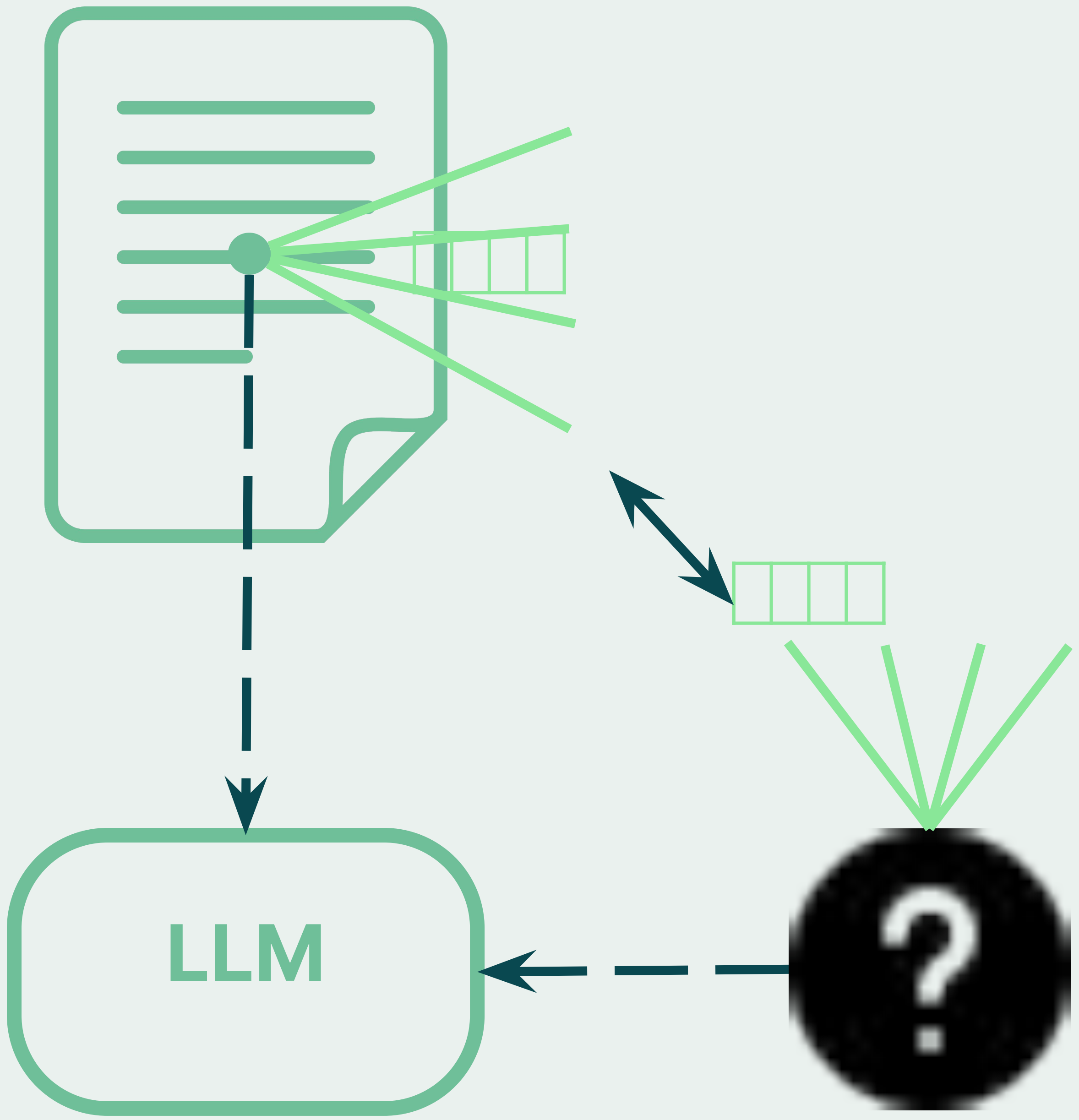
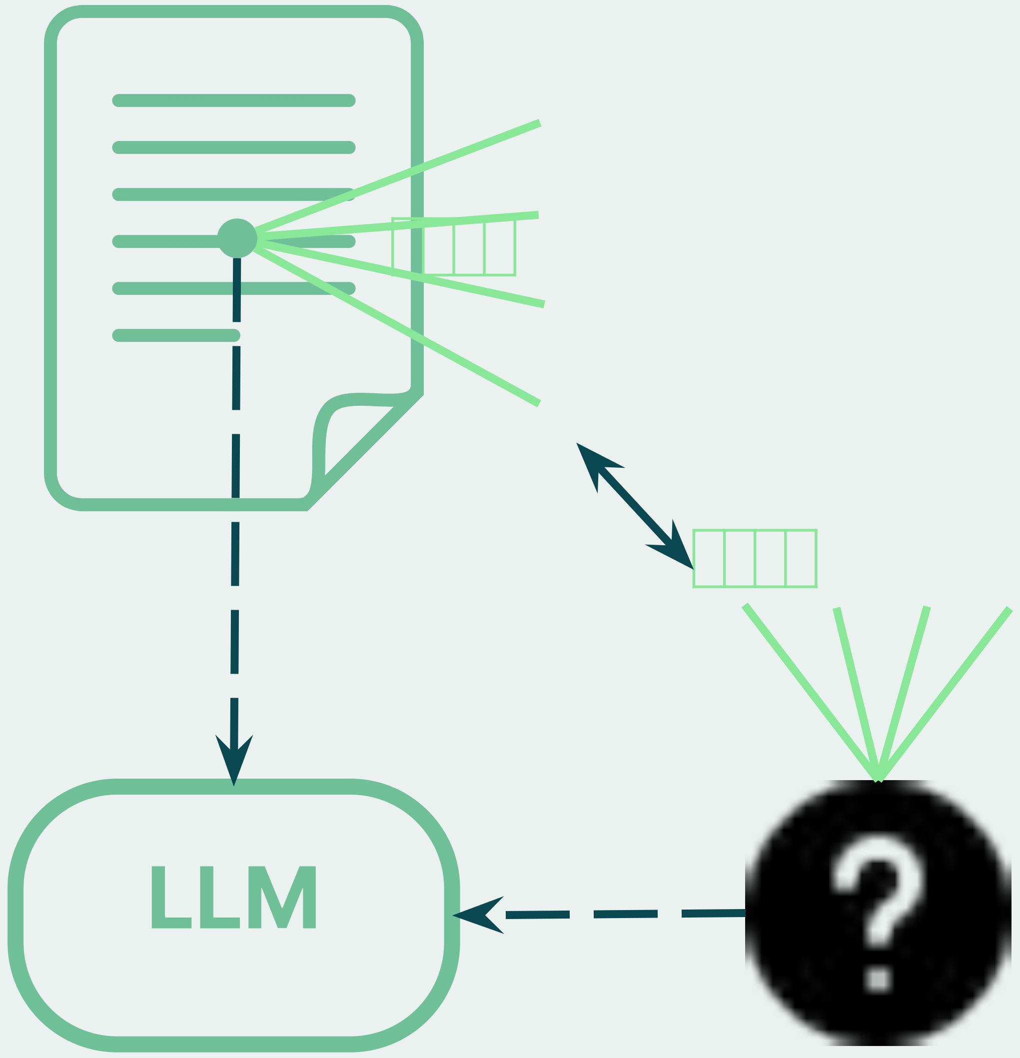
Retrieval Augmented Generation

thanks to Lucía Urcelay



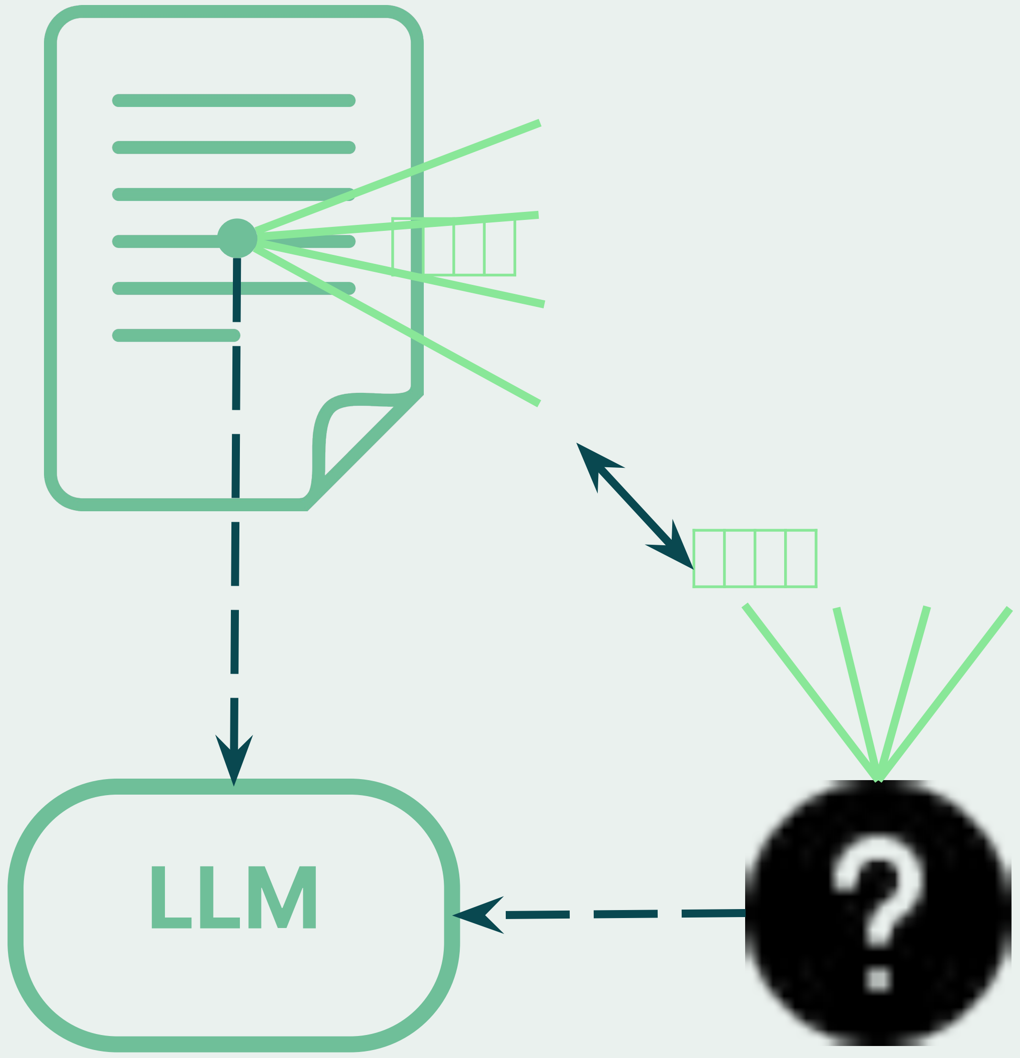
Basic idea



Integrate LLM with an information retrieval engine

- Embed documents into vectors
- Find most similar documents
- Add as context for response

Why RAG

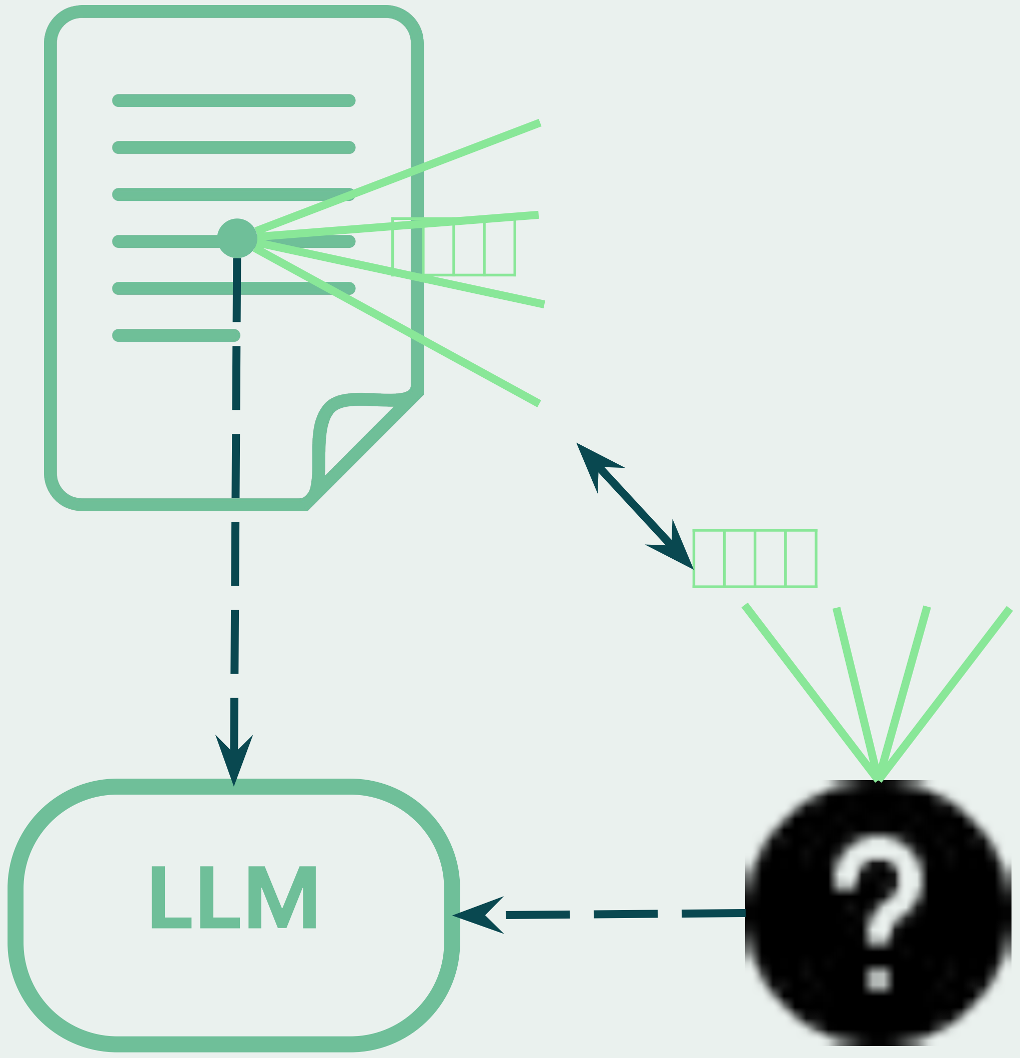


Access to up-to-date knowledge

Reduce hallucinations/Improve factuality

Reliable referencing (sources)

Why not RAG



Limited context

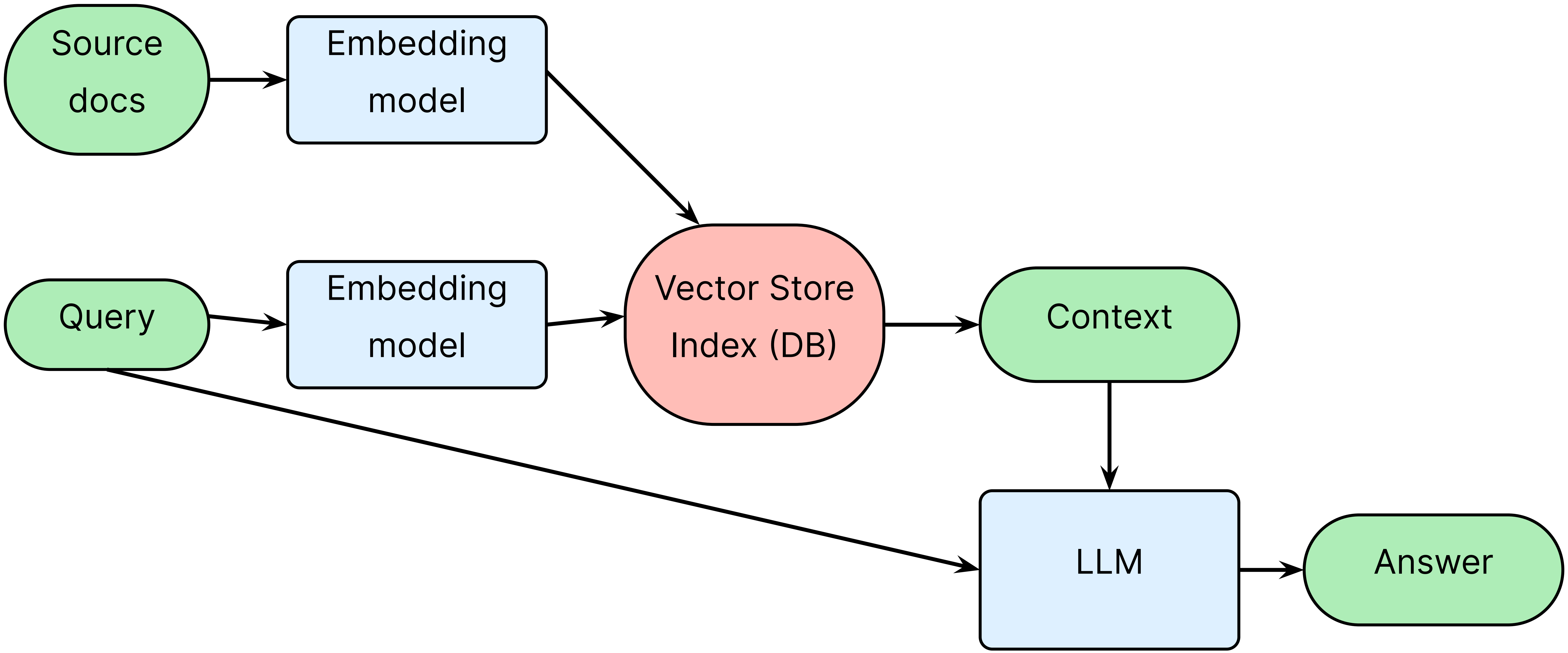
Cost of inference

Cost of persistence

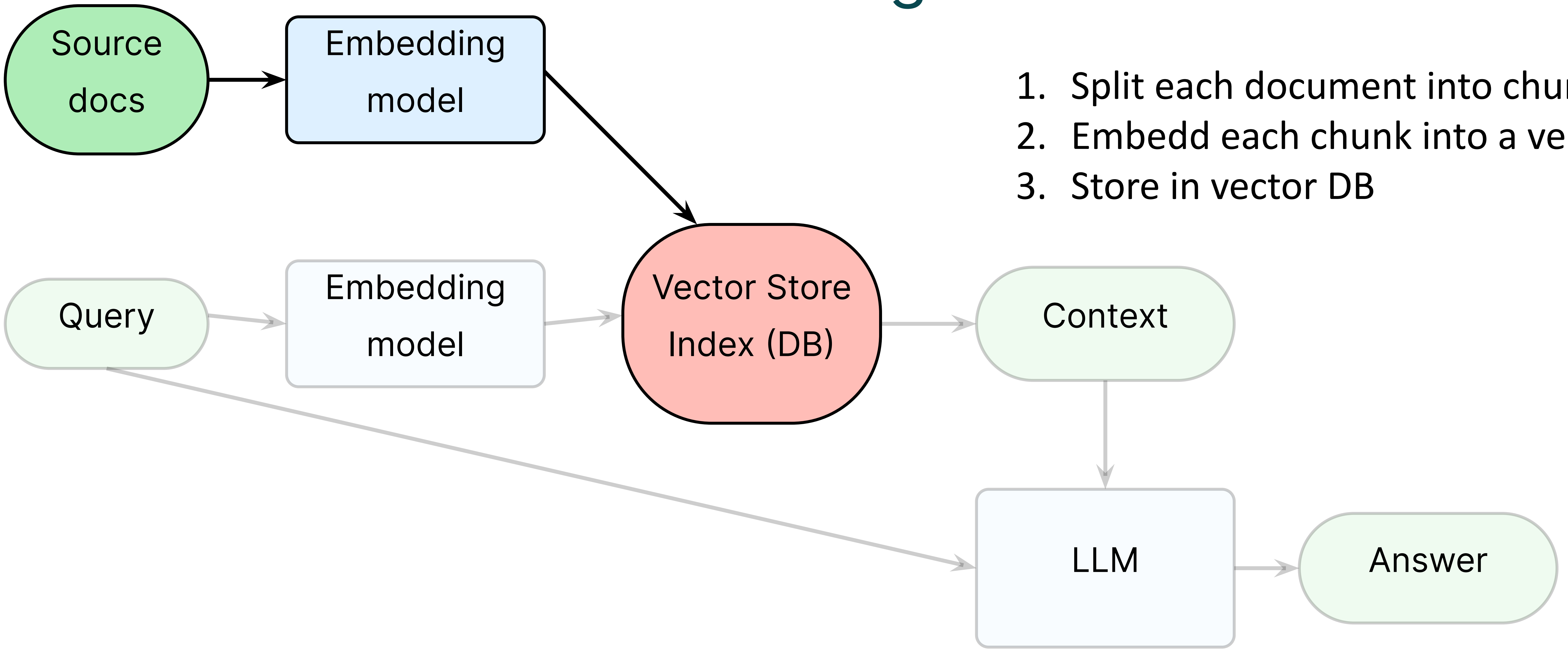
No inter-document processing

Baseline RAG

Baseline RAG

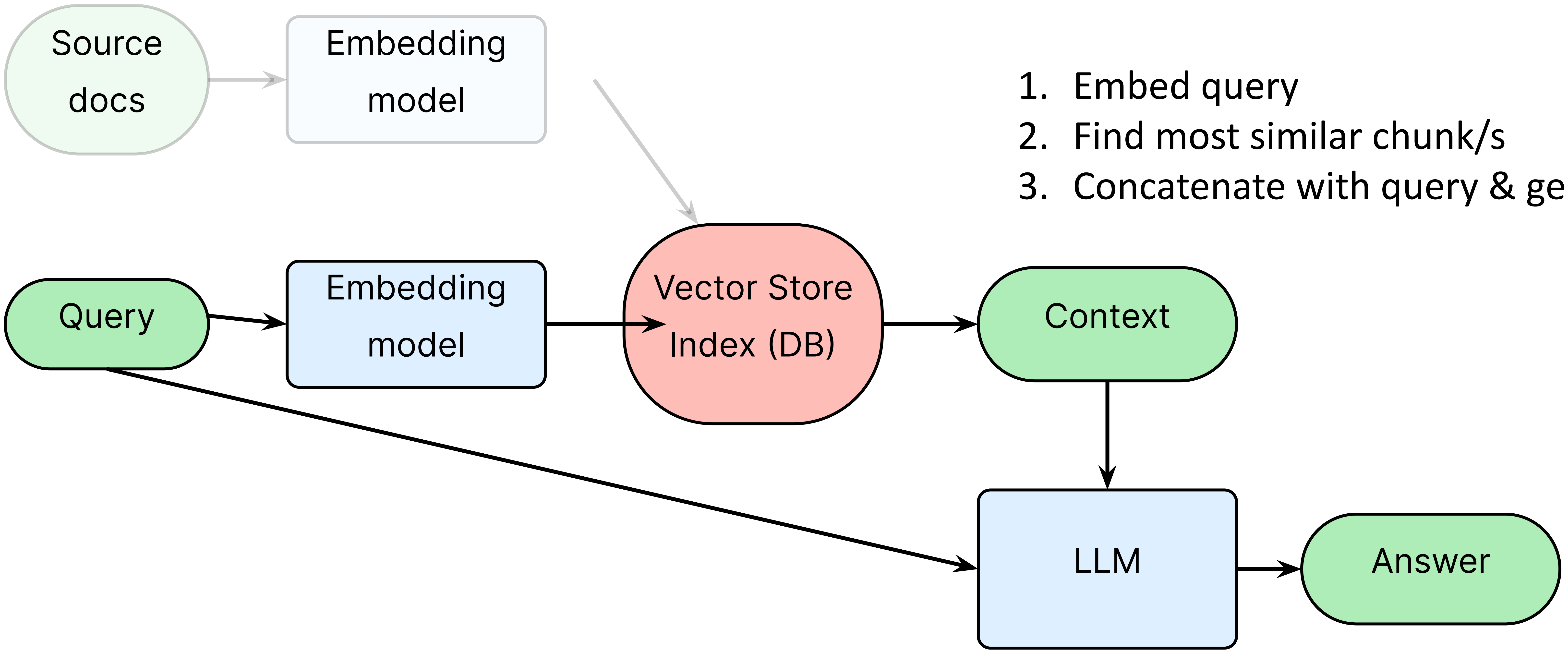


Phase 1: Document indexing



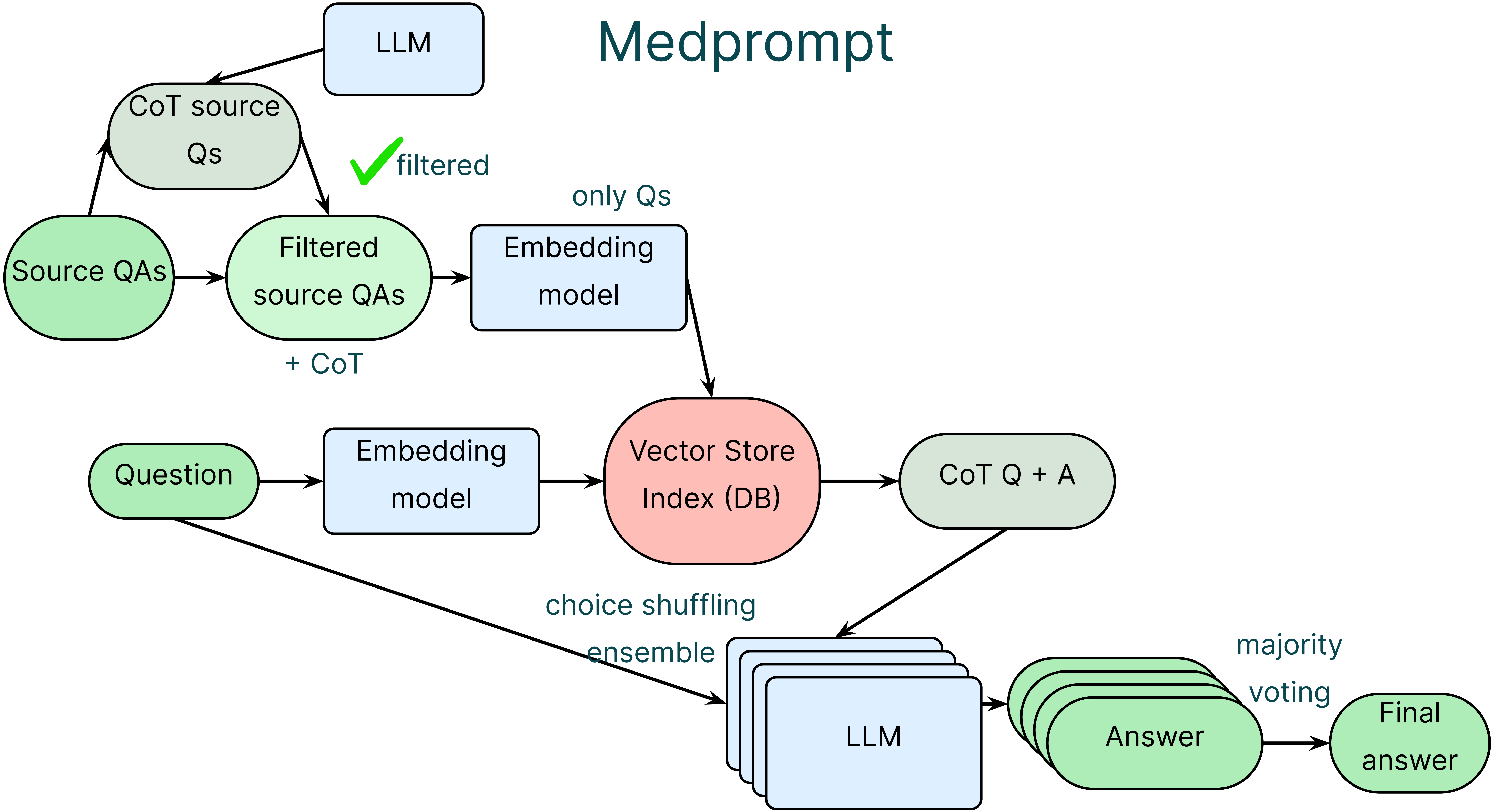
1. Split each document into chunks
2. Embedd each chunk into a vector
3. Store in vector DB

Phase 2: Inference

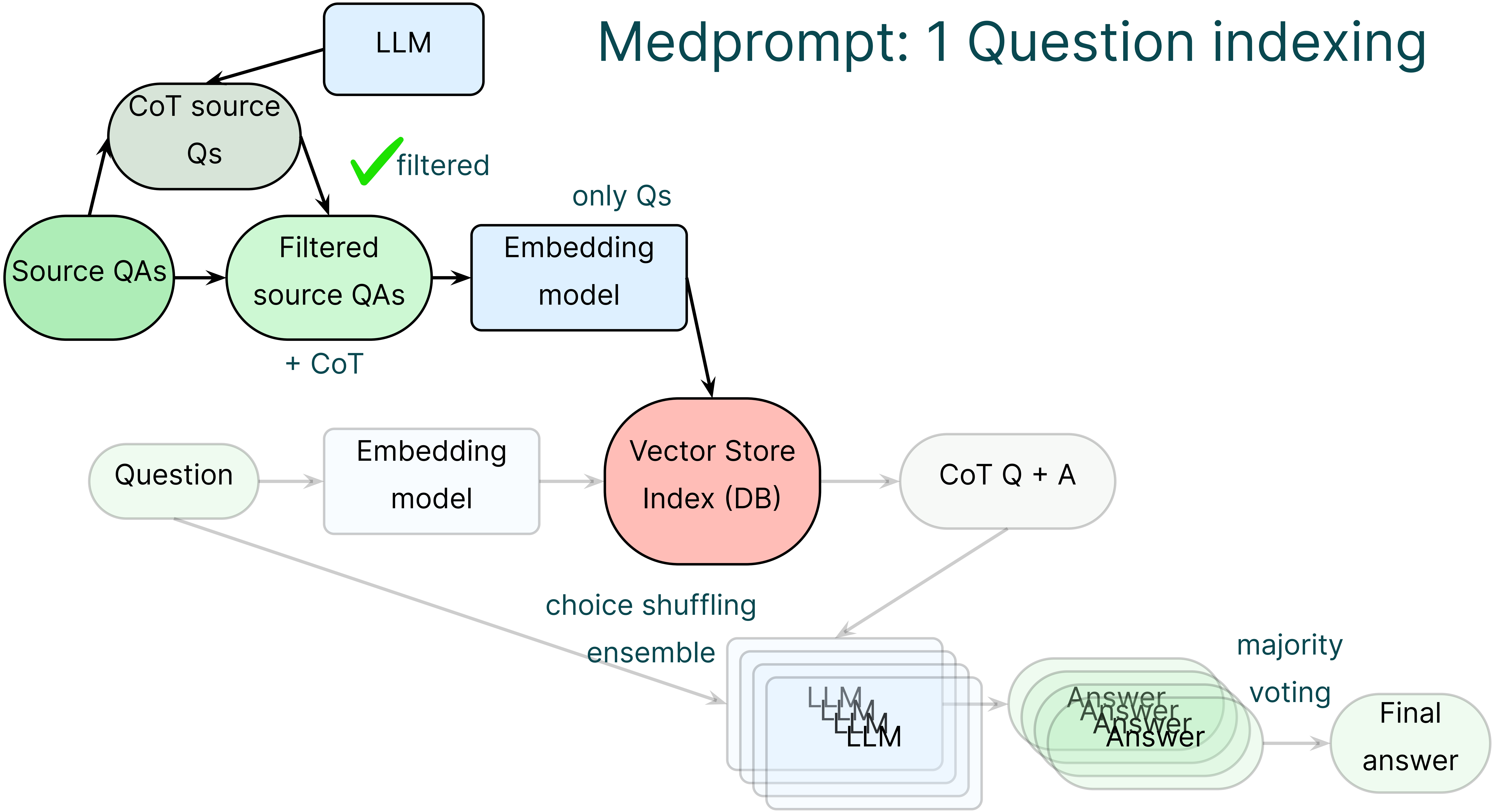


1. Embed query
2. Find most similar chunk/s
3. Concatenate with query & generate

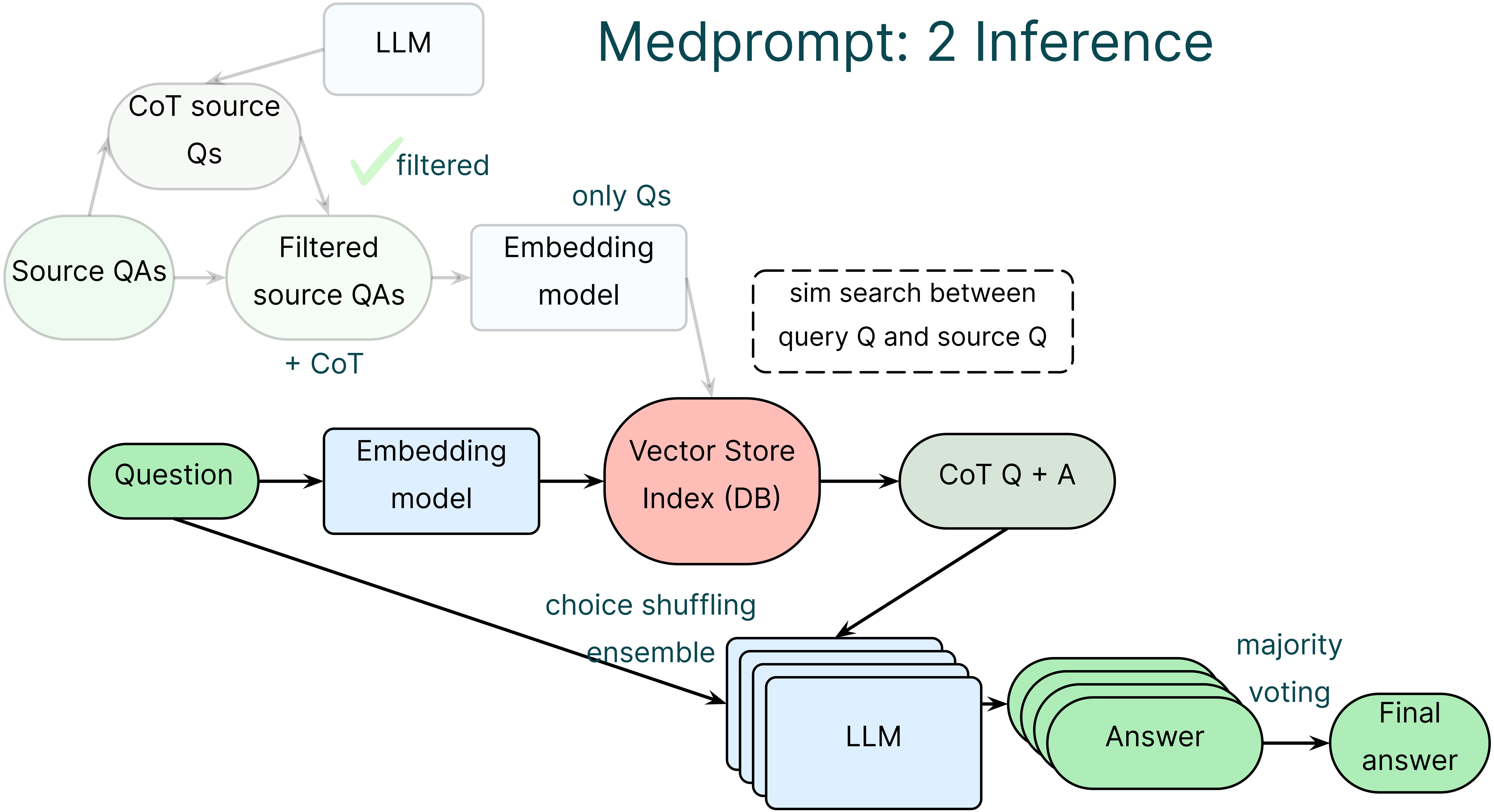
Medprompt


Medprompt
LLM
CoT source Qs
filtered
only Qs
Source QAs
Filtered source QAs
+ CoT
Embedding model
Question
Embedding model
Vector Store Index (DB)
CoT Q + A
choice shuffling ensemble
LLM
Answer
majority voting
Final answer

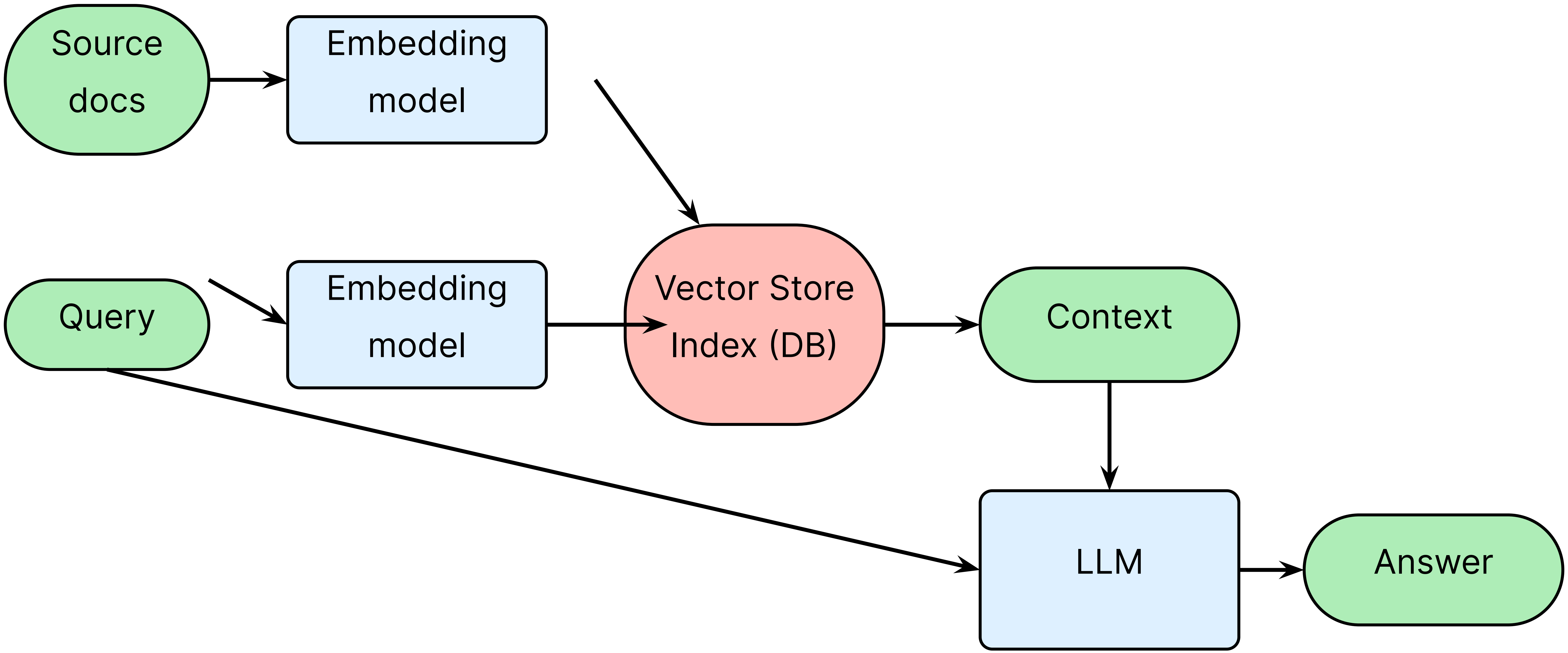
Medprompt: 1 Question indexing



Medprompt: 2 Inference



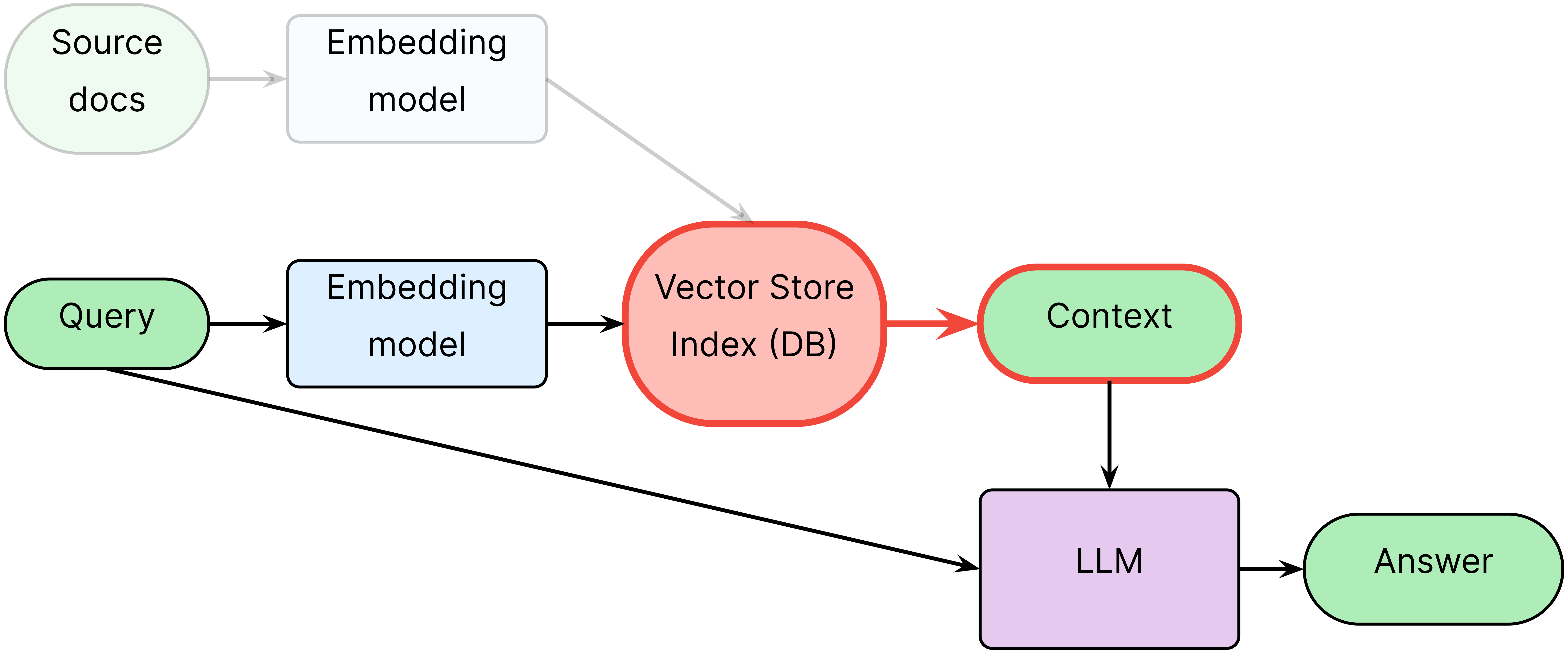
Back to baseline RAG



RAG

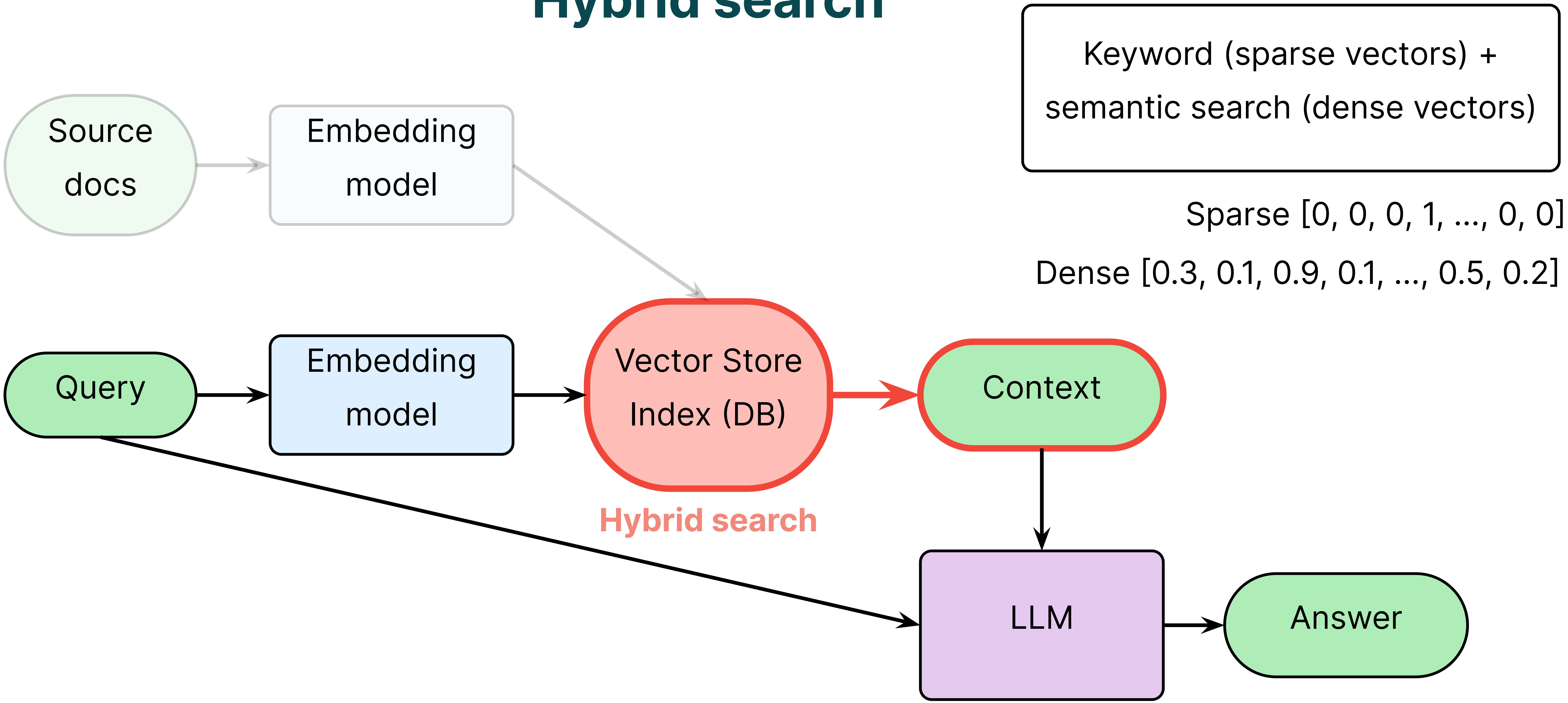
Painpoints

Retrieval of low-relevance docs/chunks (1)

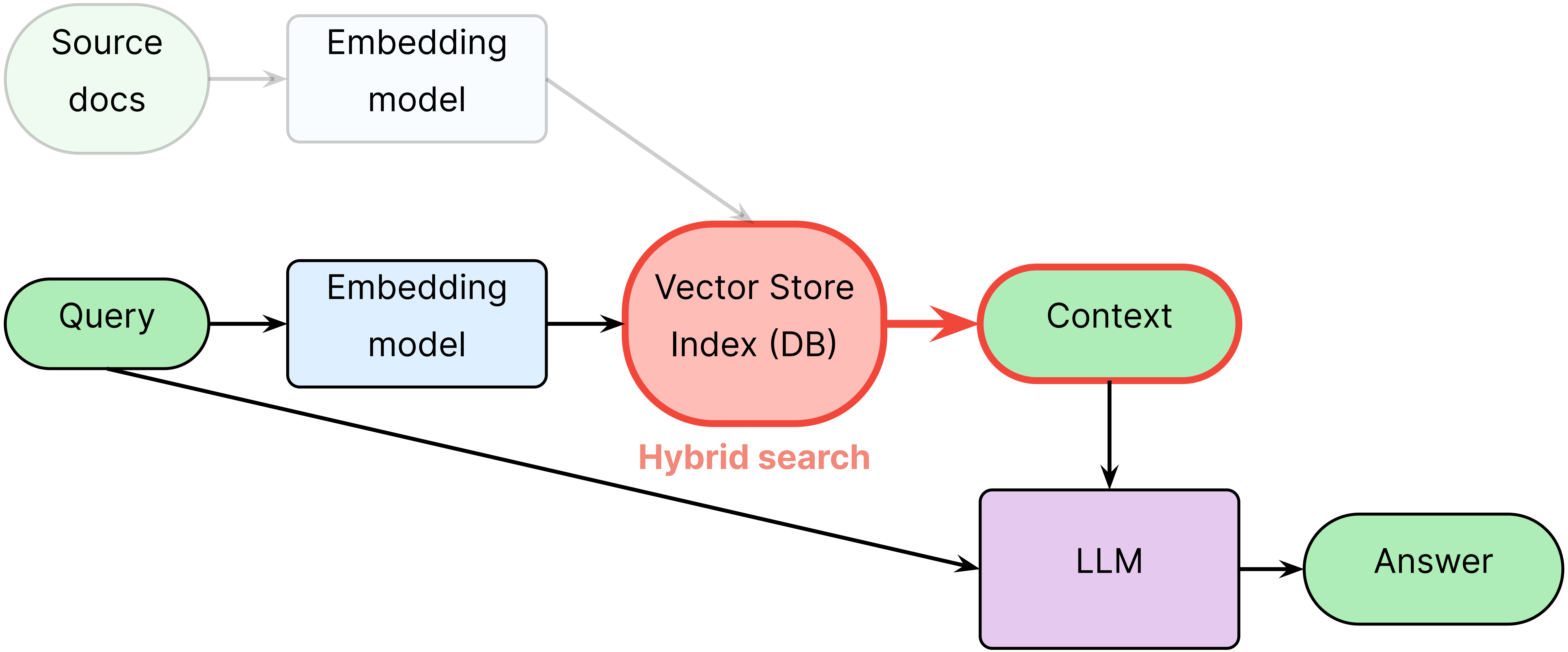


Retrieval of low-relevance docs/chunks (1)

Hybrid search

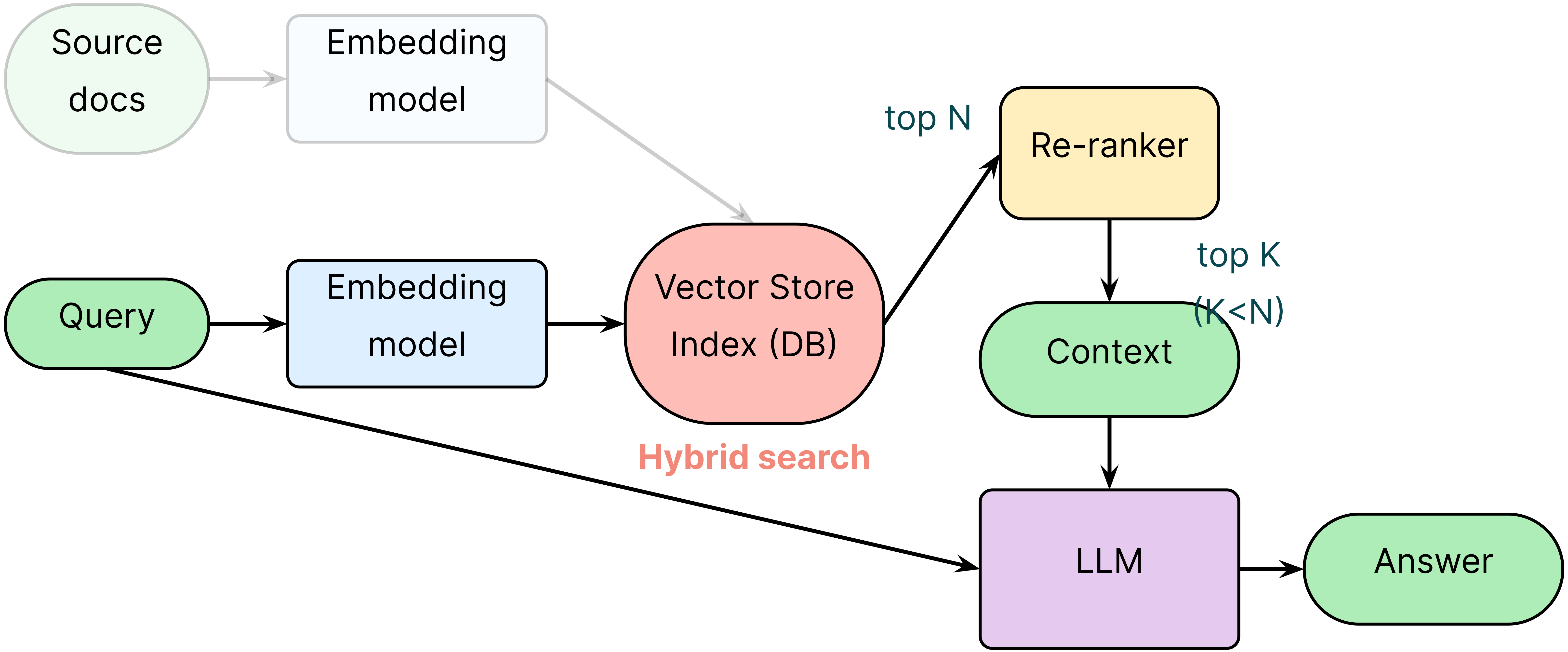


Retrieval of low-relevance docs/chunks (2)



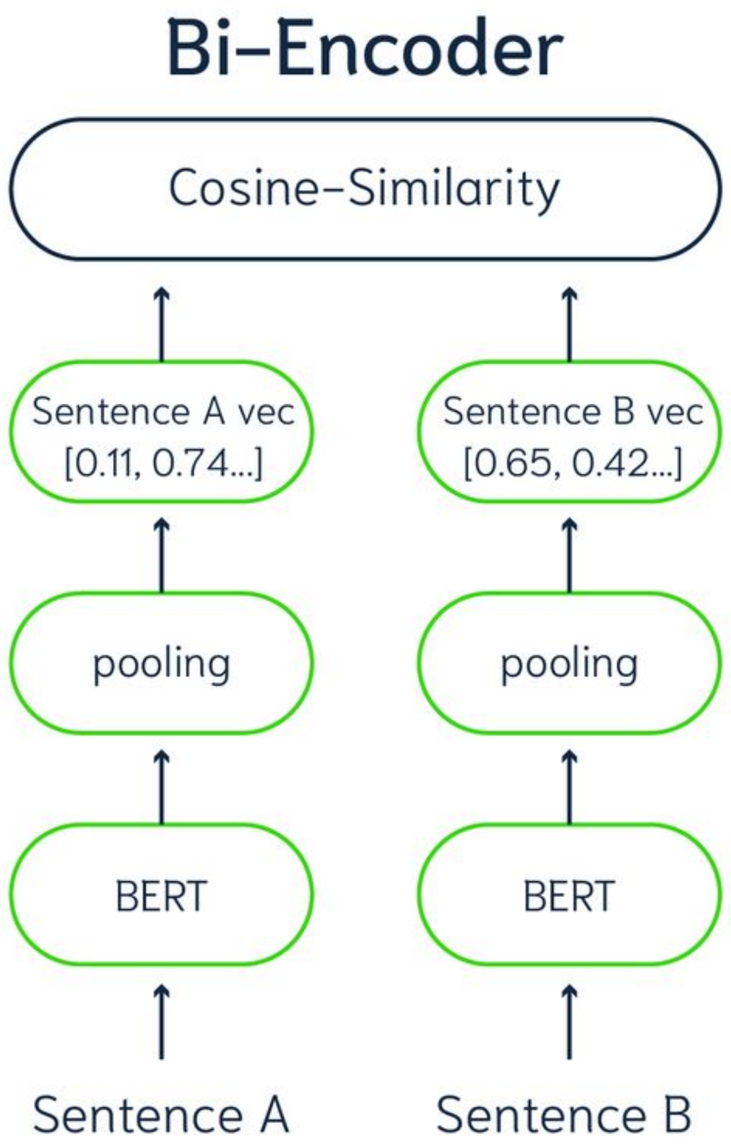
Retrieval of low-relevance docs/chunks (2)

Re-ranking



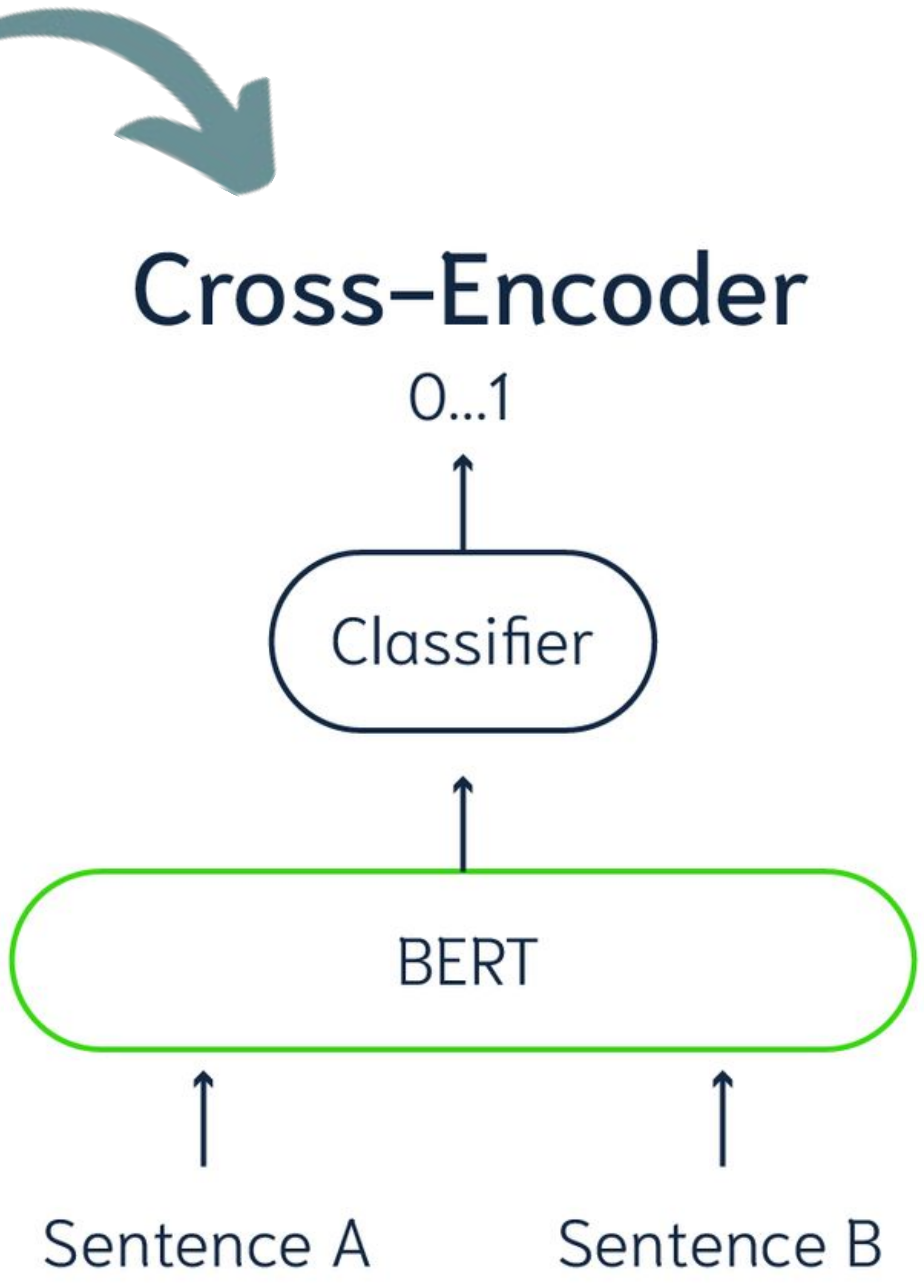
Retrieval of low-relevance docs/chunks (2)

Re-ranking



fast, but less accurate

(used in first stage retrieval)



more accurate due to the cross-attention computation between query and article tokens, but slow

(can be trained using datasets as MS-MARCO)

https://weaviate.io/blog/cross-encoders-as-reranke
https://cookbook.openai.com/examples/search_reranking_with_cross-encoder

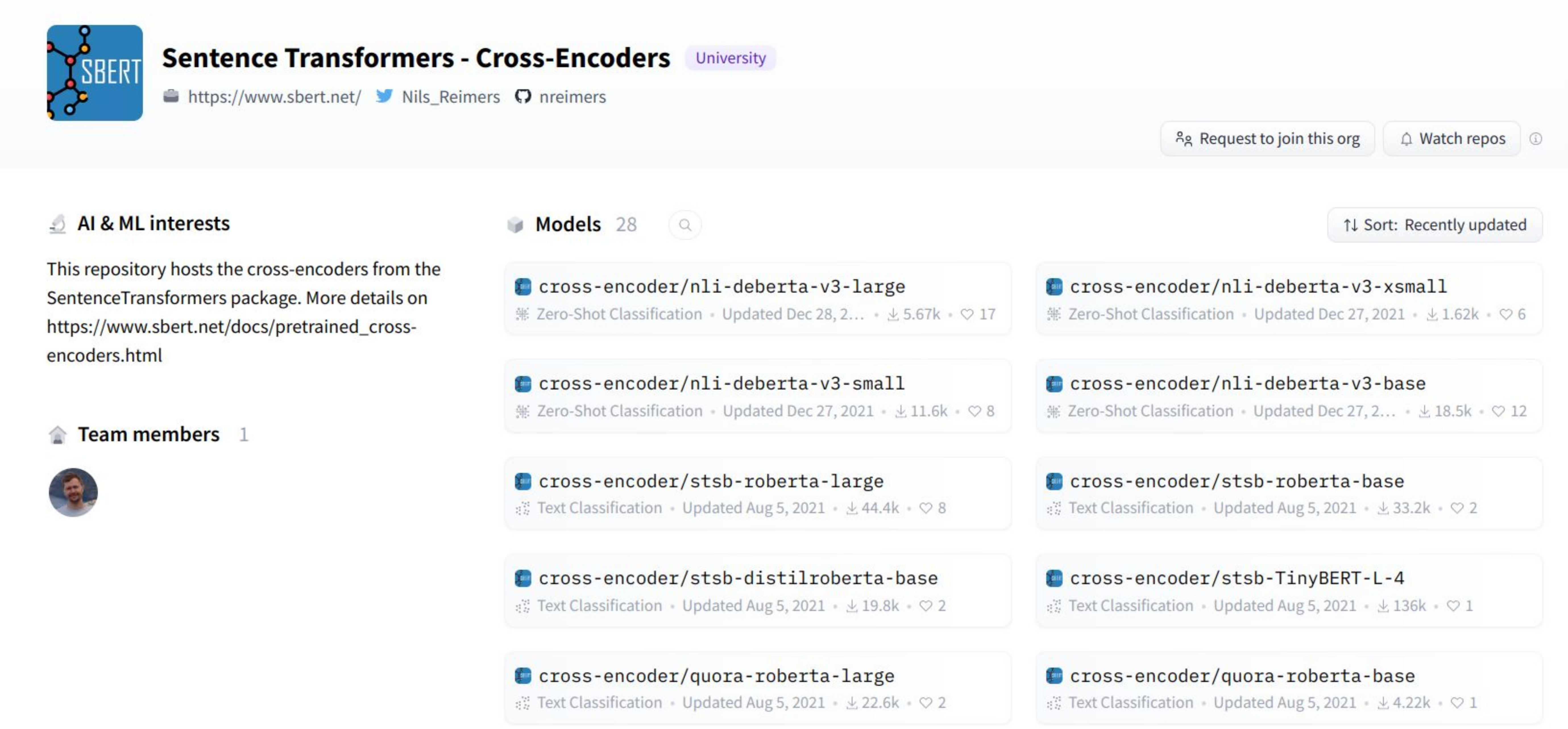
Retrieval of low-relevance docs/chunks (2)

Re-ranking

https://weaviate.io/blog/cross-encoders-as-reranker

Retrieval of low-relevance docs/chunks (2)

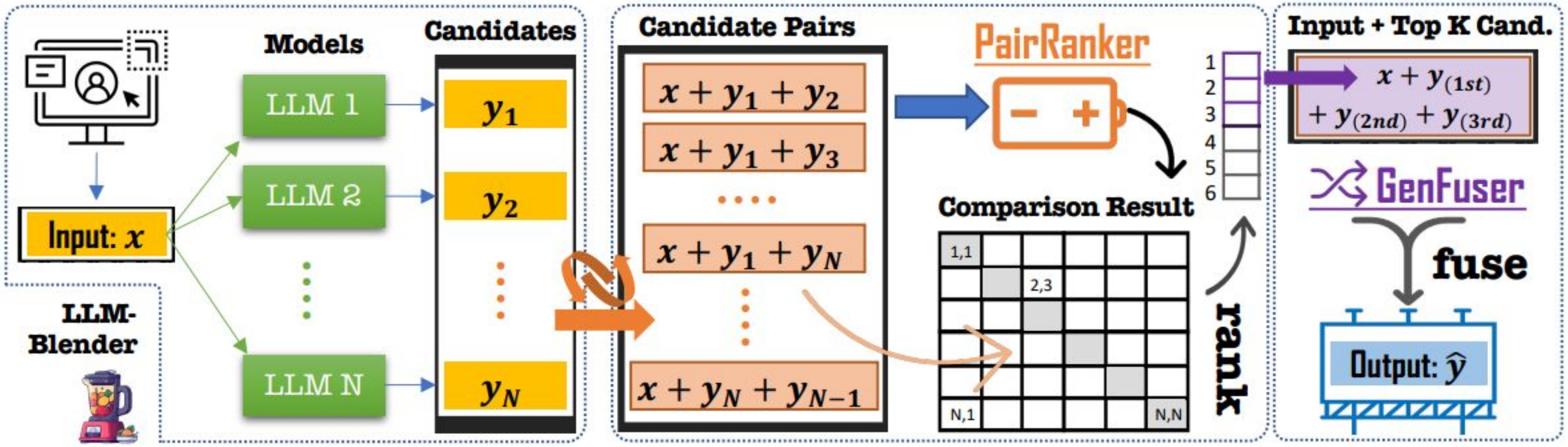
Re-ranking



https://huggingface.co/cross-encoder

Retrieval of low-relevance docs/chunks (2)

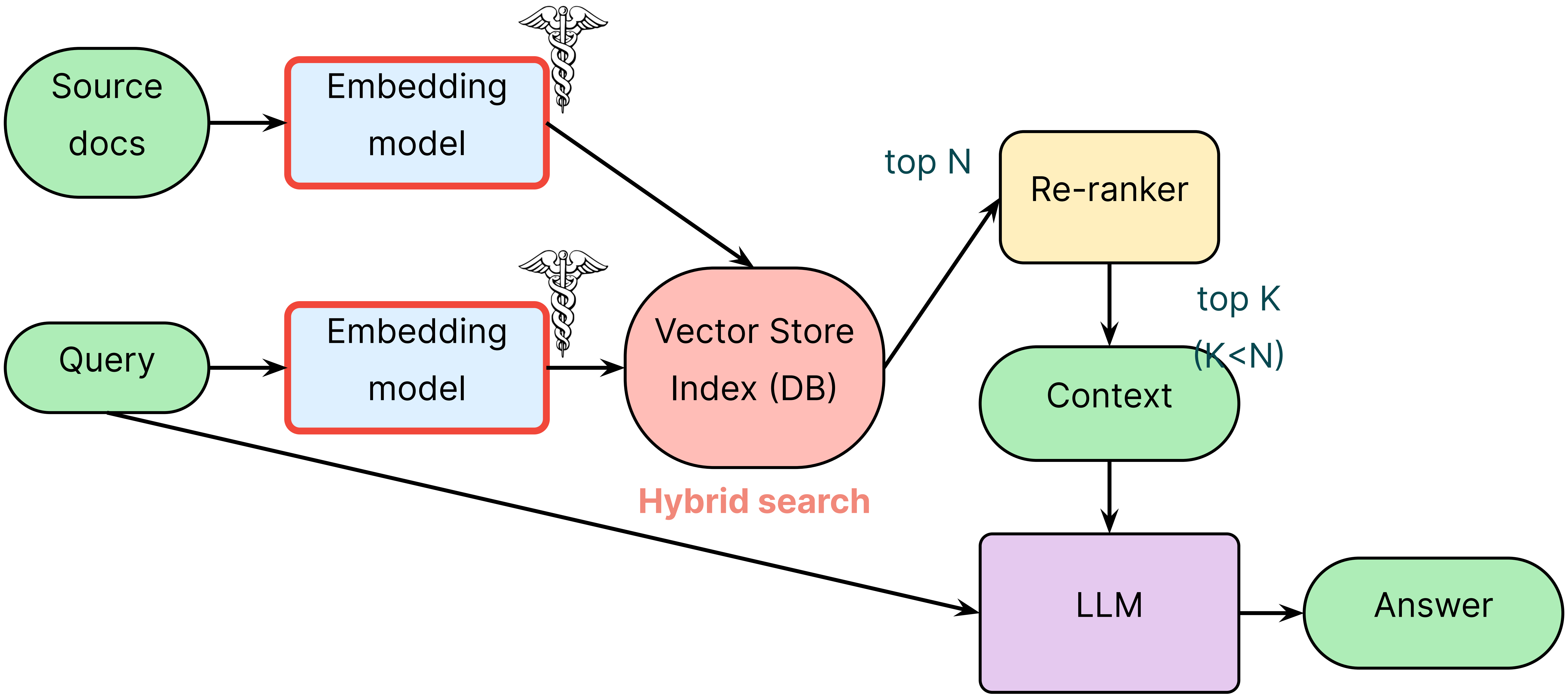
Pair RM



- based on microsoft/deberta-v3-large
- achieves superior performance by mixing the outputs of multiple LLMs

https://arxiv.org/pdf/2306.02561.pdf

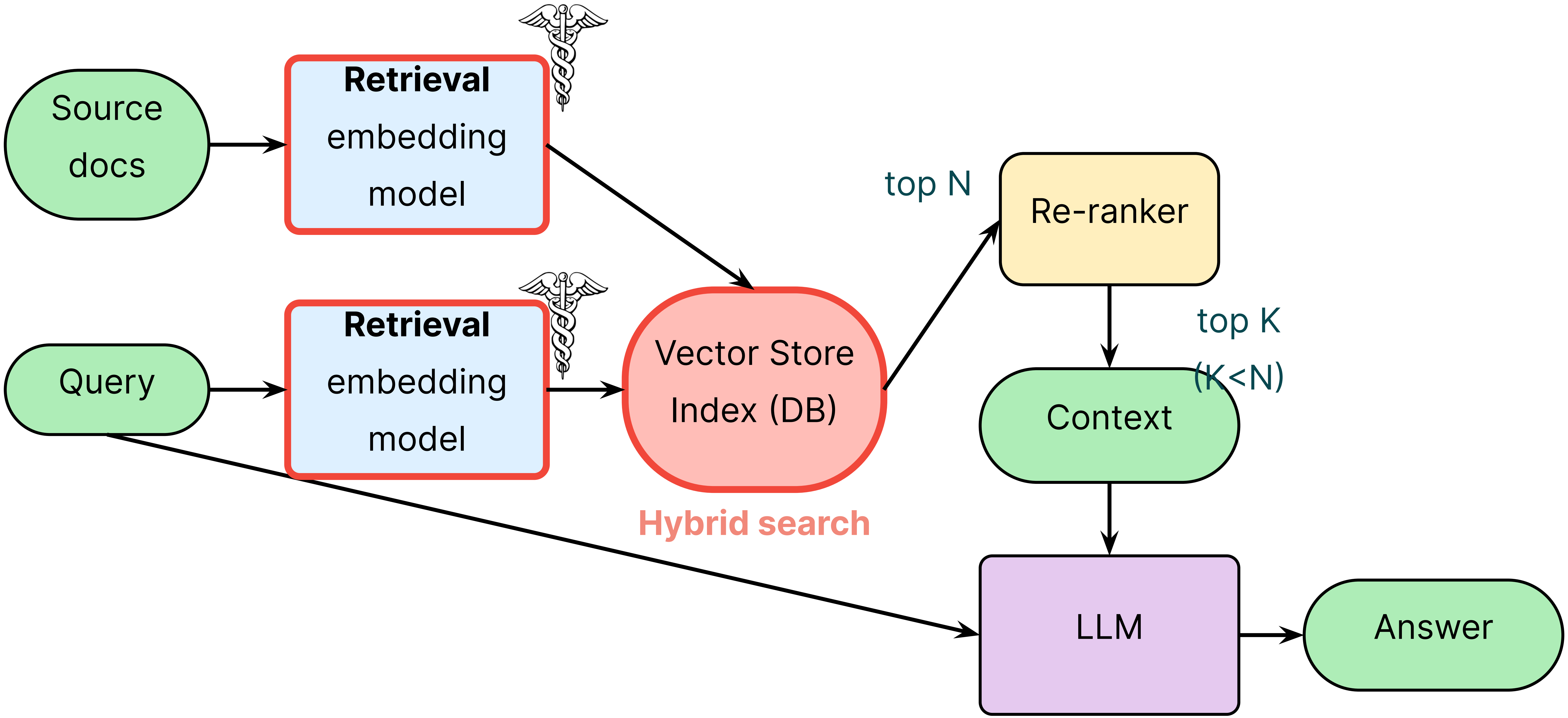
Non specialised embeddings (for medicine)



Non specialised embeddings (for medicine)

Medical Embedding Models

Non specialised embeddings (for medical retrieval)



Non specialised embeddings (for medical retrieval)

Q: What is the primary function of the spleen?

A1: The primary function of the spleen is unknown

A2: The spleen plays a key role in filtering and removing old or damaged red blood cells from the bloodstream

<u>Non specialised</u> embeddings (for <u>medical retrieval</u>)

Q: What is the primary function of the spleen?

A1: The primary function of the spleen is unknown

Sim=0.913

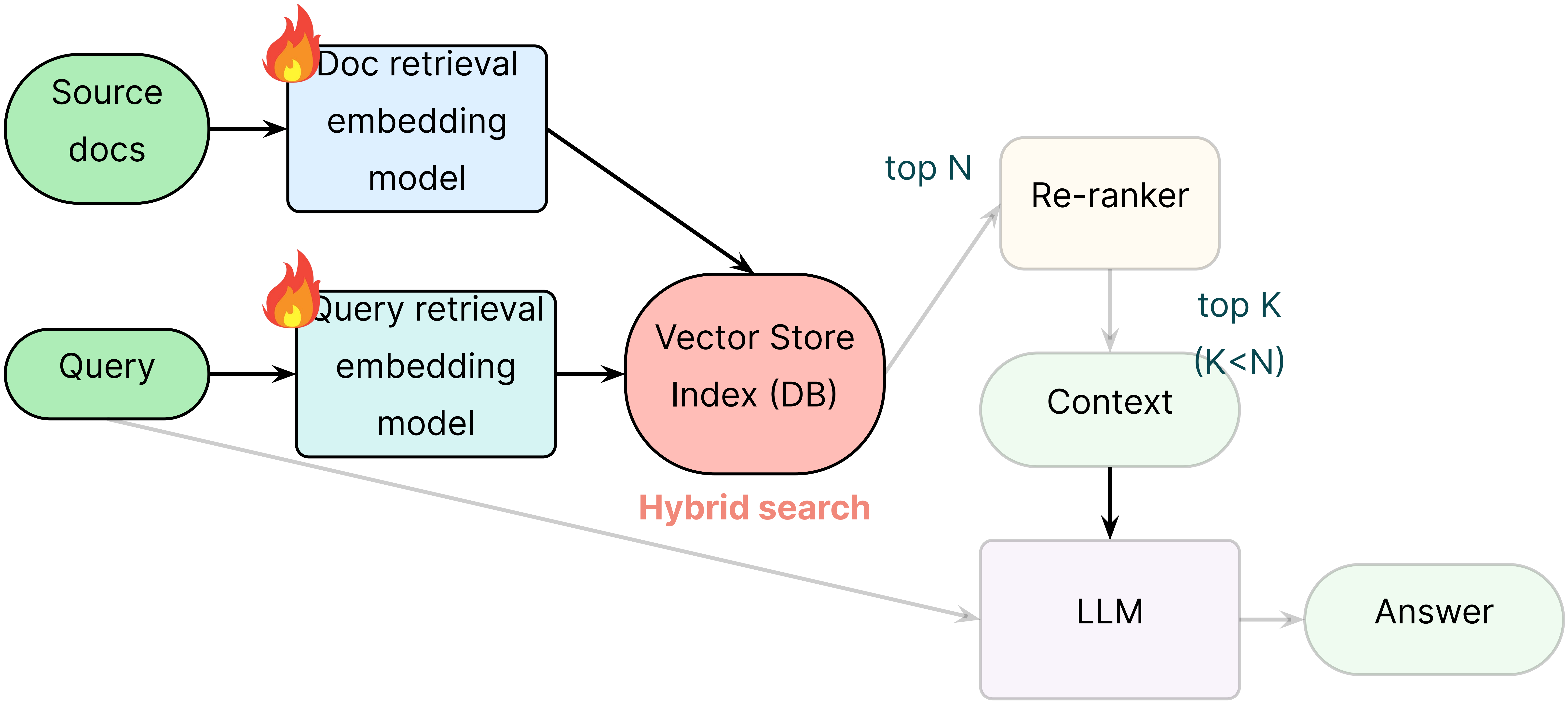
A2: The spleen plays a key role in filtering and removing old or damaged red blood cells from the bloodstream

Sim=0.667

similar but NOT relevant

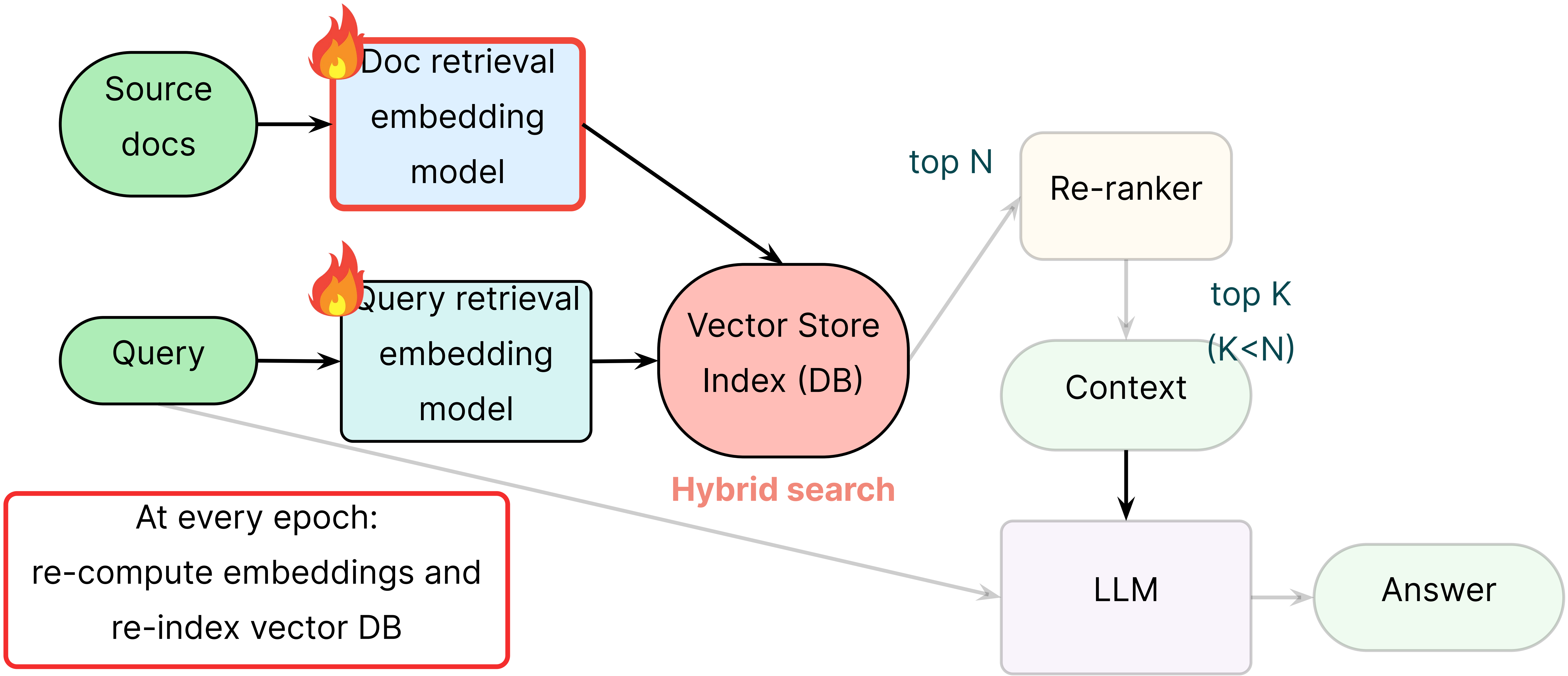
Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models



Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models



Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models

MedCPT: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot Biomedical Information Retrieval

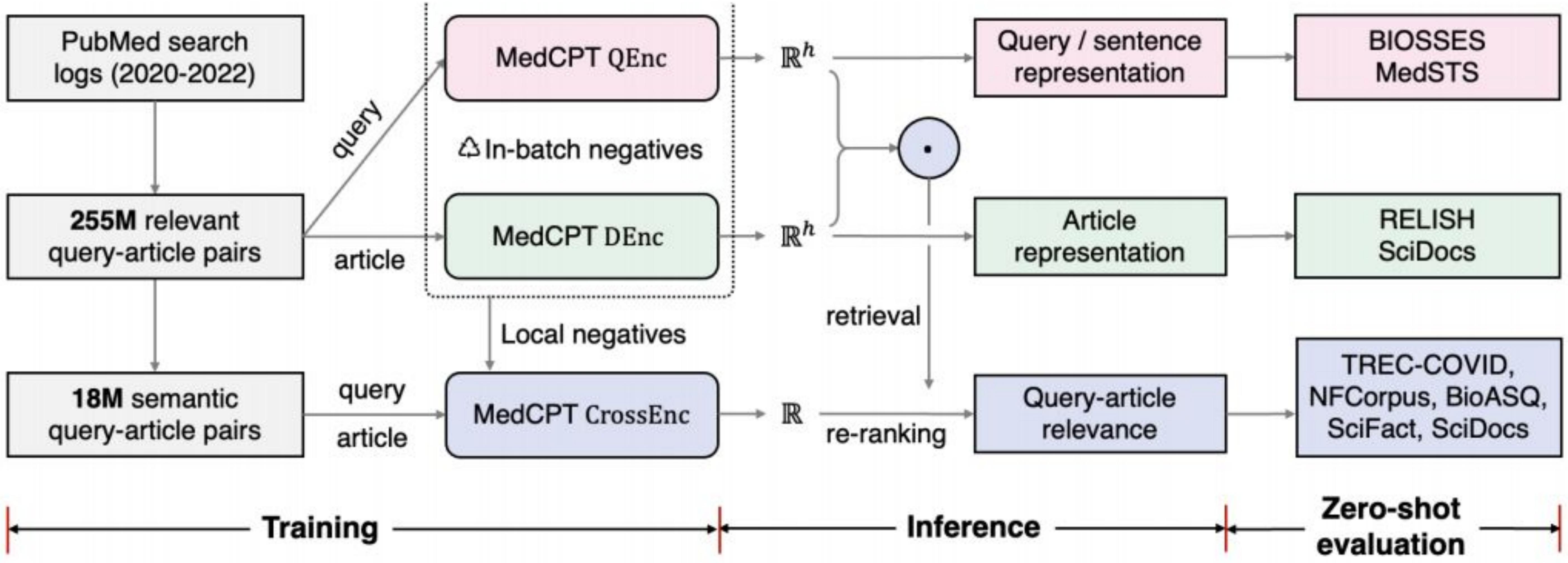
Qiao Jin[1], Won Kim[1], Qingyu Chen[1], Donald C. Comeau[1], Lana Yeganova[1], W. John Wilbur[1], Zhiyong Lu[1]
[1]National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH)
Correspondence: zhiyong.lu@nih.gov

https://arxiv.org/abs/2307.00589

MedCPT: Query Encoder + Article Encoder + re-ranker cross-encoder

Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models

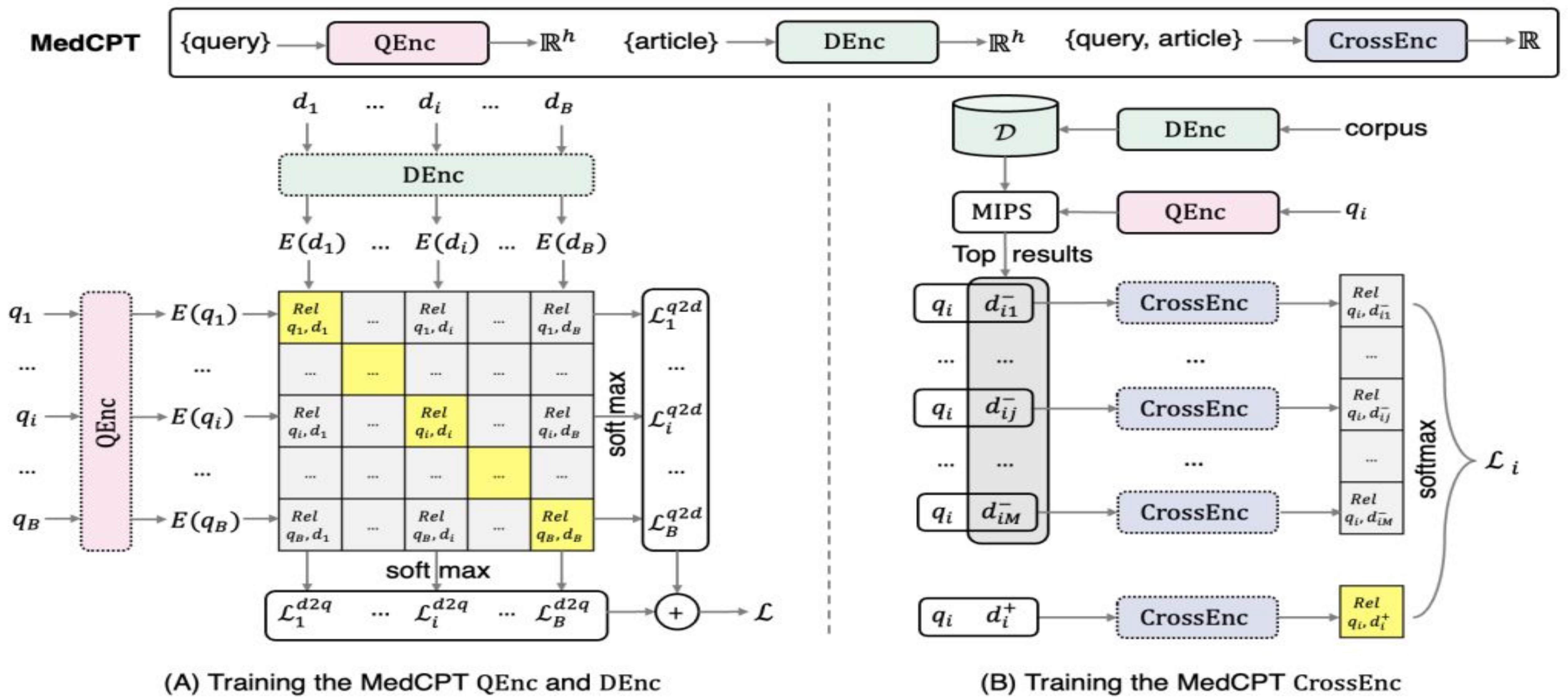


https://arxiv.org/abs/2307.00589

Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models

Initialized from PubMedBERT



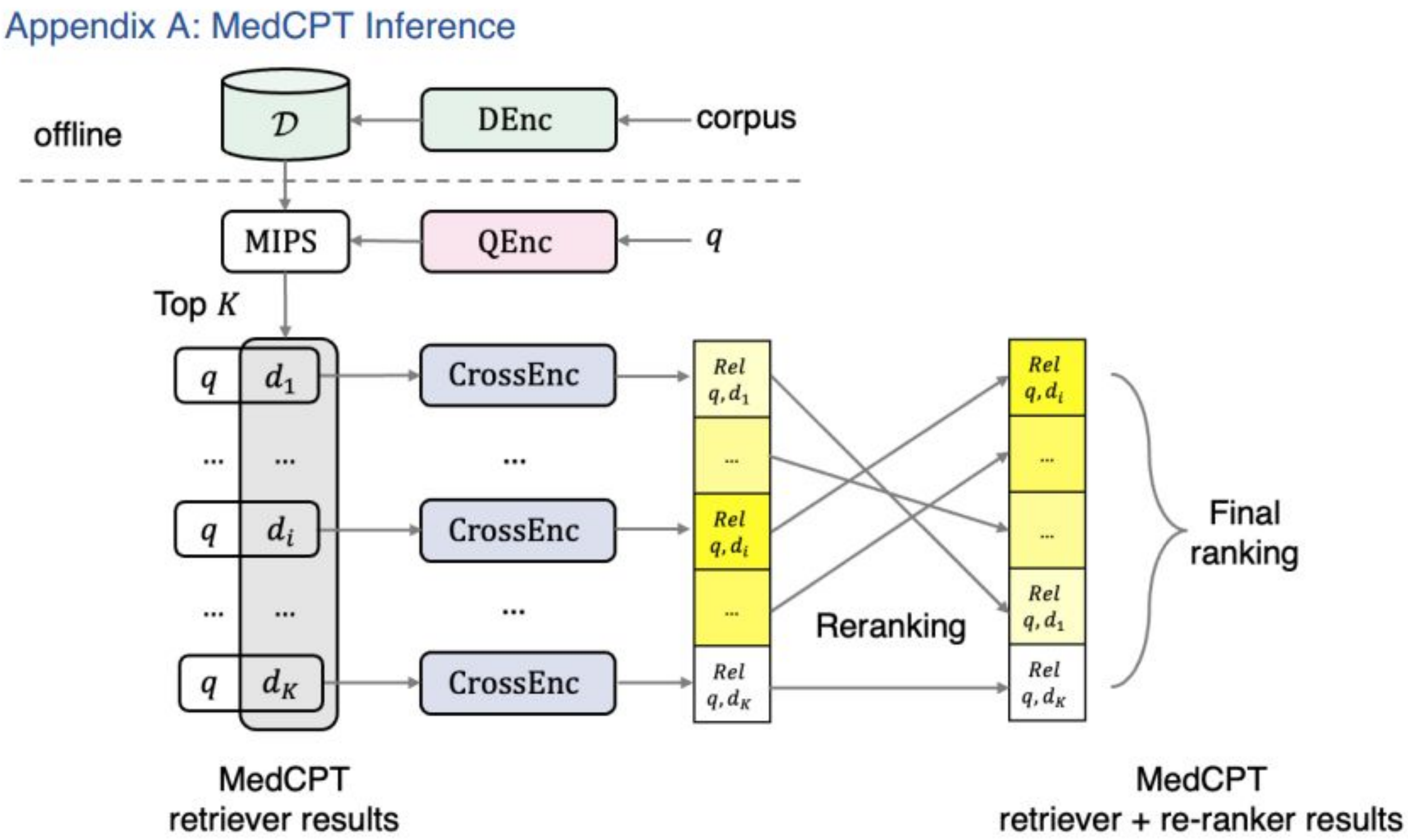
Initialized from PubMedBERT

Figure 2. Overview of the MedCPT training process. (A) Training the MedCPT query encoder (QEnc) and document encoder (DEnc) using a contrastive loss with query-document pairs and in-batch negatives; (B) Training the MedCPT cross-encoder

https://arxiv.org/abs/2307.00589

Non specialised embeddings (for medical retrieval)

Medical Retrieval Embedding Models



https://arxiv.org/abs/2307.00589

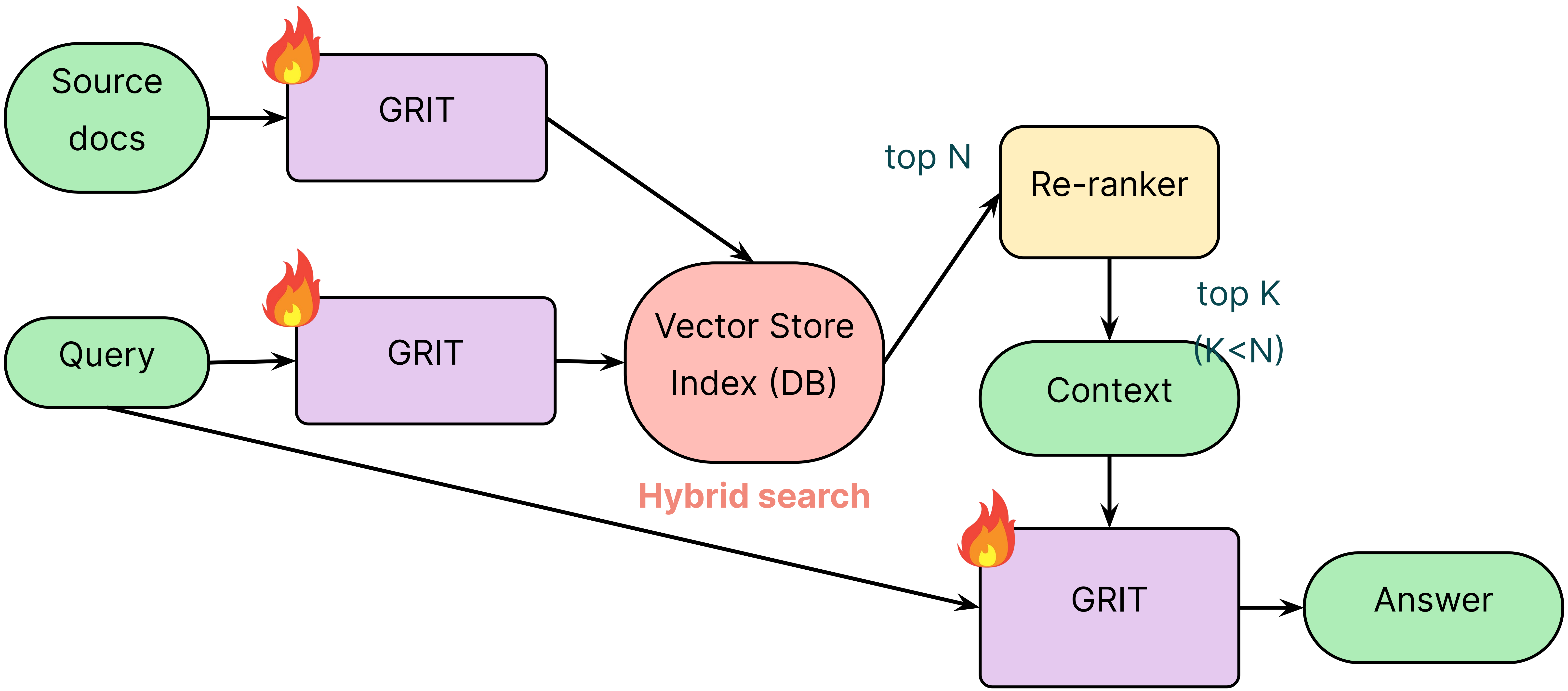
Medical Retrieval Embedding Models

Results

- **Biomedical Information Retrieval (BEIR benchmark)**
  - SotA on 3/5 biomedical tasks
  - Improves its initialization PubMedBERT by huge margins
- **Biomedical article representations (RELISH article similarity benchmark)**
  - MedCPT article encoder (DEnc) outperforms all other models
  - MedCPT article encoder improves PubMedBERT initialiationby over 10%
- **Biomedical sentence representations (BIOSSES and MedSTS benchmarks)**
  - On BIOSSES, MedCPT performs the best among all compared models
  - On the MedSTS dataset, MedCPT ranks the second and the performance is

Other

Generative Representational Instruction Tuning



Unified text embedding & generation model: GRIT

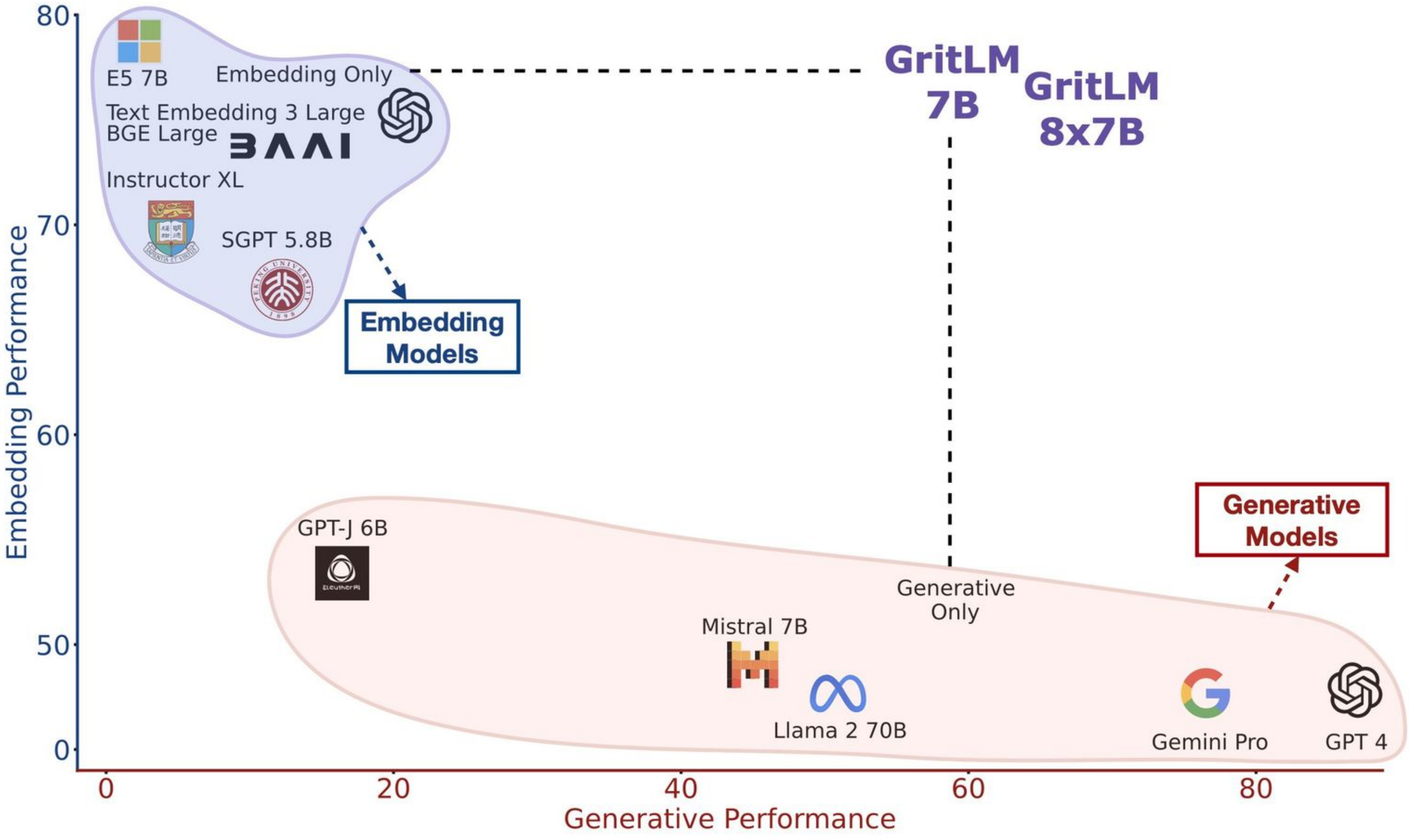


Figure 1: **Performance of various models on text representation (embedding) and generation tasks.** GRITLM is the first model to perform best-in-class at both types of tasks simultaneously.

https://arxiv.org/abs/2402.09906
https://twitter.com/Muennighoff/status/1758307967802224770

Unified text embedding & generation model: GRIT

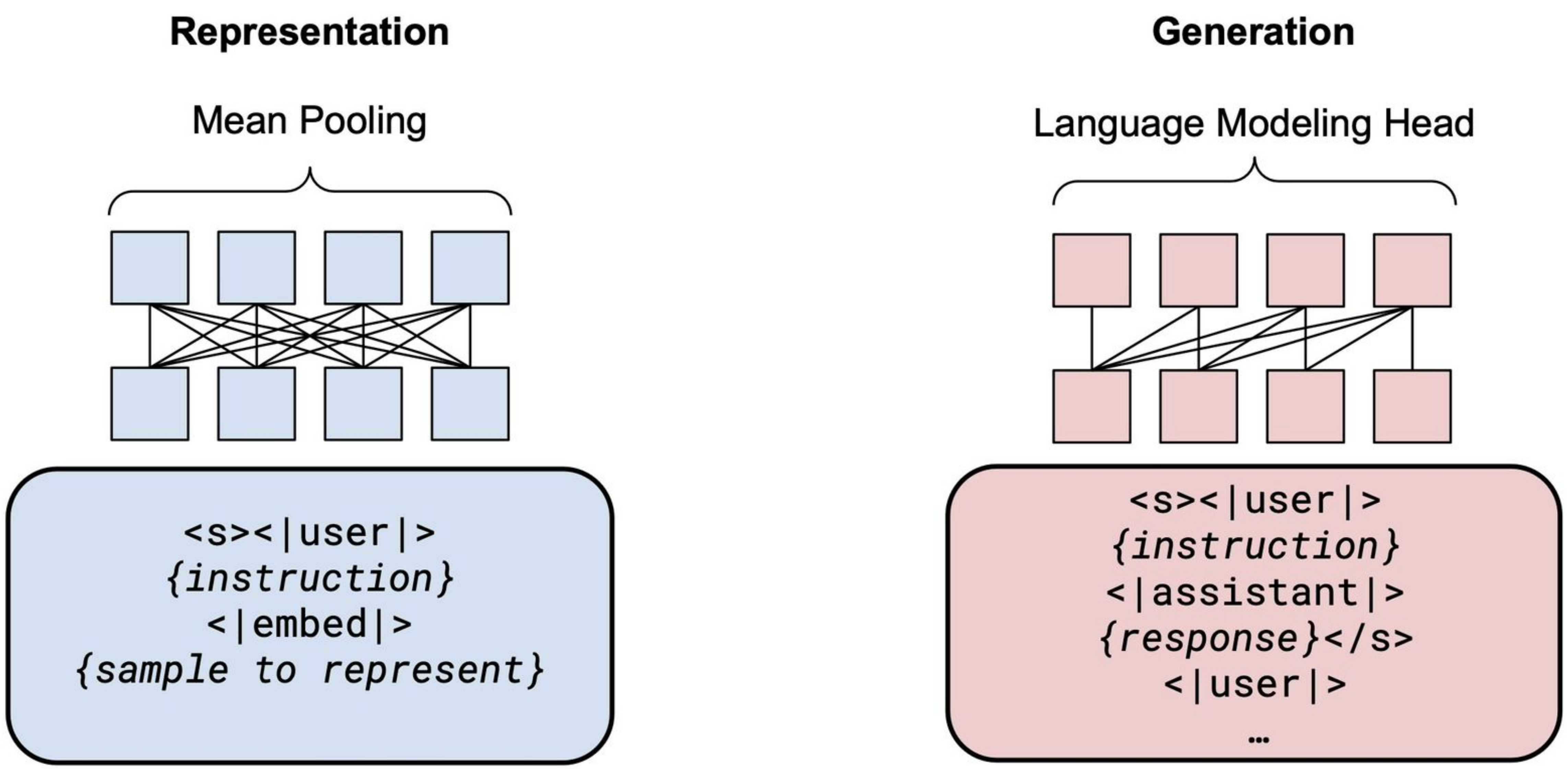


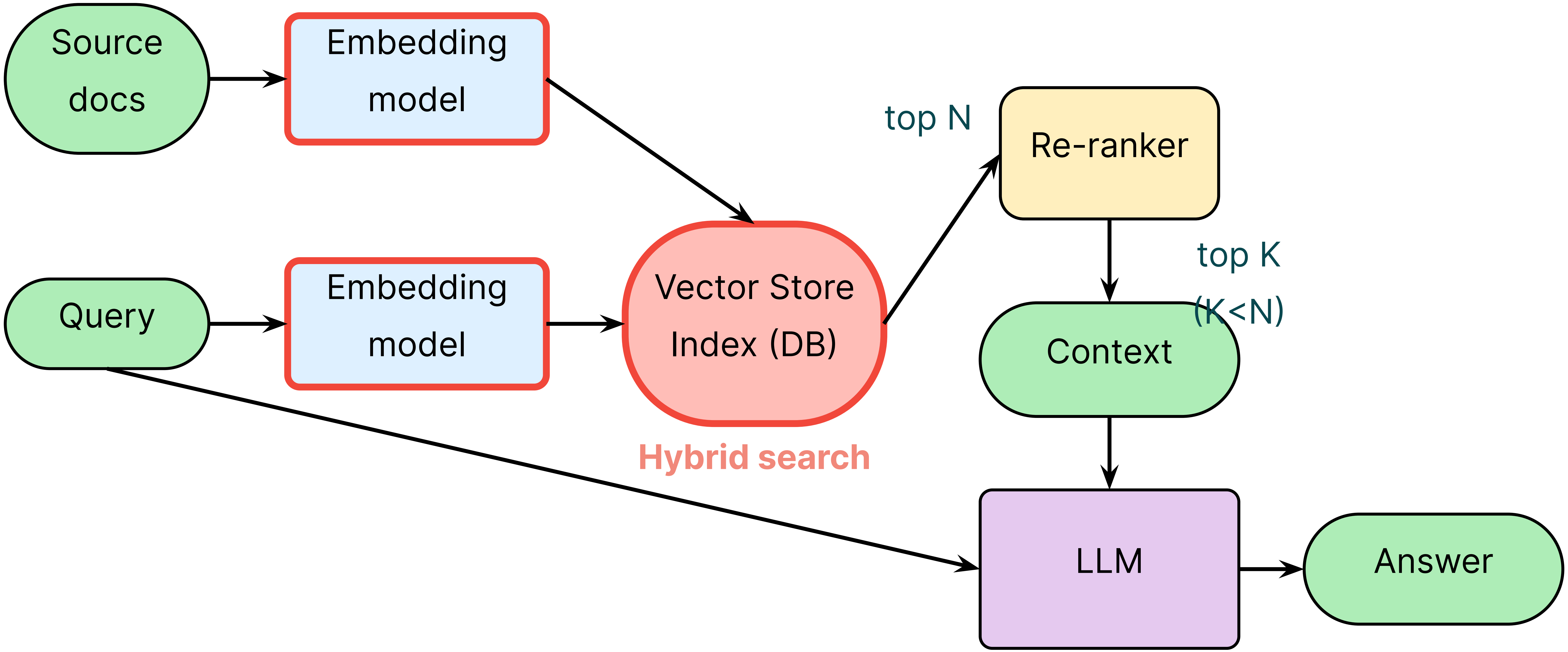
Figure 3: **GritLM architecture and format.** *Left:* GritLM uses bidirectional attention over the input for embedding tasks. Mean pooling is applied over the final hidden state to yield the final representation. *Right:* GritLM uses causal attention over the input for generative tasks. A language modeling head on top of the hidden states predicts the next tokens. The format supports conversations with multiple turns (indicated with "...").

Unified text embedding & generation model: GRIT

Use the same model as both embedder & reranker

Boosts perf on 15/16 Retrieval dsets

Large embeddings (for retrieval)

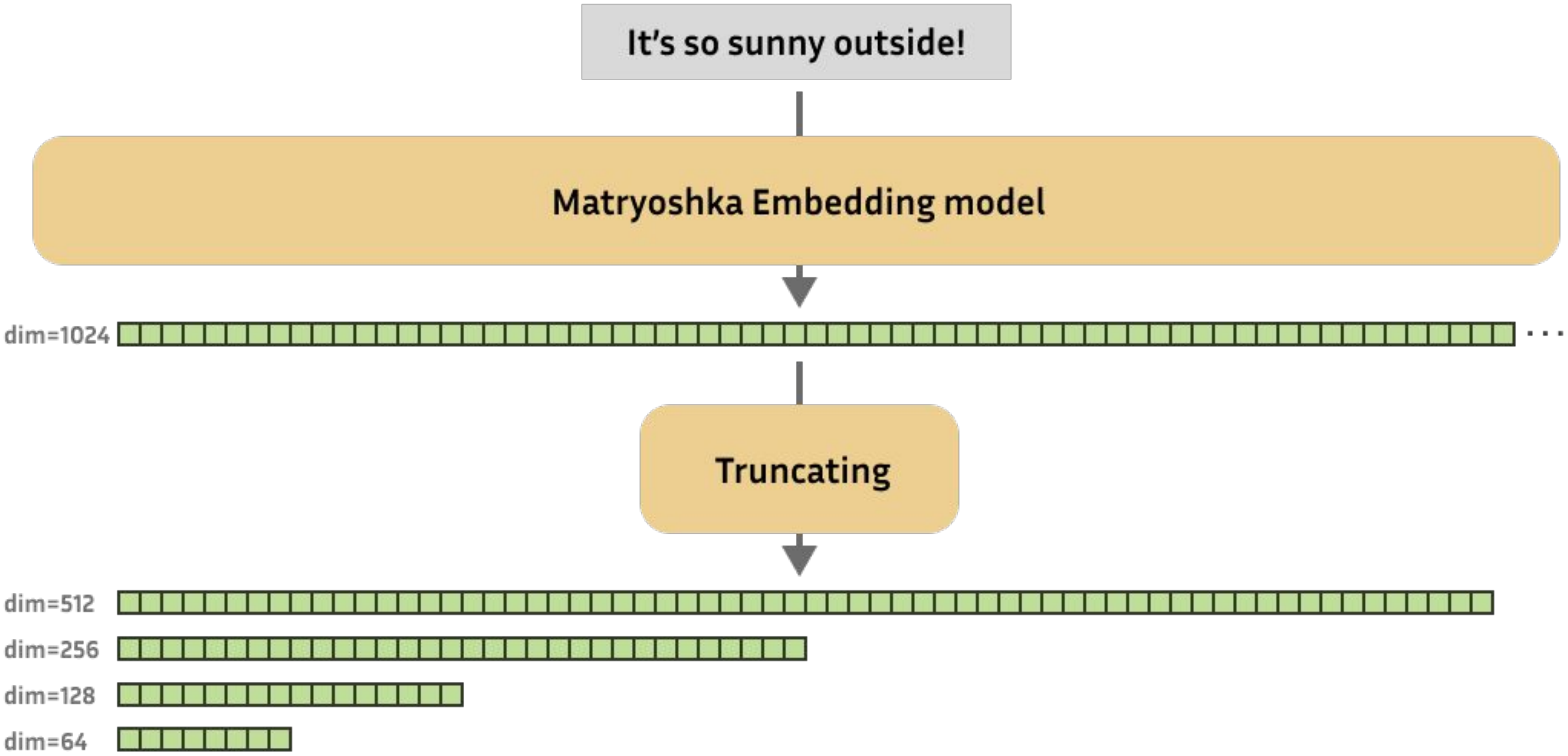


Large embeddings (for retrieval)

Better embeddings → more dimensions → less efficient search

Large embeddings (for retrieval)

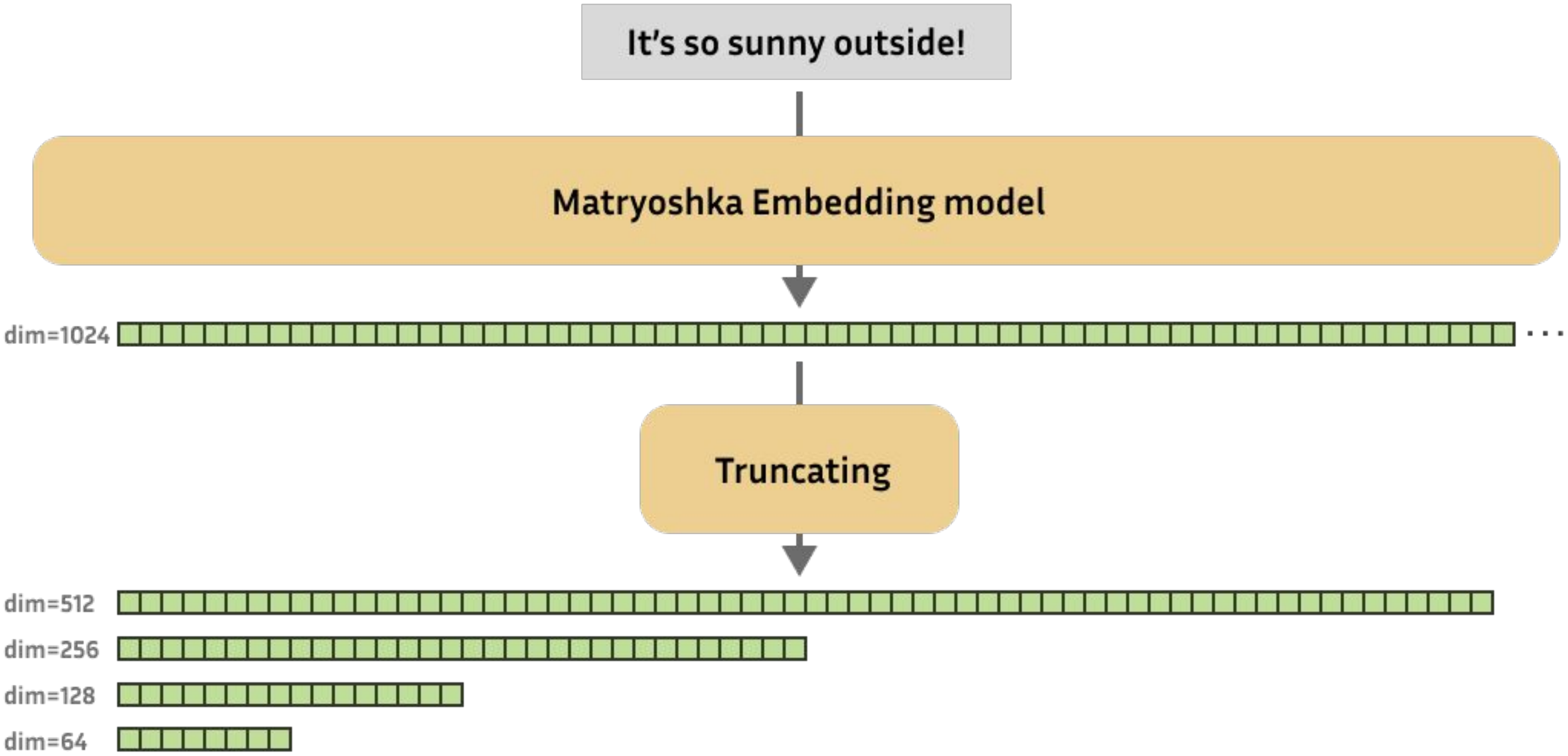
Matryoshka Embedding Models



https://arxiv.org/abs/2205.13147
https://huggingface.co/blog/matryoshka

Large embeddings (for retrieval)

Matryoshka Embedding Models



The **loss values** for each dimensionality are **added** together, resulting in a final loss value.

SotA RAG for medicine

Benchmarking Retrieval-Augmented Generation for Medicine

dec 2023

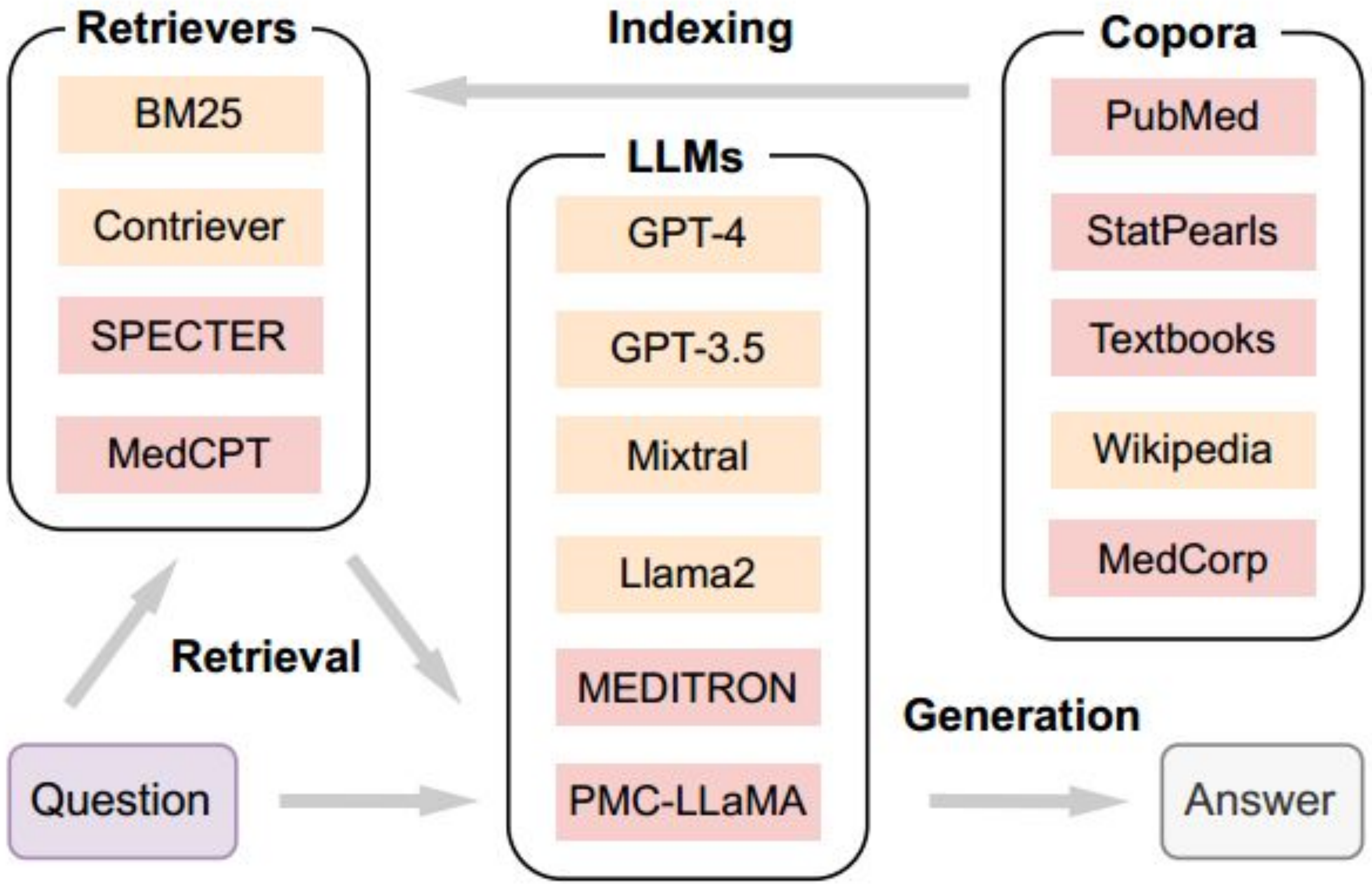


Figure 2: Component overview of the MEDRAG toolkit.

https://arxiv.org/pdf/2402.13178.pdf

Benchmarking Retrieval-Augmented Generation for Medicine

LLM	Method	MIRAGE Benchmark Dataset					Avg.
		MMLU-Med	MedQA-US	MedMCQA	PubMedQA*	BioASQ-Y/N	
GPT-4 (-32k-0613)	CoT	89.44 ± 0.93	83.97 ± 1.03	69.88 ± 0.71	39.60 ± 2.19	84.30 ± 1.46	73.44
	MEDRAG	87.24 ± 1.01	82.80 ± 1.06	66.65 ± 0.73	70.60 ± 2.04	92.56 ± 1.06	79.97
GPT-3.5 (-16k-0613)	CoT	72.91 ± 1.35	65.04 ± 1.34	55.25 ± 0.77	36.00 ± 2.15	74.27 ± 1.76	60.69
	MEDRAG	75.48 ± 1.30	66.61 ± 1.32	58.04 ± 0.76	67.40 ± 2.10	90.29 ± 1.19	71.57
Mixtral (8×7B)	CoT	74.01 ± 1.33	64.10 ± 1.34	56.28 ± 0.77	35.20 ± 2.14	77.51 ± 1.68	61.42
	MEDRAG	75.85 ± 1.30	60.02 ± 1.37	56.42 ± 0.77	67.60 ± 2.09	87.54 ± 1.33	69.48
Llama2 (70B)	CoT	57.39 ± 1.50	47.84 ± 1.40	42.60 ± 0.76	42.20 ± 2.21	61.17 ± 1.96	50.24
	MEDRAG	54.55 ± 1.51	44.93 ± 1.39	43.08 ± 0.77	50.40 ± 2.24	73.95 ± 1.77	53.38
MEDITRON (70B)	CoT	64.92 ± 1.45	51.69 ± 1.40	46.74 ± 0.77	53.40 ± 2.23	68.45 ± 1.87	57.04
	MEDRAG	65.38 ± 1.44	49.57 ± 1.40	52.67 ± 0.77	56.40 ± 2.22	76.86 ± 1.70	60.18
PMC-LLaMA (13B)	CoT	52.16 ± 1.51	44.38 ± 1.39	46.55 ± 0.77	55.80 ± 2.22	63.11 ± 1.94	52.40
	MEDRAG	52.53 ± 1.51	42.58 ± 1.39	48.29 ± 0.77	56.00 ± 2.22	65.21 ± 1.92	52.92

Table 6: Benchmark results of different backbone LLMs on MIRAGE. All numbers are accuracy in percentages.

Full corpus + fusion of 4 retrievers

https://arxiv.org/pdf/2402.13178.pdf

Benchmarking Retrieval-Augmented Generation for Medicine

dec 2023

Corpus	Retriever	MIRAGE Benchmark Dataset					Average
		MMLU-Med	MedQA-US	MedMCQA	PubMedQA*	BioASQ-Y/N	
None	None	72.91 ± 1.35	65.04 ± 1.34	55.25 ± 0.77	36.00 ± 2.15	74.27 ± 1.76	60.69
PubMed (23.9M)	BM25	72.27 ± 1.36	63.71 ± 1.35	55.49 ± 0.77	66.20 ± 2.12	88.51 ± 1.28	69.23
	Contriever	71.72 ± 1.36	63.94 ± 1.35	54.29 ± 0.77	65.60 ± 2.12	85.44 ± 1.42	68.20
	SPECTER	73.19 ± 1.34	65.20 ± 1.34	53.12 ± 0.77	54.80 ± 2.23	75.73 ± 1.72	64.41
	MedCPT	73.09 ± 1.34	66.69 ± 1.32	54.94 ± 0.77	66.40 ± 2.11	85.76 ± 1.41	69.38
	RRF-2	75.57 ± 1.30	64.34 ± 1.34	55.34 ± 0.77	69.00 ± 2.07	87.06 ± 1.35	70.26
	RRF-4	73.37 ± 1.34	64.73 ± 1.34	54.75 ± 0.77	67.20 ± 2.10	88.51 ± 1.28	69.71
StatPearls (301.2k)	BM25	71.63 ± 1.37	65.67 ± 1.33	54.89 ± 0.77	27.60 ± 2.00	60.36 ± 1.97	56.03
	Contriever	73.28 ± 1.34	67.48 ± 1.31	54.24 ± 0.77	28.80 ± 2.03	58.41 ± 1.98	56.44
	SPECTER	73.74 ± 1.33	64.73 ± 1.34	52.83 ± 0.77	23.20 ± 1.89	57.77 ± 1.99	54.45
	MedCPT	72.82 ± 1.35	64.89 ± 1.34	54.17 ± 0.77	27.60 ± 2.00	60.68 ± 1.96	56.03
	RRF-2	72.64 ± 1.35	65.67 ± 1.33	54.63 ± 0.77	30.00 ± 2.05	61.17 ± 1.96	56.82
	RRF-4	73.83 ± 1.33	65.12 ± 1.34	53.81 ± 0.77	30.60 ± 2.06	59.71 ± 1.97	56.61
Textbooks (125.8k)	BM25	74.66 ± 1.32	66.54 ± 1.32	54.05 ± 0.77	30.20 ± 2.05	60.03 ± 1.97	57.10
	Contriever	74.10 ± 1.33	67.16 ± 1.32	54.53 ± 0.77	26.60 ± 1.98	60.19 ± 1.97	56.52
	SPECTER	72.82 ± 1.35	67.40 ± 1.31	53.29 ± 0.77	25.60 ± 1.95	55.50 ± 2.00	54.92
	MedCPT	74.93 ± 1.31	66.22 ± 1.33	54.41 ± 0.77	29.20 ± 2.03	61.33 ± 1.96	57.22
	RRF-2	76.68 ± 1.28	65.91 ± 1.33	54.79 ± 0.77	31.00 ± 2.07	59.39 ± 1.98	57.55
	RRF-4	75.76 ± 1.30	66.06 ± 1.33	55.56 ± 0.77	30.40 ± 2.06	60.68 ± 1.96	57.69
Wikipedia (29.9M)	BM25	73.37 ± 1.34	63.47 ± 1.35	54.10 ± 0.77	26.40 ± 1.97	71.36 ± 1.82	57.74
	Contriever	74.10 ± 1.33	65.99 ± 1.33	54.03 ± 0.77	26.40 ± 1.97	69.90 ± 1.85	58.08
	SPECTER	72.18 ± 1.36	63.63 ± 1.35	52.71 ± 0.77	22.20 ± 1.86	66.83 ± 1.89	55.51
	MedCPT	71.99 ± 1.36	65.12 ± 1.34	55.15 ± 0.77	29.00 ± 2.03	73.46 ± 1.78	58.95
	RRF-2	74.20 ± 1.33	64.57 ± 1.34	54.72 ± 0.77	31.00 ± 2.07	76.21 ± 1.71	60.14
	RRF-4	73.19 ± 1.34	64.96 ± 1.34	54.53 ± 0.77	31.00 ± 2.07	72.01 ± 1.81	59.14
MedCorp (65.3M)	BM25	73.65 ± 1.34	65.91 ± 1.33	56.78 ± 0.77	66.20 ± 2.12	87.70 ± 1.32	70.05
	Contriever	75.48 ± 1.30	64.10 ± 1.34	56.11 ± 0.77	62.40 ± 2.17	84.95 ± 1.44	68.61
	SPECTER	74.38 ± 1.32	65.44 ± 1.33	54.41 ± 0.77	55.80 ± 2.22	73.14 ± 1.78	64.63
	MedCPT	74.75 ± 1.32	67.40 ± 1.31	55.85 ± 0.77	66.40 ± 2.11	85.92 ± 1.40	70.06
	RRF-2	73.74 ± 1.33	67.24 ± 1.32	56.08 ± 0.77	67.80 ± 2.09	88.19 ± 1.30	70.61
	RRF-4	75.48 ± 1.30	66.61 ± 1.32	58.04 ± 0.76	67.40 ± 2.10	90.29 ± 1.19	71.57

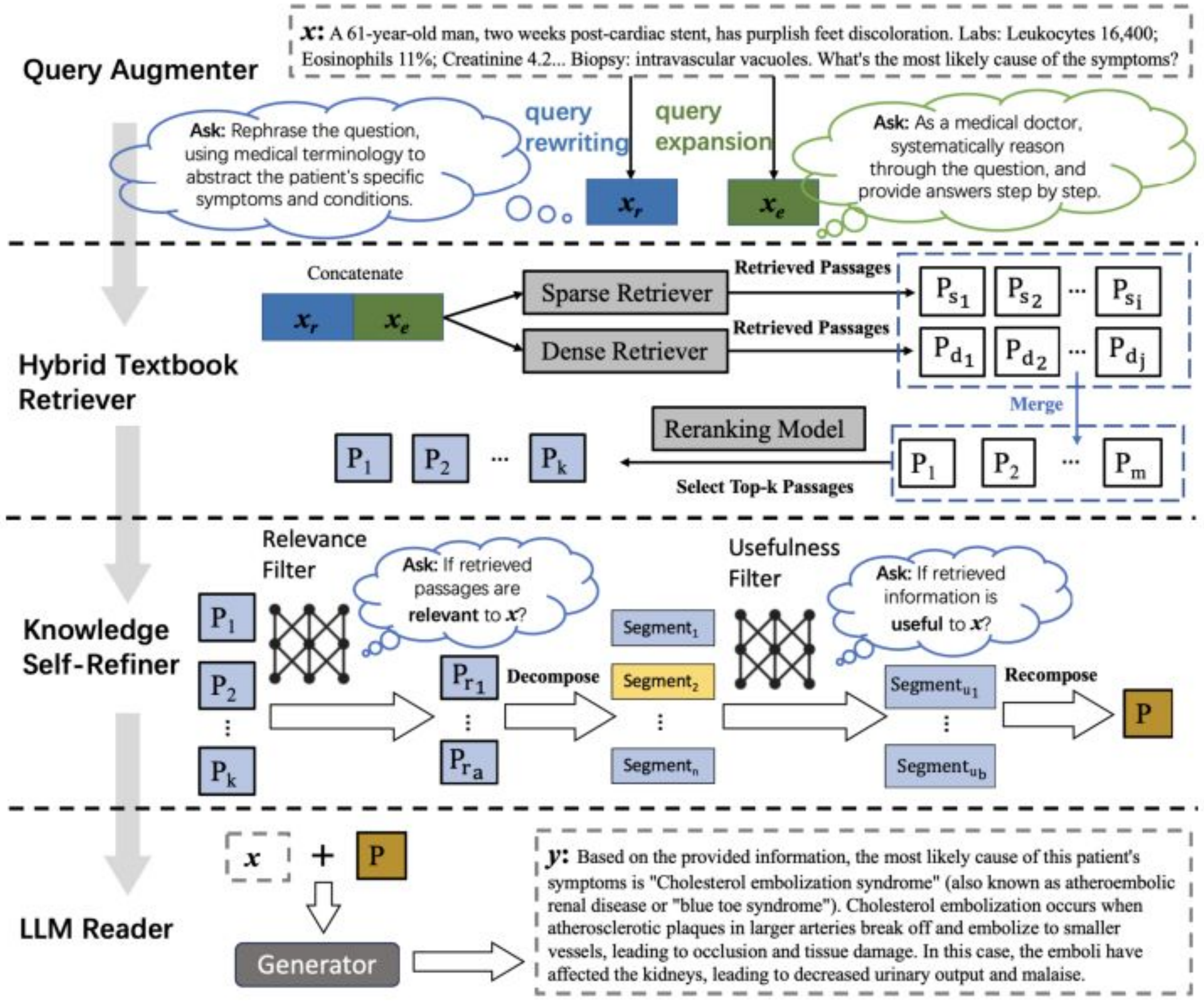
Table 7: Accuracy (%) of GPT-3.5 (MEDRAG) with different corpora and retrievers on MIRAGE. Red and green denote performance decreases and increases compared to CoT (first row). The shade reflects the relative change.

- Performance in specific tasks is strongly related to the used corpus
- Using a combination of all corpora provides highest performance
- Hybrid search yields better performance than dense search
- Retrievers show best performance when retrieving data from corpora within the same domain on which they have been trained

*RRF-2 (fusion of BM25 and MedCPT)

feb 2024

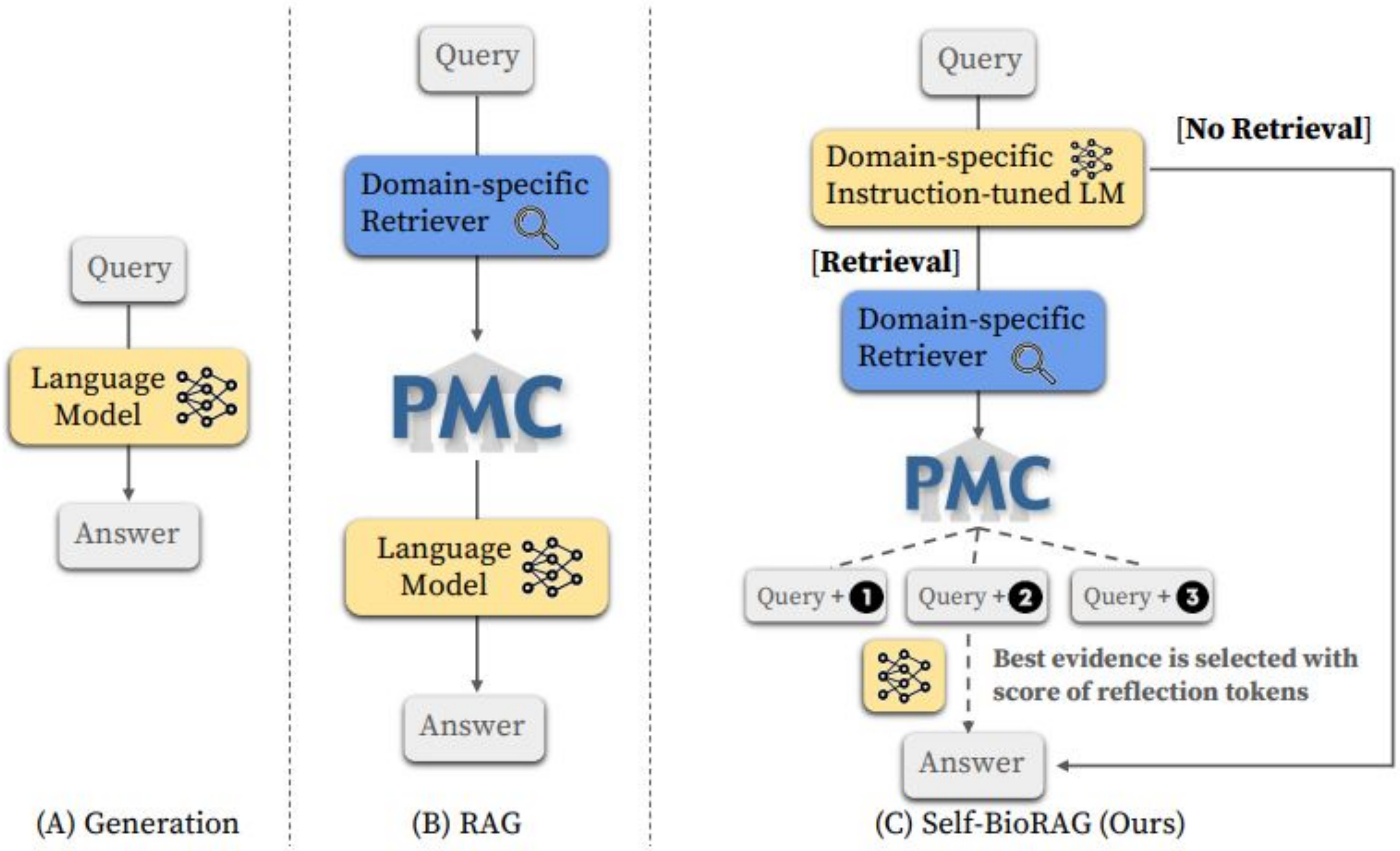
Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering



https://arxiv.org/abs/2309.02233

jan 2024

Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models



https://arxiv.org/abs/2401.15269

Almanac—Retrieval-Augmented Language Models for Clinical Medicine

PubMed, UpToDate, BMJ Best Practices

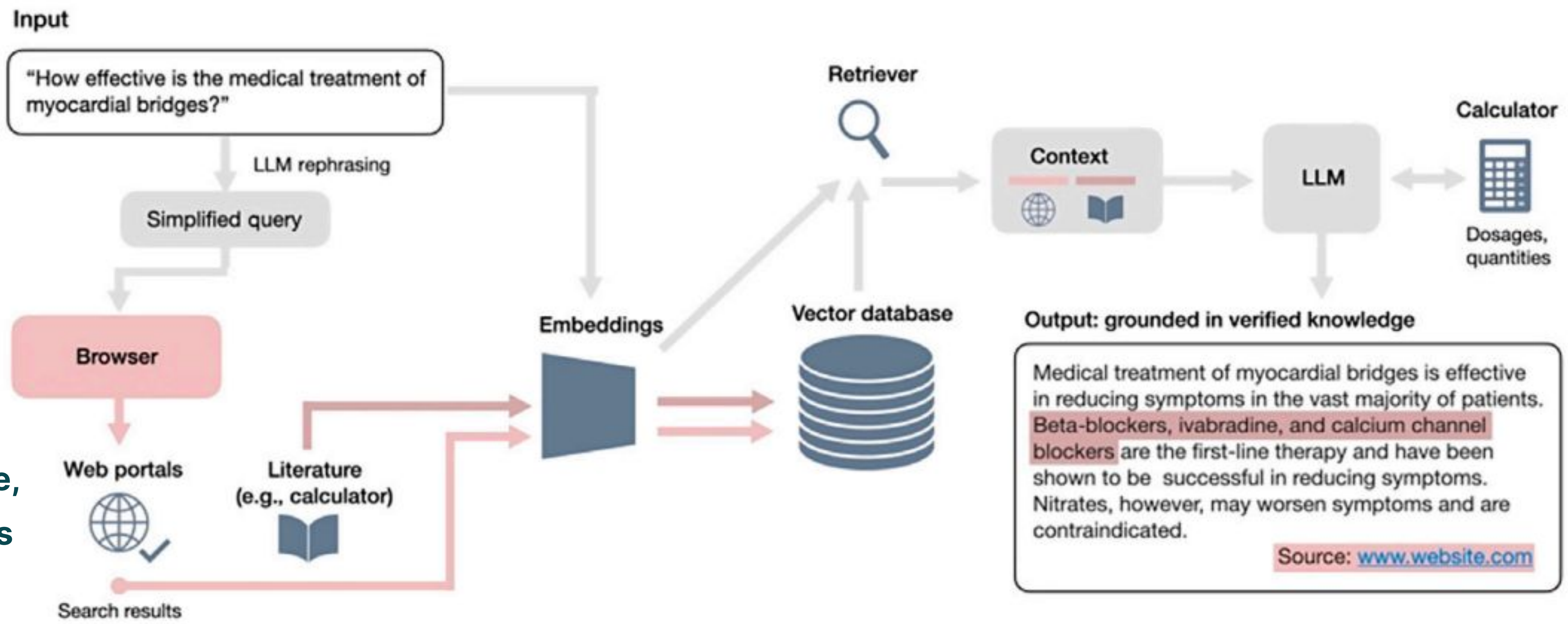
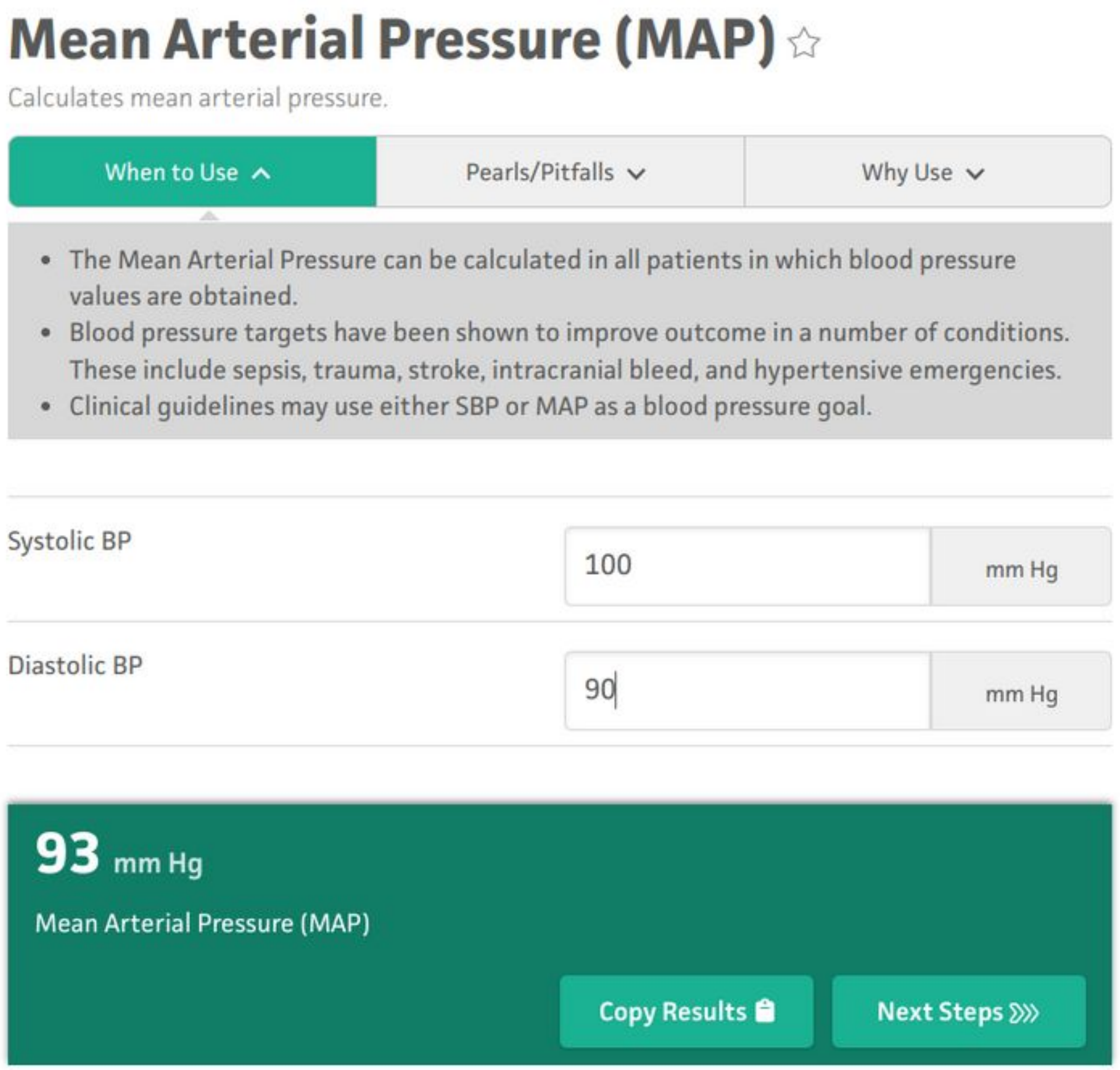


Figure 1. Almanac Overview.

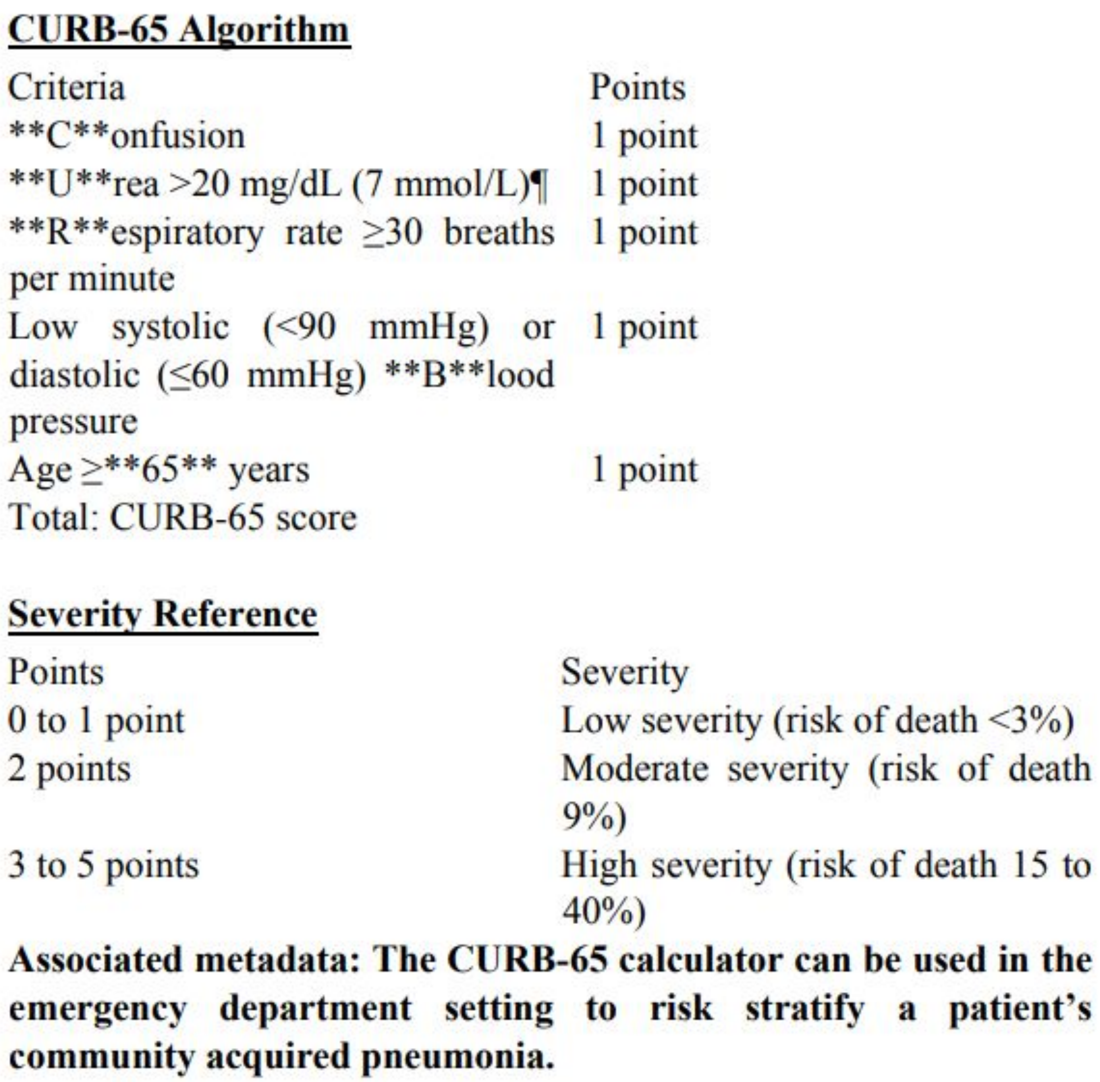
When presented with a query, Almanac uses external tools to retrieve relevant information before synthesizing a response with citations referencing source material. With this framework, large language model (LLM) outputs remain grounded in truth while providing a reliable way of fact-checking.

https://ai.nejm.org/doi/pdf/10.1056/AIoa2300068

Almanac—Retrieval-Augmented Language Models for Clinical Medicine



Medical Calculators
(MedCalc)



CURB-65 Algorithm

Criteria	Points
Confusion	1 point
Urea >20 mg/dL (7 mmol/L)¶	1 point
Respiratory rate ≥30 breaths per minute	1 point
Low systolic (<90 mmHg) or diastolic (≤60 mmHg) **B**lood pressure	1 point
Age ≥**65** years	1 point
Total: CURB-65 score	

Severity Reference

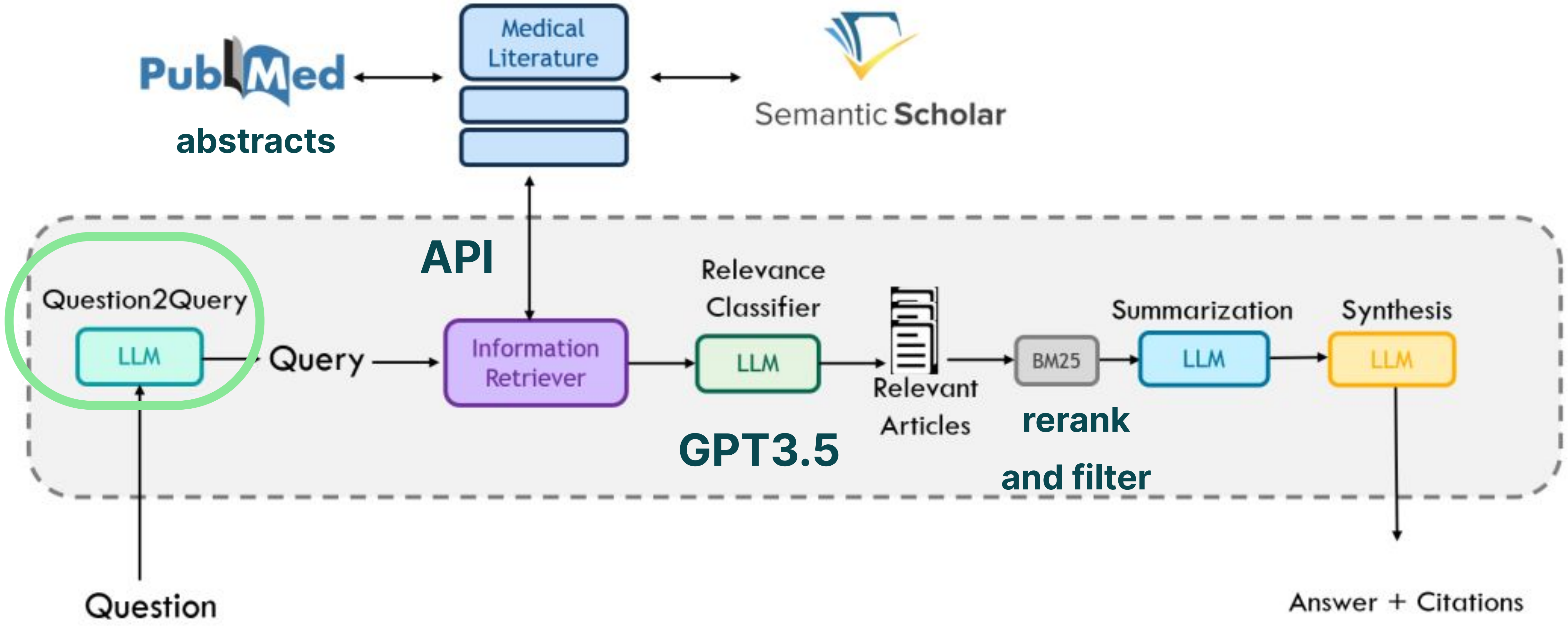
Points	Severity
0 to 1 point	Low severity (risk of death <3%)
2 points	Moderate severity (risk of death 9%)
3 to 5 points	High severity (risk of death 15 to 40%)

Associated metadata: The CURB-65 calculator can be used in the emergency department setting to risk stratify a patient's community acquired pneumonia.

MedCalc in Almanac
(embedded in Qdrant)

feb 2024

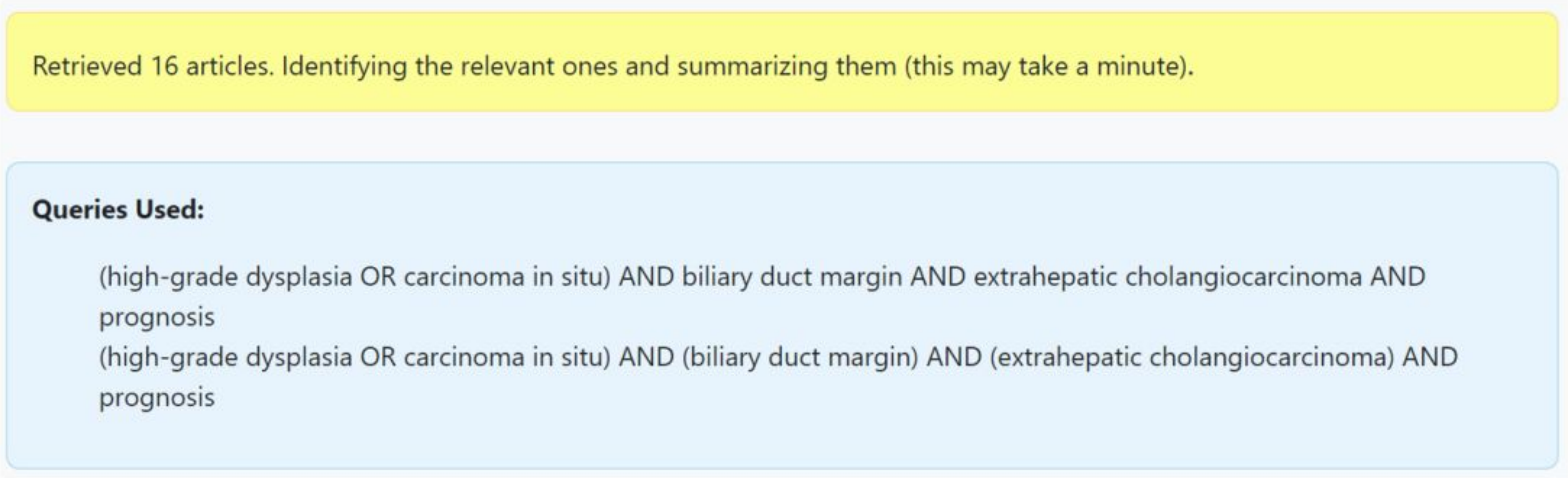
Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature



https://www.worldscientific.com/doi/abs/10.1142/9789811286421_0002

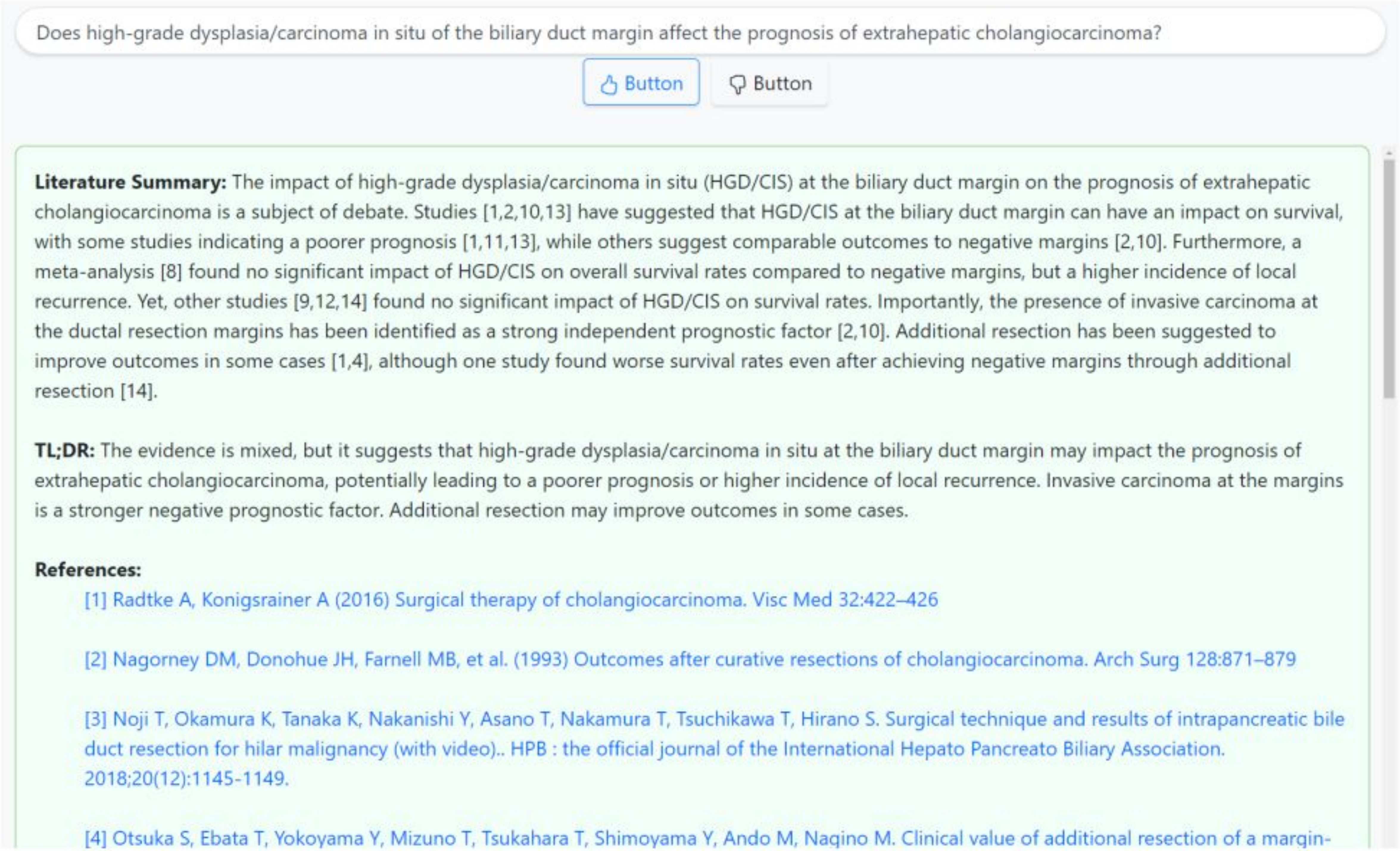
Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature

Question2Query



https://www.worldscientific.com/doi/abs/10.1142/9789811286421_0002

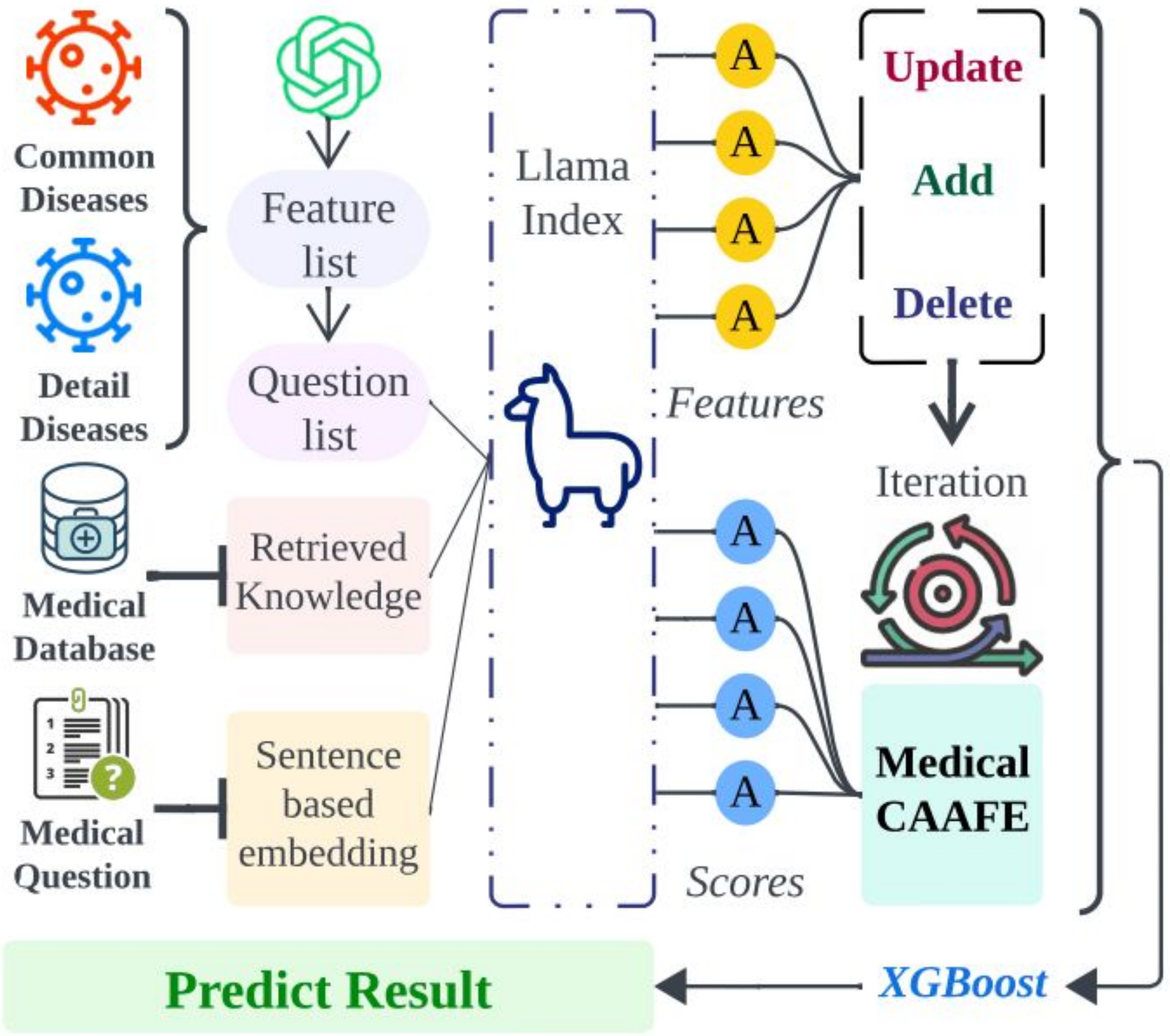
Clinfo.ai: An Open-Source Retrieval-Augmented Large Language Model System for Answering Medical Questions using Scientific Literature



https://www.worldscientific.com/doi/abs/10.1142/9789811286421_0002

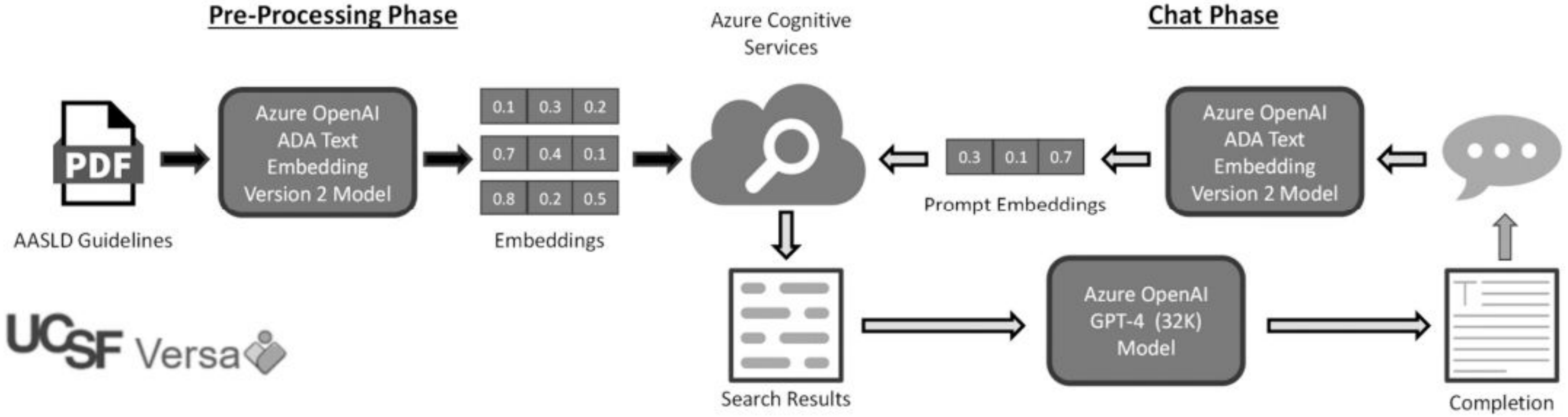
feb 2024

Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model



https://arxiv.org/abs/2402.00746

Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation



Disease specific!

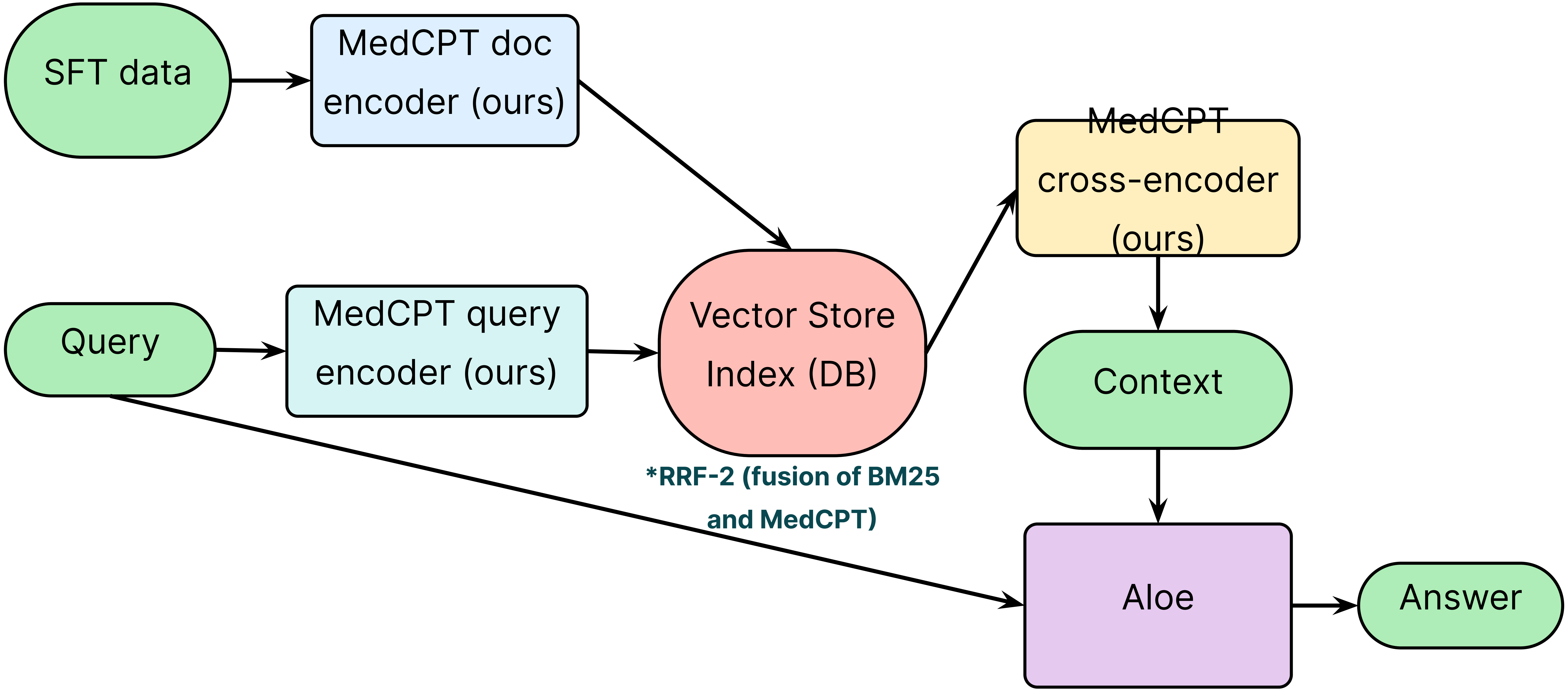
30 publicly available American Association for the Study of Liver Diseases (AASLD) guidelines

https://www.medrxiv.org/content/10.1101/2023.11.10.23298364v1

Reca
p

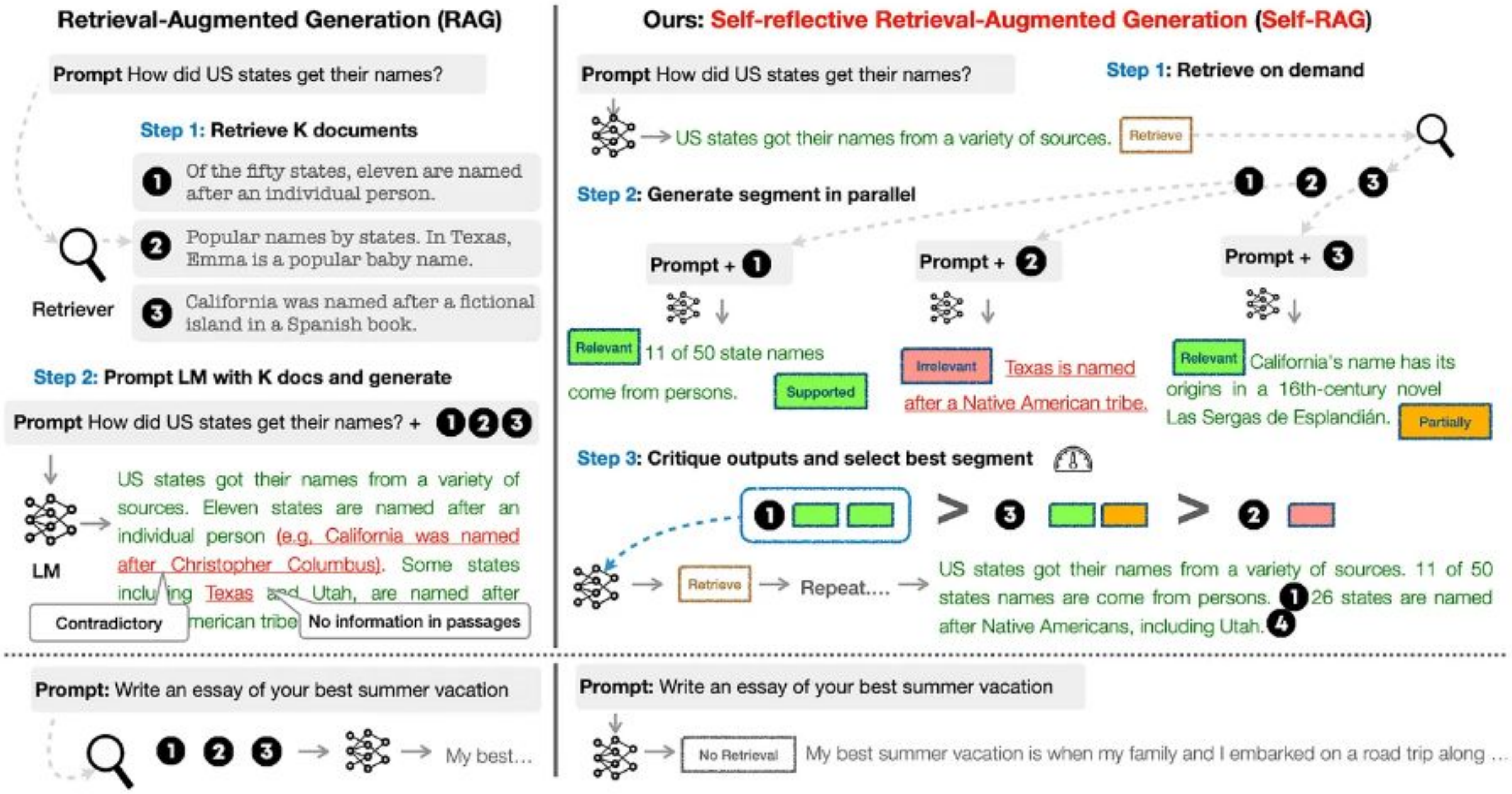
- Hybrid search replacing dense search
- Reranker model to rerank and filter retrieved elements
- Variable length embedding models
- Medical embedding model
- Medical embedding model fine-tuned for retrieval
- Single model for embedding and text generation

Proposal


SFT data
MedCPT doc encoder (ours)
Query
MedCPT query encoder (ours)
Vector Store Index (DB)
*RRF-2 (fusion of BM25 and MedCPT)
MedCPT cross-encoder (ours)
Context
Aloe
Answer

Extra stuff

Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection



https://arxiv.org/abs/2310.11511

Evaluatio

n