



# Retrieval-Augmented Generation & applications in archives

Theory Meets Practice: Harnessing AI for practical  
implementations in digital archiving, 7th November 2024

*Péter Körösi-Szabó, Gábor Kovács, Domonkos Czipra*

# LLMs introduction

- LLM is an artificial intelligence program, that can recognize and generate text, among wide range of tasks
- Pre-trained on large amount of data
- Transformer architecture : encoder and decoder with self attention capabilities
- Question answering, summarization, translating languages, complete sentences

Llama 3.1 : 8B and 70B parameters, 128K maximum prompt size

# LLM's limitations

- Not up-to-date information
  - Re-train could take long time, and expensive
- Hallucinations
- No domain-specific information
- Citation / Explainability

# RAG Introduction

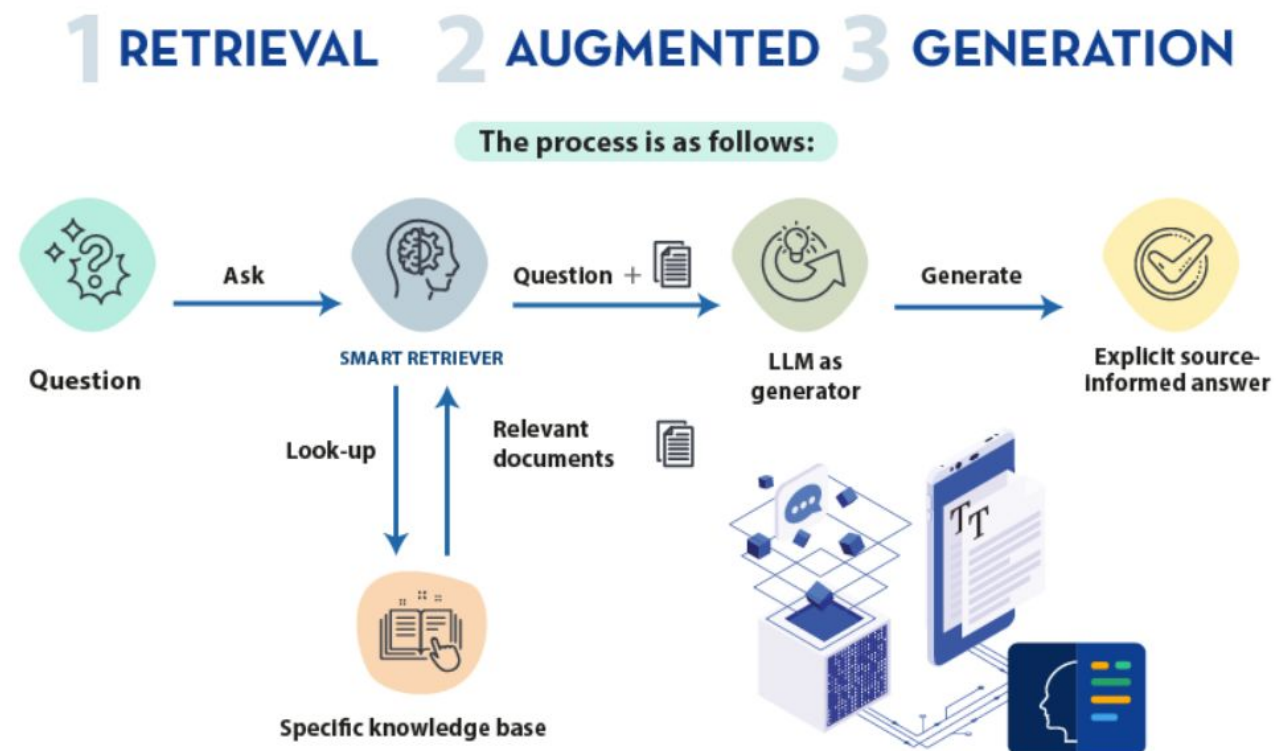
- Retrieval-Augmented Generation
- Advanced technique in Q&A systems, and Chatbots
- Combines retrieval and generation process to enhance the capabilities of LLMs
- Retrieve relevant information from a knowledge base
- Retrieved information is used generate coherent and contextually relevant responses

⇒ Empowers an LLM with accurate and up-to-date informations

# RAG - mechanism

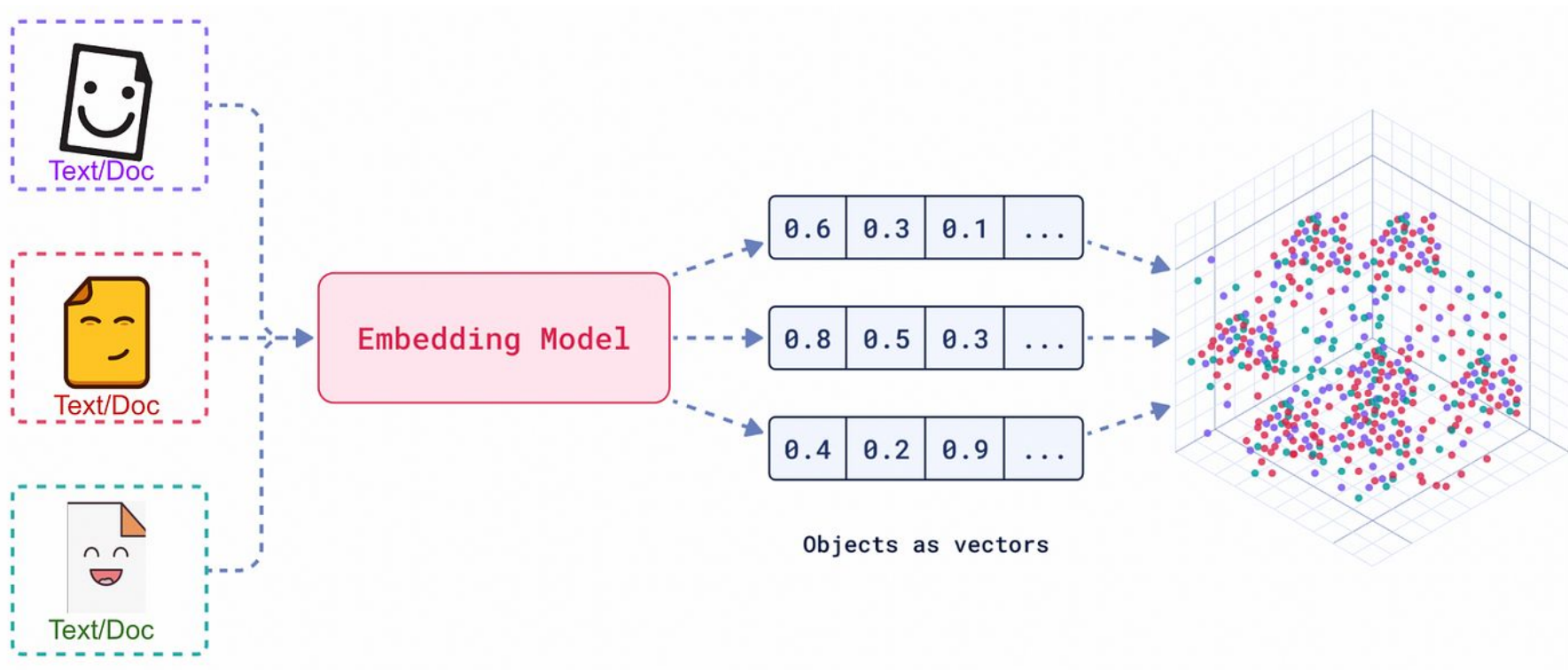
## Motivation

- accelerate search, and question answering
- Q&A with natural language
- use external resource
- update knowledge dynamically
- explainability, citations

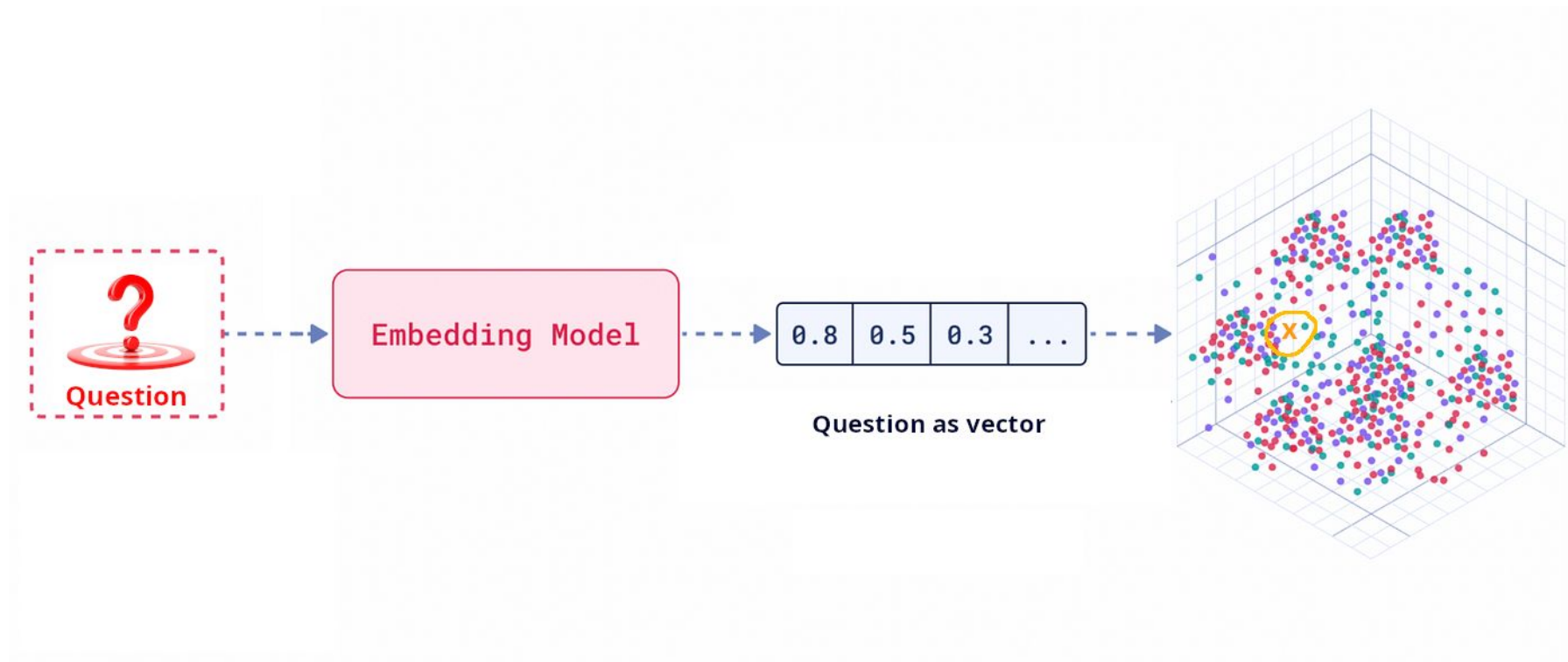


Source: <https://www.ml6.eu/blogpost/leveraging-llms-on-your-domain-specific-knowledge-base>

# RAG - retrieval part

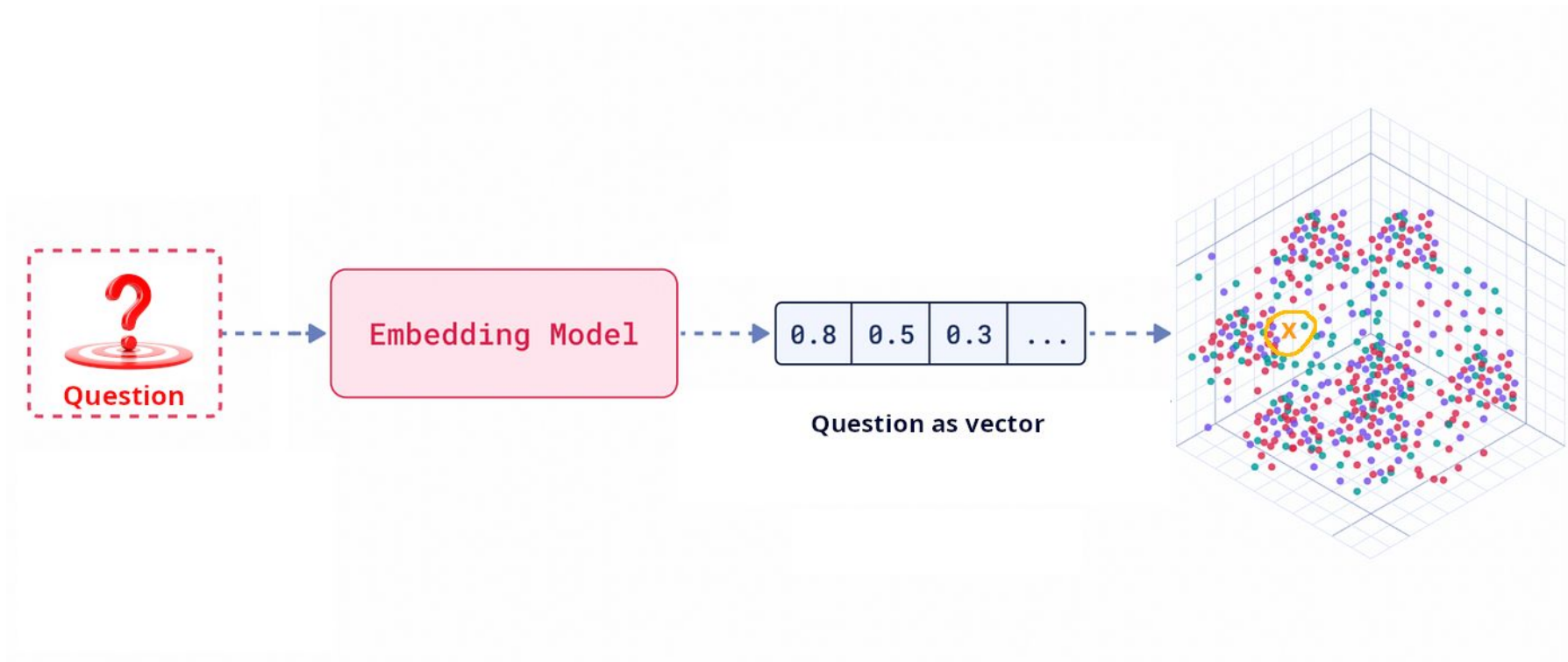


# RAG - retrieval part



# RAG - retrieval part

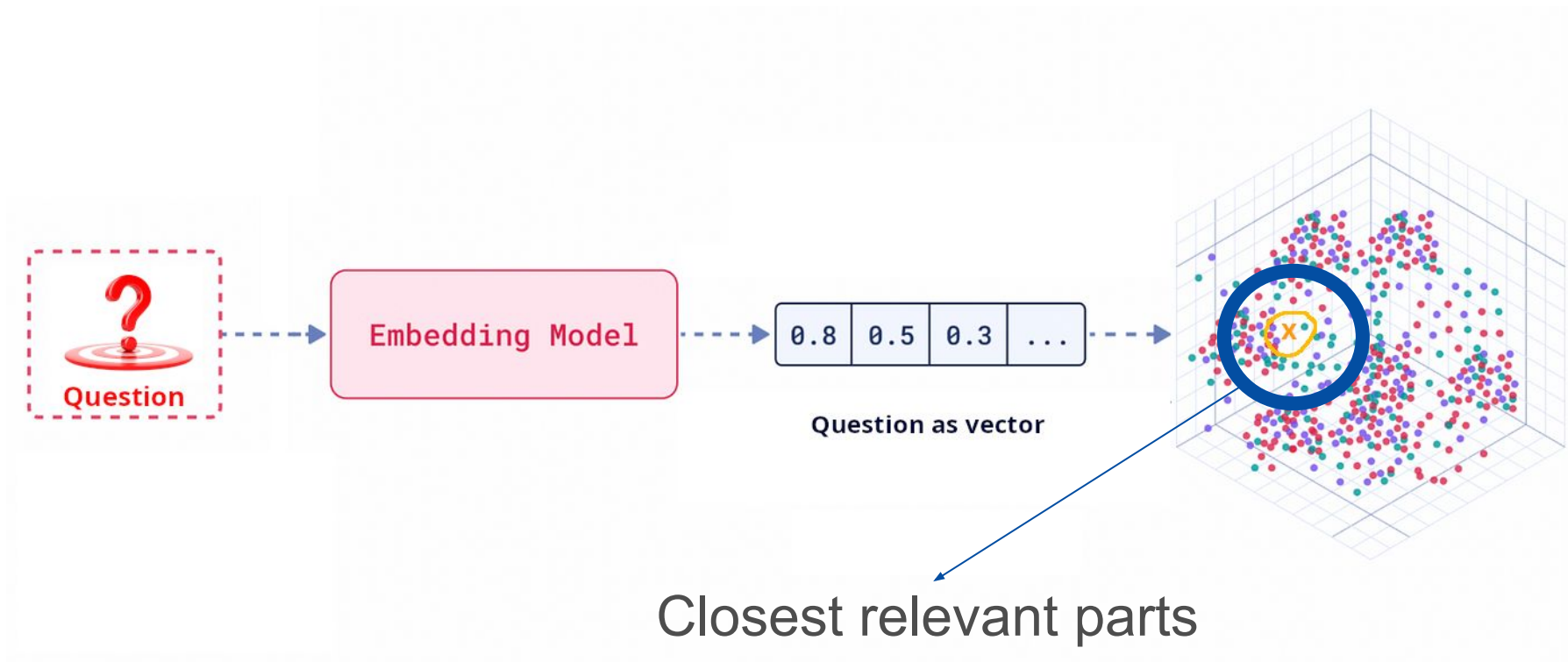
- Challenges
  - noisy (OCR) data
  - Hungarian language





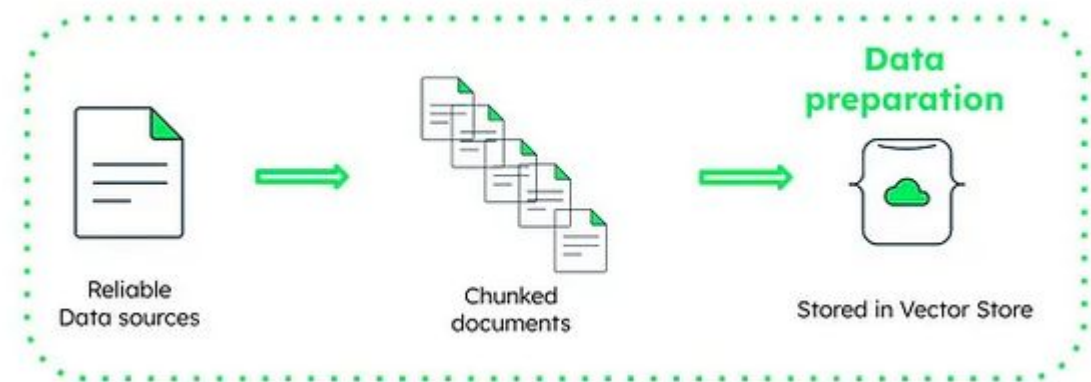
# RAG - retrieval part

- Challenges
  - noisy (OCR) data
  - Hungarian language



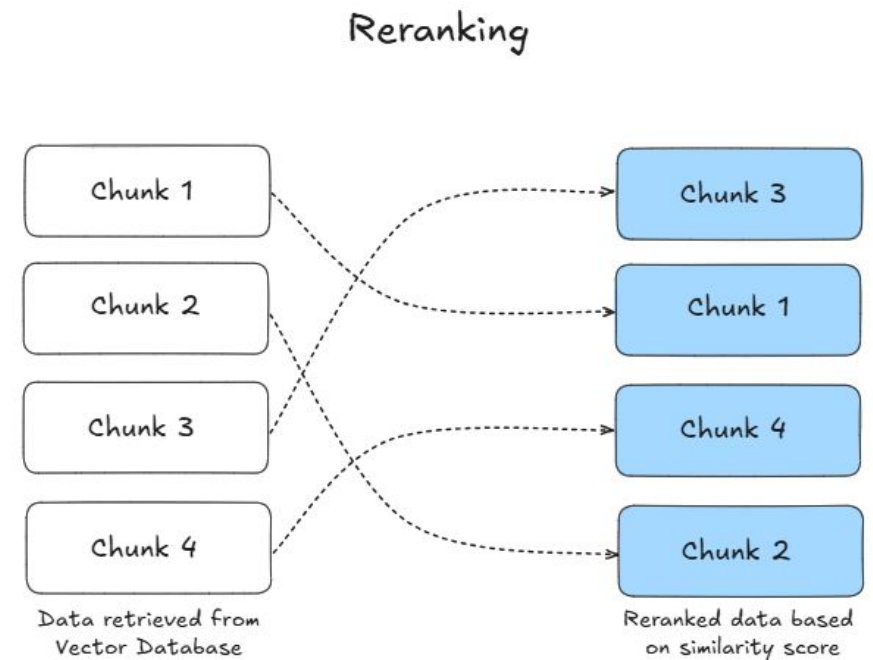
# Retrieval - Preprocessing

- Problems with embedding
  - too large documents → performance reduction
- Chunking
  - split large documents → nodes
- Enhance retrieval performance
  - handle overlapping information
  - rerank retrieved nodes
  - improve with preprocessing the OCR'd text quality



# Retrieval: reranking

- Motivation
  - combine embedding models / retrieval methods (also with classical retrieval methods, like BM25)
  - use more expensive model
- Commercial solutions
  - Cohere, Colbert, etc.
- LLM based solutions
  - expensive, but ideal for postprocessing

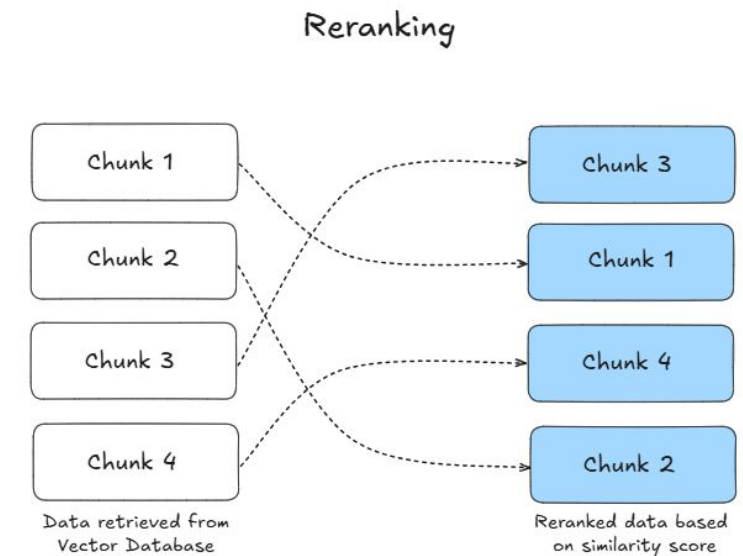


# Retrieval: reranking

- Motivation
  - combine embedding models / retrieval methods (also with classical retrieval methods, like BM25)
  - use more expensive model
- Commercial solutions
  - Cohere, Colbert, etc.
- LLM based solutions
  - expensive, but ideal for postprocessing

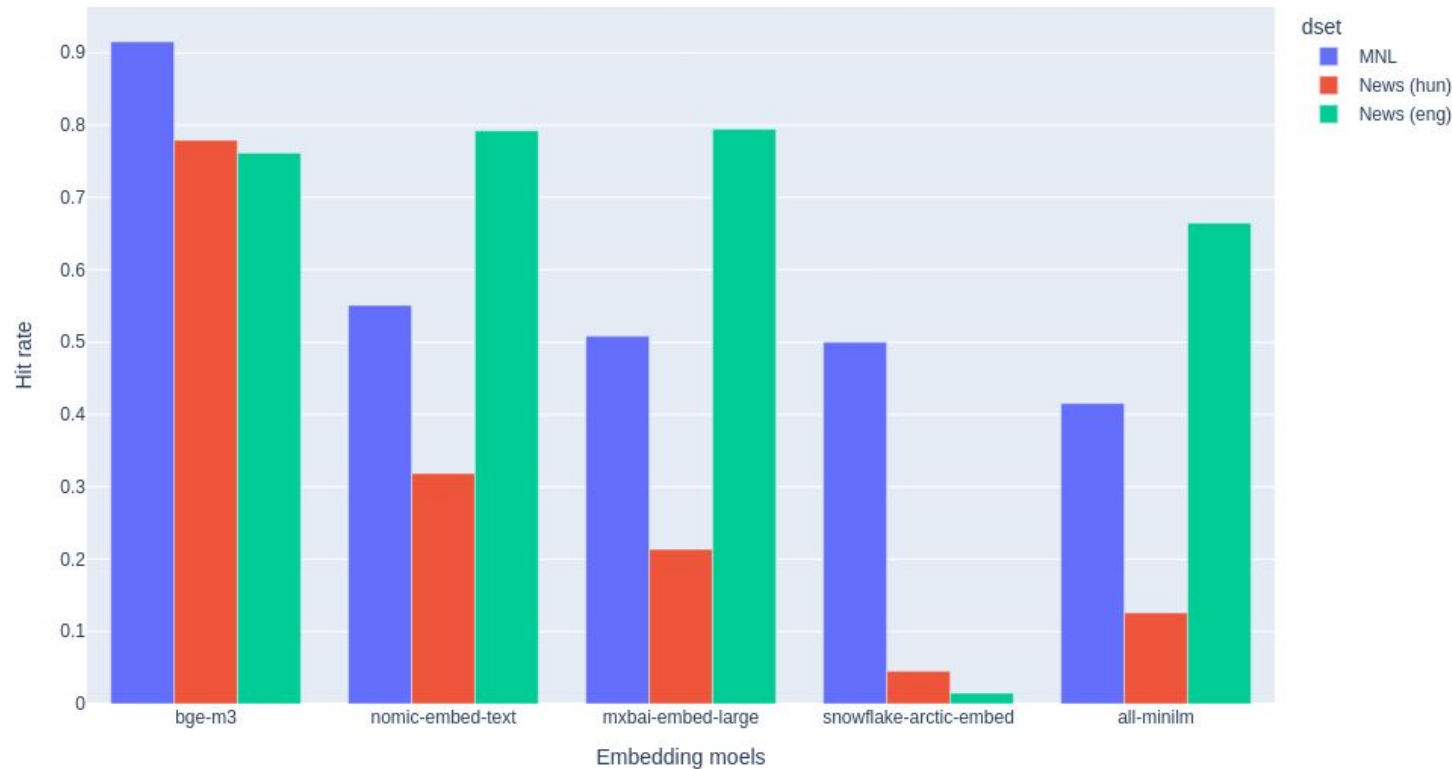
## Advanced retrieval

- Store document hierarchies
- Metadata filters + auto retrieval



# Retrieval benchmark

Retrieval performance of different embedding models



## MNL (National Archives of Hungary)

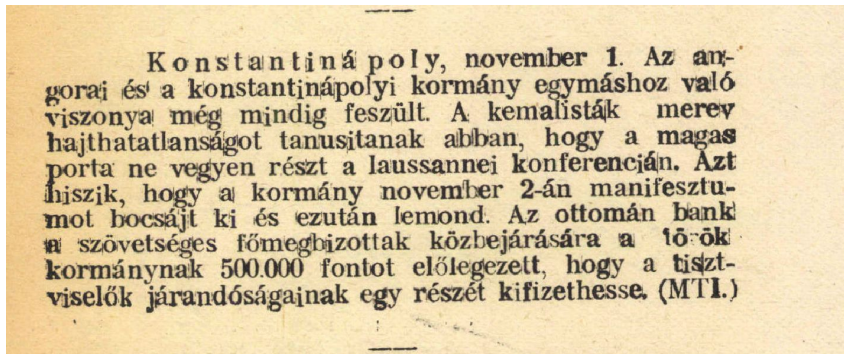
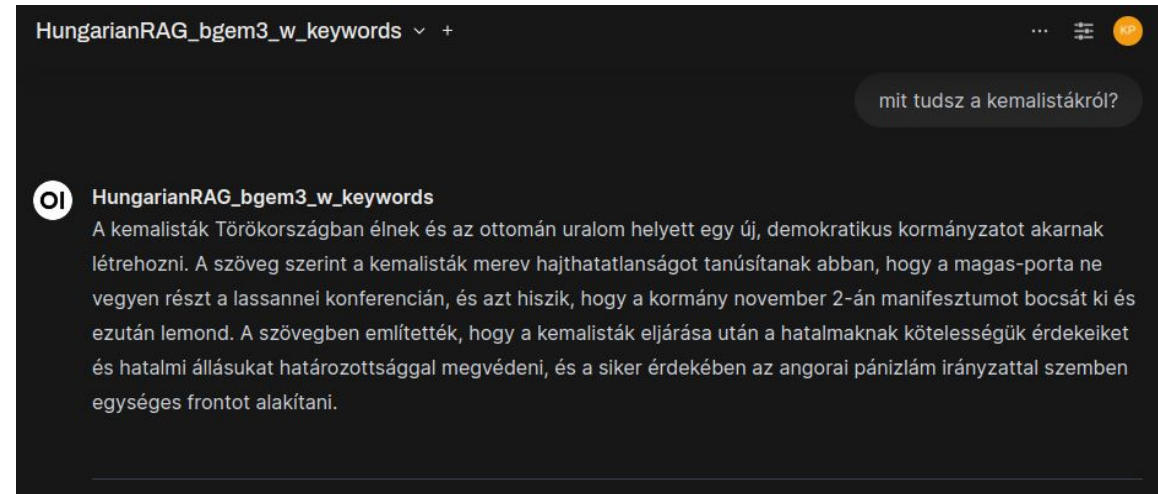
- 20 documents
- 59 chunks
- 60 questions

## News (hun/eng)

- 636 articles
- 1272 chunks
- 2533 questions

# Our current setup and work in progress...

- retrieval
  - bge-m3 modell (576m parameter)
  - embedding size: 1024
- augmented
  - with top 5
- generation
  - llama3.1 (8b parameter)
- data
  - daily news from 1922 november
  - ~300 documents ~2100 news
  - ~600 chunk (with chunk size 1024)



Ollama hosts models



LlamaIndex



Open WebUI web interface

kiket választottak be a londoni alsóházba, és mikor?

**oi** HungarianRAG\_nomicembed 5:03 PM

A megismételt szöveg:

A szövegből nem derül ki, hogy kiket választottak be a londoni alsóházba. A szöveg említi, hogy Lord Derby az új hadügyminiszter, és intézkedett a Strafel (úgy gondolom, ez egy név lehet) ői tér levő saját lakásának kiürítéséről, de nem tartalmaz információt a londoni alsóház választásairól.

### Used sources:

**Source 1 - 0.7816803179643415**

[Excerpt from document]

file\_path:

/shared\_data/archives/data/rag\_example/mnl\_napi\_hirek\_1922-11\_1\_1014/1922-11-10\_003\_text\_die.txt

document\_title: A Szovjetunió külpolitikája és gazdaságának helyzete 1922-ben.

section\_summary: A szöveg részlet beszámol egy dokumentum részleteiről, amely a Szovjetunió külpolitikájáról és gazdaságának helyzetéről szól 1922-ben. A lapok arról is informálják az olvasót, hogy az olasz és az orosz küldöttek részt vennének a Lausannei értekezleten, amelynek célja a tengersizorsok kérdésének rendezése. Az orosz kormány hangoztatja, hogy nem lehet elkülöníteni a tengersizorsok kérdését a Közel-Kelet általános rendezésétől, ezért részt akarnak venni az értekezlet egész folyamában. A szövegben említett másik fontos esemény egy jegyzék átadása a szövetségesek részéről az angorai kormány képviselőinek, amelyben

**oi** HungarianRAG\_bgem3\_wo\_metadata 5:03 PM

A szövegrészlet szerint a következőket választották be az alsóházba:

- 42 konzervatív jelöltet (nem szerepel időpont)
- 32 unionista jelöltet (nem szerepel időpont)
- 5 nemzeti liberális jelöltet (nem szerepel időpont)
- 4 munkáspárti jelöltet (nem szerepel időpont)
- 10-úr unionista és 1 nacionalista jelöltet (nem szerepel időpont)

Azonban a szövegrészletben van egy másik rész, amely szerint a következőket választották be az alsóházba:

- 42 konzervatív, 5 nemzeti liberális, 4 munkáspárti, 10-úr unionista és 1 nacionalista jelöltet (London, november 5. napján)

Ebből a részből következtethetünk arra, hogy a válasz az alsóházba bejutott jelöltek listájára:

- London, november 5.

A válasz tehát: London, november 5.

### Used sources:

**Source 1 - 0.5314969704678477**

file\_path:

/shared\_data/archives/data/rag\_example/mnl\_napi\_hirek\_1922-11\_1\_1014/1922-11-14\_001\_text\_die.txt

a: London, november 13. (Wolfi.) A Times jelentése szerint az új alsóházban, ötvenegy eddig kétes mandálműtől eltekintve, a pártok a következőké csoportosulnak: 322 unionista, 95 munkáspárti, 75 liberális. 63 nemzeti liberális, 5 függetlenségi, 3 ir 'nacionalista. Ezek a számok az unjonistáknak a ház valamennyi pártjával szemben nyolcvan szavazatnyi többséget jelentenek. Még ha mind az ötvenegy kétes mandátum az ellenzéki pártoknak iutna is. az

# Compare embedding models

**Question:** "Who was elected to the House of Commons in London, and when?"

## *Nomic-embed-text*

- High confidence (0.78)
- Answer: "The text does not provide any information about who was elected to the House of Commons in London and when."

## *BGE-M3*

- Finds the related part
- Answer: "The text reveals that the following were elected to the House of Commons in London:
  - 42 Conservative candidates
  - 27 Liberal candidates

... "

mit tudsz Sir W. Horwoodról? és ki volt ő?

**HungarianRAG\_bgem3\_w\_keywords** 12:00 AM  
A szövegből kiderül, hogy Sir W. Horwood a londoni rendőrség feje volt. A szöveg szerint ő csütörtökön este rosszul lett egy lakoma előkészületei közben, és emiatt kórházba kellett szállítani. A rendőrfőnököt ismeretlen tettesek meg akarták mérgezni. Állapota komoly volt.

Sir W. Horwoodról azonban a szövegben nem található több információ, például nem tudomásunkra jutott, hogy ki volt ő vagy hogyan került a londoni rendőrség élére.

#### Used sources:

**Source 1 - 0.49828491161151156**

[Excerpt from document]

file\_path:  
/shared\_data/archives/data/rag\_example/mnl\_napi\_hirek\_1922-11\_1\_1014/1922-11-11\_011\_text\_die.txt

excerpt\_keywords: politika, választások, szenátus, demokrácia, parlamentarizmus, republikánusok, demokrata párt, parasztok, londoni rendőrség, Sir W. Horwood, Mérgezés, tisztviselők, Kórház, Konstantinápoly, minisztertanács, MTI, Magyar Távirati Iroda, nyomdászat

Excerpt:

:: London, november 11. A Timesnek jelentik Washingtonból: A választások végleges eredménye még nem ismeretes, de valószínű, hogy a szenátusban 53 republikánus, 42 demokrata, 1 parasztúriánusok, eszerint az fényi többséggel rendelkeznek. (Mánusok,

**HungarianRAG\_bgem3\_w\_summary** 12:00 AM  
A szövegrészletben nem található említés Sir W. Horwoodról, így nincs elérhető információ a kérdésre adott válaszhoz.

#### Used sources:

**Source 1 - 0.35552905519010014**

[Excerpt from document]

file\_path:  
/shared\_data/archives/data/rag\_example/mnl\_napi\_hirek\_1922-11\_1\_1014/1922-11-14\_008\_text\_die.txt  
document\_title: Törökország békekonferenciája, ellenpontos olasz-magyar egyeztetés.

section\_summary: A szöveg részlet egy híradó kiemelését tartalmazza a Magyar Távirati Iroda archívumából. A hír szerint az olasz kormány tagadja, hogy bármilyen egyezmény lenne Mussolini és Horthy között. Ugyanakkor a Chicago Tribune jelentése szerint Mussolini della Torretta márki Londonba kinevezett követet utasítja arra, hogy azonnal Londonba utazzon és vegyen részt a Curzon Lord és Poincaré közti tárgyalásokon. A hírt Zimmer Ferenc felelős szerkesztő és a Magyar Távirati Iroda adta ki.

excerpt\_keywords: Kínos helyzet, zajos szöveg... De nem adjunk fel!

Az alábbi 15 egyedi kulcsszónál találsz:

Bécs, Chicago, Curzon Lord, Horthy Miklós, Mussolini, Poincaré, olasz-magyar egyeztetés, Törökország békekonferenciája.

Excerpt:

[Bécs, november 14. Olasz hivatalos részről kijelentik, hogy egyik Bécsben megjelenő magyar lapnak az a híre, mely szerint Mussolini olasz miniszterelnök és Horthy Miklós kormányzó között egyezmény áll fenn,

## Compare metadat extractions

**Question:** "What do you know about Sir W. Horwood? And who was he?"

### keyword extraction

- find the related part:

"The text reveals that Sir W. Horwood was the head of the London police. According to the text, he fell ill on Thursday evening during the preparations for a banquet and had to be taken to the hospital..."

### summary extraction

- answer:

"The text excerpt does not mention Sir W. Horwood, so there is no available information to answer the question."

- noisy ocr data -> unreliable summary



# Future work

scale to larger data

- quantize models
- distill knowledge
- etc.

More complex knowledge representation

- knowledge graph
- hybrid search (keyword+embedding)
- etc.

finetune models on our data

- embedding
- LLM



Thanks for your Attention!

# Thank you



© European Union 2024

Reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

