# How to Enhance LLM with Retrieval-Augmented Generation (RAG)

Cholatid Ratanatharathorn, MD
PhD student, Data Science for Healthcare and Clinical Informatics
Department of Clinical Epidemiology and Biostatistics

# LLM Hallucination Causes

- **Data-Related Causes:**

**Flawed Data Sources** (1) **:** Poor-quality data sources can introduce misinformation and biases. This includes imitative falsehoods, duplication biases, and social biases.

**Knowledge Boundaries** (2) **:** LLMs might lack specific domain knowledge or up-to-date information, leading to incorrect or outdated responses.

**Inferior Data Utilization** (3) **:** reliance on spurious correlations or difficulties in complex knowledge.

(1) Lin et al. (2022); Lee et al. (2022a); Bender et al. (2021)
(2) Singhal et al. (2023); Katz et al. (2023); Onoe et al. (2022)
(3) Mallen et al. (2023); Zheng et al. (2023); Liu et al. (2023e)

# LLM Hallucination Causes

- **Training-Related Causes:**

**Pre-Training Issues:** During pre-training, the model might face architectural flaws. (1)

- **Inference-Related Causes:**

**Decoding Strategies:** The randomness inherent in decoding strategies, such as sampling, can introduce errors. Higher temperatures in sampling can lead to increased hallucinations. (2)

**Decoding Representation:** The top-layer representation used for predicting the next token might have limitations. Insufficient context attention may be a cause of irrelevant answers.

(1) Lewis PSH, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33: virtual.(2) Singhal et al. (2023); Katz et al. (2023); Onoe et al. (2022)
(3) Dziri N, Madotto A, Zaïane O, Bose AJ. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. Proc 2021 Conf Empir Methods Nat Lang Process. 2021;2197-2214. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

# How to improve it?

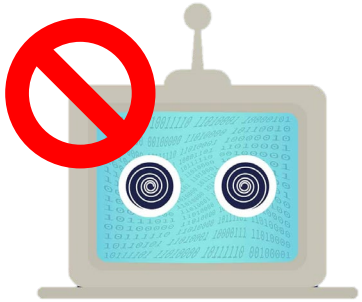
Individual or organization-level

- Better prompt
- Finetune
- RAG

# What is RAG?

- Retrieval – retrieve the data from (vector) database
- Augmentation – use the retrieved data to augment prompt for LLM
- Generation – LLM generate the output/answer

# Benefit of RAG



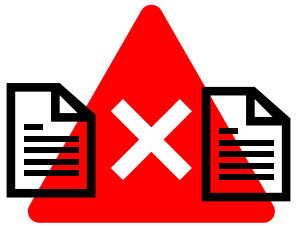
Less hallucination

Private document

Fine-tune is no need

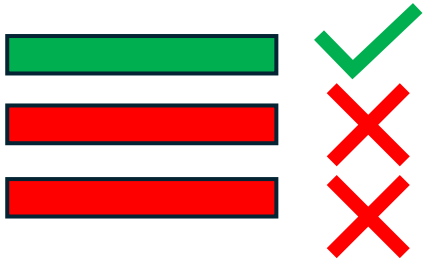
Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023 Dec 18.

# Drawdown of (naïve) RAG



Irrelevant/missing information retrieval

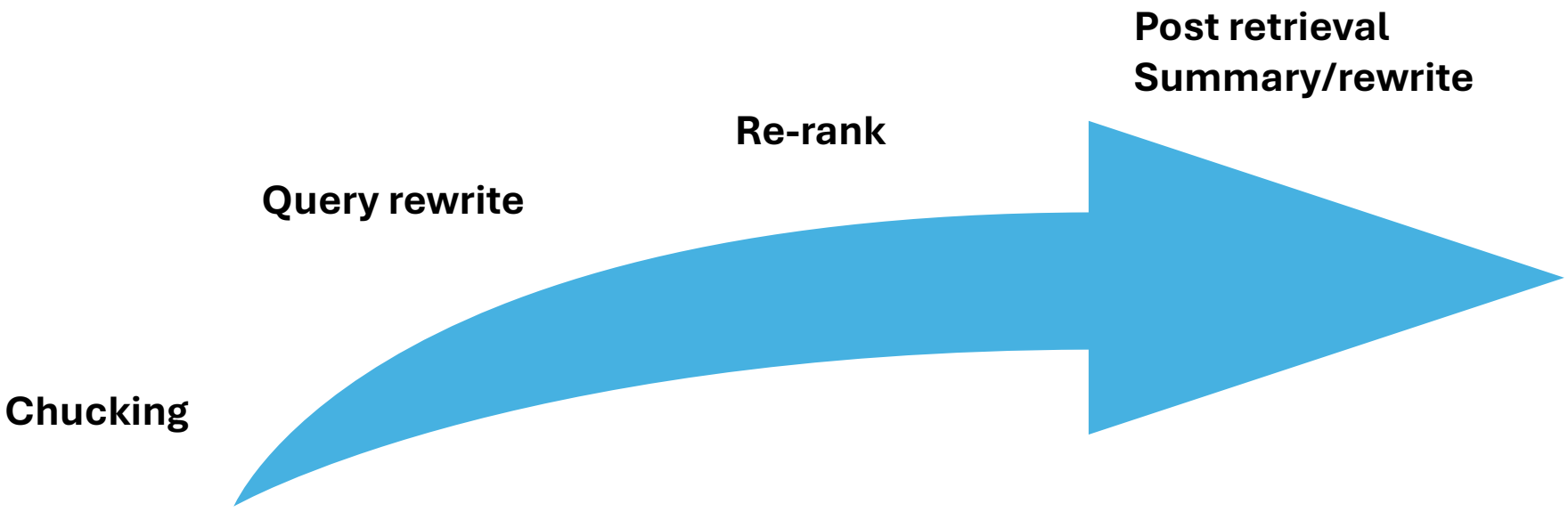
Hallucination



Single task/question

Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023 Dec 18.

# The next step of RAG




Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, Dai Y, Sun J, Wang H. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023 Dec 18.

# Example of use case

Question: How can we improve SNOMED CT coding with LLM and advance RAG?

# What is SNOMED CT?

Delivering SNOMED CT

(Systematized Nomenclature of Medicine Clinical Terms)

- **Comprehensive Coverage**: More than 300,000+ concepts which cover a wide range of clinical concepts including diseases, procedures, symptoms, and findings.
- **Interoperability**: Enhances data sharing and interoperability between different healthcare systems and providers.
- **Standardization**: Provides a standardized language for clinical terms, reducing ambiguity and improving communication.

For research, statistics, and reimbursement

https://www.snomed.org/

# How to code it nowadays?

- A coder/physician has to select a code manually.
- Some SNOMED CT code was mapped to other standards e.g. ICD10, ICD9

**Can LLM, which understands some context, help this out?**

# Let’s try with LLM alone first

# Prompt crafting

Role assign

What to do

- You are a helpful medical coder assistant. Extract clinical terms from the following medical note if it's an observation, interpret and use the correct clinical term, but send only a list of clinical terms and SNOMED CT code. Do not send any thinking process.

Example RR 40/min -> Tachypnea
Sodium 130 mg/dl-> Hyponatremia

Few-shot

Input: case pt 40 yr male came to ED with diarrhea, vomit and bloody stool BT 40 C RR 24/min PR 120/min BP 89/50 mmHg
Abdomen - soft, not tender
IMP: infective diarrhea

INPUT

List of clinical terms and SNOMED CT code:

Desired OUTPUT

## GPT 4o

1. Diarrhea - 267036007
2. Vomiting - 422400008
3. Bloody stool - 86615007
4. Fever - 386661006
5. Tachypnea - 230145002
6. Tachycardia - 3424008
7. Hypotension - 45007003
8. Infective diarrhea - 310861008

**4/8**

## GPT 4o mini

1. Diarrhea - SNOMED CT: 11768100119100
2. Vomiting - SNOMED CT: 422587008
3. Bloody stool - SNOMED CT: 431855005
4. Fever - SNOMED CT: 386725007
5. Tachypnea - SNOMED CT: 267036007
6. Tachycardia - SNOMED CT: 10200004
7. Hypotension - SNOMED CT: 271327008
8. Infective diarrhea - SNOMED CT: 13270003

**1/8**

# Example of RAG method

Extract clinical term

RAG retrieval

RAG re-rank

LLM Generation

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

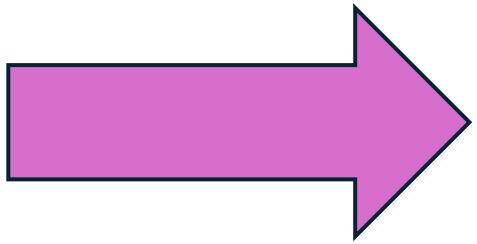
You are a helpful medical coder assistant. Extract clinical terms from the following medical note if it's an observation, interpret and use the correct clinical term, but send only a list of clinical terms and SNOMED CT code. Do not send any thinking process.
Example RR 40/min -> Tachypnea
Sodium 130 mg/dl-> Hyponatremia

Input: case pt 40 yr male came to ED with diarrhea, vomit and bloody stool BT 40 C RR 24/min PR 120/min BP 89/50 mmHg Abdomen - soft, not tender
IMP: infective diarrhea

INPUT

**Separate each clinical term by a new line.**
List of clinical terms:

**GPT 4o mini**



Diarrhea

Vomiting

Bloody stool

Fever

Tachypnea

Tachycardia

Hypotension

Infective diarrhea

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**SNOMED CT concept 400,000+**

| | |
|---|---|
| 386661006 | Fever (finding) |
| 3424008 | Tachycardia (finding) |
| 233604007 | Pneumonia |
| 38362002 | Dengue (disorder) |

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

# Database


**Bi-encoder**

**Compress <u>each book</u> into a vector**

**Embedding**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

# Vector database

**QUERY = the data that you want**

Bi-encoder

**An embedding represents query**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

# Vector database

**QUERY = the data that you want**

**Similar score = 0.222**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**Similar score = 0.929**

**Similar score = 0.829**

**Similar score = 0.729**

**Top K = 3**

**Similar score = 0.629**

**Similar score = 0.229**

**QUERY = the data that you want**

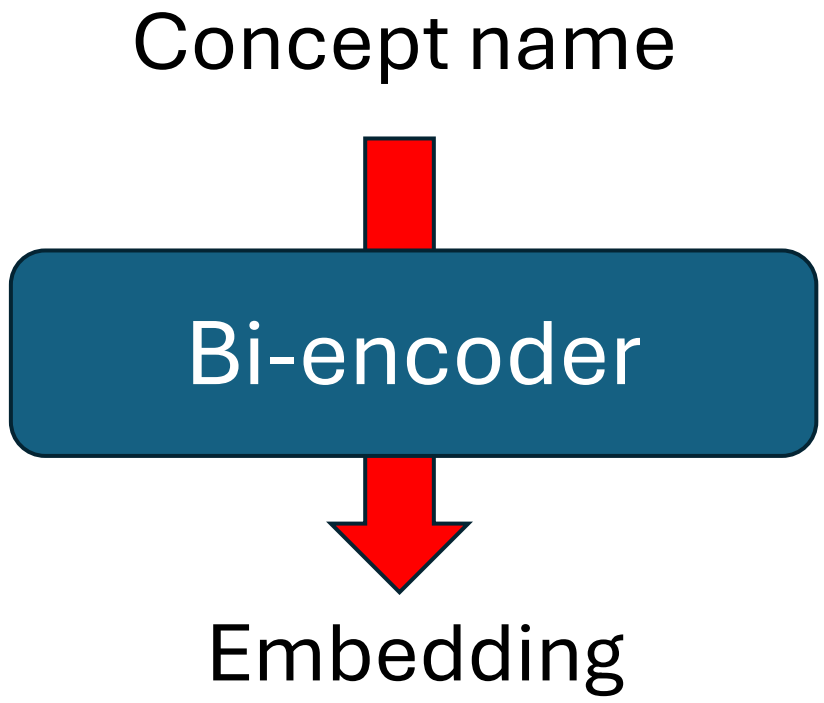
Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**QUERY = the data that you want**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**SNOMED CT concept 400,000+**

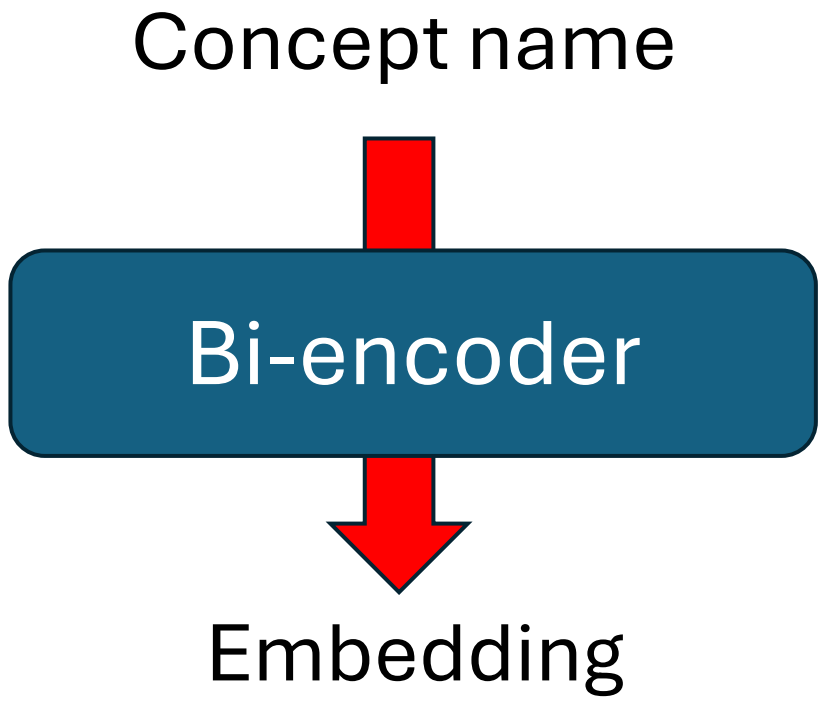
| | |
|---|---|
| 386661006 | Fever (finding) |
| 3424008 | Tachycardia (finding) |
| 233604007 | Pneumonia |
| 38362002 | Dengue (disorder) |



Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**SNOMED CT concept 400,000+**

| | |
|---|---|
| 386661006 | [0.223,0.366,0.55,…,0.14] |
| 3424008 | [0.233,0.766,0.45,…,0.67] |
| 233604007 | [0.523,0.366,0.55,…,0.69] |
| 38362002 | [0.723,0.666,0.55,…,0.86] |

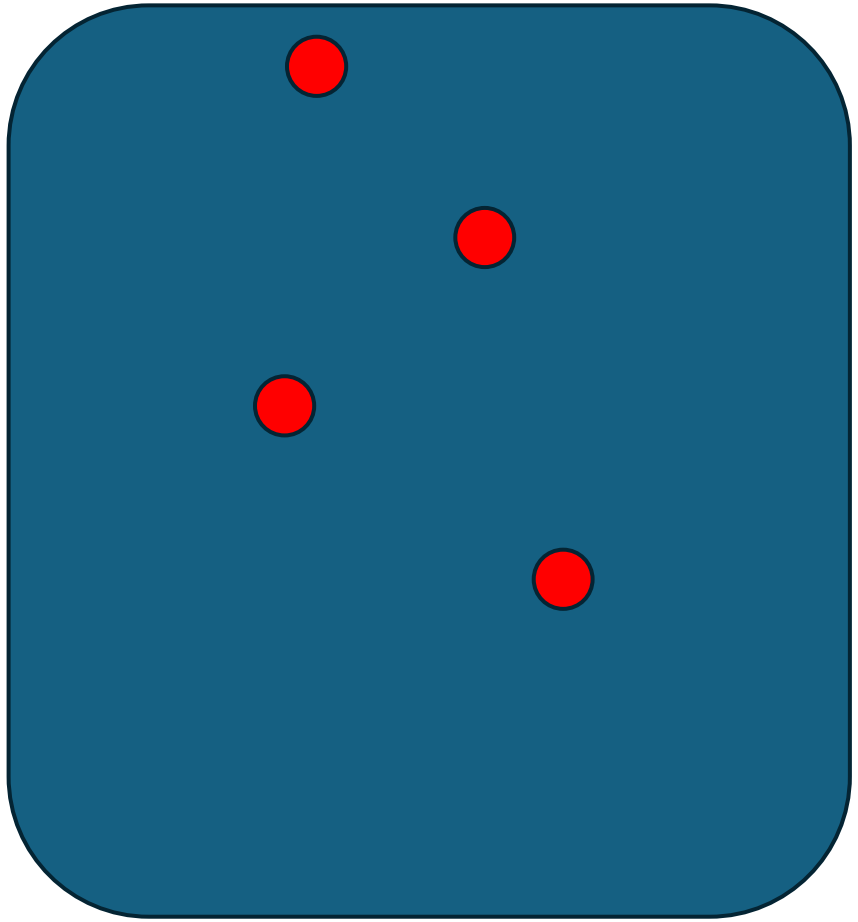


Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**SNOMED CT concept 400,000+**

| | |
|---|---|
| 386661006 | [0.223,0.366,0.55,…,0.14] |
| 3424008 | [0.233,0.766,0.45,…,0.67] |
| 233604007 | [0.523,0.366,0.55,…,0.69] |
| 38362002 | [0.723,0.666,0.55,…,0.86] |

**Vector database**



Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

Diarrhea

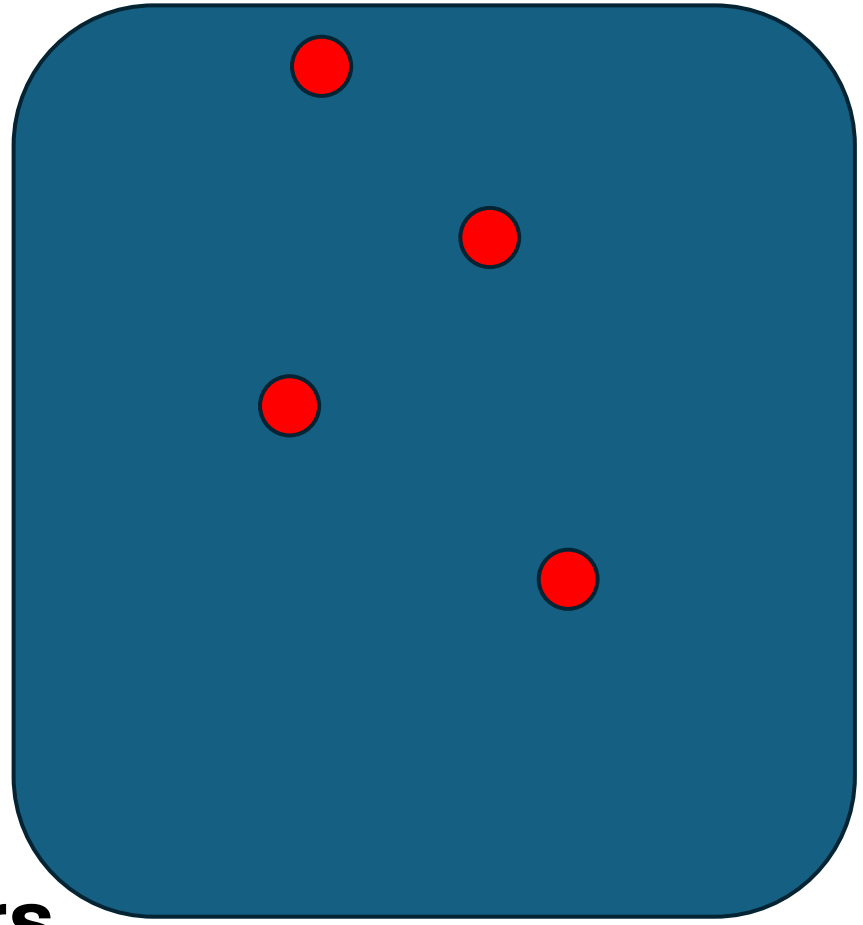
Query 1

Bi-encoder

[0.723,0.666,0.55,…,0.86]

Cosine similarity

**Vector database**



Select Top K = **25 vectors**

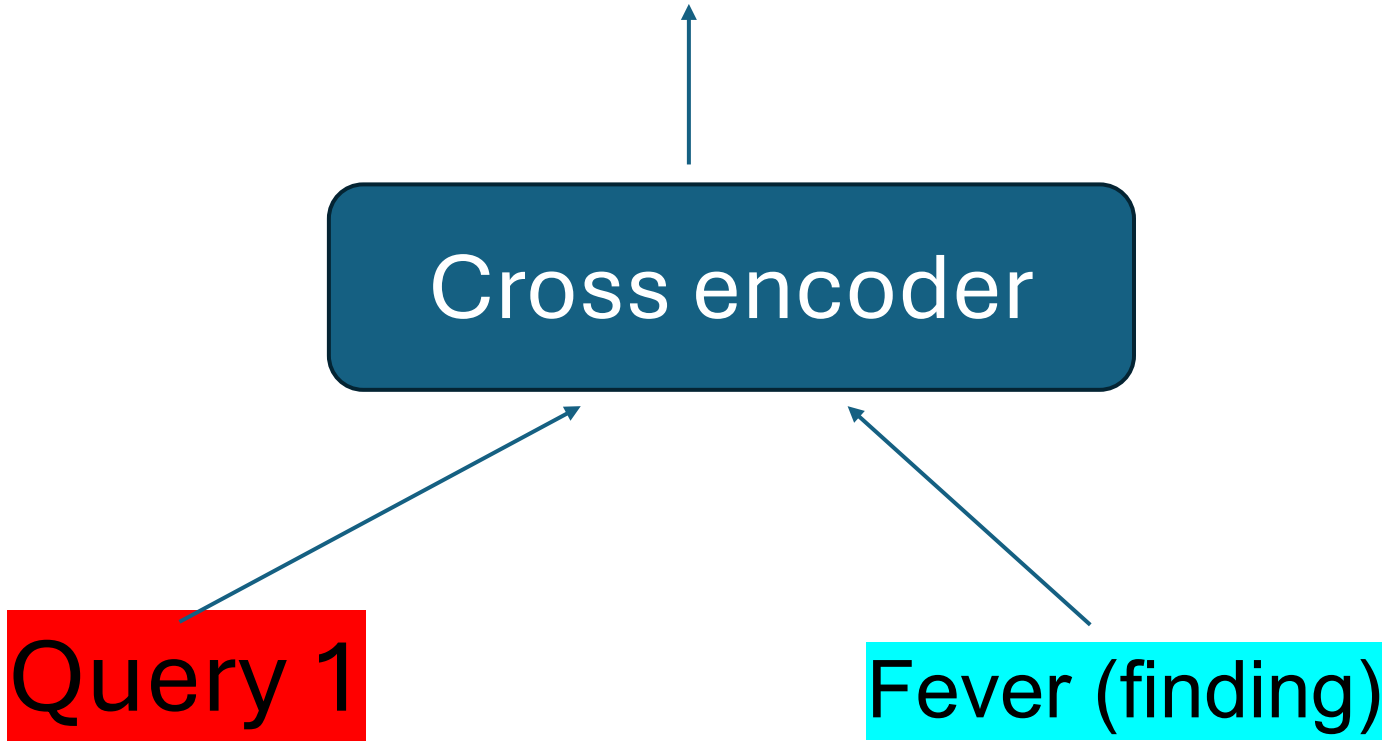
Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**QUERY = the data that you want**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

Select Top K = **25 vectors**

**Relevance score** of Query 1 and Fever (finding)

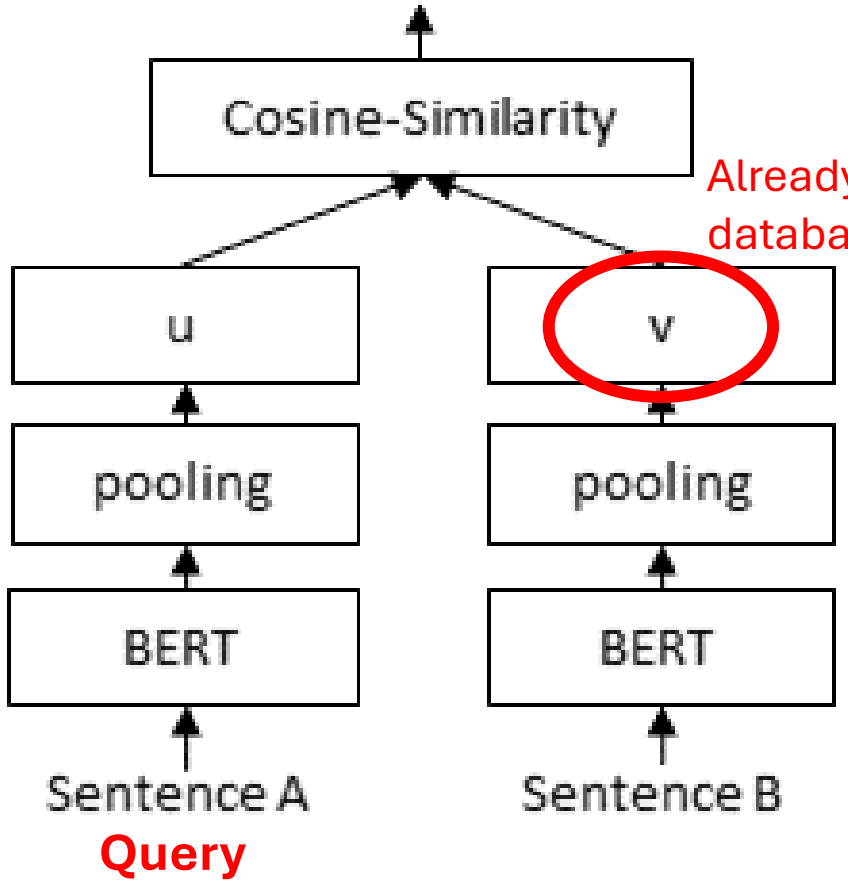


RAG retrieval

RAG re-rank

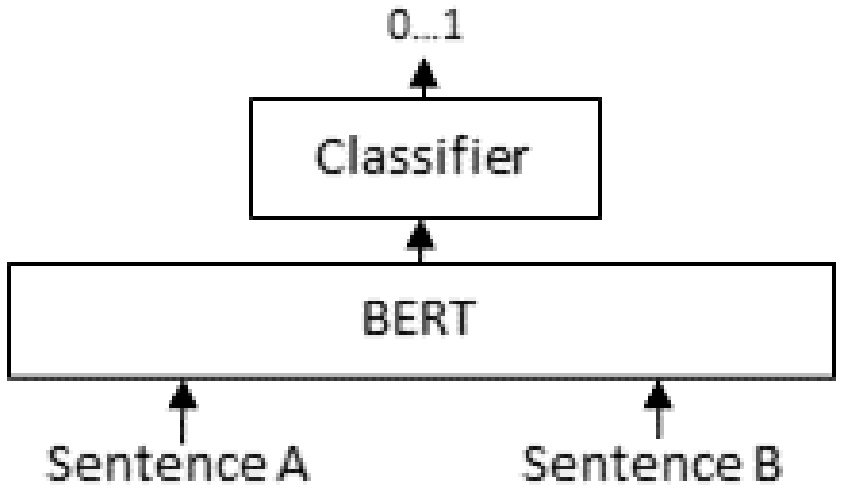
## Bi-Encoder

## Cross-Encoder



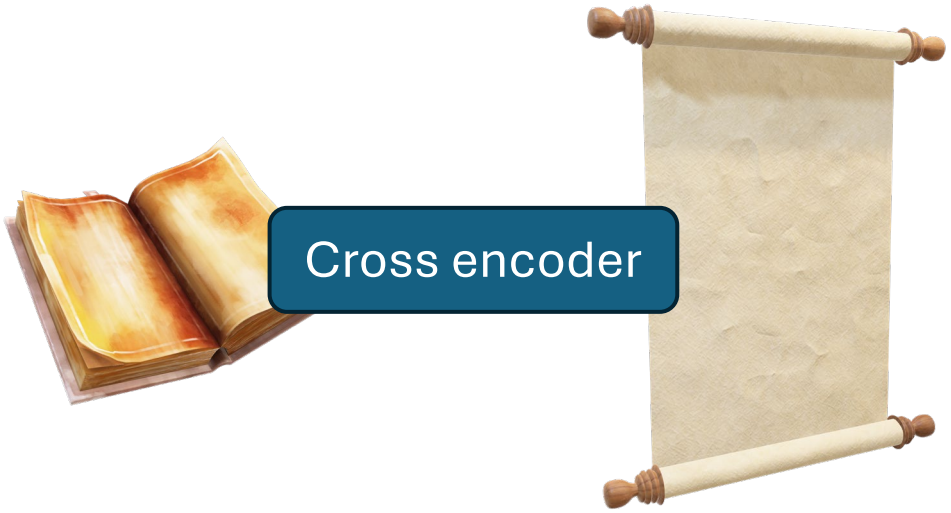

Already in the vector database

**Query**



Extract clinical term | RAG retrieval | **RAG re-rank** | LLM Generation

**QUERY = the data that you want**




**Relevance score = 0.829**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**Relevance score = 0.829**

**Relevance score = 0.729**

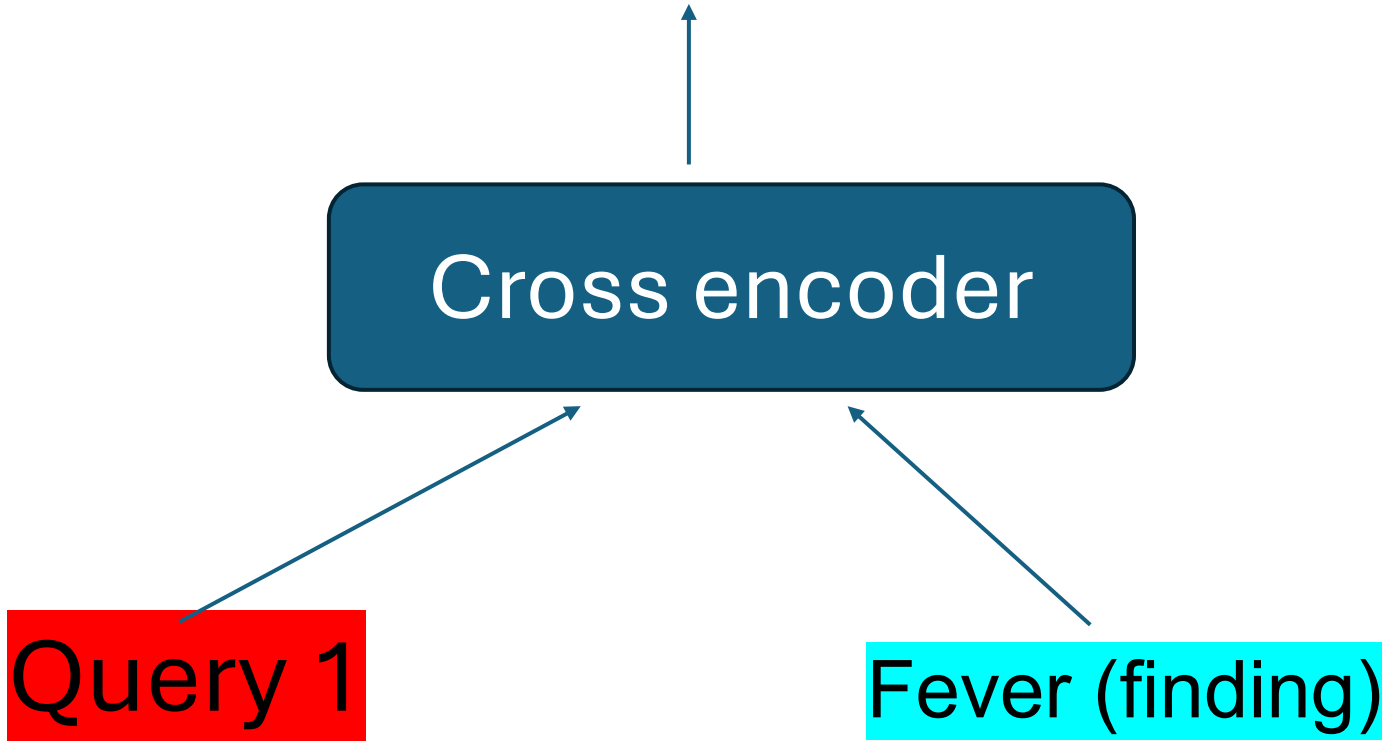
**Relevance score = 0.329**

**QUERY = the data that you want**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

Select Top K = **25 vectors**

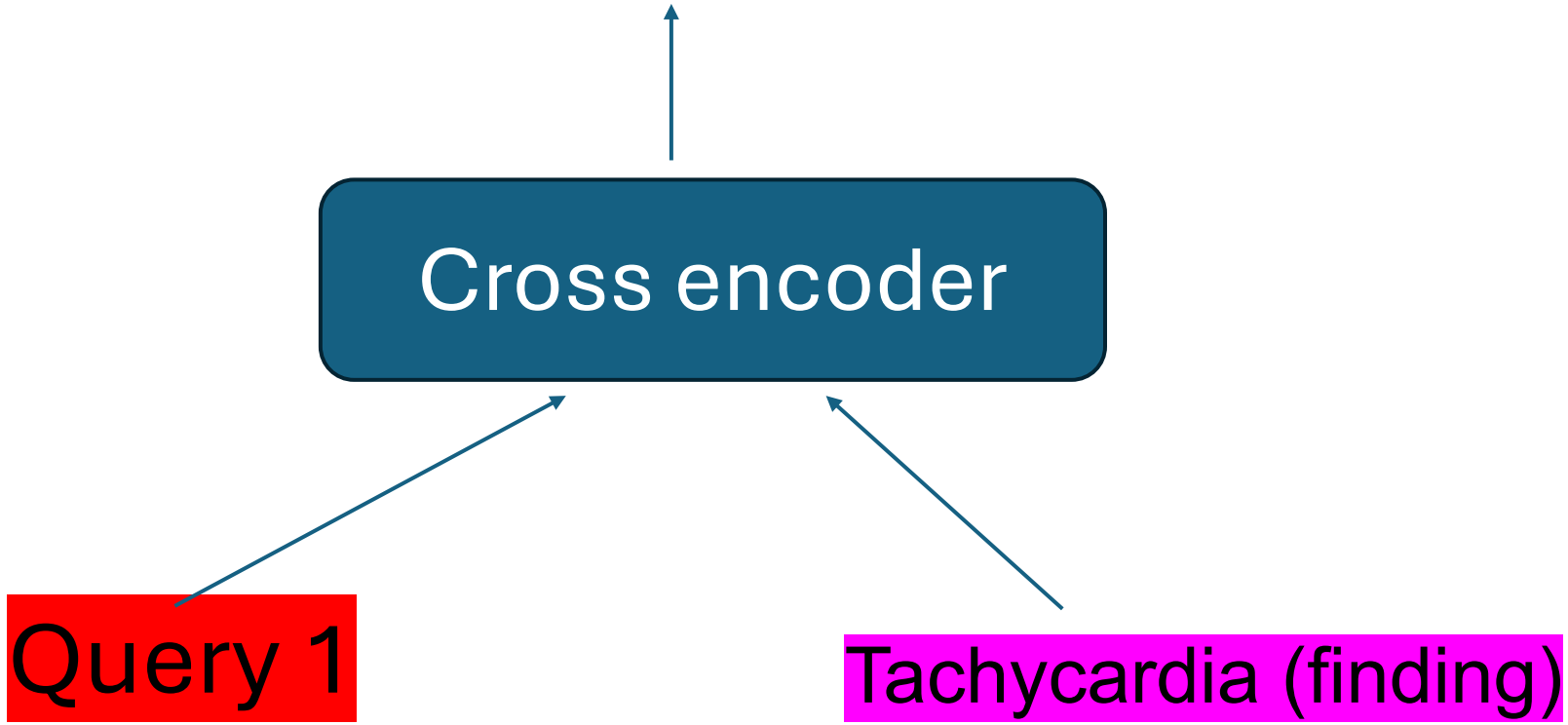
**Relevance score** of Query 1 and Fever (finding)



Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

Select Top K = **25 vectors**

**Relevance score** of Query 1 and Tachycardia (finding)



Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

0.633 **Relevance score** of Query 1 and Salmonella infection

0.133 **Relevance score** of Query 1 and Tachycardia (finding)

0.733 **Relevance score** of Query 1 and IBS

.

0.866 **Relevance score** of Query 1 and Infective diarrhea

.

0.033 **Relevance score** of Query 1 and Dengue

**25 Cosine similarity**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

0.866 **Relevance score** of Query 1 and Infective diarrhea

0.733 **Relevance score** of Query 1 and IBS

0.633 **Relevance score** of Query 1 and Salmonella infection

•

0.133 **Relevance score** of Query 1 and Tachycardia (finding)

•

0.033 **Relevance score** of Query 1 and Dengue

**25 Cosine similarity** → **Top 5 reranked**

Extract clinical term | RAG retrieval | RAG re-rank | LLM Generation

**GPT 4o mini**

```
Select the best document for the clinical term QUERY1 from the
following options:
Option 1:Code: 15223002 Description: Clinical A
Option 2:Code: 12223355 Description: Disease B
Option 3:Code: 22330000 Description: Disease C
Option 4:Code: 22555668 Description: Disease D
Option 5:Code: 55335555 Description: Laboratory A
Please choose the best option and provide the code and description.
example Output: (Option 1) 11200025556 - Clinical term
Output:
```

**Top 5 from reranked RAG**

**one-shot**

(Option 1) **3424008 : Tachycardia (finding)**

**Final answer !**

# Query for : Tachycardia

RAG retrieval

```
Querying for term: Tachycardia
1 - 276796006 : Atrial tachycardia (disorder)
2 - 74615001 : Tachycardia-bradycardia (disorder)
3 - 25569003 : Ventricular tachycardia (disorder)
4 - 82838007 : Irregular tachycardia (disorder)
5 - 6456007 : Supraventricular tachycardia (disorder)
6 - 278482008 : Atrioventricular tachycardia (disorder)
7 - 6285003 : Tachyarrhythmia (disorder)
8 - 426300009 : Tachycardia-induced cardiomyopathy (disorder)
9 - 3424008 : Tachycardia (finding)
10 - 708124001 : Recurrent ventricular tachycardia (disorder)
11 - 233894001 : Incessant atrial tachycardia (disorder)
12 - 233907003 : Induced ventricular tachycardia (disorder)
13 - 12026006 : Paroxysmal tachycardia (disorder)
14 - 233897008 : Re-entrant atrioventricular tachycardia (disorde
15 - 426525004 : Sustained ventricular tachycardia (disorder)
16 - 426761007 : Electrocardiogram: supraventricular tachycardia
17 - 233896004 : Re-entrant atrioventricular node tachycardia (di
18 - 233893007 : Re-entrant atrial tachycardia (disorder)
19 - 413342000 : Neonatal tachycardia (disorder)
20 - 164895002 : ECG: ventricular tachycardia (finding)
21 - 69730002 : Idiojunctional tachycardia (disorder)
22 - 195070000 : Paroxysmal atrioventricular tachycardia (disorde
23 - 49982000 : Multifocal atrial tachycardia (disorder)
24 - 234225006 : Pacemaker re-entrant tachycardia (disorder)
25 - 66657009 : Paroxysmal ventricular tachycardia (disorder)
```

RAG re-rank

```
== Reranked ==
1 (old rank 9) - 3424008 : Tachycardia (finding)
2 (old rank 8) - 426300009 : Tachycardia-induced cardiomyopathy
3 (old rank 2) - 74615001 : Tachycardia-bradycardia (disorder)
4 (old rank 7) - 6285003 : Tachyarrhythmia (disorder)
5 (old rank 13) - 12026006 : Paroxysmal tachycardia (disorder)
```

LLM Generation

**Final answer - 3424008 : Tachycardia (finding)**