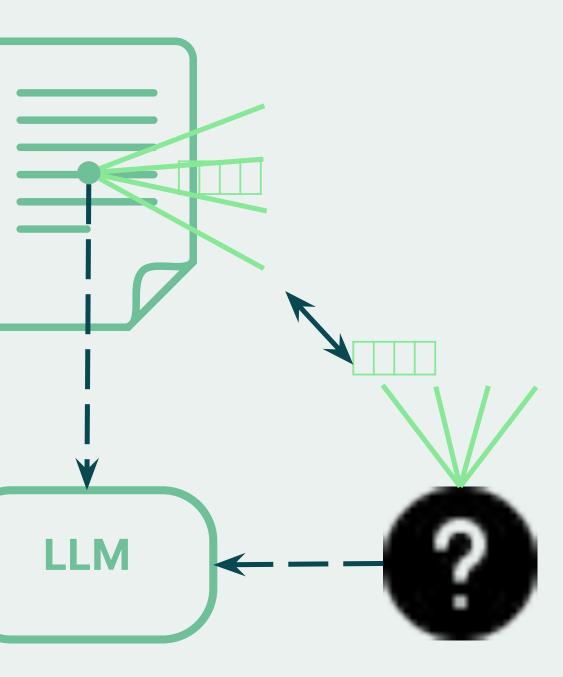
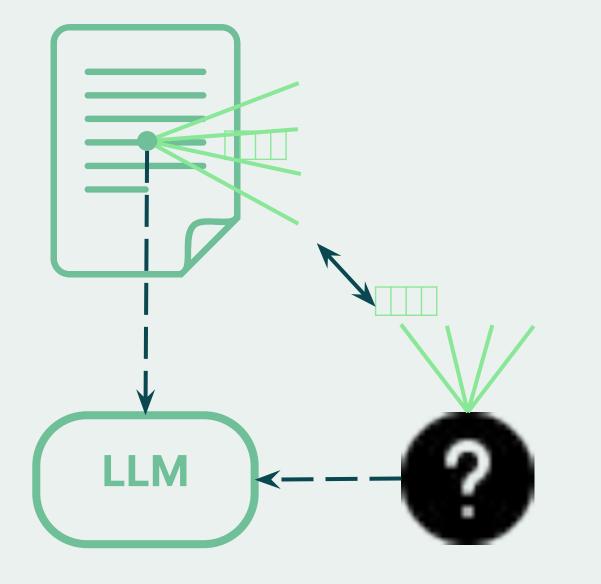
# Retrieval Augmented Generation

thanks to Lucía Urcelay



# **Basic idea**



retrieval engine

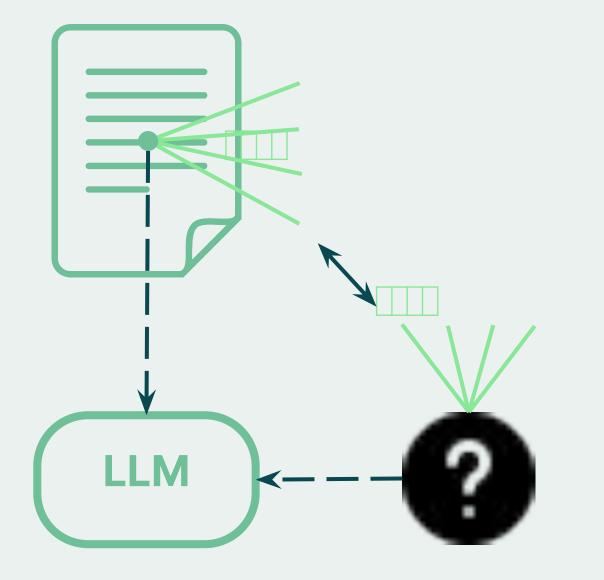
# Integrate LLM with an information

## • Embed documents into vectors

# • Find most similar documents

# Add as context for response



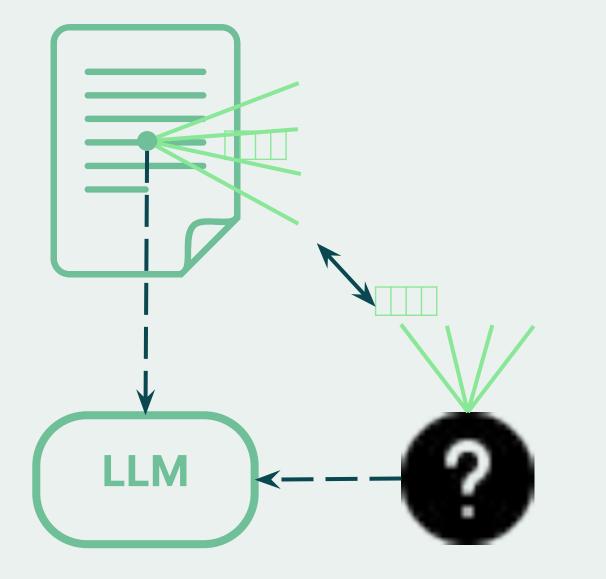


Access to up-to-date knowledge

Reduce hallucinations/Improve factuality

Reliable referencing (sources)





Limited context

Cost of inference

Cost of persistence

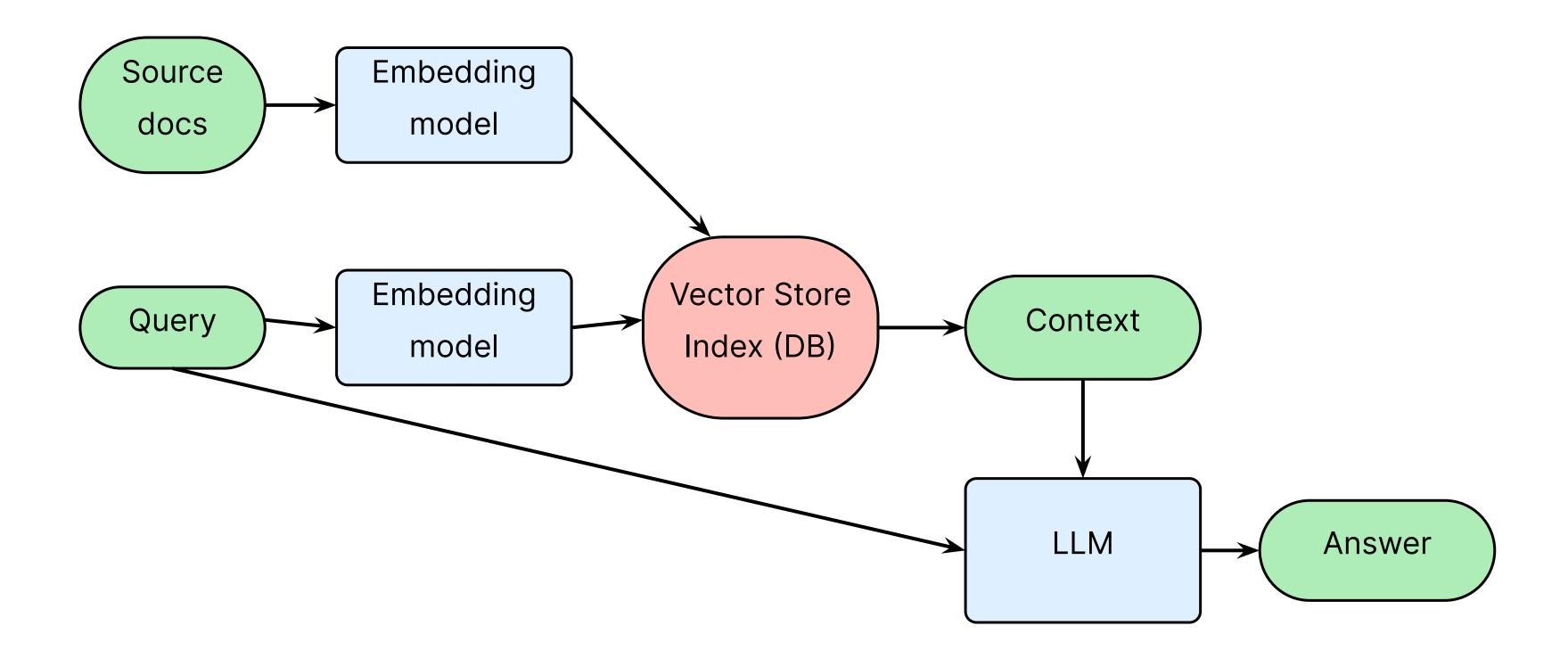
No inter-document processing

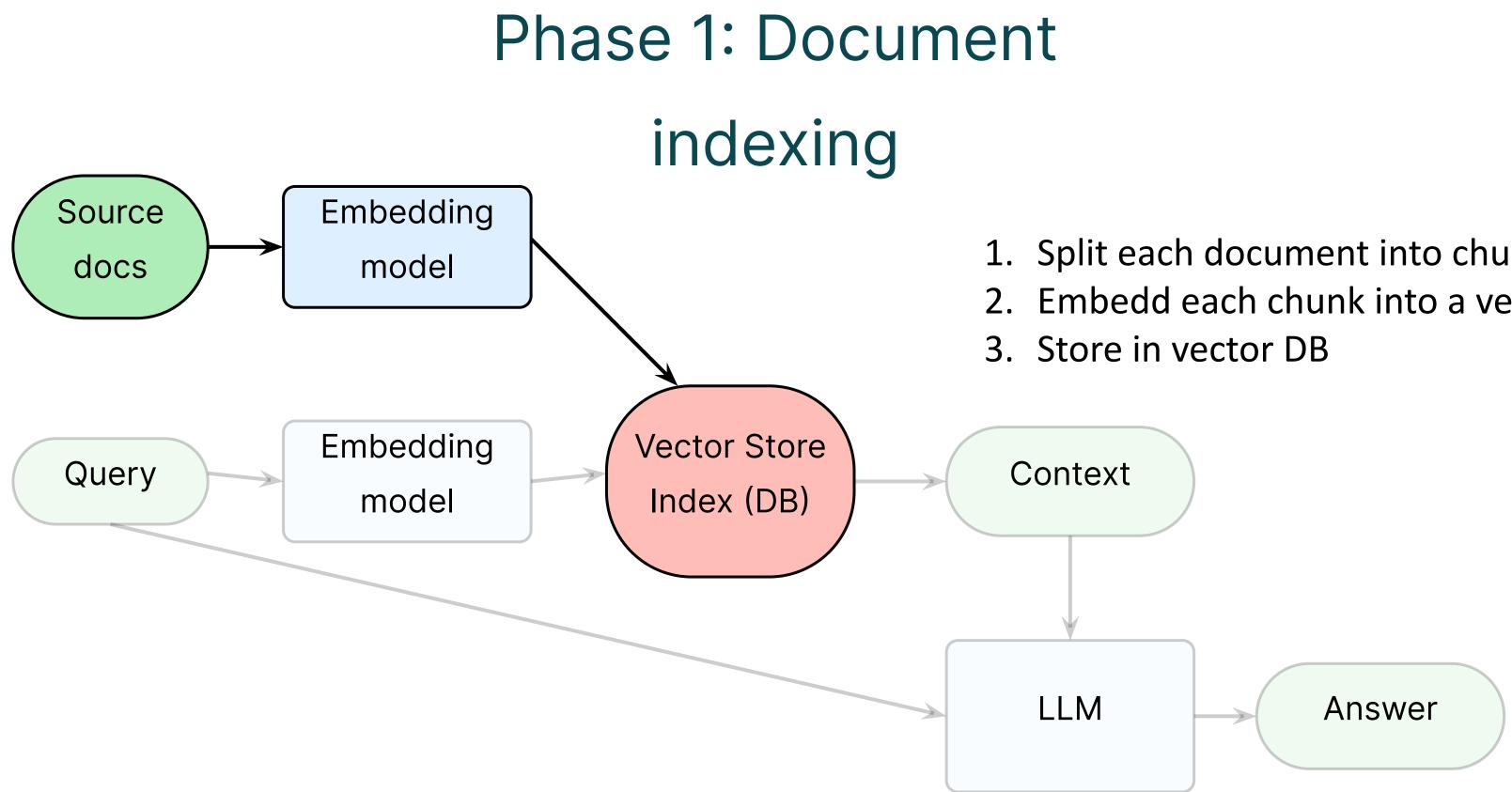
Baseline

RAG



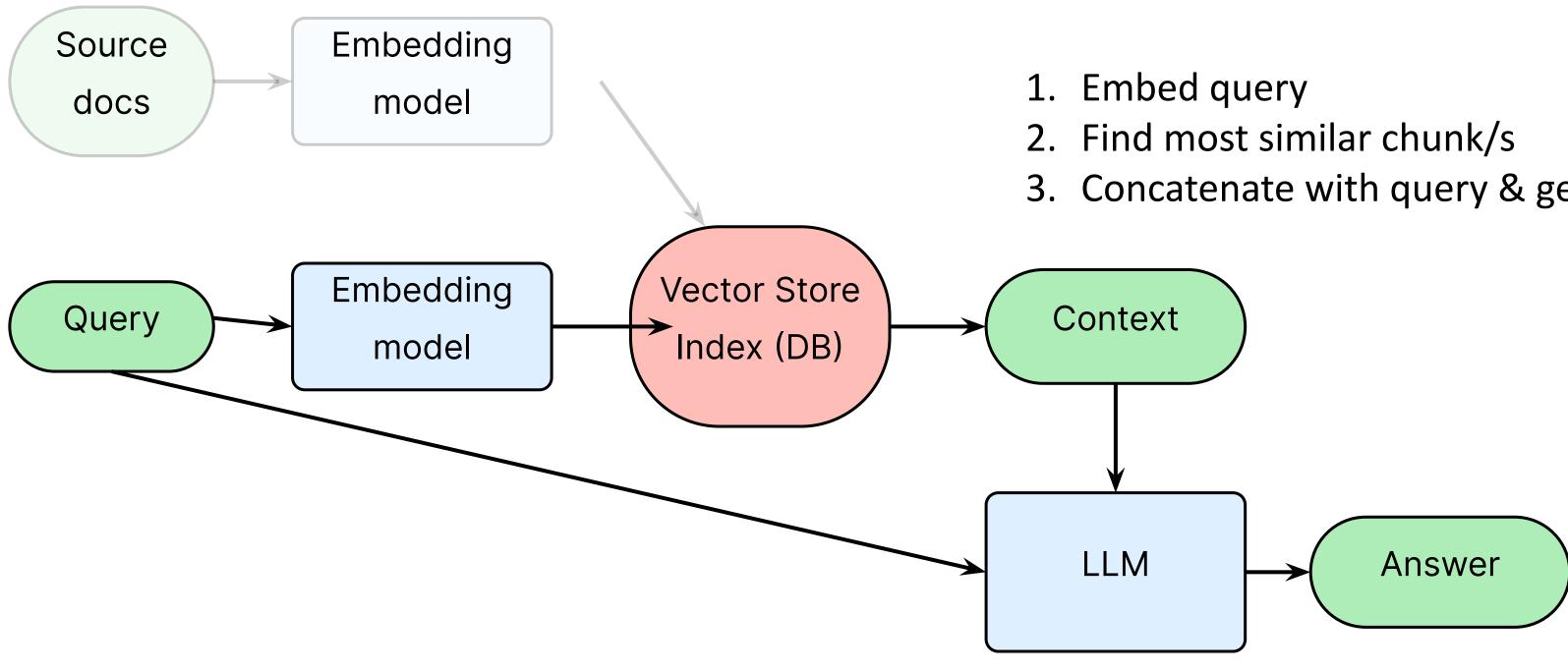
# Baseline RAG





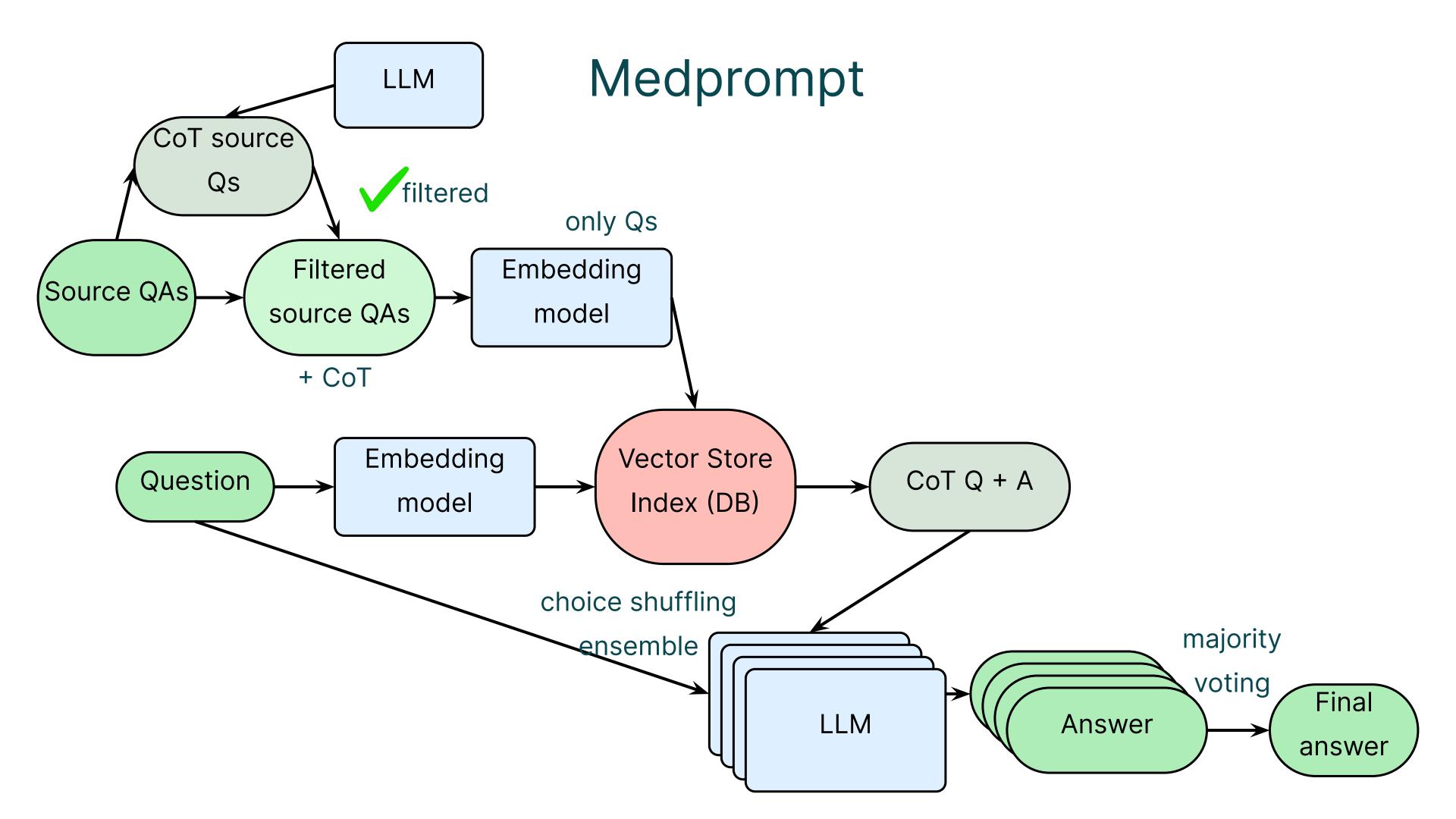
- 1. Split each document into chunks
- 2. Embedd each chunk into a vector

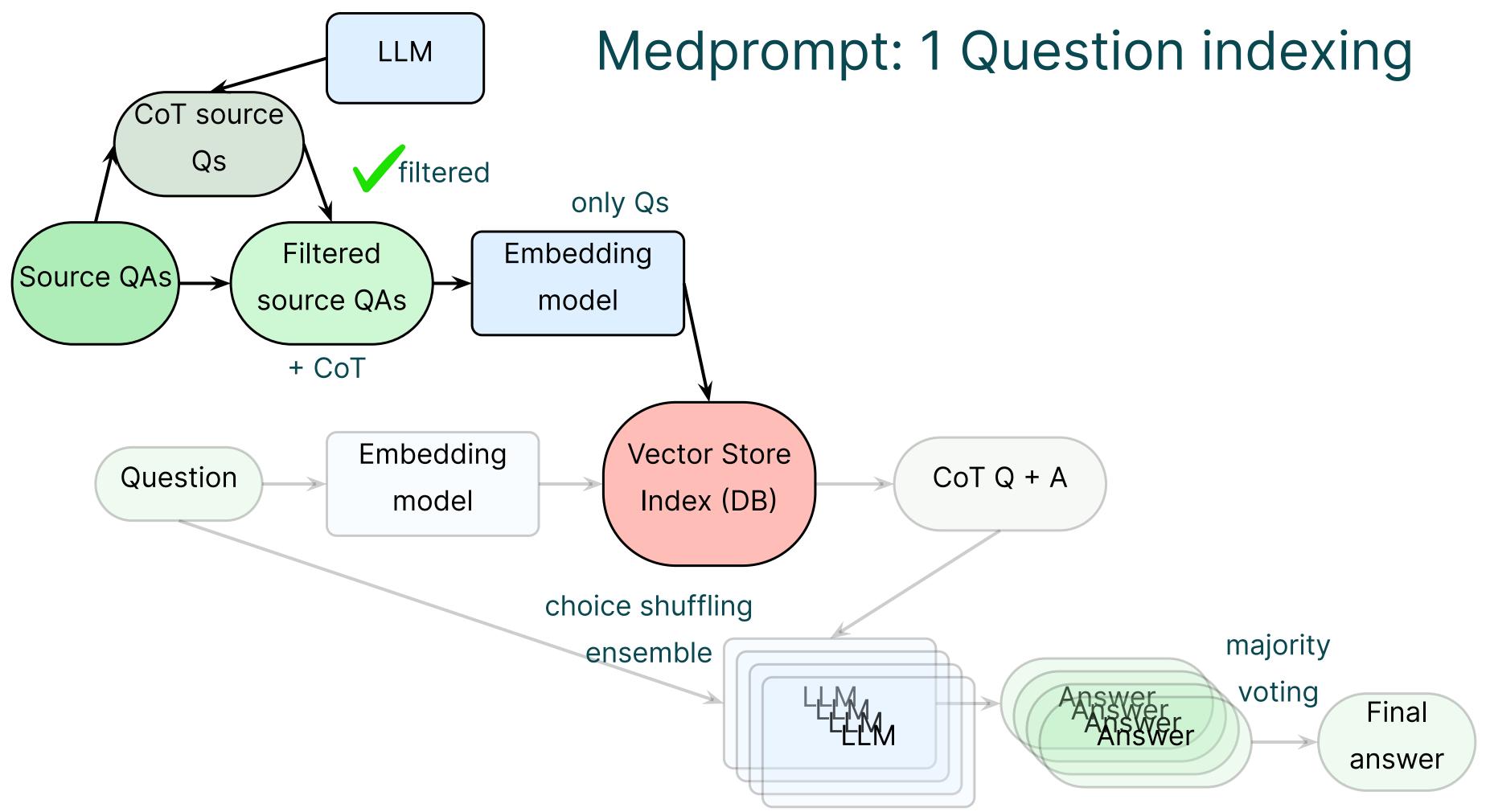
# Phase 2: Inference

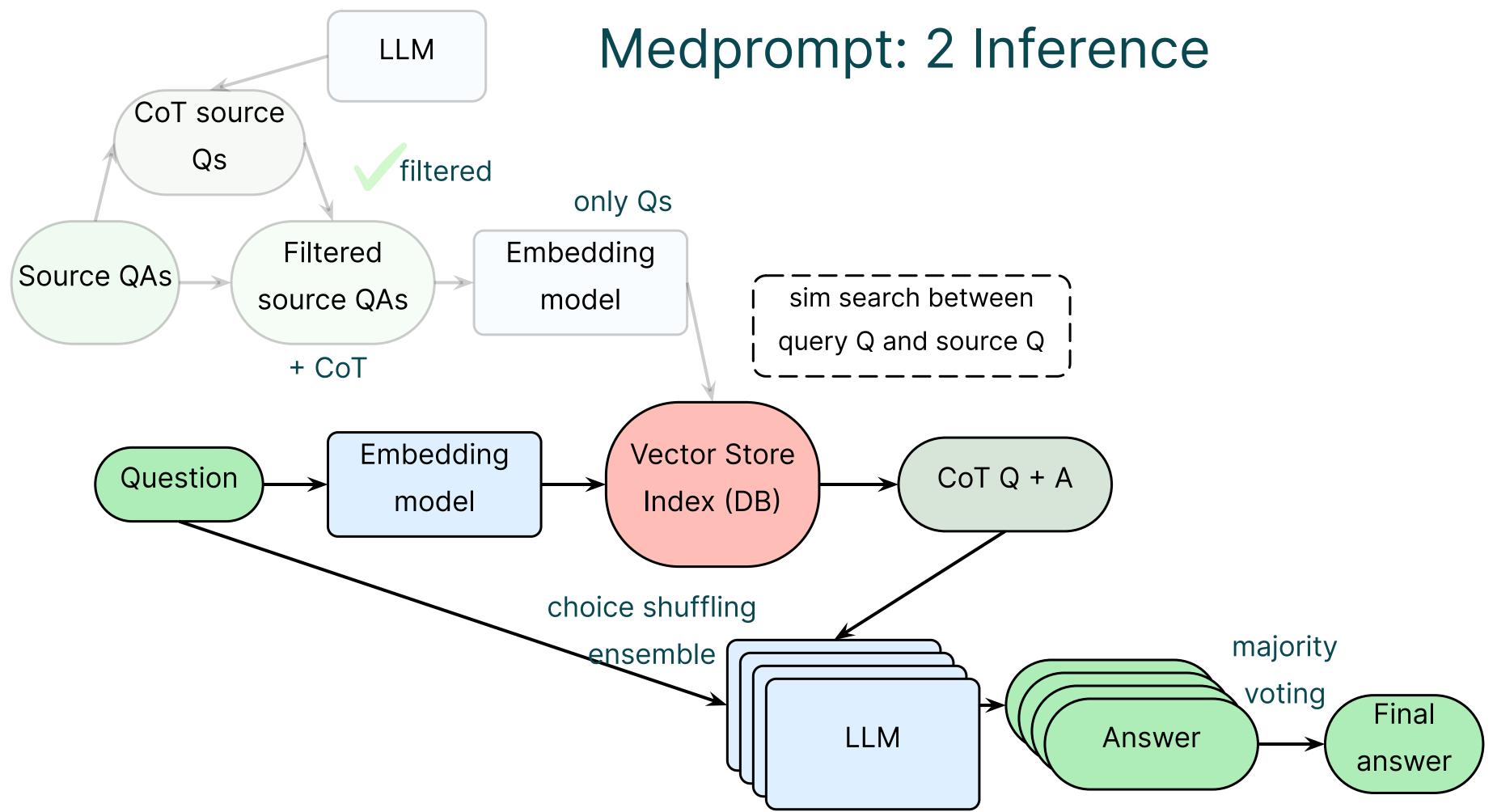


- 3. Concatenate with query & generate

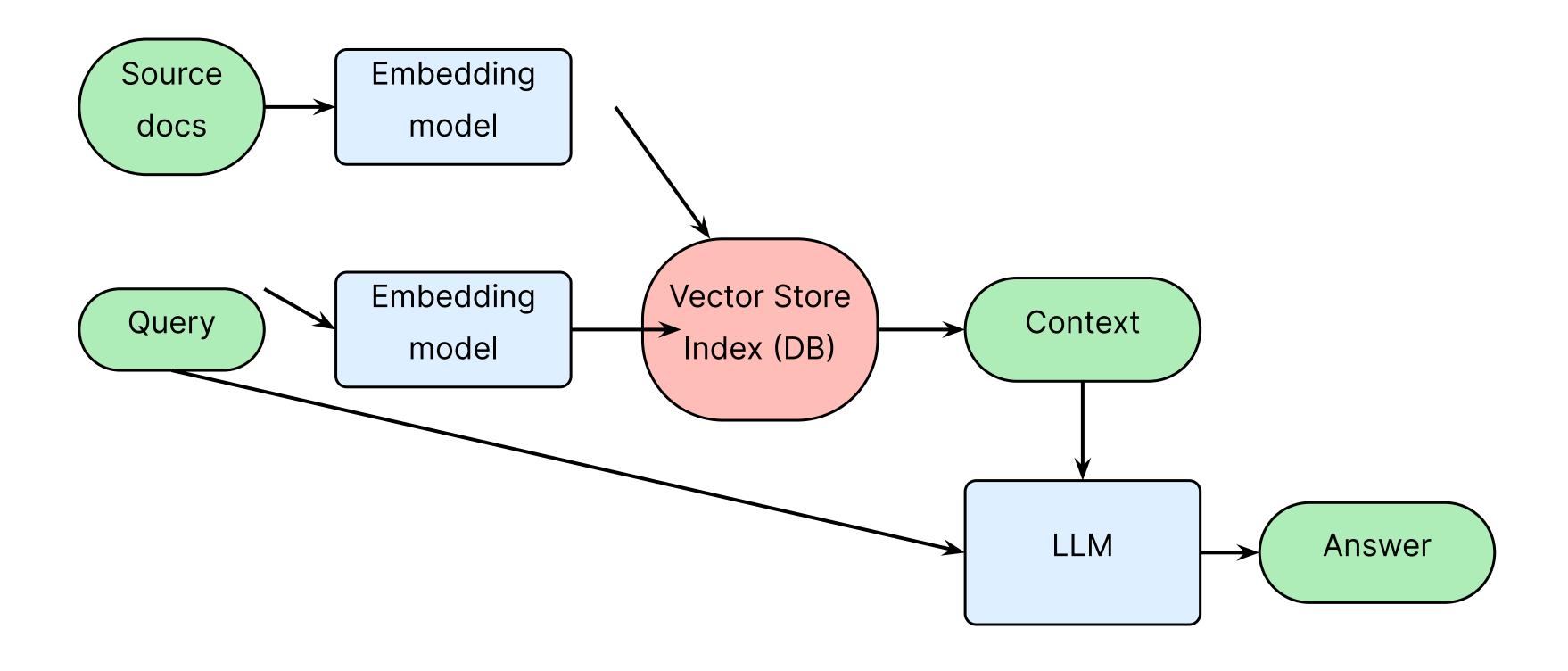
# Medprompt







# Back to baseline RAG

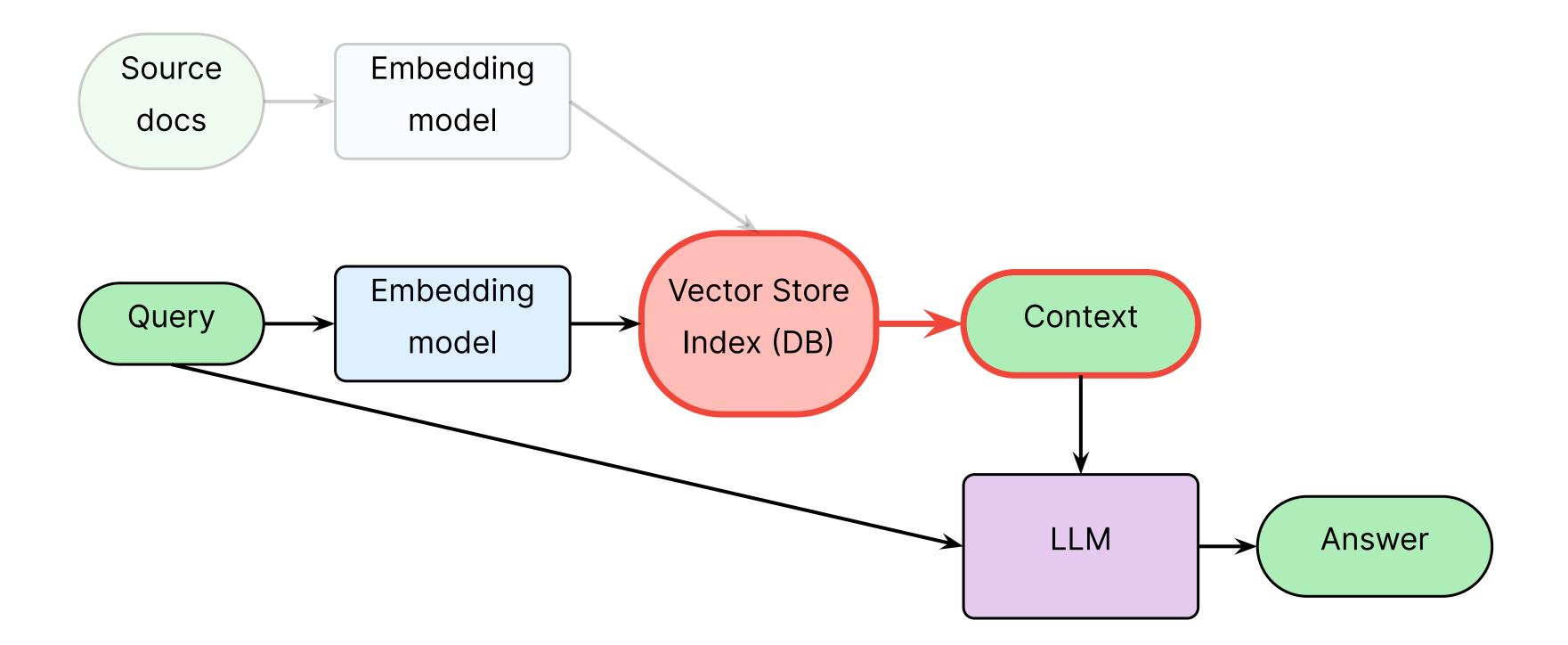




# RAG Painpoints

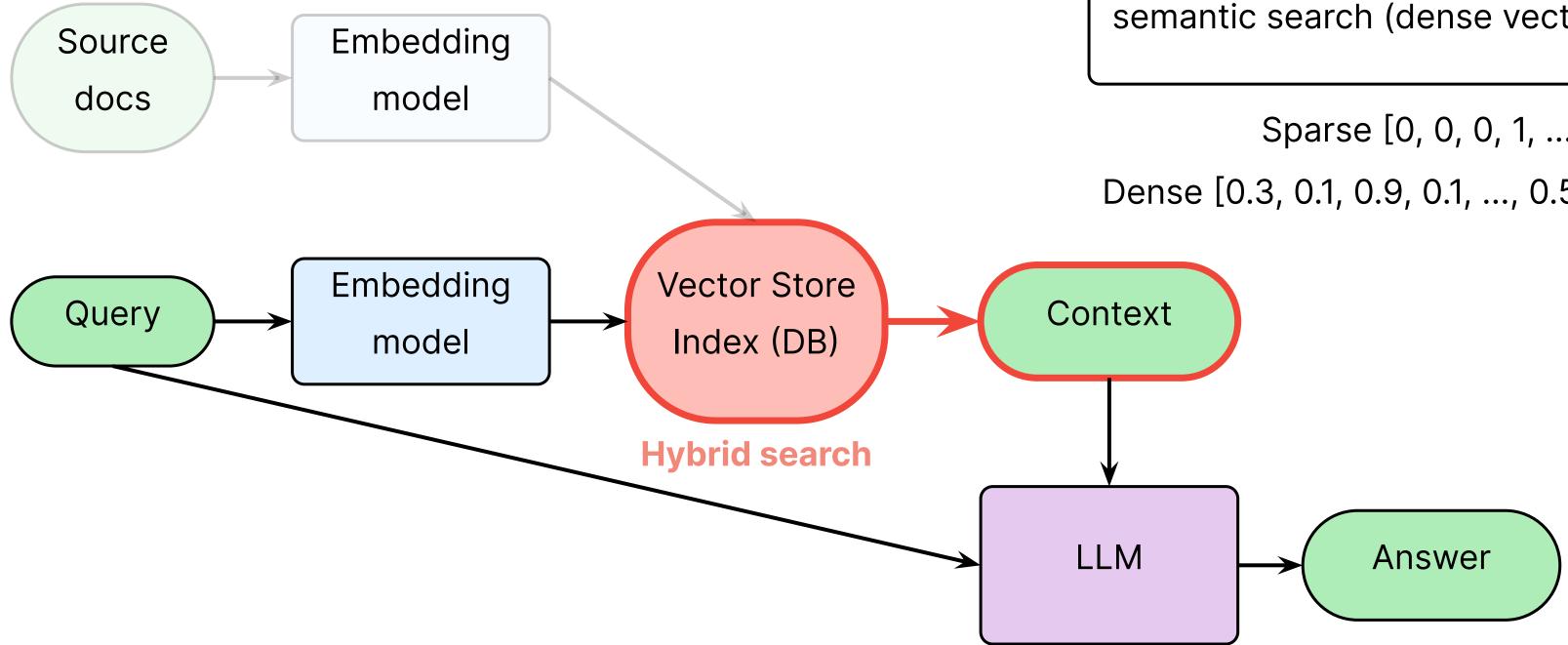


# Retrieval of <u>low-relevance</u> docs/chunks (1)



# Retrieval of <u>low-relevance</u> docs/chunks (1)

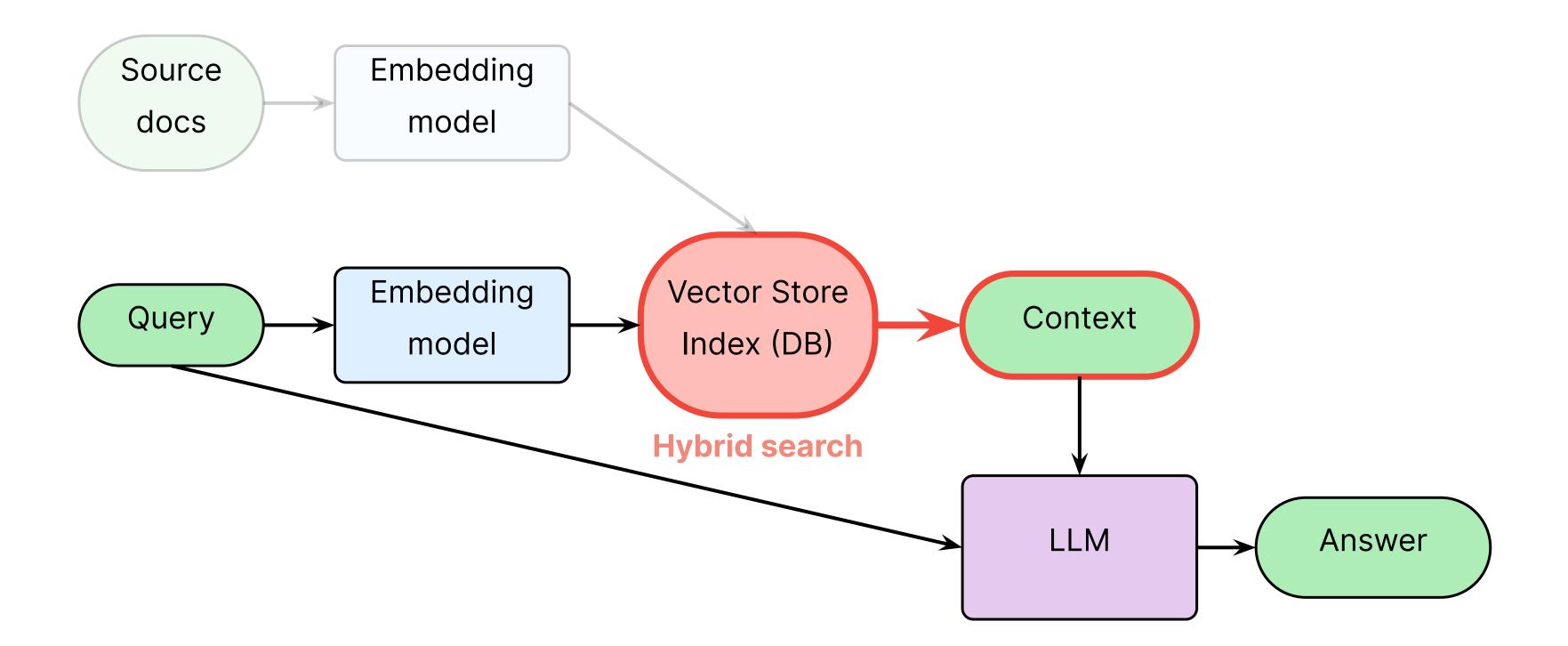
# **Hybrid search**



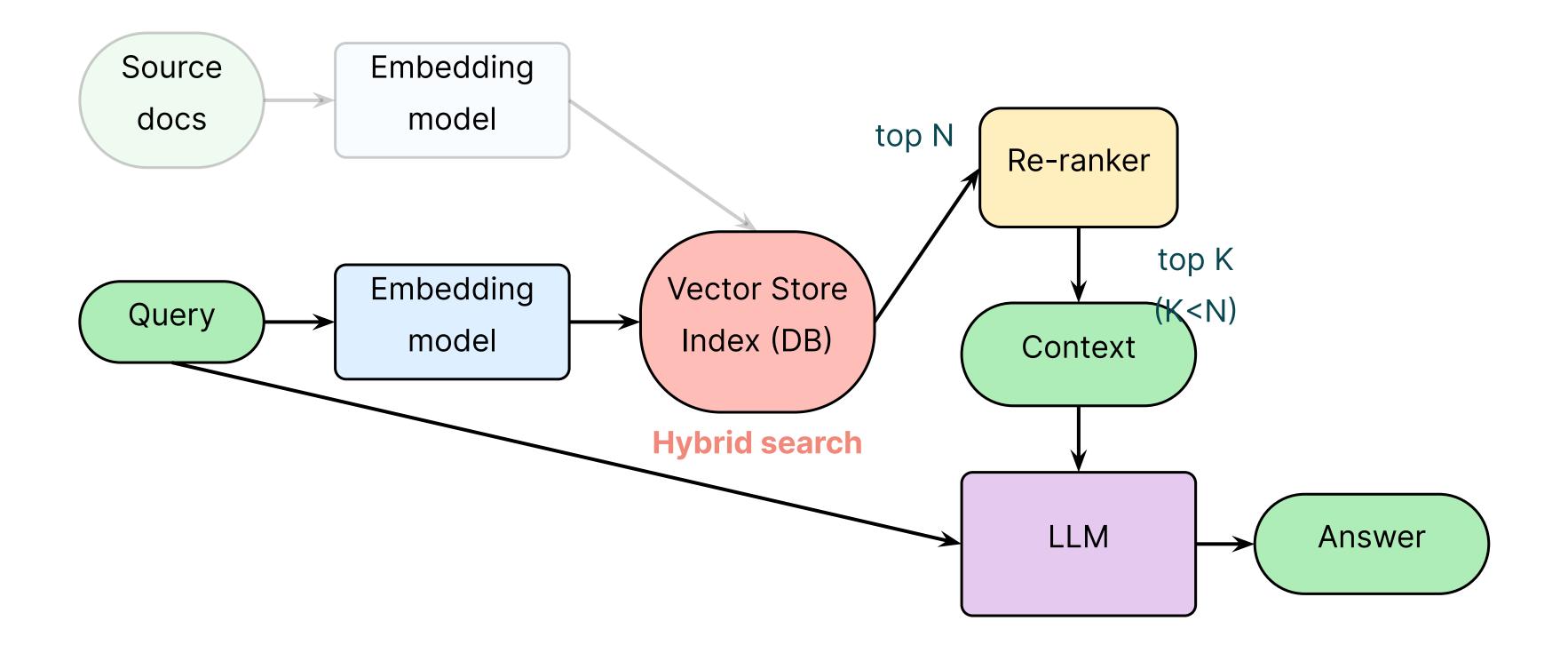
Keyword (sparse vectors) + semantic search (dense vectors)

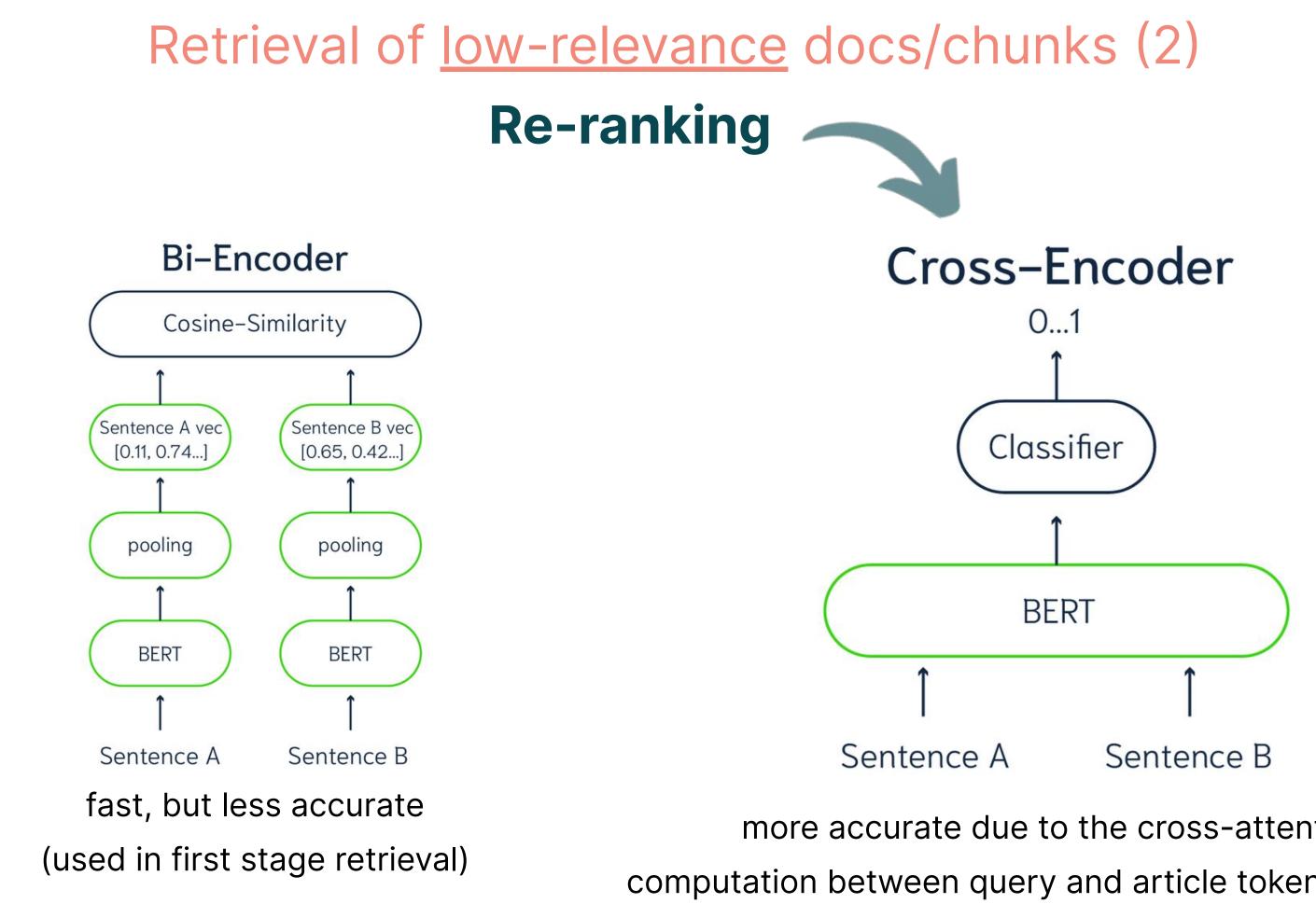
Sparse [0, 0, 0, 1, ..., 0, 0] Dense [0.3, 0.1, 0.9, 0.1, ..., 0.5, 0.2]

# Retrieval of <u>low-relevance</u> docs/chunks (2)



# Retrieval of <u>low-relevance</u> docs/chunks (2) **Re-ranking**





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

more accurate due to the cross-attention computation between query and article tokens, but slow (can be trained using datasets as MS-MARCO)

# Retrieval of <u>low-relevance</u> docs/chunks (2) **Re-ranking**



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

# Retrieval of <u>low-relevance</u> docs/chunks (2) **Re-ranking**



### Sentence Transformers - Cross-Encoders University

https://www.sbert.net/ 🔰 Nils\_Reimers 🖸 nreimers

### AI & ML interests

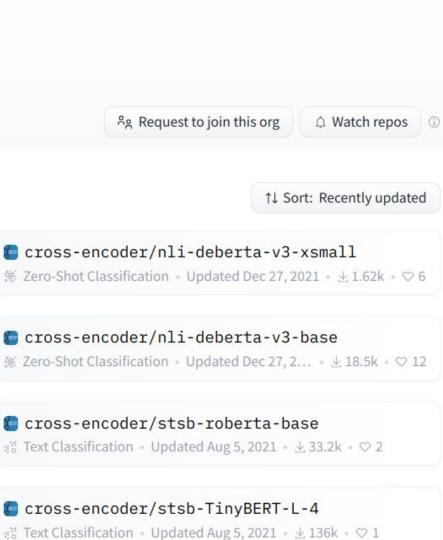
This repository hosts the cross-encoders from the SentenceTransformers package. More details on https://www.sbert.net/docs/pretrained\_crossencoders.html

Team members 1



HIR.	cross-encoder/nli-deberta-v3-large
10 10 10 10 10	Zero-Shot Classification • Updated Dec 28, 2 • ± 5.67k • ♡ 17
Ð	cross-encoder/nli-deberta-v3-small
<u>814'</u> 212	Zero-Shot Classification → Updated Dec 27, 2021 → ± 11.6k → ♡ 8
<b>1</b>	cross-encoder/stsb-roberta-large
	Text Classification → Updated Aug 5, 2021 → ± 44.4k → ♡ 8
	cross-encoder/stsb-distilroberta-base
	Text Classification • Updated Aug 5, 2021 • ± 19.8k • ♡ 2

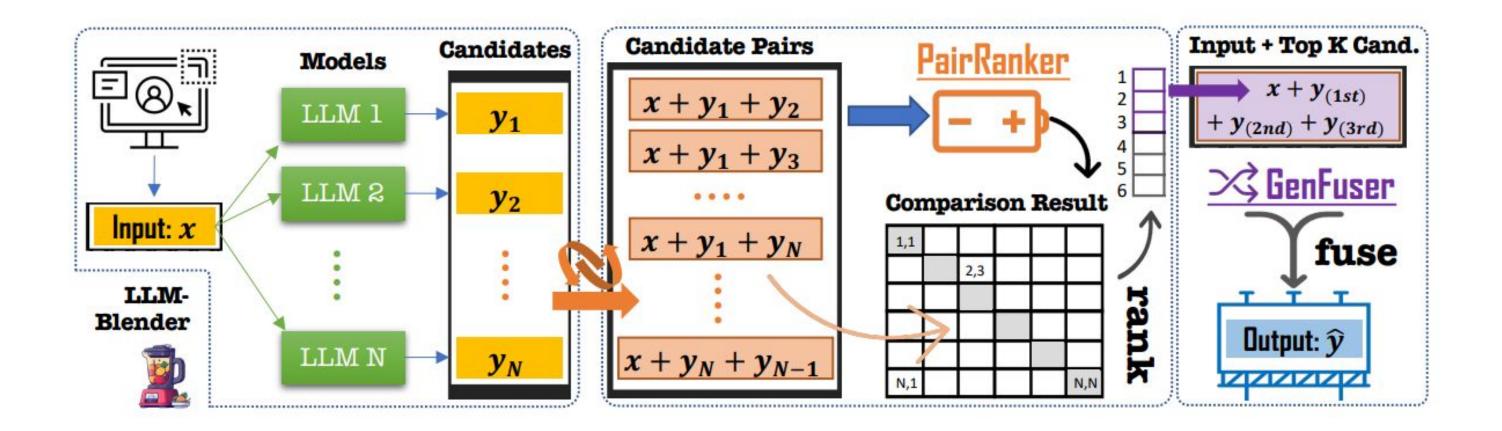
### https://huggingface.co/cross-encode



### cross-encoder/quora-roberta-base Text Classification • Updated Aug 5, 2021 • ± 4.22k • ♡ 1

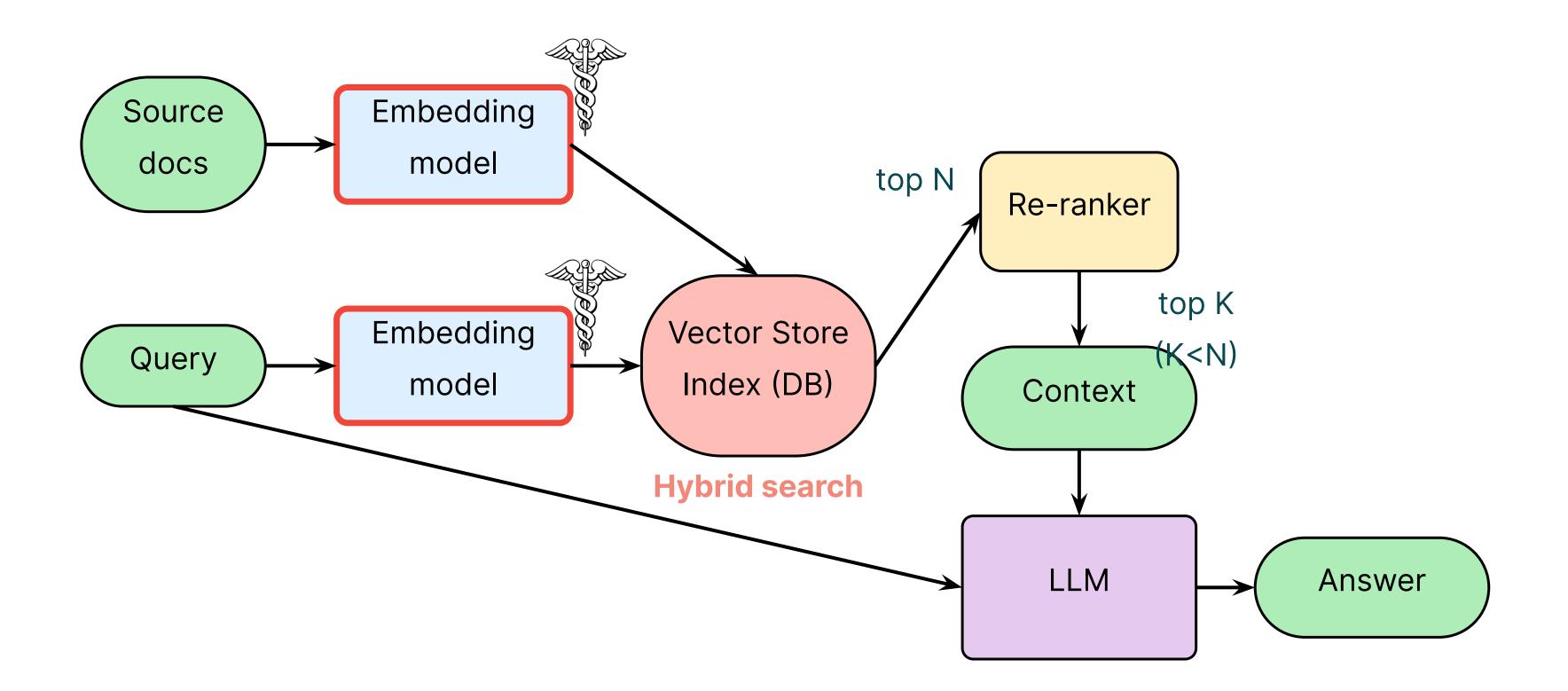
# Retrieval of <u>low-relevance</u> docs/chunks (2)

# Pair RM



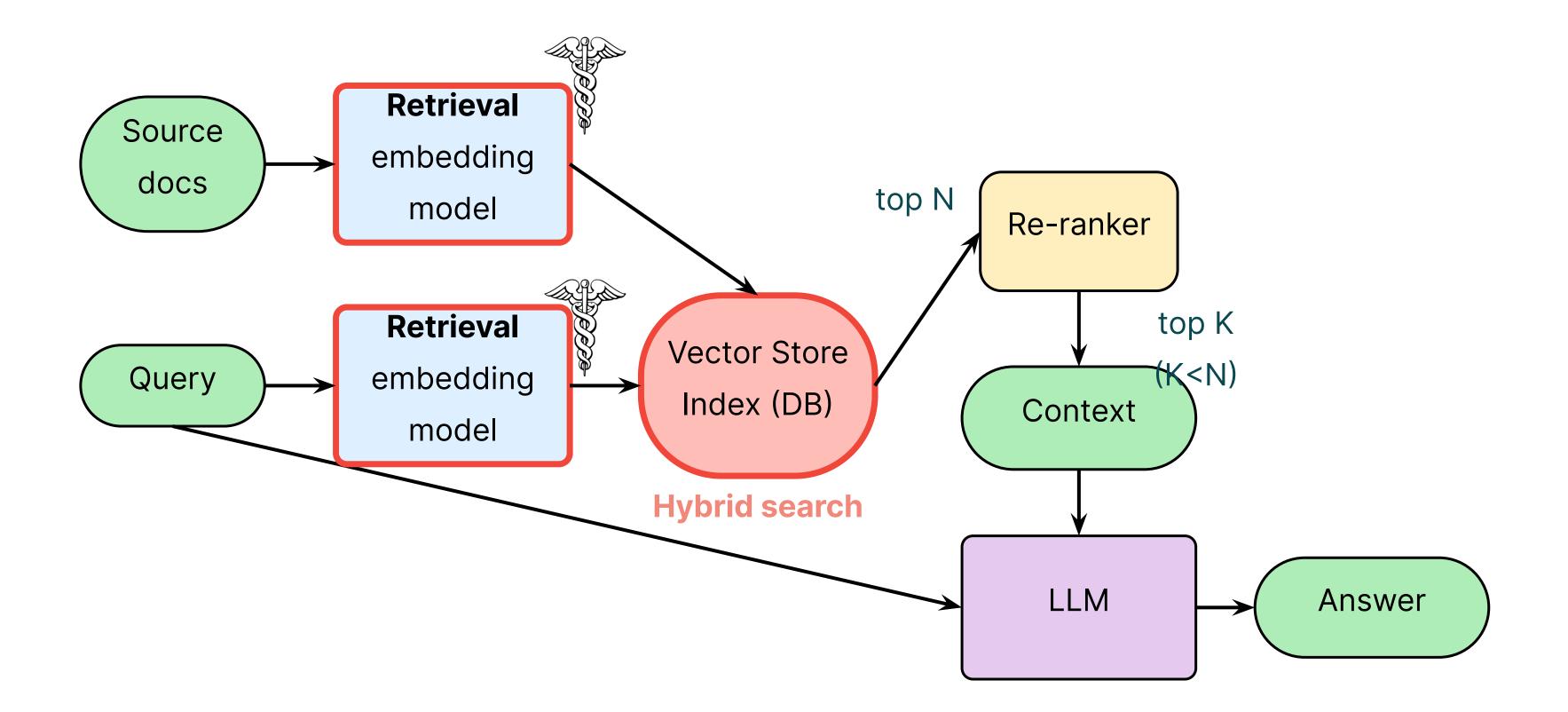
- based on microsoft/deberta-v3-large
- achieves superior performance by mixing the outputs of multiple LLMs

# Non specialised embeddings (for medicine)



# Non specialised embeddings (for medicine)

# **Medical Embedding Models**



### **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

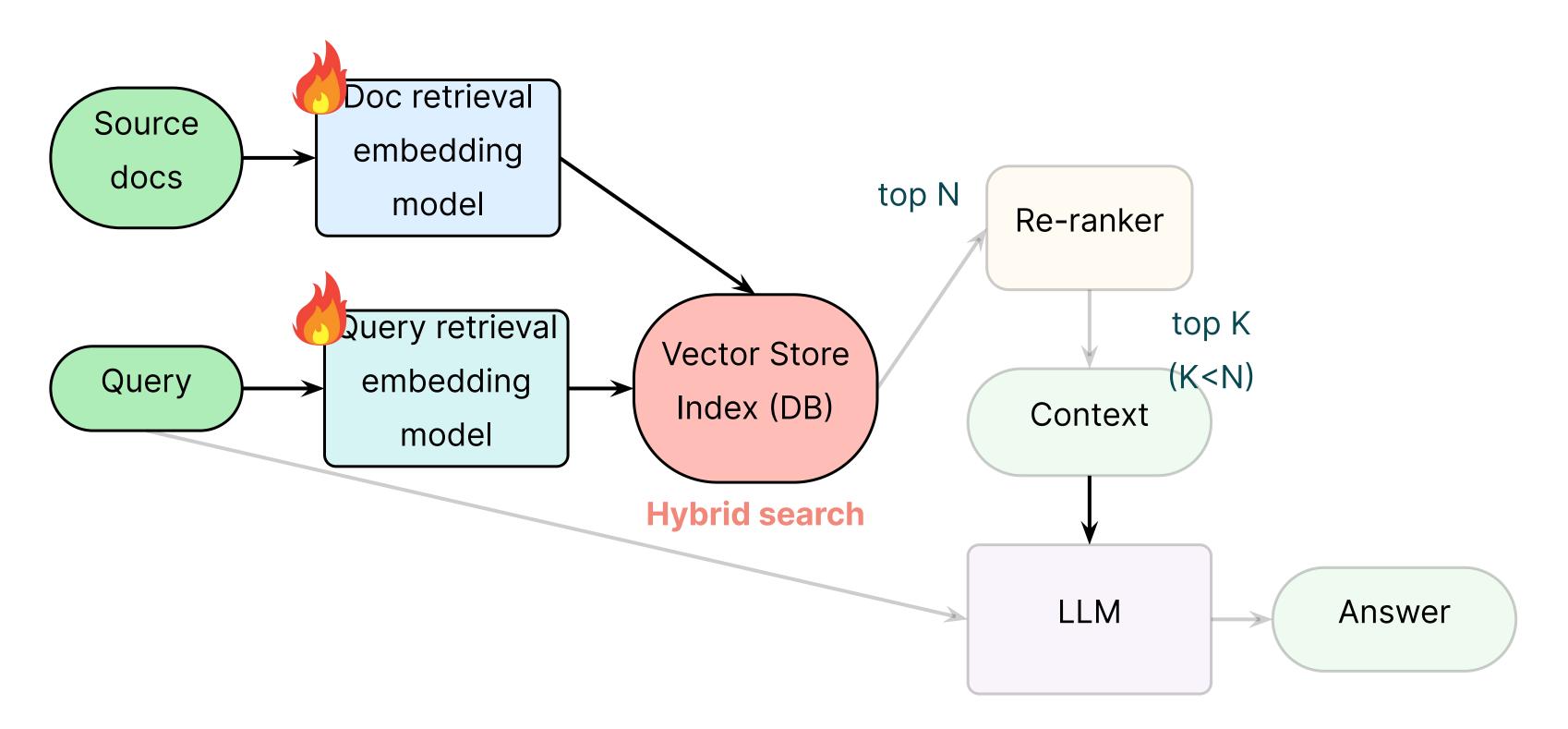
### Q: What is the primary function of the spleen?

A1: The primary function of the spleen is unknown Sim=0.913

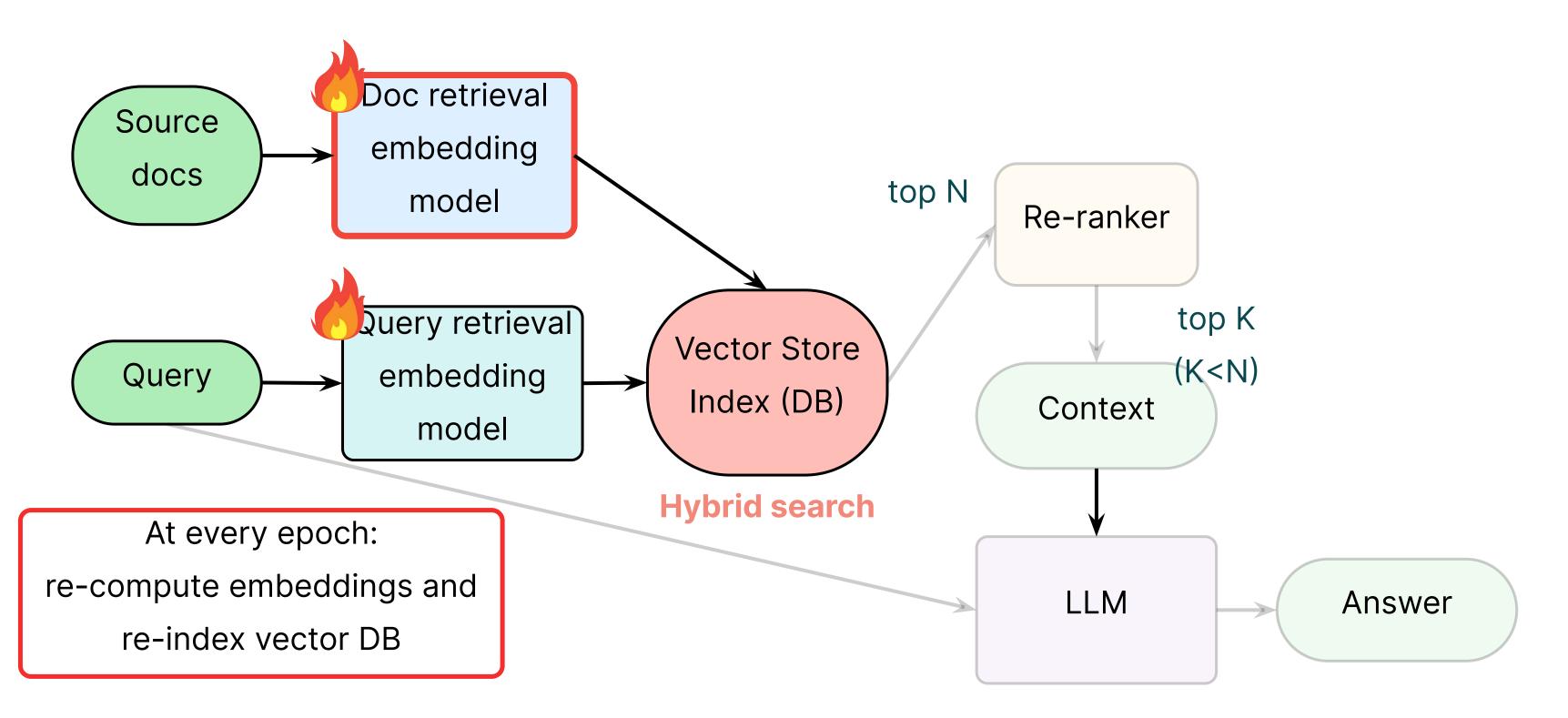
A2: The spleen plays a key role in filtering and removing old or damaged red blood cells from the bloodstream Sim=0.667

similar but NOT relevant

# **Medical Retrieval Embedding Models**



# **Medical Retrieval Embedding Models**



# Non specialised embeddings (for medical retrieval) **Medical Retrieval Embedding Models**

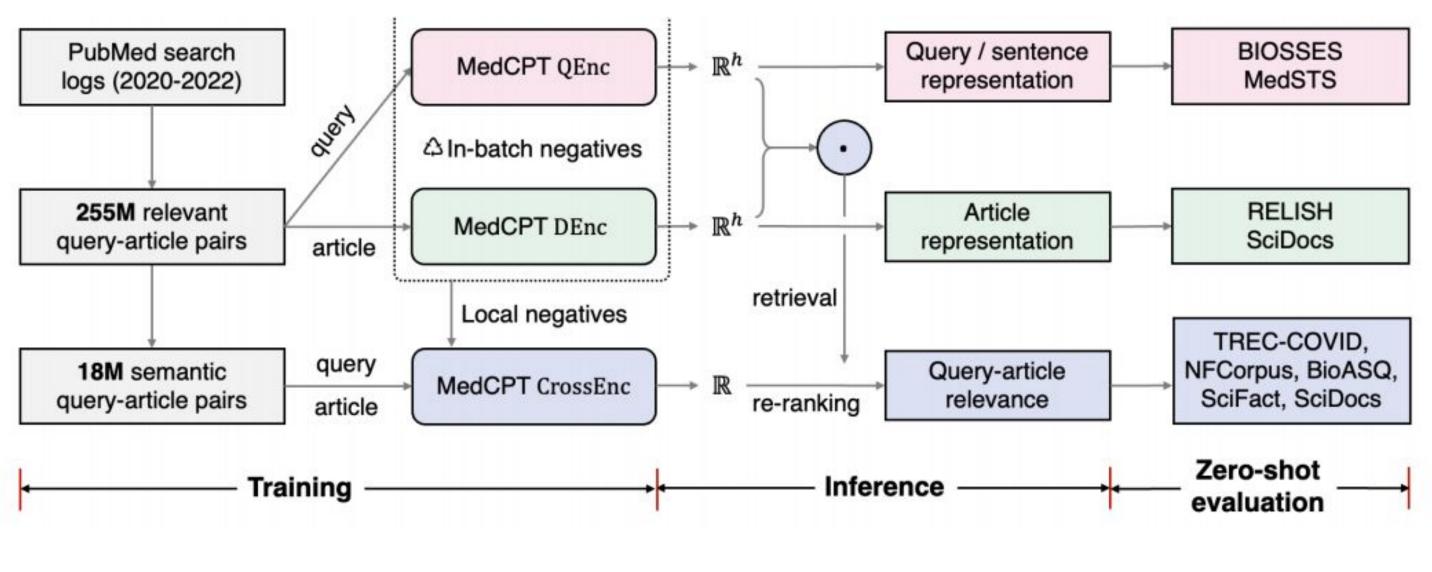
### MedCPT: Contrastive Pre-trained Transformers with Large-scale PubMed Search Logs for Zero-shot **Biomedical Information Retrieval**

Qiao Jin<sup>1</sup>, Won Kim<sup>1</sup>, Qingyu Chen<sup>1</sup>, Donald C. Comeau<sup>1</sup>, Lana Yeganova<sup>1</sup>, W. John Wilbur<sup>1</sup>, Zhiyong Lu<sup>1</sup> <sup>1</sup>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

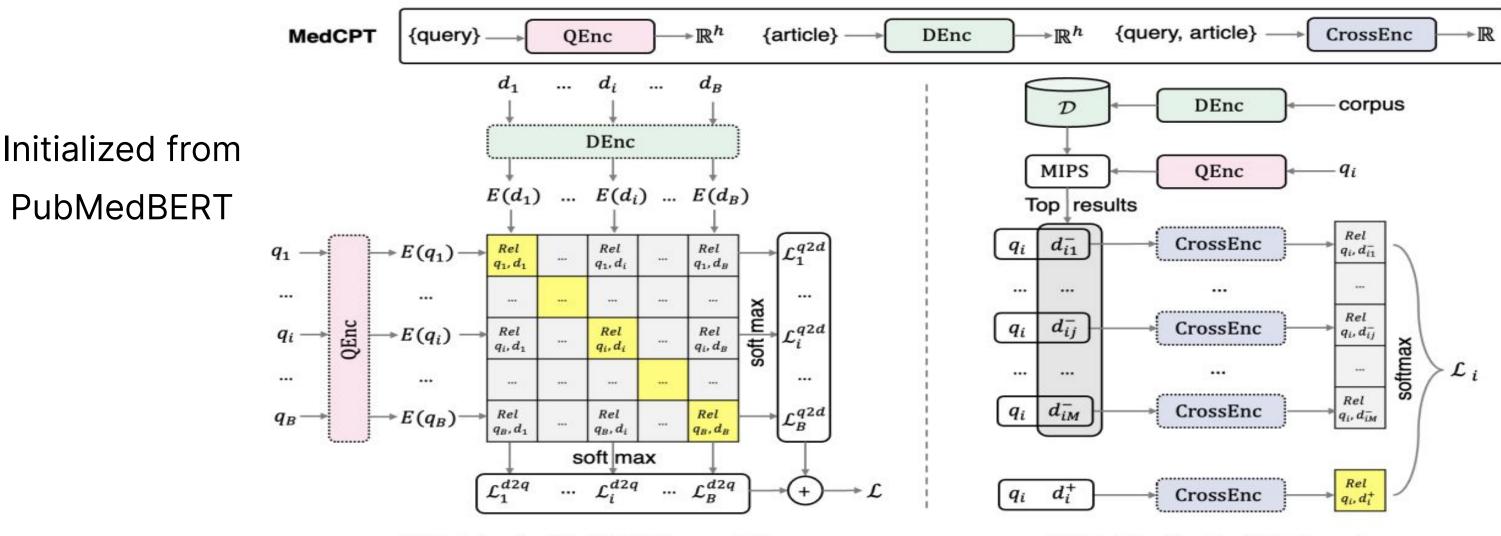
MedCPT: Query Encoder + Article Encoder + re-ranker cross-encoder

# Non specialised embeddings (for medical retrieval) **Medical Retrieval Embedding Models**



https://arxiv.org/abs/2307.00589

# <u>Non specialised embeddings (for medical retrieval)</u> **Medical Retrieval Embedding Models**



(A) Training the MedCPT QEnc and DEnc

(B) Training the MedCPT CrossEnc

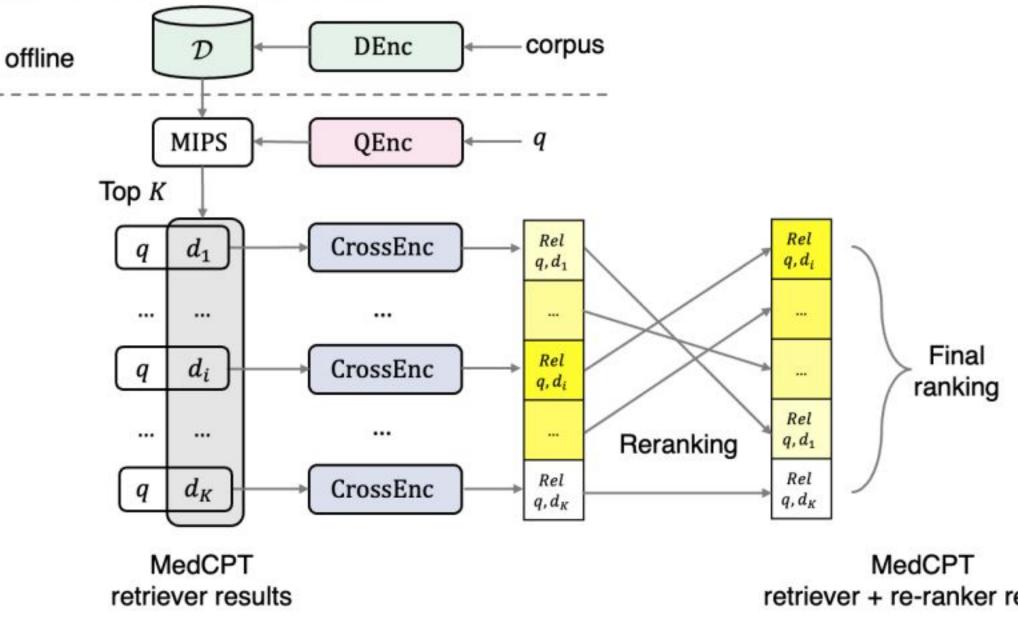
Figure 2. Overview of the MedCPT training process. (A) Training the MedCPT query encoder (QEnc) and document encoder (DEnc) using a contrastive loss with querydocument pairs and in-batch negatives; (B) Training the MedCPT cross-encoder

https://arxiv.org/abs/2307.00589

### Initialized from PubMedBERT

# **Medical Retrieval Embedding Models**

### Appendix A: MedCPT Inference



### https://arxiv.org/abs/2307.00589

retriever + re-ranker results

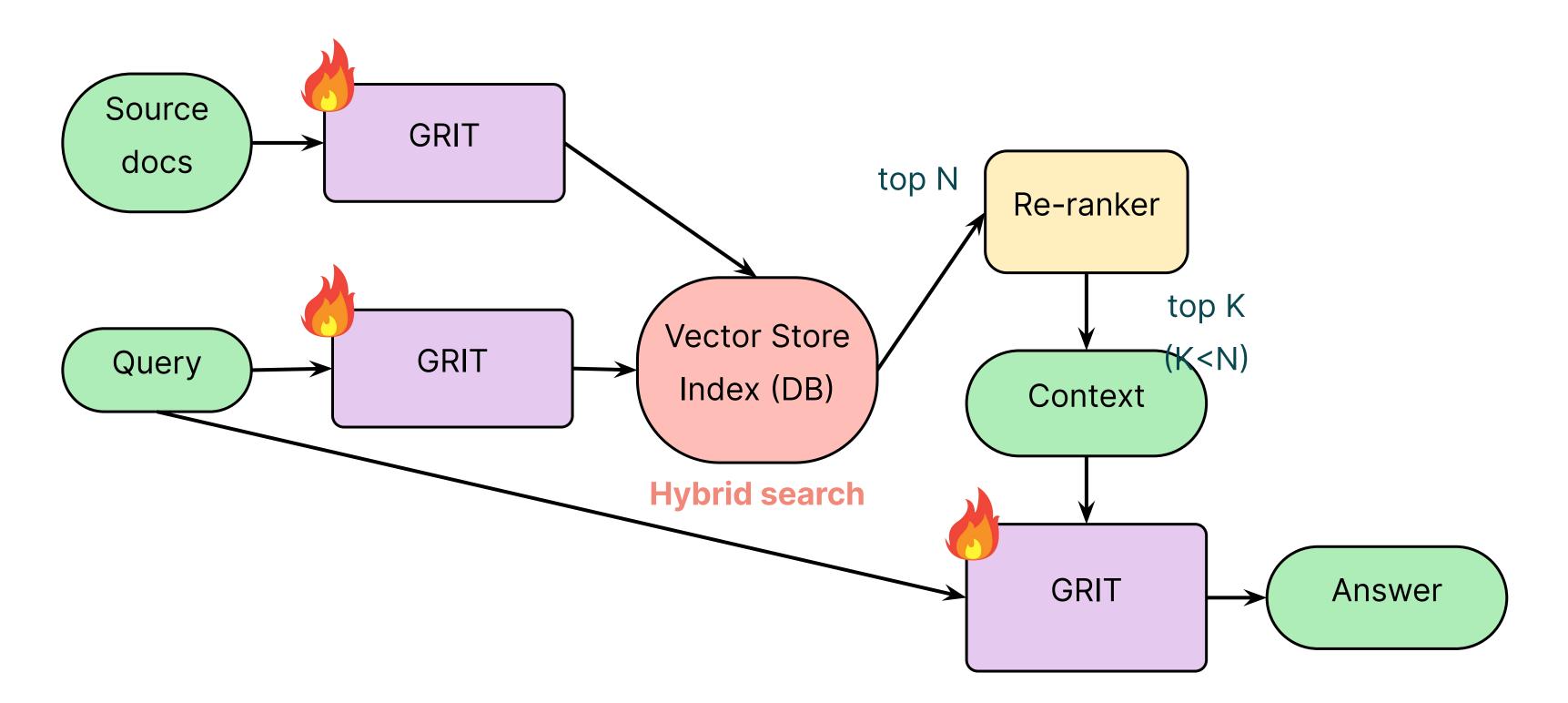
# Non specialised embeddings (for medical retrieval) **Medical Retrieval Embedding Models**

### Results

- Biomedical Information Retrieval (BEIR benchmark)
  - SotA on 3/5 biomedical tasks
  - Improves its initialization PubMedBERT by huge margins Ο
- Biomedical article representations (RELISH article similarity benchmark)
  - MedCPT article encoder (DEnc) outperforms all other models Ο
  - MedCPT article encoder improves PubMedBERT initialiation by over 10% Ο
- 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 Ο

### <u>Other</u>

# **Generative Representational Instruction Tuning**



# **Unified text embedding & generation model: GRIT**

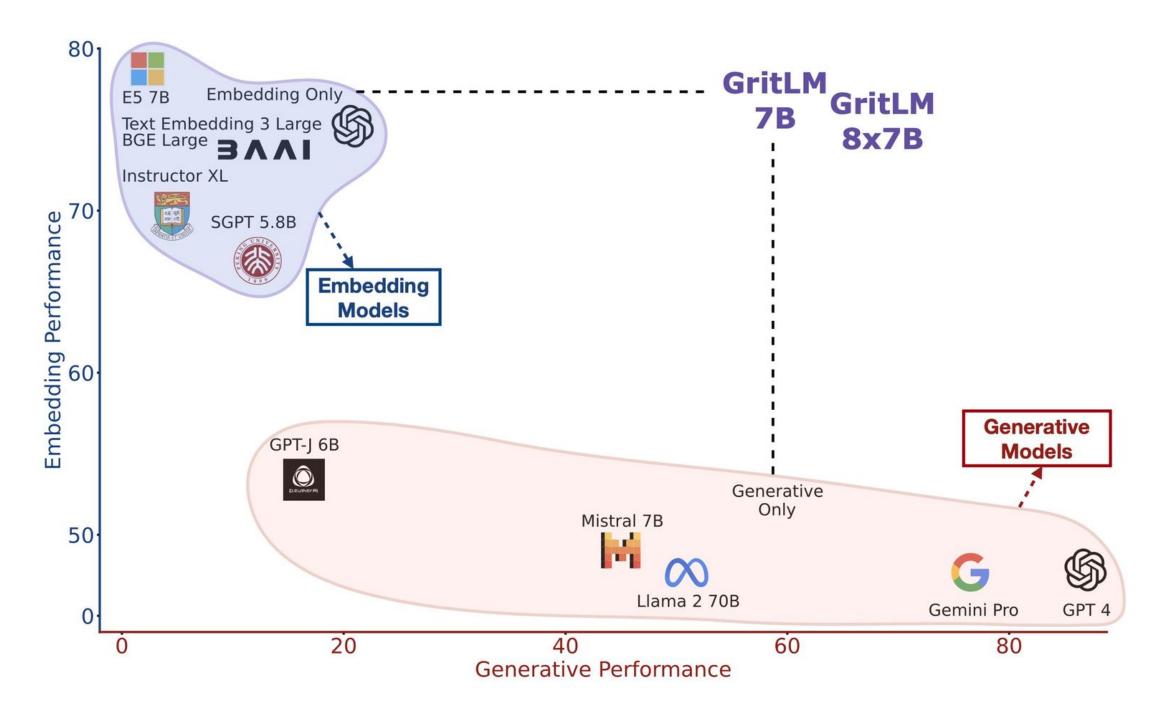


Figure 1: Performance of various models on text representation (embedding) and generation tasks. GRITLM is the first model to perform best-in-class at both types of tasks simultaneously.

https://arxiv.org/abs/2402.09906

https://twitter.com/Muennighoff/status/1758307967802224770

# **Unified text embedding & generation model: GRIT**

### Representation

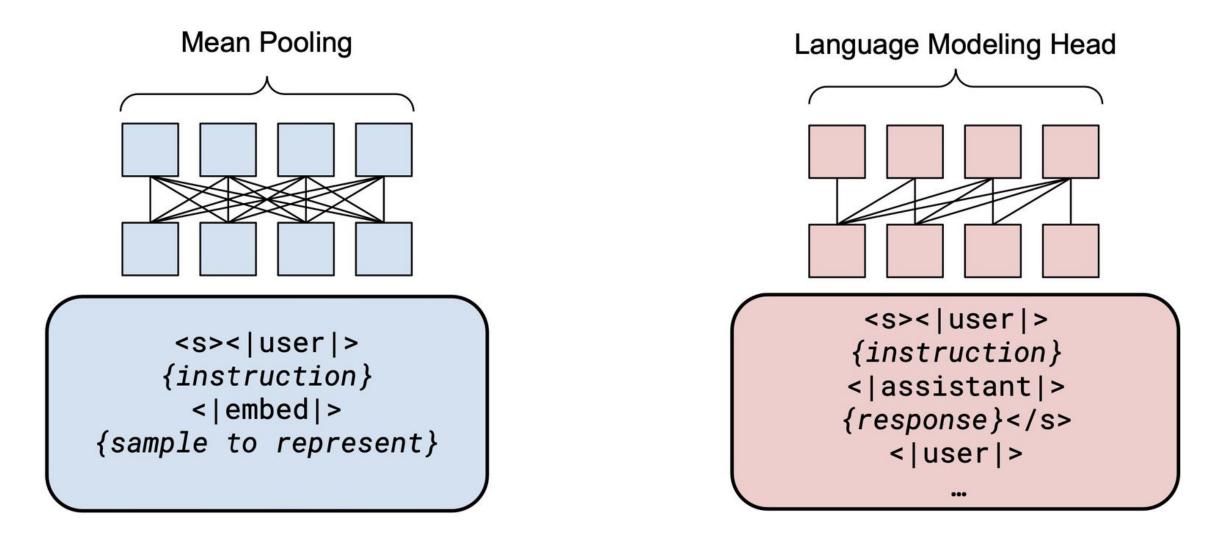


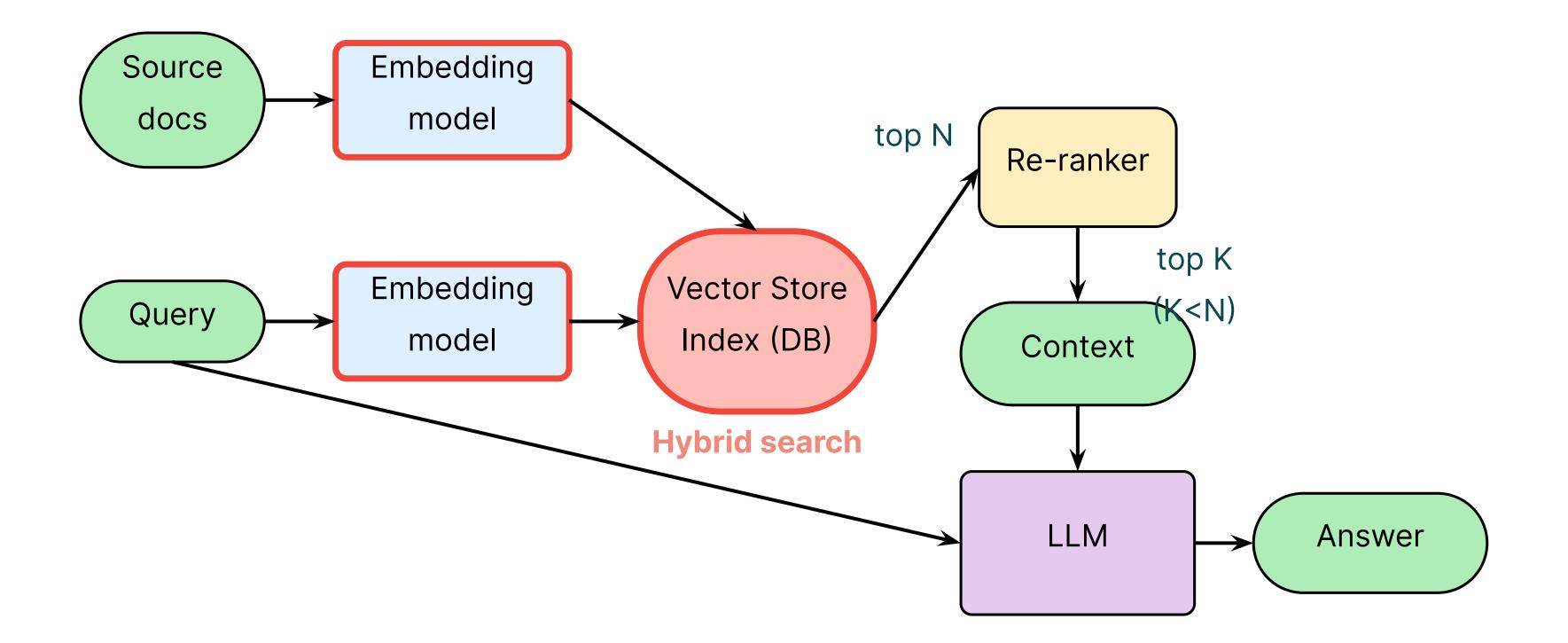
Figure 3: GRITLM architecture and format. Left: GRITLM uses bidirectional attention over the input for embedding tasks. Mean pooling is applied over the final hidden state to yield the final representation. Right: GRITLM uses causal attention over the input for generative tasks. A language modeling head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

### Generation

# **Unified text embedding & generation model: GRIT**

Use the same model as both embedder & reranker

Boosts perf on 15/16 Retrieval dsets

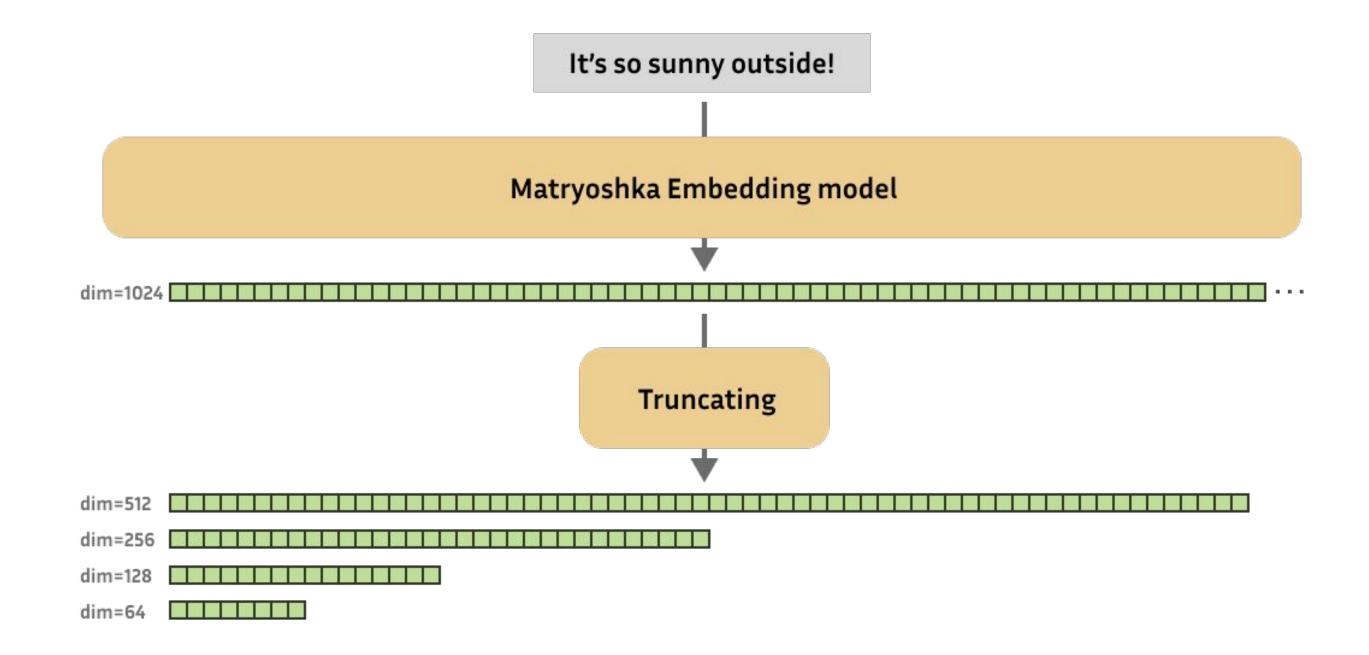




# Better embeddings $\rightarrow$ more dimensions $\rightarrow$ less efficient search



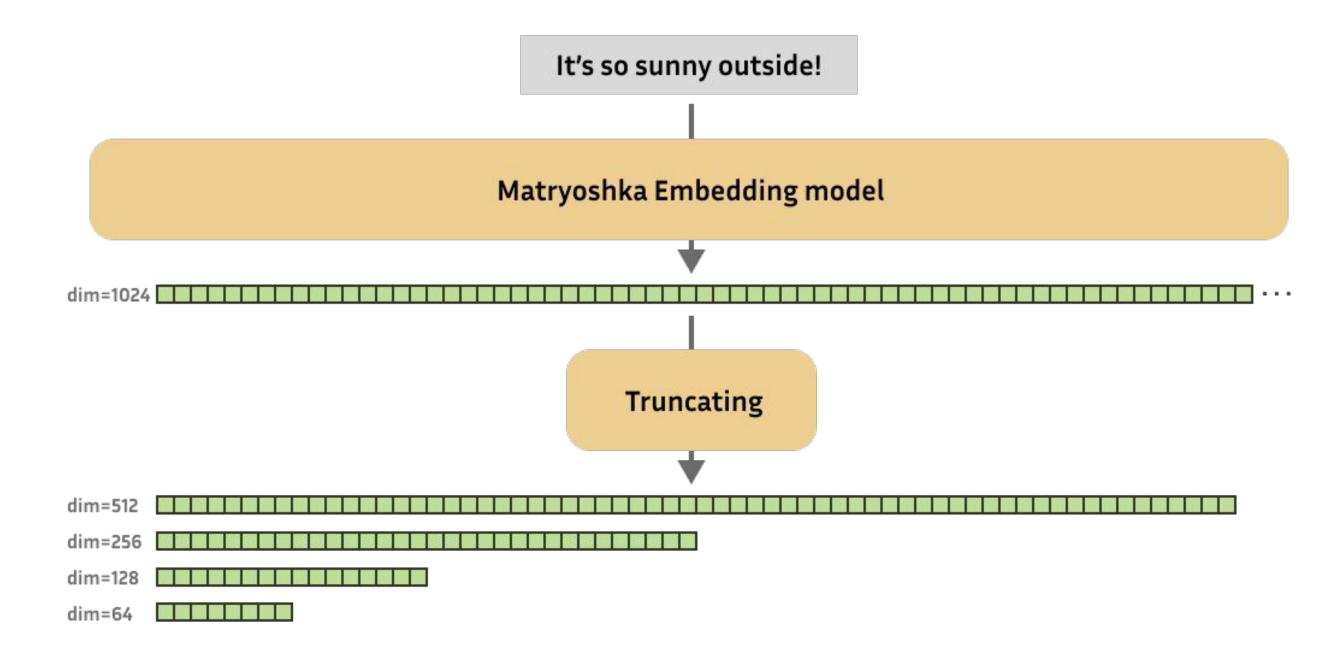
# Matryoshka Embedding Models



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



# 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**

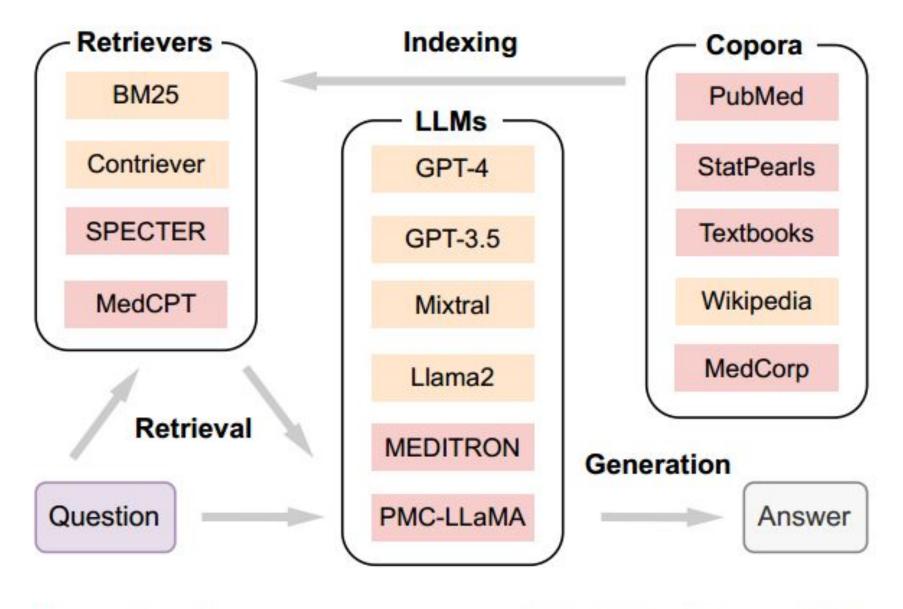


Figure 2: Component overview of the MEDRAG toolkit.

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

### dec 2023

### **Benchmarking Retrieval-Augmented Generation for Medicine**

LLM	Method	MIRAGE Benchmark Dataset					83
		MMLU-Med	MedQA-US	MedMCQA	PubMedQA*	BioASQ-Y/N	Avg.
<b>GPT-4</b> (-32k-0613)	CoT MEDRAG	$\begin{array}{r} \textbf{89.44} \pm 0.93 \\ \textbf{87.24} \pm 1.01 \end{array}$	$\frac{83.97 \pm 1.03}{82.80 \pm 1.06}$	$\begin{array}{c} 69.88 \pm 0.71 \\ 66.65 \pm 0.73 \end{array}$	$\begin{array}{c} \textbf{39.60} \pm \textbf{2.19} \\ \textbf{70.60} \pm \textbf{2.04} \end{array}$	$\begin{array}{r} 84.30 \pm 1.46 \\ 92.56 \pm 1.06 \end{array}$	73.44 79.97
<b>GPT-3.5</b> (-16k-0613)	CoT MEDRAG	$\begin{array}{c} \textbf{72.91} \pm 1.35 \\ \textbf{75.48} \pm 1.30 \end{array}$	$\frac{65.04 \pm 1.34}{66.61 \pm 1.32}$	$\begin{array}{c} \textbf{55.25} \pm 0.77 \\ \textbf{58.04} \pm 0.76 \end{array}$	$\begin{array}{c} \textbf{36.00} \pm \textbf{2.15} \\ \textbf{67.40} \pm \textbf{2.10} \end{array}$	$\begin{array}{c} 74.27 \pm 1.76 \\ 90.29 \pm 1.19 \end{array}$	60.69 71.57
Mixtral (8×7B)	CoT MEDRAG	$\begin{array}{c} \textbf{74.01} \pm 1.33 \\ \textbf{75.85} \pm 1.30 \end{array}$	$\begin{array}{c} \textbf{64.10} \pm 1.34 \\ \textbf{60.02} \pm 1.37 \end{array}$	$\frac{56.28 \pm 0.77}{56.42 \pm 0.77}$	$\begin{array}{r} \textbf{35.20} \pm \textbf{2.14} \\ \textbf{67.60} \pm \textbf{2.09} \end{array}$	$\begin{array}{c} 77.51 \pm 1.68 \\ 87.54 \pm 1.33 \end{array}$	61.42 69.48
Llama2 (70B)	CoT MEDRAG	$\begin{array}{c} \textbf{57.39} \pm 1.50 \\ \textbf{54.55} \pm 1.51 \end{array}$	$\begin{array}{c} \textbf{47.84} \pm 1.40 \\ \textbf{44.93} \pm 1.39 \end{array}$	$\frac{\textbf{42.60} \pm 0.76}{\textbf{43.08} \pm 0.77}$	$\begin{array}{c} 42.20 \pm 2.21 \\ 50.40 \pm 2.24 \end{array}$	$\begin{array}{c} 61.17 \pm 1.96 \\ 73.95 \pm 1.77 \end{array}$	50.24 53.38
MEDITRON (70B)	CoT MEDRAG	$\begin{array}{c} \textbf{64.92} \pm 1.45 \\ \textbf{65.38} \pm 1.44 \end{array}$	$\frac{51.69 \pm 1.40}{49.57 \pm 1.40}$	$\frac{46.74 \pm 0.77}{52.67 \pm 0.77}$	$\begin{array}{c} 53.40 \pm 2.23 \\ 56.40 \pm 2.22 \end{array}$	$\begin{array}{c} 68.45 \pm 1.87 \\ \textbf{76.86} \pm 1.70 \end{array}$	57.04 60.18
PMC-LLaMA (13B)	CoT MEDRAG	$\frac{52.16 \pm 1.51}{52.53 \pm 1.51}$	$\begin{array}{c} 44.38 \pm 1.39 \\ 42.58 \pm 1.39 \end{array}$	$\begin{array}{r} \textbf{46.55} \pm 0.77 \\ \textbf{48.29} \pm 0.77 \end{array}$	$\begin{array}{c} 55.80 \pm 2.22 \\ 56.00 \pm 2.22 \end{array}$	$\begin{array}{c} 63.11 \pm 1.94 \\ 65.21 \pm 1.92 \end{array}$	52.40 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

a	S	e	t	
-	~	~	-	

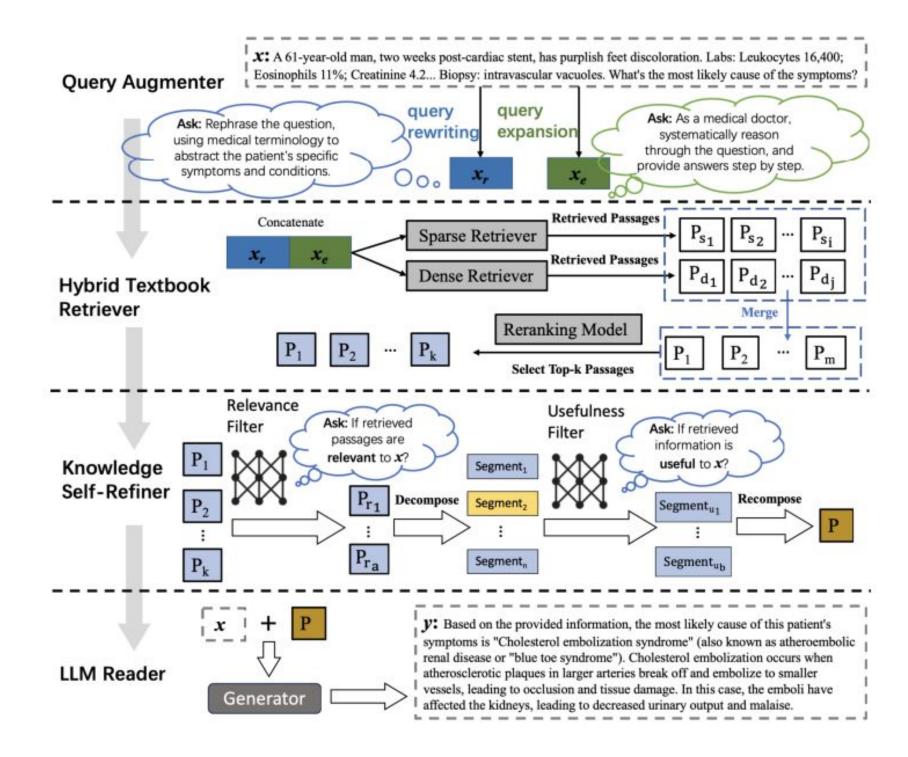
### **Benchmarking Retrieval-Augmented Generation for Medicine**

Corpus	Retriever	MIRAGE Benchmark Dataset					
		MMLU-Med	MedQA-US	MedMCQA	PubMedQA*	BioASQ-Y/N	Average
None	None	72.91 ± 1.35	$\textbf{65.04} \pm 1.34$	$55.25 \pm 0.77$	$\textbf{36.00} \pm 2.15$	$\textbf{74.27} \pm 1.76$	60.69
	BM25	$72.27 \pm 1.36$	63.71 + 1.85	$\textbf{55.49} \pm 0.77$	$66.20 \pm 2.12$	$88.51 \pm 1.28$	69.23
PubMed (23.9M)	Contriever	71.72 ± 1.3%	$63.94 \pm 1.35$	$54.29 \pm 0.77$	$65.60 \pm 2.12$	$\textbf{85.44} \pm 1.42$	68.20
	SPECTER	$\textbf{73.19} \pm 1.34$	$65.20 \pm 1.34$	53.12 ± 4.77	$54.80 \pm 2.23$	75.73 ± 1.72	64.41
	MedCPT	$\textbf{73.09} \pm 1.34$	$66.69 \pm 1.32$	$54.94 \pm 0.77$	$66.40 \pm 2.11$	$85.76 \pm 1.41$	69.38
	RRF-2	$75.57 \pm 1.30$	$64.34 \pm 1.34$	$55.34 \pm 0.77$	$69.00 \pm 2.07$	$87.06 \pm 1.35$	70.26
	RRF-4	73.37 ± 1.34	$64.73 \pm 1.34$	$54.75 \pm 0.77$	$67.20 \pm 2.10$	$88.51 \pm 1.28$	69.71
	BM25	71.63 ± 1.37	65.67 ± 1.33	$54.89 \pm 0.77$	$27.60 \pm 2.00$	60.36 ± 1.97	56.03
	Contriever	$\textbf{73.28} \pm 1.34$	$67.48 \pm 1.31$	$54.24 \pm 0.77$	$\textbf{28.80} \pm \textbf{2.03}$	58.41 ± 1.48	56.44
StatPearls	SPECTER	73.74 ± 1.33	$64.73 \pm 1.34$	52.83 ± 0.77	23.20 ± 1.89	57.77 ± 1.09	54.45
(301.2k)	MedCPT	$72.82 \pm 1.35$	$64.89 \pm 1.34$	$54.17 \pm 0.77$	$\textbf{27.60} \pm 2.00$	$60.68 \pm 1.96$	56.03
	RRF-2	$72.64 \pm 1.35$	$65.67 \pm 1.33$	$54.63 \pm 0.77$	$30.00 \pm 2.05$	$61.17 \pm 1.96$	56.82
	RRF-4	$\textbf{73.83} \pm 1.33$	$65.12 \pm 1.34$	$53.81 \pm 0.77$	$\textbf{30.60} \pm 2.06$	59.71 ± 1.97	56.61
	BM25	$74.66 \pm 1.32$	$66.54 \pm 1.32$	$54.05 \pm 0.77$	$\textbf{30.20} \pm 2.05$	60.03 ± 1.97	57.10
	Contriever	$\textbf{74.10} \pm 1.33$	$67.16 \pm 1.32$	$54.53 \pm 0.77$	$26.60 \pm 1.98$	$60.19 \pm 1.97$	56.52
Textbooks	SPECTER	$72.82 \pm 1.35$	$67.40 \pm 1.31$	$53.29 \pm 0.77$	$25.60 \pm 1.95$	<b>55.50</b> ± 2.00	54.92
(125.8k)	MedCPT	$\textbf{74.93} \pm \textbf{1.31}$	$66.22 \pm 1.33$	$54.41 \pm 0.77$	$29.20 \pm 2.03$	$61.33 \pm 1.96$	57.22
	RRF-2	$76.68 \pm 1.28$	$65.91 \pm 1.33$	$54.79 \pm 0.77$	$31.00 \pm 2.07$	<b>59.39</b> ± 1.98	57.55
	RRF-4	$75.76 \pm 1.30$	<b>66.06</b> ± 1.33	<b>55.56</b> ± 0.77	$\textbf{30.40} \pm 2.06$	$60.68 \pm 1.96$	57.69
	BM25	<b>73.37</b> ± 1.34	63.47 ± 1.35	$54.10 \pm 0.77$	$26.40 \pm 1.97$	$\textbf{71.36} \pm 1.82$	57.74
	Contriever	$74.10 \pm 1.33$	$65.99 \pm 1.33$	$54.03 \pm 0.77$	$26.40 \pm 1.97$	$69.90 \pm 1.85$	58.08
Wikipedia	SPECTER	$72.18 \pm 1.36$	63.63 ± 1.88	$52.71 \pm 0.77$	$22.20 \pm 1.86$	$66.83 \pm 1.89$	55.51
(29.9M)	MedCPT	$71.99 \pm 1.36$	$65.12 \pm 1.34$	$55.15 \pm 0.77$	$29.00 \pm 2.03$	$73.46 \pm 1.78$	58.95
	RRF-2	$\textbf{74.20} \pm 1.33$	$64.57 \pm 1.34$	$54.72 \pm 0.77$	$\textbf{31.00} \pm 2.07$	$76.21 \pm 1.71$	60.14
	RRF-4	$\textbf{73.19} \pm 1.34$	$64.96 \pm 1.34$	$54.53 \pm 0.77$	$\textbf{31.00} \pm 2.07$	$72.01 \pm 1.81$	59.14
MedCorp (65.3M)	BM25	73.65 ± 1.34	65.91 ± 1.33	$56.78 \pm 0.77$	$66.20 \pm 2.12$	$87.70 \pm 1.32$	70.05
	Contriever	$\textbf{75.48} \pm 1.30$	$64.10 \pm 1.34$	$56.11 \pm 0.77$	$\textbf{62.40} \pm 2.17$	$84.95 \pm 1.44$	68.61
	SPECTER	$\textbf{74.38} \pm 1.32$	$65.44 \pm 1.33$	$54.41 \pm 0.77$	$\textbf{55.80} \pm \textbf{2.22}$	$\textbf{73.14} \pm 1.78$	64.63
	MedCPT	$74.75 \pm 1.32$	$67.40 \pm 1.31$	$\textbf{55.85} \pm 0.77$	$66.40 \pm 2.11$	$85.92 \pm 1.40$	70.06
	RRF-2	$73.74 \pm 1.33$	$67.24 \pm 1.32$	$\textbf{56.08} \pm 0.77$	$67.80 \pm 2.09$	$88.19 \pm 1.30$	70.61
	RRF-4	$\textbf{75.48} \pm 1.30$	$66.61 \pm 1.32$	$58.04 \pm 0.76$	$67.40 \pm 2.10$	$90.29 \pm 1.19$	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 BM25 and MedCPT)

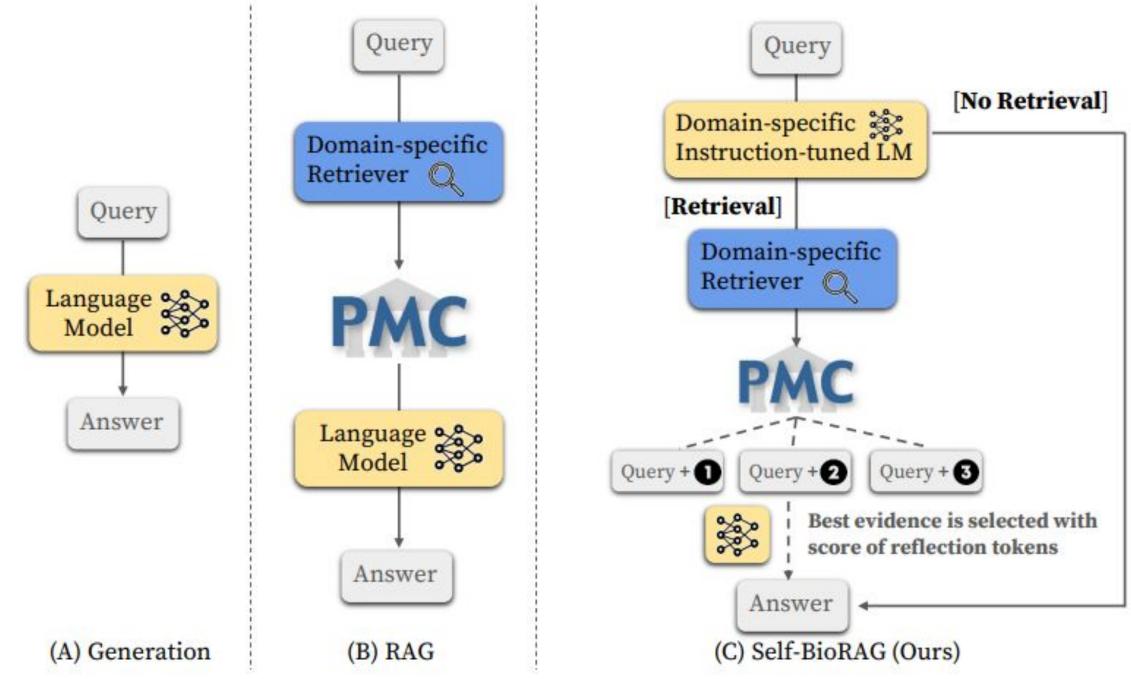
# Augmenting Black-box LLMs with Medical Textbooks for Clinical **Question Answering**



https://arxiv.org/abs/2309.02233

### feb 2024

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



https://arxiv.org/abs/2401.15269

### jan 2024

## Almanac—Retrieval-Augmented Language Models for Clinical Medicine

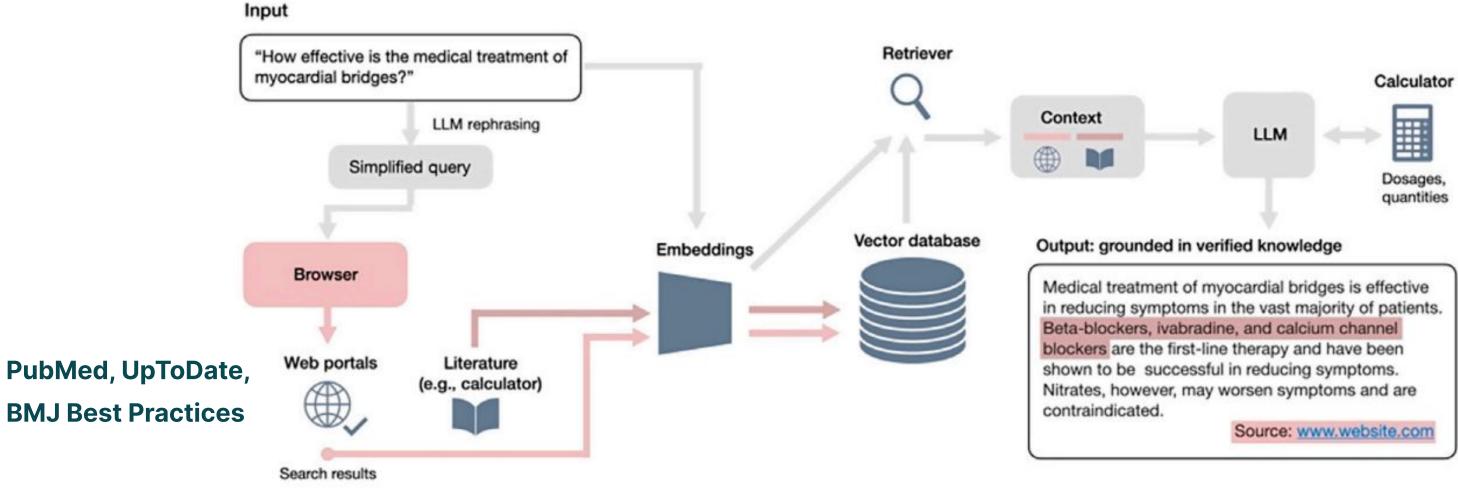


Figure 1. Almanac Overview.

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/Aloa230006

### aug 2023

# Almanac—Retrieval-Augmented Language Models for Clinical Medicine

### Mean Arterial Pressure (MAP)

Calculates mean arterial pressure.

A	Pearls/Pitfalls 🗸	Why Use 🗸
<ul> <li>values are obtained.</li> <li>Blood pressure targets has These include sepsis, trau</li> </ul>	e can be calculated in all patients ir ve been shown to improve outcome ma, stroke, intracranial bleed, and h e either SBP or MAP as a blood pres	in a number of conditions. hypertensive emergencies.
stolic BP	100	mm Hg
astolic BP	90	mm Hg

# Medical Calculators

(MedCalc)

### **CURB-65** Algorithm

Criteria \*\*C\*\*onfusion \*\*U\*\*rea >20 mg \*\*R\*\*espiratory per minute Low systolic (< diastolic (≤60 mn pressure Age ≥\*\*65\*\* year Total: CURB-65 score

### **Severity Reference**

Points 0 to 1 point 2 points

3 to 5 points

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

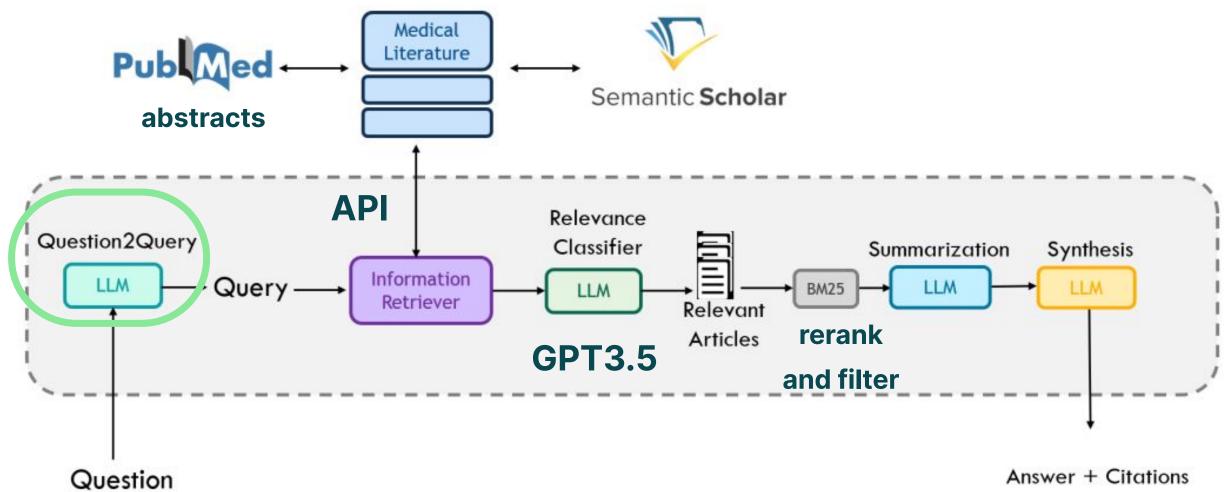
	Points
	1 point
g/dL (7 mmol/L)¶	1 point
rate $\geq 30$ breaths	1 point
<90 mmHg) or mHg) **B**lood	1 point
rs	1 point
score	

Severity Low severity (risk of death <3%) Moderate severity (risk of death 9%) High severity (risk of death 15 to 40%)

### MedCalc in Almanac

(embedded in Qdrant)

# 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

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

### Question2Query

Retrieved 16 articles. Identifying the relevant ones and summarizing them (this may take a minute).

### **Queries Used:**

(high-grade dysplasia OR carcinoma in situ) AND biliary duct margin AND extrahepatic cholangiocarcinoma AND prognosis

(high-grade dysplasia OR carcinoma in situ) AND (biliary duct margin) AND (extrahepatic cholangiocarcinoma) AND prognosis

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

Does high-grade dysplasia/carcinoma in situ of the biliary duct margin affect the prognosis of extrahepatic cholangiocarcinoma?

∆ Button 🛛 🖓 Button

**Literature Summary:** The impact of high-grade dysplasia/carcinoma in situ (HGD/CIS) at the biliary duct margin on the prognosis of extrahepatic cholangiocarcinoma is a subject of debate. Studies [1,2,10,13] have suggested that HGD/CIS at the biliary duct margin can have an impact on survival, with some studies indicating a poorer prognosis [1,11,13], while others suggest comparable outcomes to negative margins [2,10]. Furthermore, a meta-analysis [8] found no significant impact of HGD/CIS on overall survival rates compared to negative margins, but a higher incidence of local recurrence. Yet, other studies [9,12,14] found no significant impact of HGD/CIS on survival rates. Importantly, the presence of invasive carcinoma at the ductal resection margins has been identified as a strong independent prognostic factor [2,10]. Additional resection has been suggested to improve outcomes in some cases [1,4], although one study found worse survival rates even after achieving negative margins through additional resection [14].

**TL;DR:** The evidence is mixed, but it suggests that high-grade dysplasia/carcinoma in situ at the biliary duct margin may impact the prognosis of extrahepatic cholangiocarcinoma, potentially leading to a poorer prognosis or higher incidence of local recurrence. Invasive carcinoma at the margins is a stronger negative prognostic factor. Additional resection may improve outcomes in some cases.

### **References:**

[1] Radtke A, Konigsrainer A (2016) Surgical therapy of cholangiocarcinoma. Visc Med 32:422-426

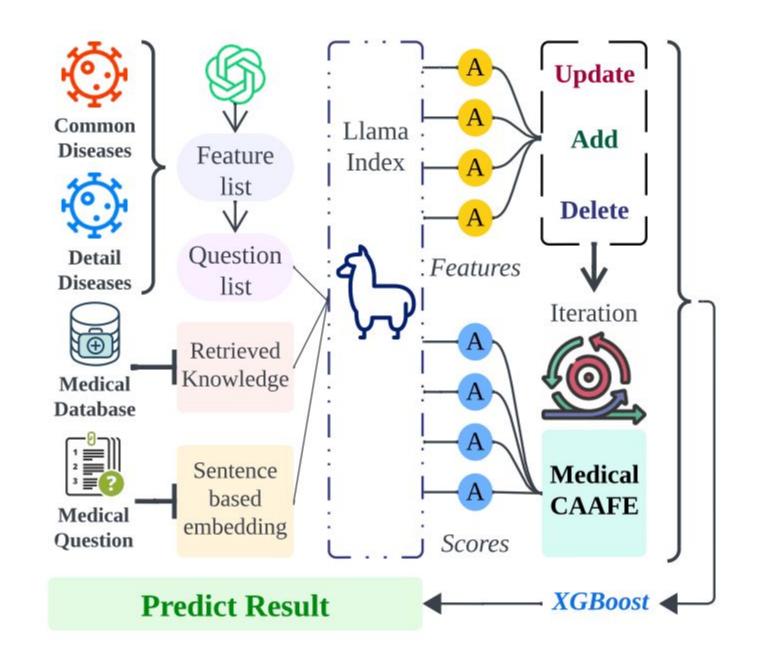
[2] Nagorney DM, Donohue JH, Farnell MB, et al. (1993) Outcomes after curative resections of cholangiocarcinoma. Arch Surg 128:871-879

[3] Noji T, Okamura K, Tanaka K, Nakanishi Y, Asano T, Nakamura T, Tsuchikawa T, Hirano S. Surgical technique and results of intrapancreatic bile duct resection for hilar malignancy (with video).. HPB : the official journal of the International Hepato Pancreato Biliary Association. 2018;20(12):1145-1149.

[4] Otsuka S, Ebata T, Yokoyama Y, Mizuno T, Tsukahara T, Shimoyama Y, Ando M, Nagino M. Clinical value of additional resection of a margin-

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

# **Health-LLM: Personalized Retrieval-Augmented Disease Prediction** Model

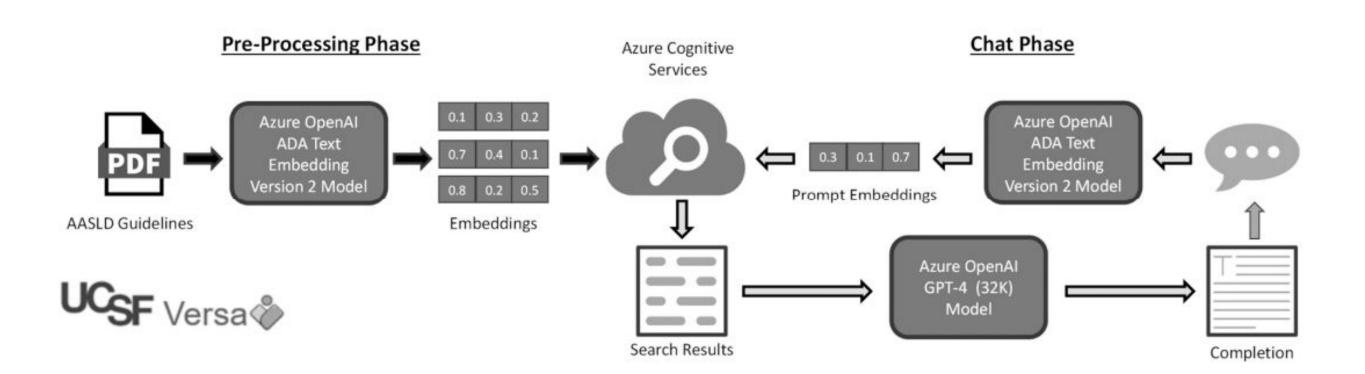


https://arxiv.org/abs/2402.00746

### feb 2024



# **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

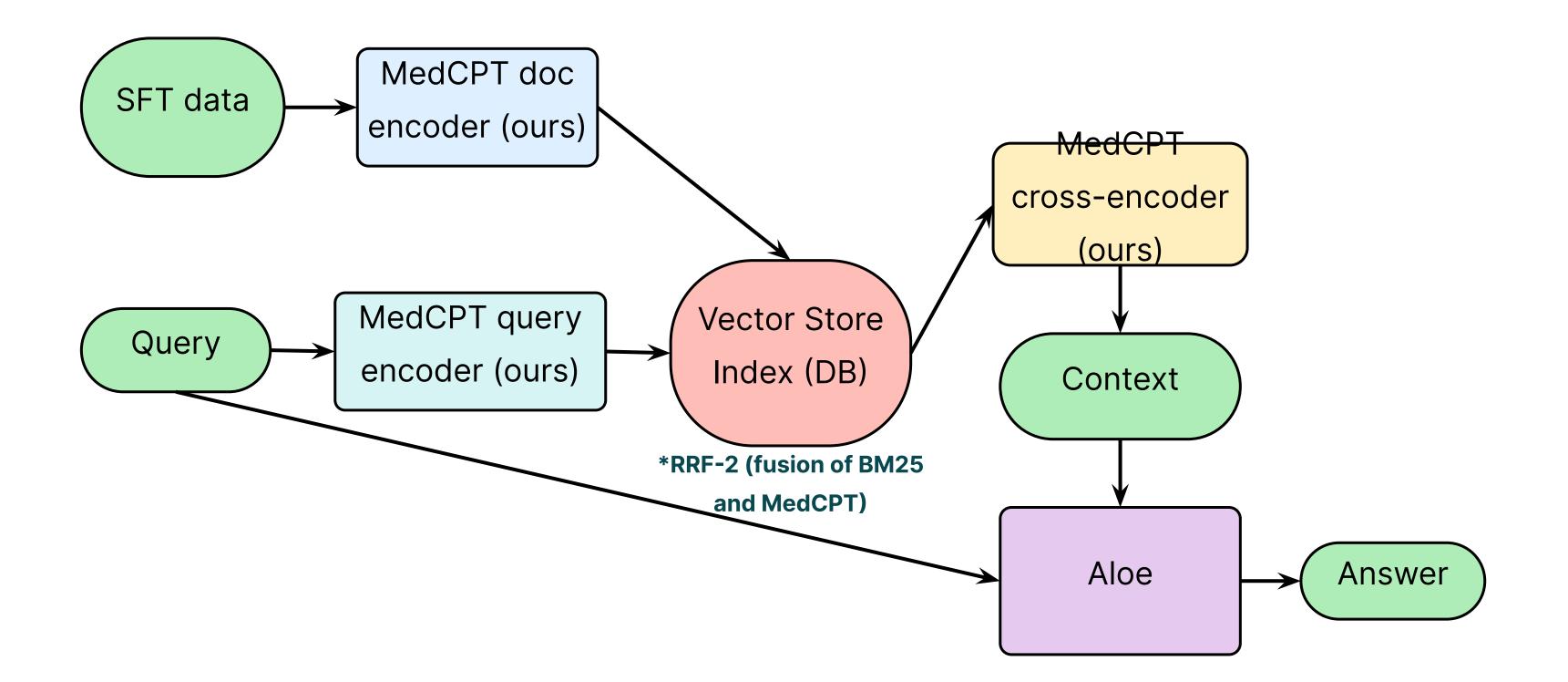
### nov 2023

Reca

р

- 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

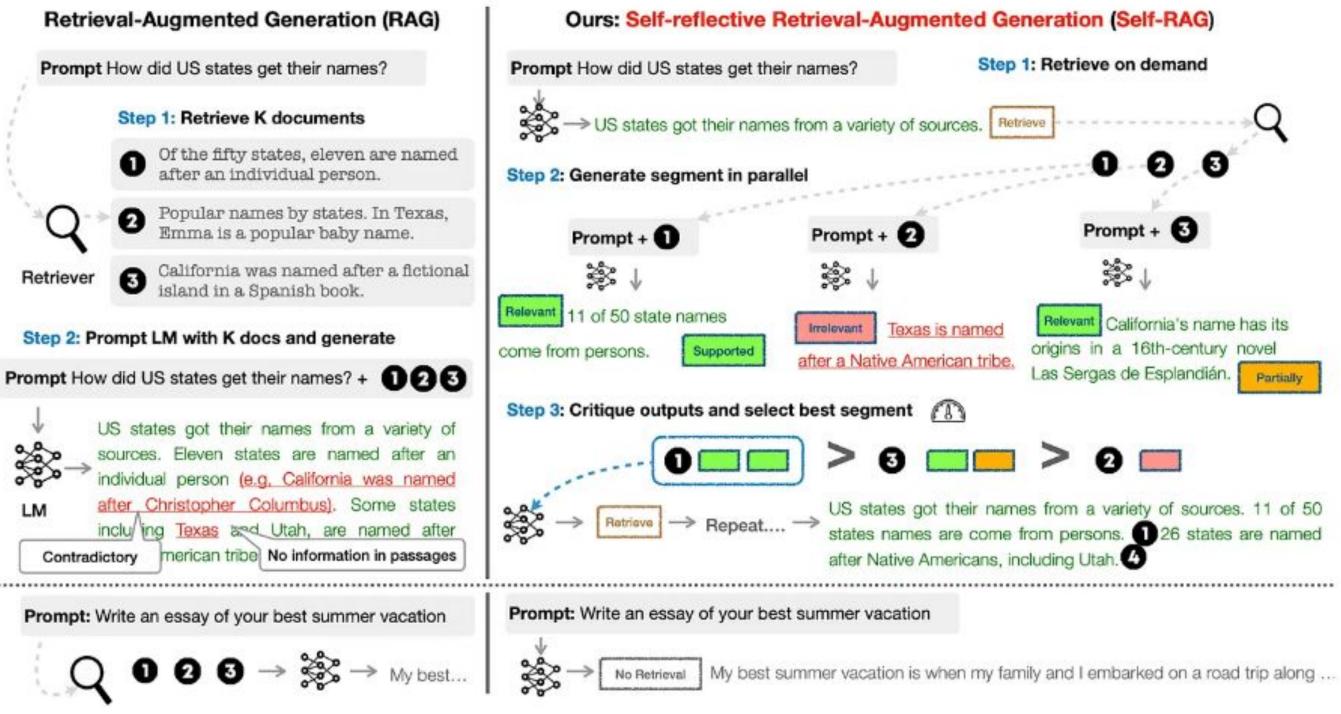
### **Proposal**



Extra

stuff

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



### https://arxiv.org/abs/2310.11511

# Evaluatio

n

