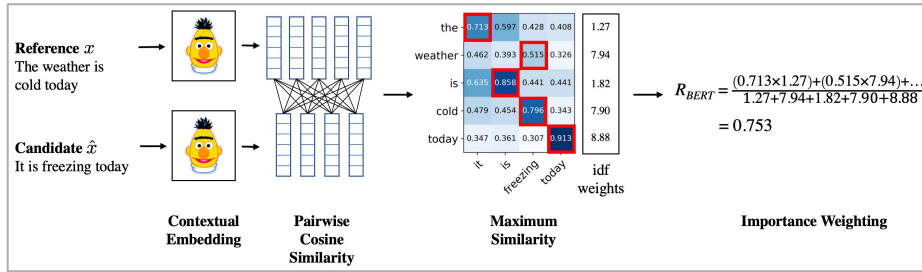
ROUGE (2004)

Metric for Evaluation of Translation with Explicit Ordering

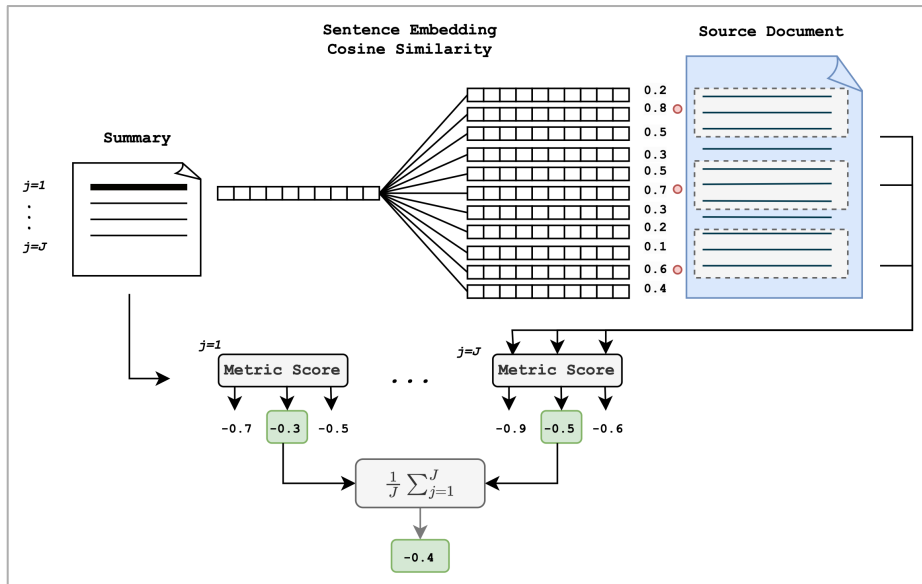
Based on the harmonic mean of **unigram precision** and **recall**, with recall weighted higher than precision

METEOR (2005)

Summarization Metrics



BERT



LDFACTS

BERT (2020)

- Leverages the pre-trained **contextual embeddings** from BERT
- Matches words in candidate and reference sentences by **cosine similarity**
- Correlate with **human judgment** on sentence-level and system-level evaluation

BART (2021)

- Evaluates Generated Text as Text Generation
- Leverages BART's pre-trained **contextual embeddings**
- Measures either the faithfulness, precision, recall, or F-score response of the main text-generation model
- **Recall** quantifies how easily a gold reference could be generated by the hypothesis
- Suitable for pyramid-based evaluation (semantic coverage) in **summarization** task since pyramid score measures fine-grained Semantic Content Units covered by system-generated text

LDFACTS (2024)

- Evaluates the factuality of Long Document Abstractive Summarization
- A reference-free framework for assessing **factual consistency**
- Uses **sentence embeddings** to calculate similarity between source document sentences and predicted summary sentences
- Applies metrics to the **highest similarity text snippets**
- The scores per sentence in the predicted summary averaged to give one score per predicted summary

Part 5: First Results

Example Summaries

The 31-year-old half-back joined Widnes until the end of the season after being told he did not feature in Tigers' plans for the rest of 2017. The former England playmaker was dropped at Cas for three games in March after an "internal investigation". Chase has swapped table toppers Castleford for bottom-of-the-table Widnes, who have won twice in 13 games. "It is a great signing for us and I'm excited we've managed to pick up someone of Rangi's quality to come into the side," Betts told BBC Radio Merseyside. "I know he's excited about playing as he wants to be playing regularly - he wants to start games and play 80 minutes. "He comes here, he's got a clean slate and he wants to go again. "He's up for the challenge as he knows we've got our backs against the wall, he knows the situation and he's been told that. He's up for the fight which was really exciting."

Original Text

Extractive: A 31-year-old half-back named Rangi Chase has joined Widnes until the end of the season after being informed that he was not part of Castleford Tigers' plans for the remainder of 2017. Widnes coach Betts expressed enthusiasm about the signing, highlighting Chase's quality and readiness for the challenge ahead. Chase, who was dropped by Castleford due to an internal investigation, has moved from the top team to the bottom team in the league, Widnes, who have only won twice in 13 games

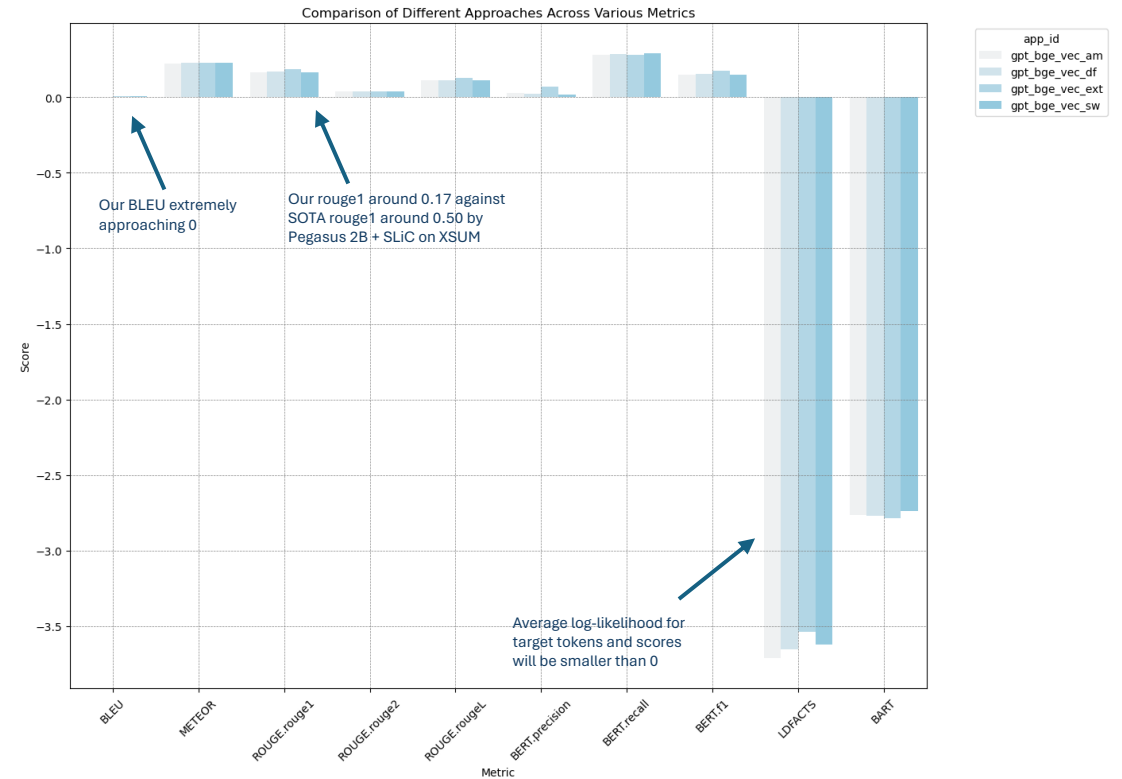
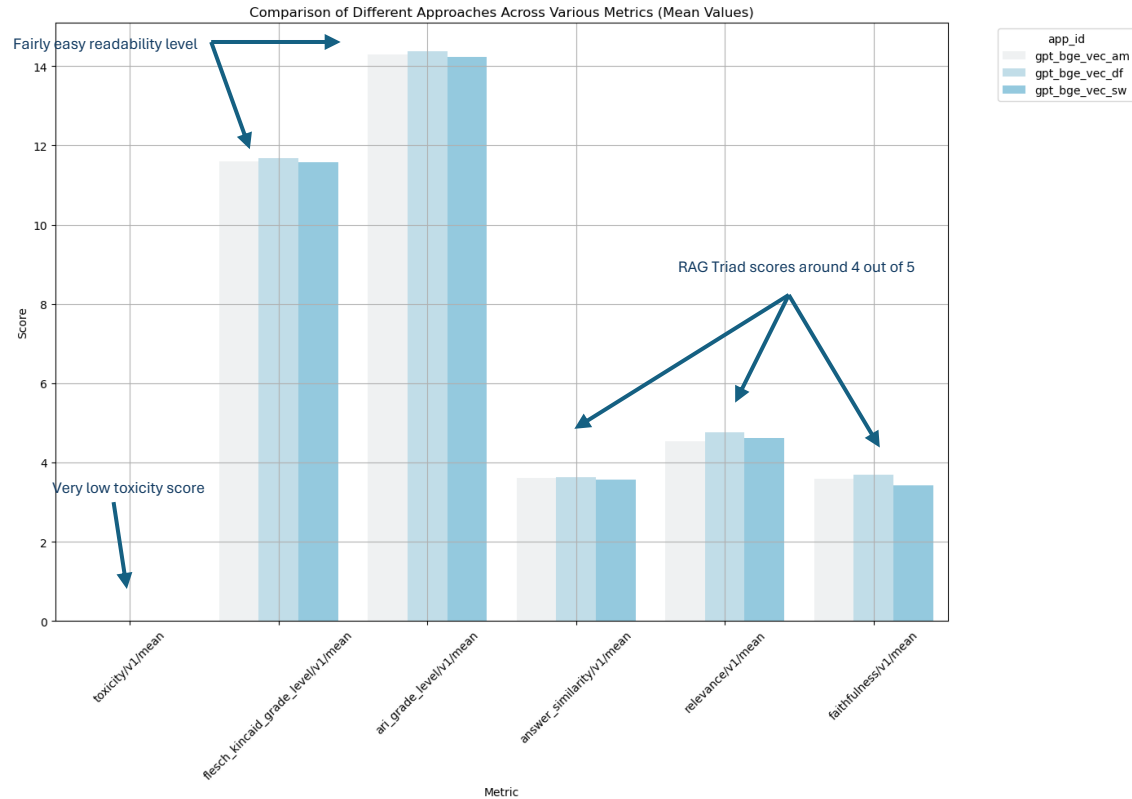
Standard RAG: The document discusses the signing of 31-year-old half-back Rangi Chase by Widnes until the end of the season. Chase was dropped by Castleford after an internal investigation and has now joined Widnes, who are struggling at the bottom of the table. Widnes coach Betts expressed excitement about the signing, highlighting Chase's quality and eagerness to play regularly and contribute to the team. Chase is motivated by the challenge and is ready to fight for the team despite their difficult situation.

Sentence-Window: A 31-year-old half-back named Rangi Chase joined Widnes until the end of the season after being dropped from Castleford's plans for the rest of 2017. Chase was involved in an internal investigation at Castleford and subsequently moved to Widnes, who are currently at the bottom of the table with two wins in 13 games. Widnes coach Betts expressed excitement about signing Chase, highlighting his quality and eagerness to play regularly and contribute to the team. Chase is motivated by the challenge and opportunity to start games and play full matches for Widnes.

Auto-Merging: The 31-year-old half-back, who was dropped from his previous team after an internal investigation, has joined Widnes until the end of the season. Widnes, currently at the bottom of the table, sees this signing as a positive move and is excited to have Rangi Chase on board. Chase is eager to play regularly, start games, and contribute fully to the team's efforts, showing enthusiasm for the challenge ahead.

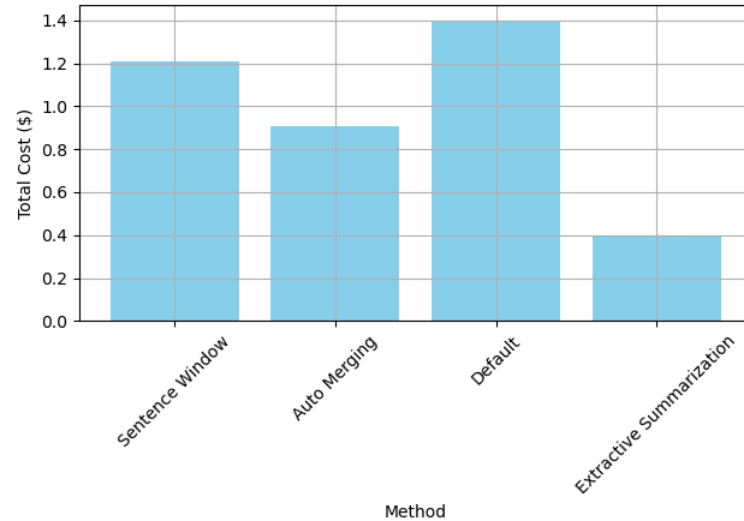
Reference: New Widnes loan signing Rangi Chase has a "clean slate" start at the Vikings, says head coach Denis Betts

Performance Analysis

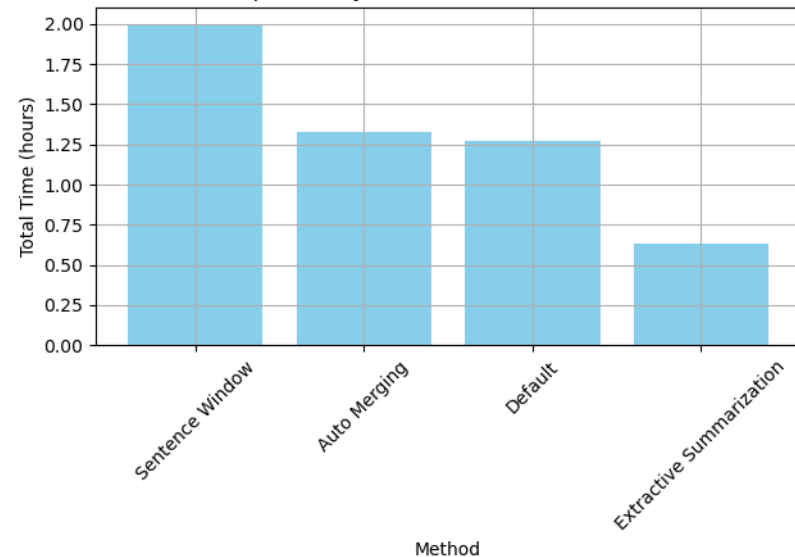


Cost and Time Analysis

Total OpenAI API Cost Comparison by Method on 281 Queries of XSUM Dataset



Total Time Comparison by Method on 281 Queries of XSUM Dataset



Current Issues

- 1 OpenAI Rate Limit
- 2 Too Slow Retrieval with Topic Extraction
- 3 Out of GPU Memory when loading Index + LLM

Part 6: Next Steps

Next Steps

- 1 Implement Knowledge Graph Index
- 2 Semantic Chunker Retrieval Strategy
- 3 Tune hyperparameters (chunk sizes, prompts etc.)

Methodological

- 1 Large (Single-) Document Benchmarks
- 2 Migrate from MLFlow to RAGAS
- 3 Use Multi-News Dataset

Evaluation

Part 7: Contribution Clarification

Contribution Clarification

- 1 Ingestion Part of RAG Pipeline
- 2 Extractive Summarization Implementation
- 3 Refactoring / Designing Pipeline structure, Version Management
- 4 Report / Presentation

Göçmen, Berke

- 1 Evaluation Part of RAG Pipeline
- 2 Benchmarking
- 3 Code Documentation
- 4 Report / Presentation

Quan, Guangyao

- 1 Retrieval Part of RAG Pipeline
- 2 Topic Retrieval Implementation
- 3 Literature Review and Research
- 4 Report / Presentation

Wiehe, Luca

Part 8: References

References



Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tianyi Zhang* and Varsha Kishore* and Felix Wu* and Kilian Q. Weinberger and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *International Conference on Learning Representations*.

Yuan, Weizhe and Neubig, Graham and Liu, Pengfei. 2021. BARTScore: Evaluating Generated Text as Text Generation. *NEURIPS2021*.

Jennifer A Bishop, Qianqian Xie, Sophia Ananiadou. 2024. LongDocFACTScore: Evaluating the Factuality of Long Document Abstractive Summarisation. *LREC-COLING*.

Yao Zhao and Misha Khalman and Rishabh Joshi and Shashi Narayan and Mohammad Saleh and Peter J. Liu. 2022. Calibrating Sequence likelihood Improves Conditional Language Generation.