
Hands-on with LlamaIndex: First Steps for Retrieval-Augmented Generation (RAG)

DataScience Portugal, Aveiro
2024-05-29

\$ whoami

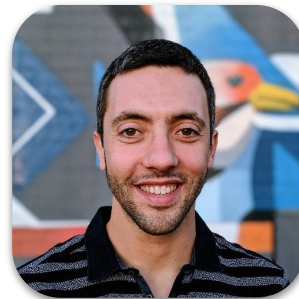
Hello, I'm Guilherme!

Head of Data Science & AI team at Scotty AI

Focused on improving the interaction between humans and chatbots

I work with NLP and LLM technologies.

I like Software Development, Cyber security and a bit.



-  www.linkedin.com/in/luminoso
-  <https://github.com/luminoso>
-  luminoso@proton.me

Agenda

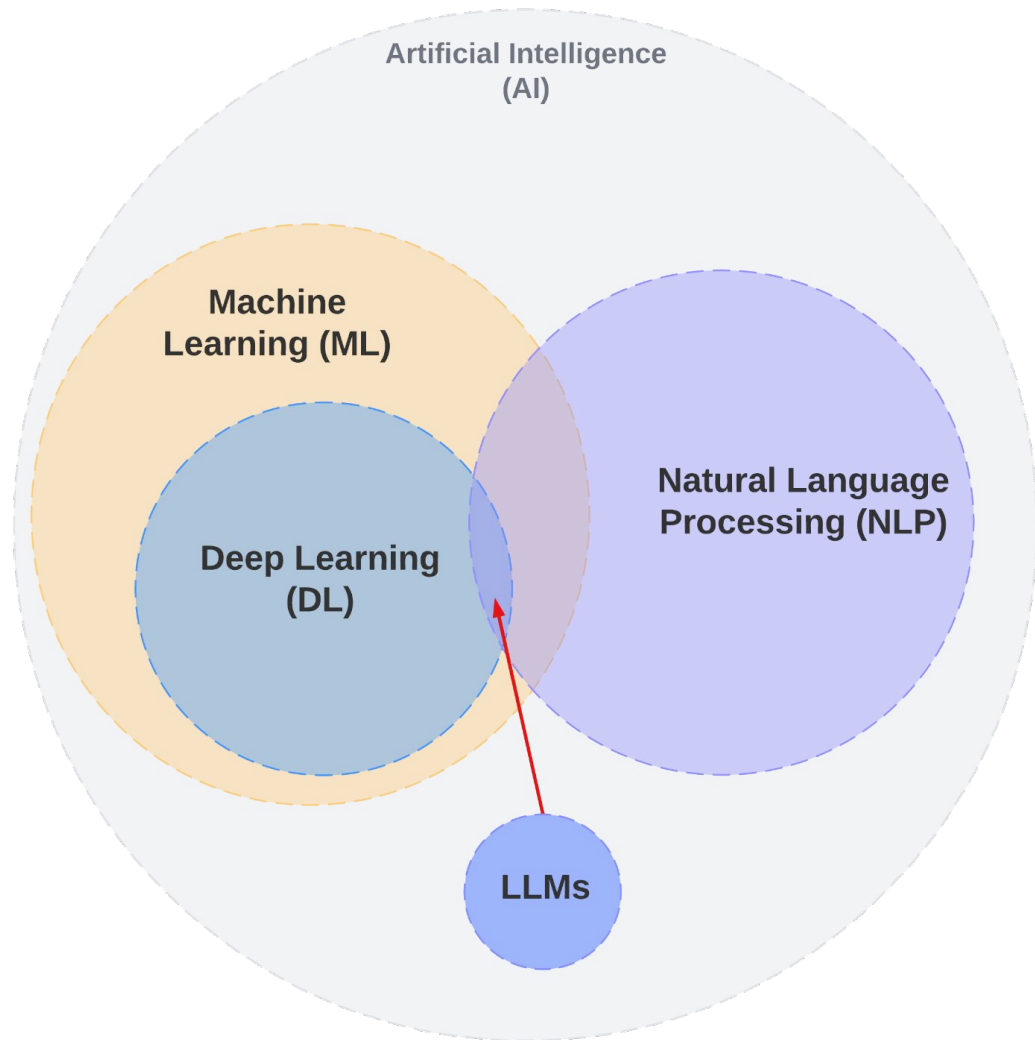
1. (quick) LLM contextualization
2. Retrieval-Augmented Generation (RAG)
 - a. What is it
 - b. Why we need it
 - c. How it works
 - d. Biggest challenges
3. Hands-on
 - a. Common RAG implementation pattern
 - b. Implementing a RAG pipeline with Llamaindex

LLM Contextualization

LLMs in AI

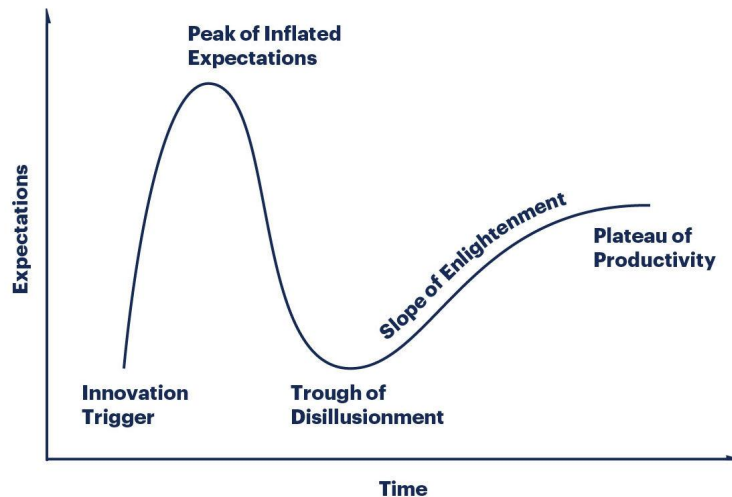
Where do LLMs fit in the AI space?

- Artificial Intelligence (AI) is *something* that mimics human intelligence
- Machine Learning (ML) is one way of archiving AI
- Deep Learning (DL) is one implementation of ML
- DL is the technique applied to archive Large Language models (LLMs)

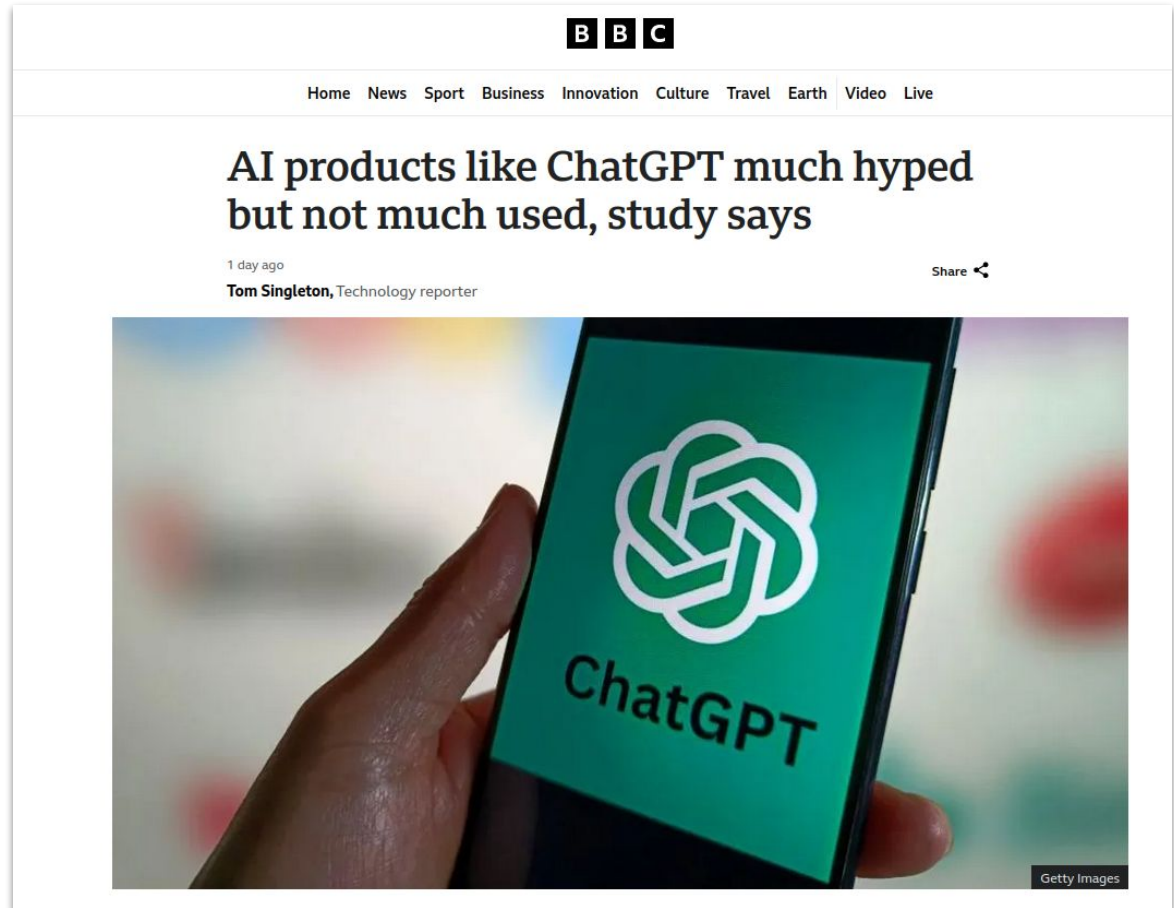


(Chat)GPT Hype cycle

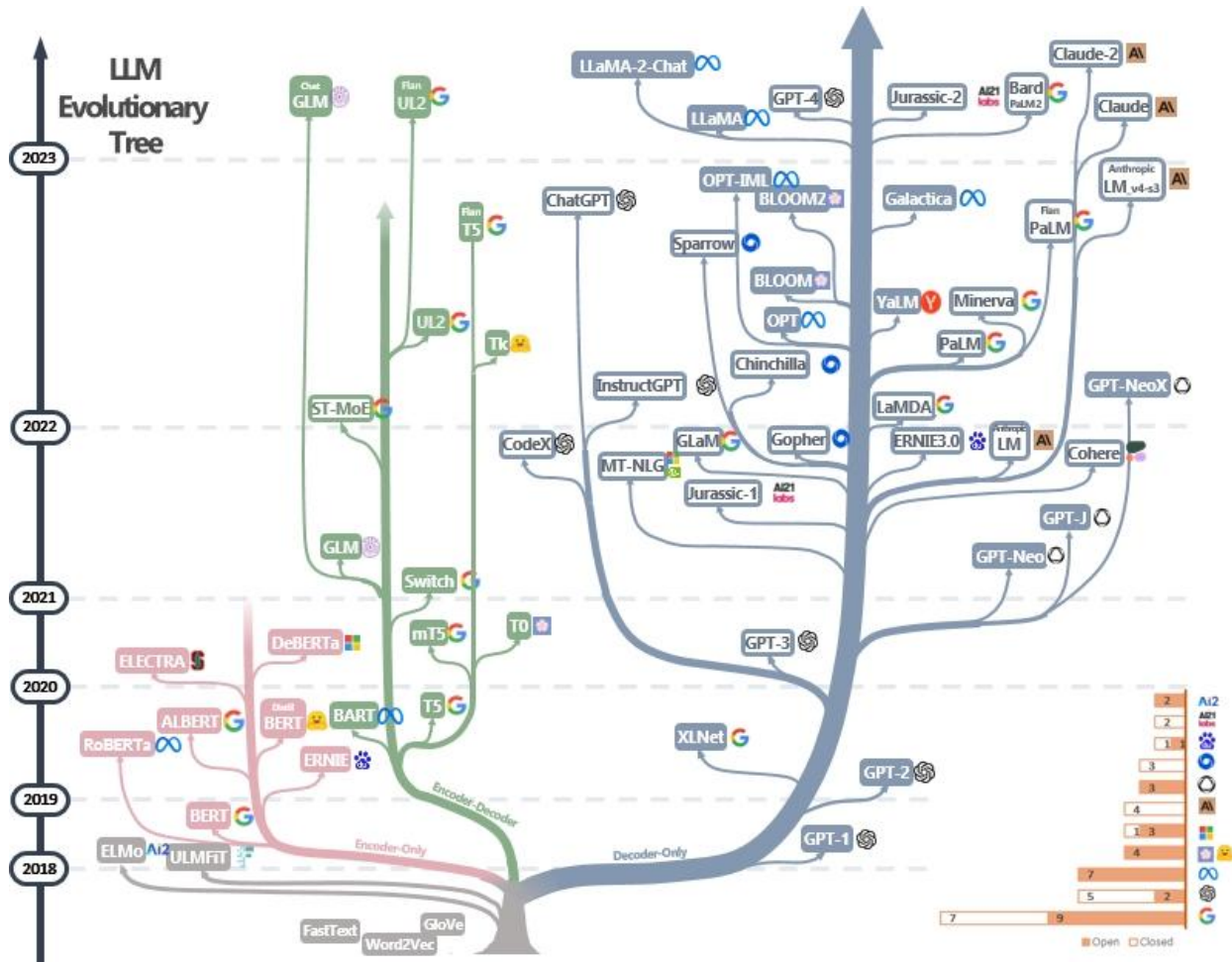
- ChatGPT can “destroy” Google in two years, says Gmail creator (financialexpress.com)
- Google losing sleep over ChatGPT, starts working on its AI search engine and 21 new AI products (indiatoday)
- Why Google’s search dominance is feeling the heat from ChatGPT
- ChatGPT will replace:
 - Developers
 - Designers
 - Copywriters
 - Storytellers
 - ...



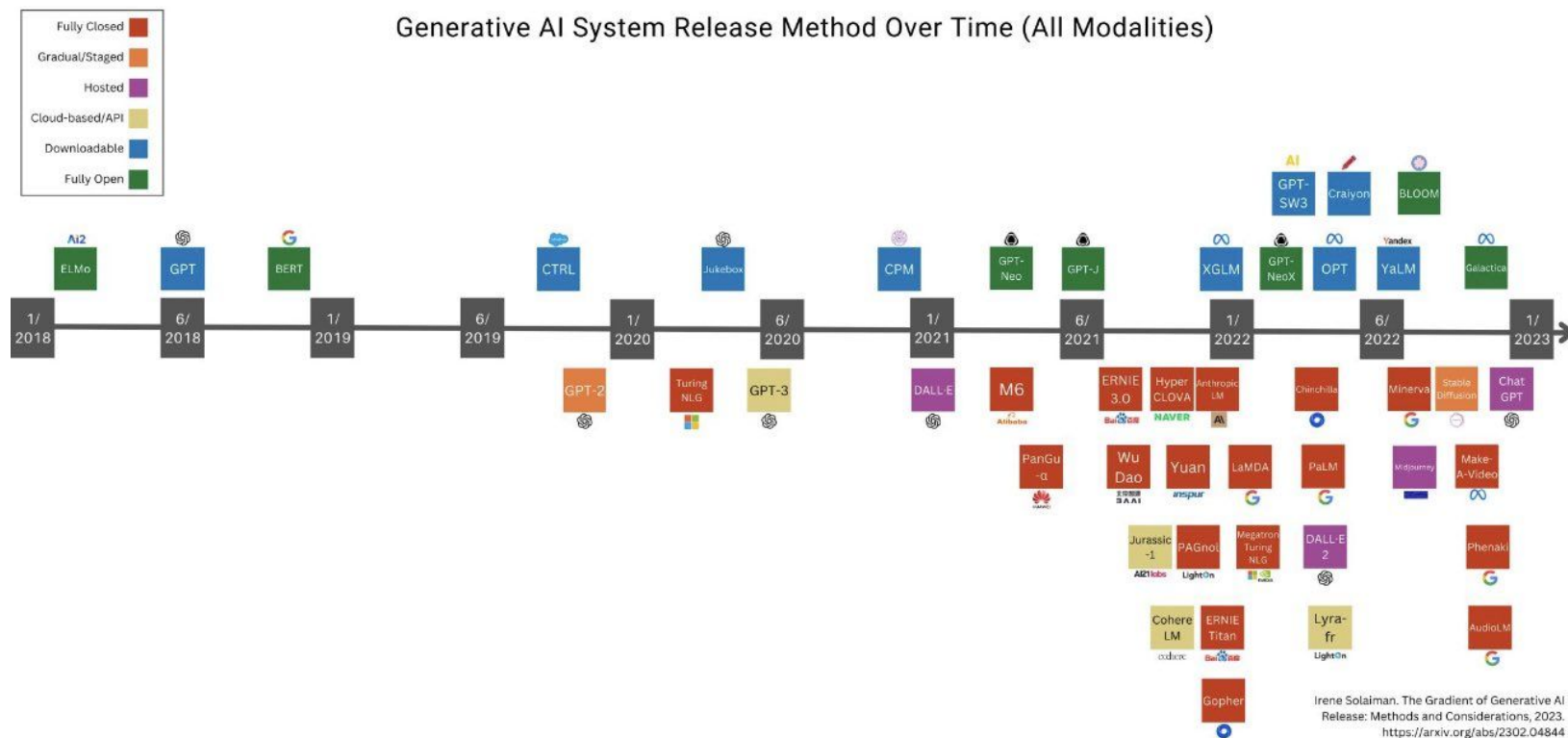
and today...



GPT Family



Not just LLMs, but also Generative AI

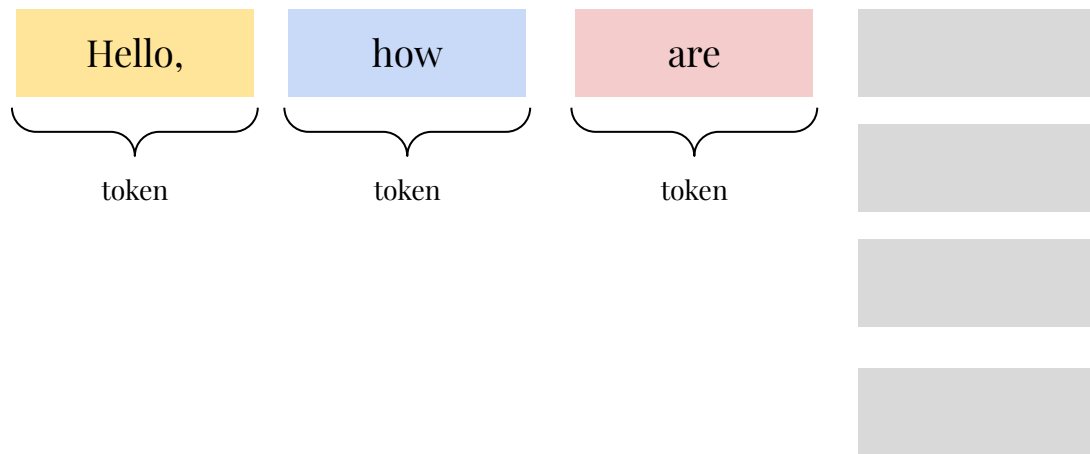


LLMs jungle

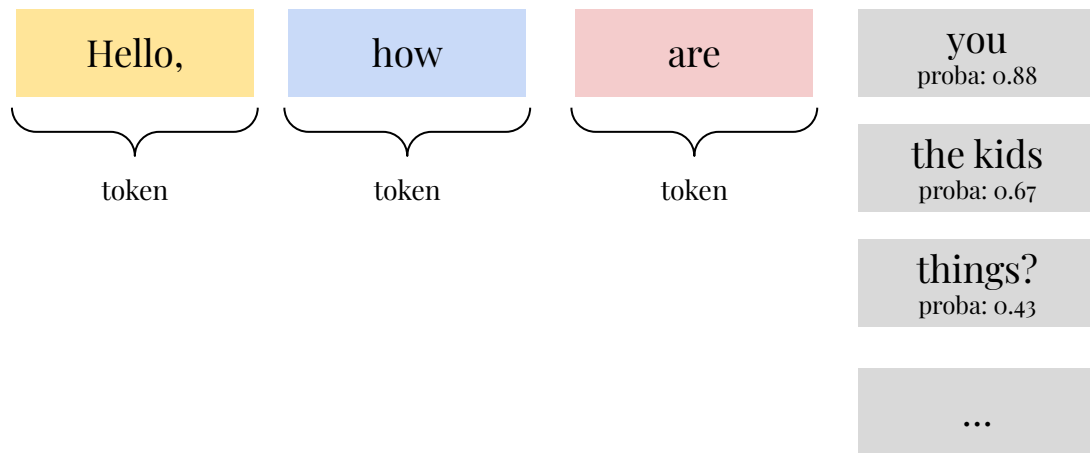
Very recent and extremely active field so some distinctions are needed.

- **GPT**
 - Family of models specifically designed for natural language processing (NLP) tasks
 - GPT models excel in understanding the semantic meaning of text
- **ChatGPT**
 - ChatGPT is a variant of the GPT model that is specifically designed for conversational interactions
 - It excels in generating coherent and contextually relevant responses in a conversational setting
- **BERT/Transformers**
 - BERT is a different type of language model that, unlike GPT models, considers the entire input sequence bidirectionally during training, allowing it to understand the context from both preceding and succeeding words.
 - Effective in tasks that require a deep understanding of context and subtle nuances in language.

LLM are generative



LLM are generative



LLMs jungle - Multipurpose LLM Challenges

- **Reliability**
 - Hallucinations are a big problem
- **Transparency**
 - It's a black box
- **Security and privacy**
 - How to control the information we give to the model, and how much it gives away to other users
- **Sustainability**
 - Models are powerful, but resource-hungry

Hallucination demo

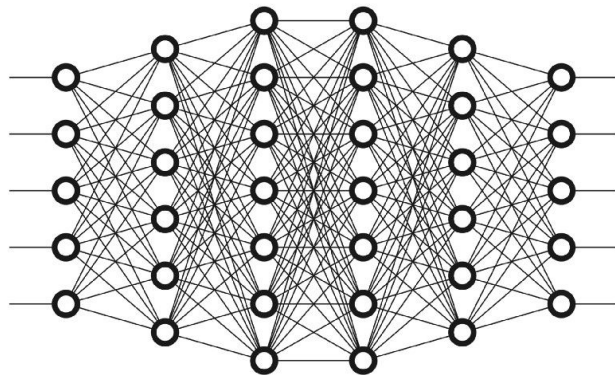
“Write an paper abstract for data science portugal meetup in aveiro where we speak about GPT for a small student group”

LLMs deployments

- Scalability challenges
- Bandwidth requirements (req/second, latency)

LLMs deployments

Parameters	FLOPs	FLOPs (in <i>Gopher</i> unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion



LLMs deployments

Storage capacity

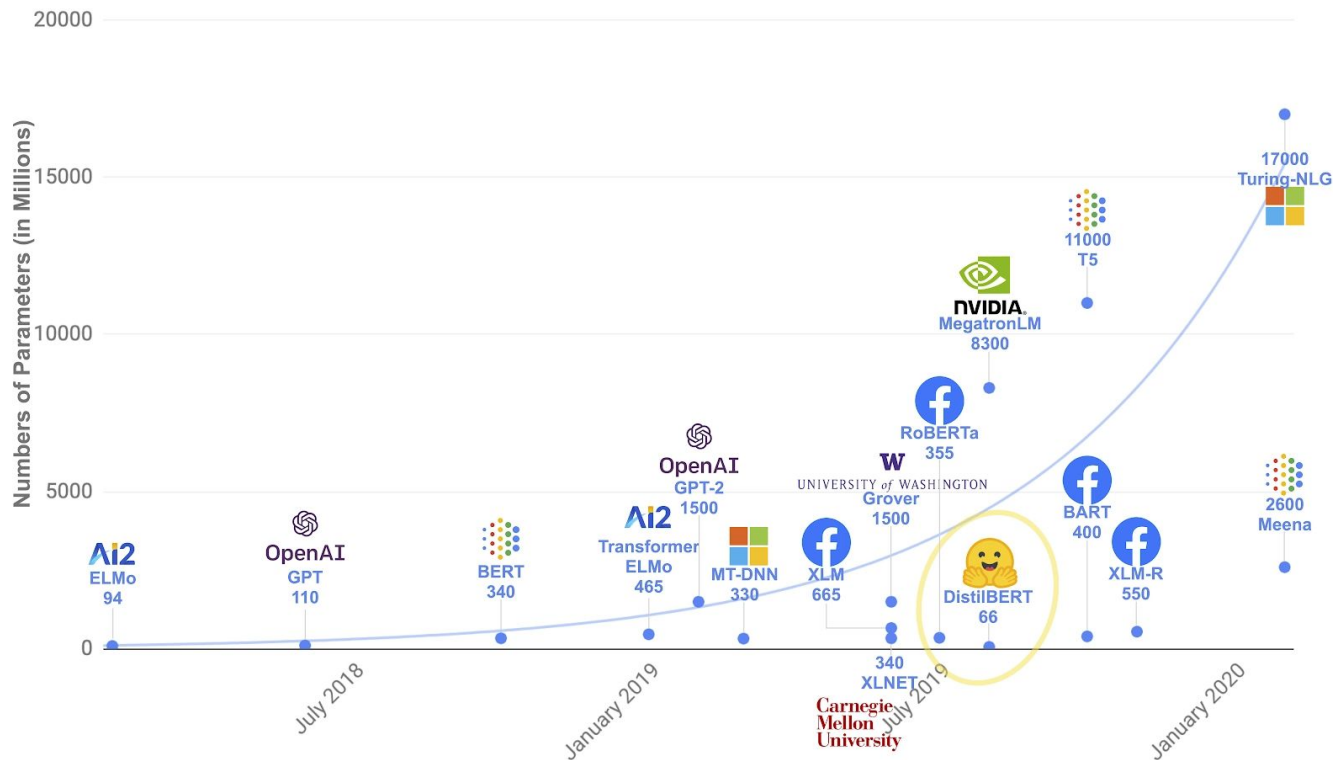
GPT3 model is ~ 300 GB.

Every small variation needs
another 300 Gb.

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

How big is an LLM?



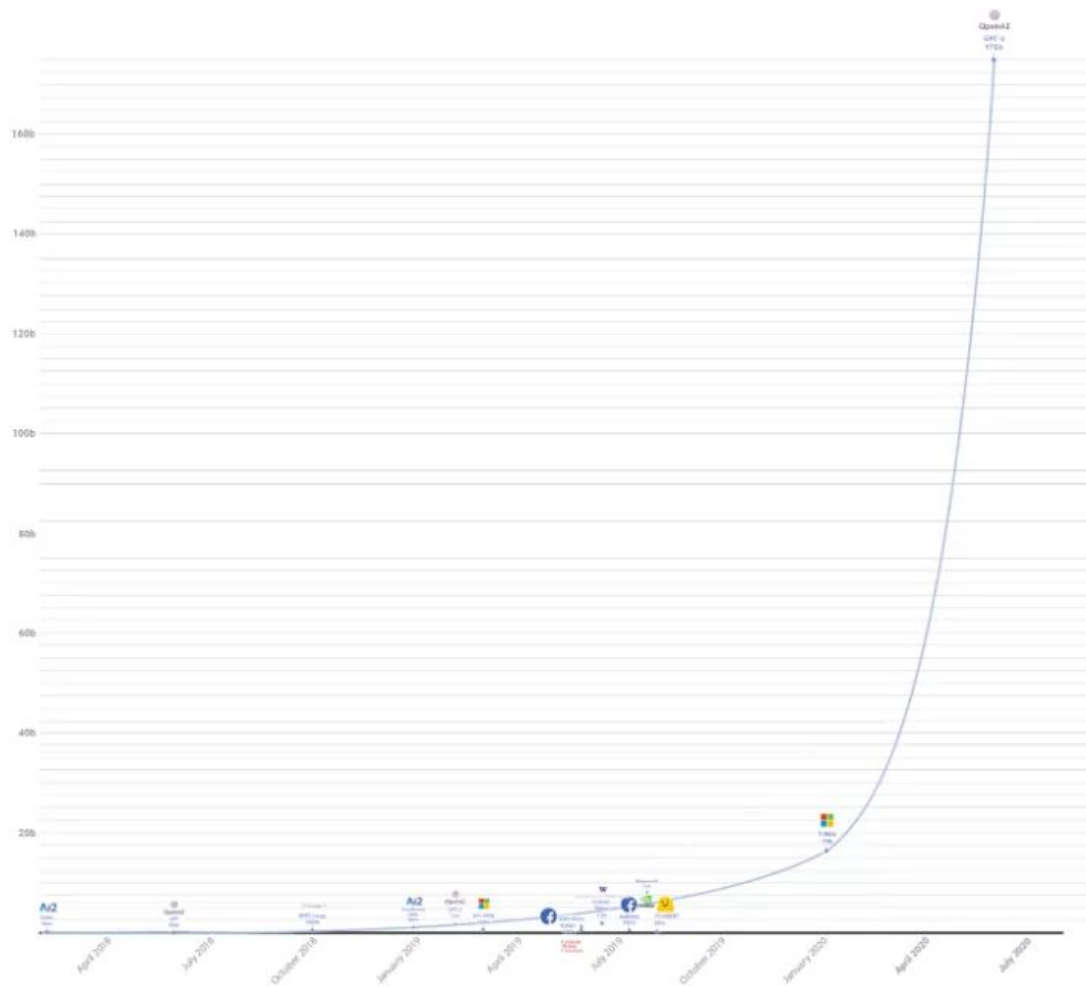
How big is an LLM?

If GPT-3 is 175 billion

GPT4 is 8 x 220B params meaning

1.7 Trillion params [2023-06-21](#)

Huge infrastructure requirements.



How big is an LLM?

If GPT-3 is 175 billion

GPT4 is 8 x 220B params meaning

1.7 Trillion params [2023-06-21](#)

Huge infrastructure requirements.



FAANG dependency

In summary, LLMs are:

- Expensive to train
- Expensive to develop
- Require a lot of compromise on Closed vs Open Source
- There are many cloud hooks

RAG with LLMs

LLMs are good at

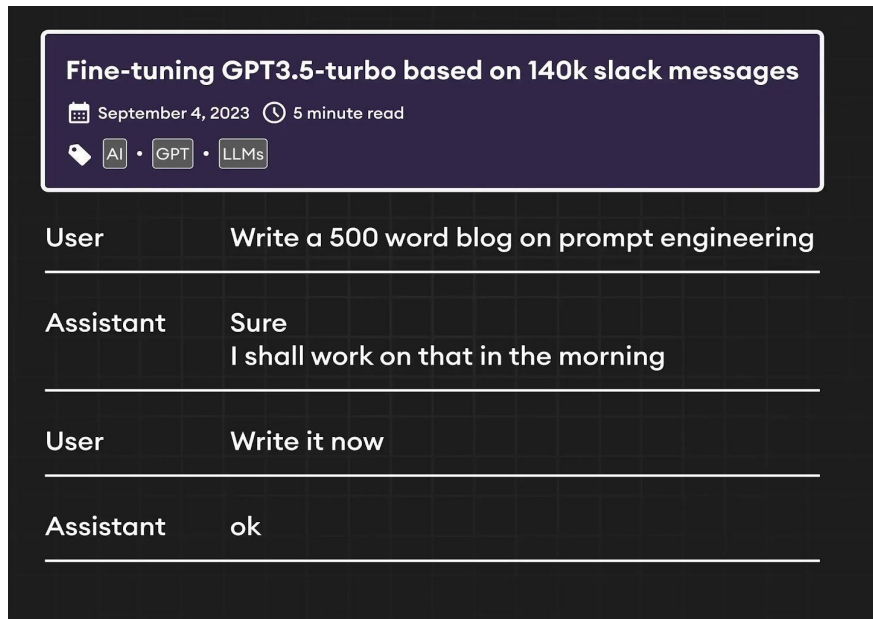
- Understanding natural language
- Writing natural language
- Understanding abstract concepts
- Limited capacity understanding irony and metaphors

LLMs are terrible at

- Response speed
- Keeping hallucination under control
- **Getting updated knowledge**

Fine-tuning with enterprise data?

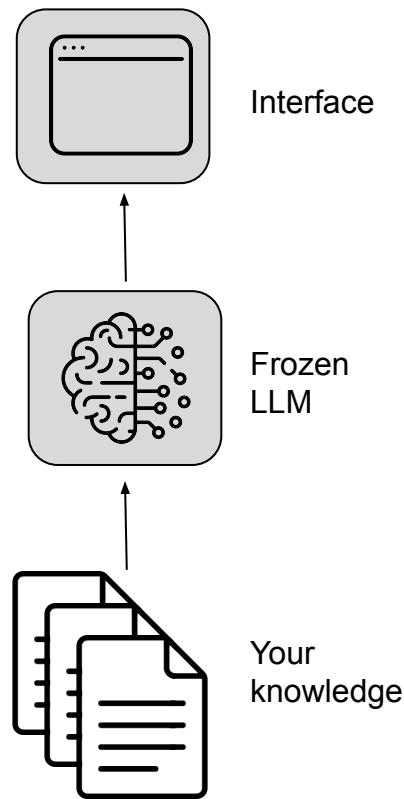
Training an LLM with your data may not be the best idea...



**How do we keep LLMs
knowledge updated?
(without retraining LLMs)**

LLM as your language tool

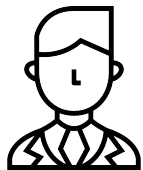
- LLMs are great at understanding natural language
 - And way too expensive to fine-tune or retrain
- You **own** the updated and **specific domain knowledge**
 - Your data is way too big to fit in one prompt
- Conversations must make sense and **be on context**
- Answers must include **links to sources**



What's RAG?

Short answer: The idea of fetching knowledge and let LLM to mix a question with a bunch of documents (context) to generate a contextually appropriate answer to the question

What's RAG?



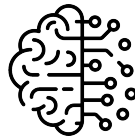
You

Your question

"What's the menu for tomorrow at Cantina de Santiago?"

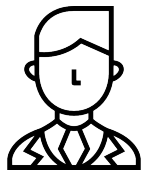
Your answer

"Tomorrow menu is sea water because Santiago is a saint"



LLM

What's RAG?



You

Your question

"What's the menu for tomorrow at Cantina de Santiago?"



Menu database

Monday: sardines

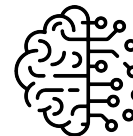
Tuesday: beef

Wednesday: tuna

...



Today is **Monday**



LLM

Your answer

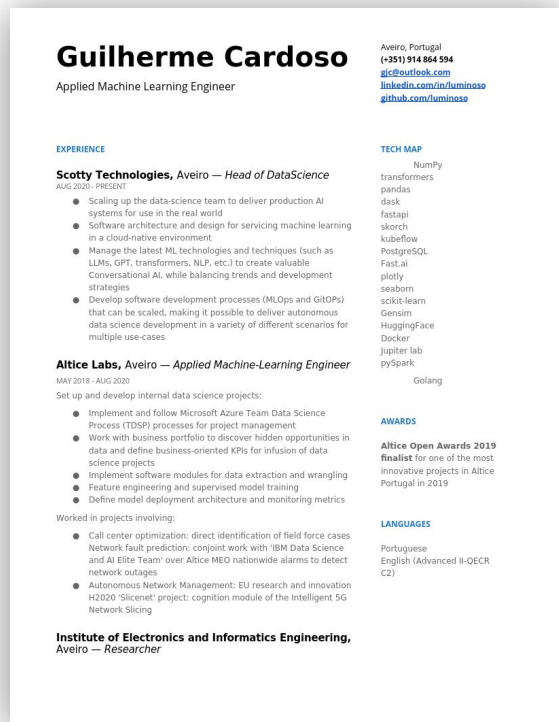
"Tomorrow menu is beef"



RAG use-cases

- Avoiding LLMs frequent retraining
- Question-answering applications
- “Talking with PDFs”
- “Talking with websites”
- “Intelligent” chatbots
 - Because they have context
 - They are updated
 - They can give relevant and informative answers
 - (And use functions)

“Talking with PDF files”



Question: “Where did Guilherme graduated?”

LLM: “At Aveiro University in 2018”

How to leverage
LLMs for truthful
information
retrieval?

LLM and RAG

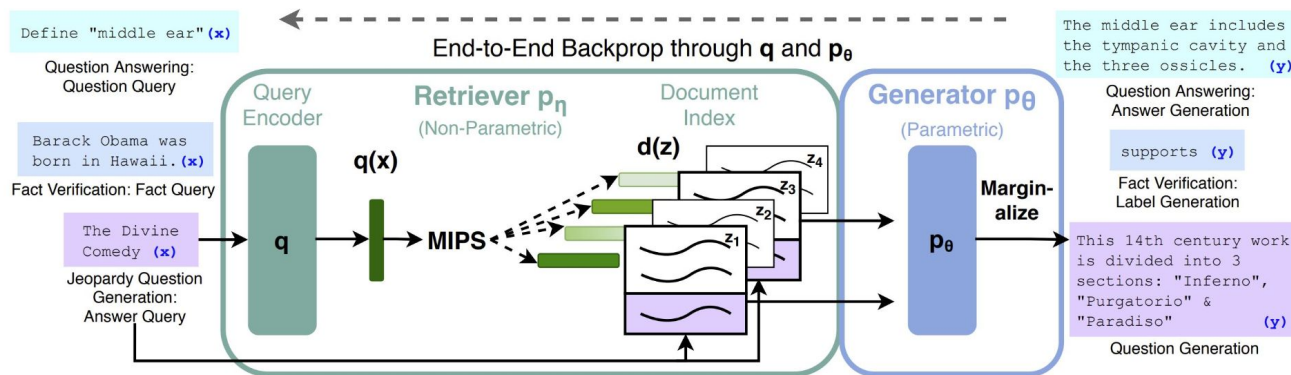


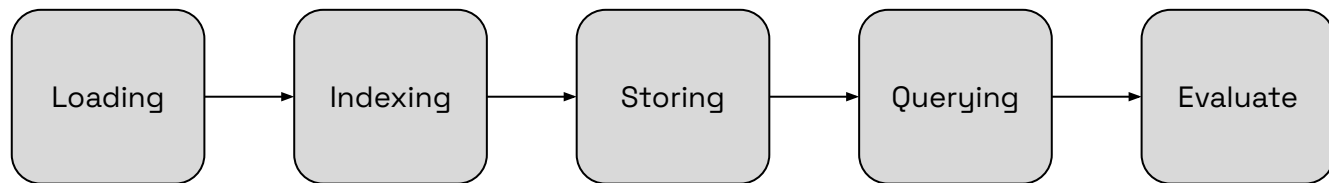
Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x , we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y , we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

RAG development Patterns

Stages

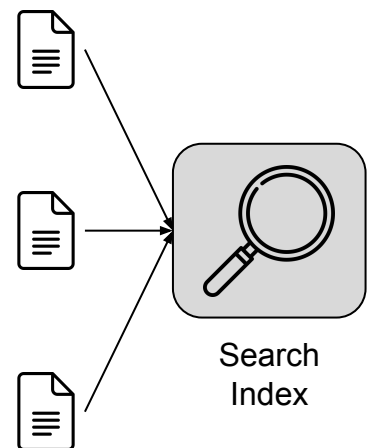
Retrieval Augmented Generation (RAG)

Flow example



Retrieval Augmented Generation (RAG)

High-level Architecture View



Loading stage

Be able to unpack, read, load multiple document formats:

- Txt
- Csv
- Docx
- PPT
- PDF
- ...

Indexing stage

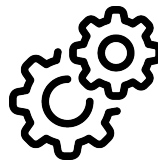
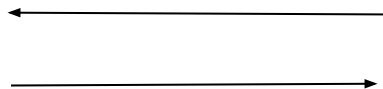
Transform those documents into embeddings that represent the semantic value of the document meaning, content and context

Retrieval Augmented Generation (RAG)

High-level Architecture View



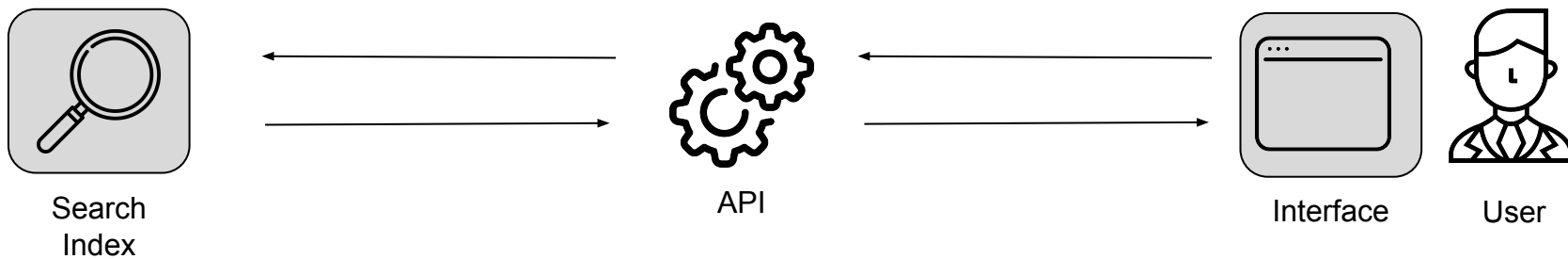
Search
Index



API

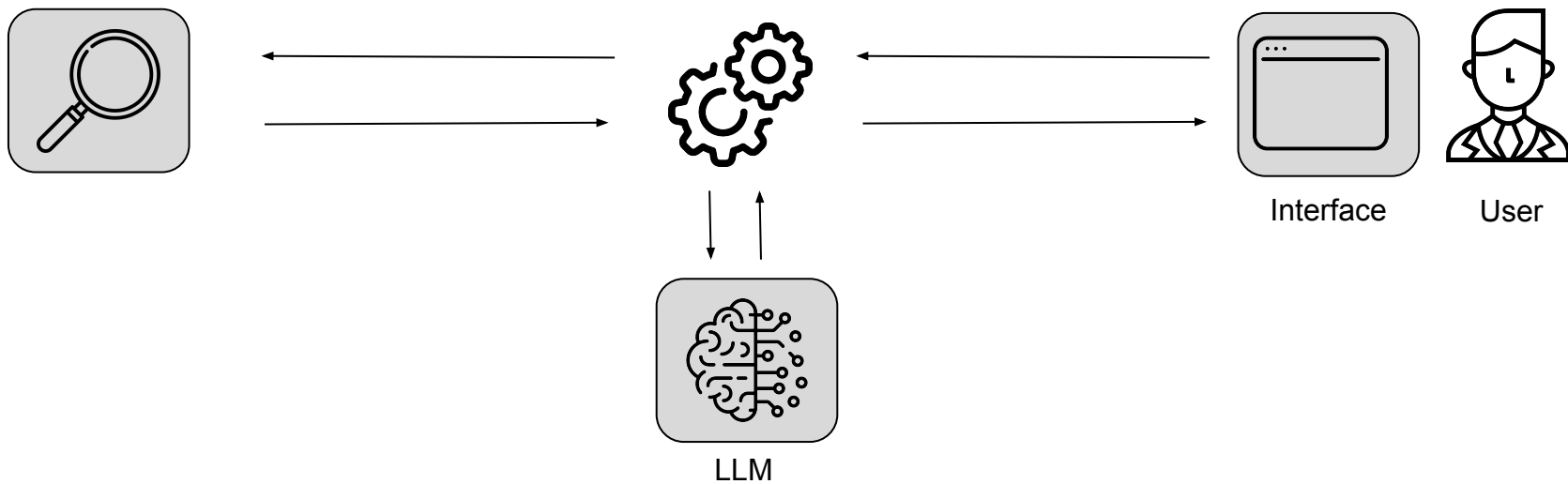
Retrieval Augmented Generation (RAG)

High-level Architecture View



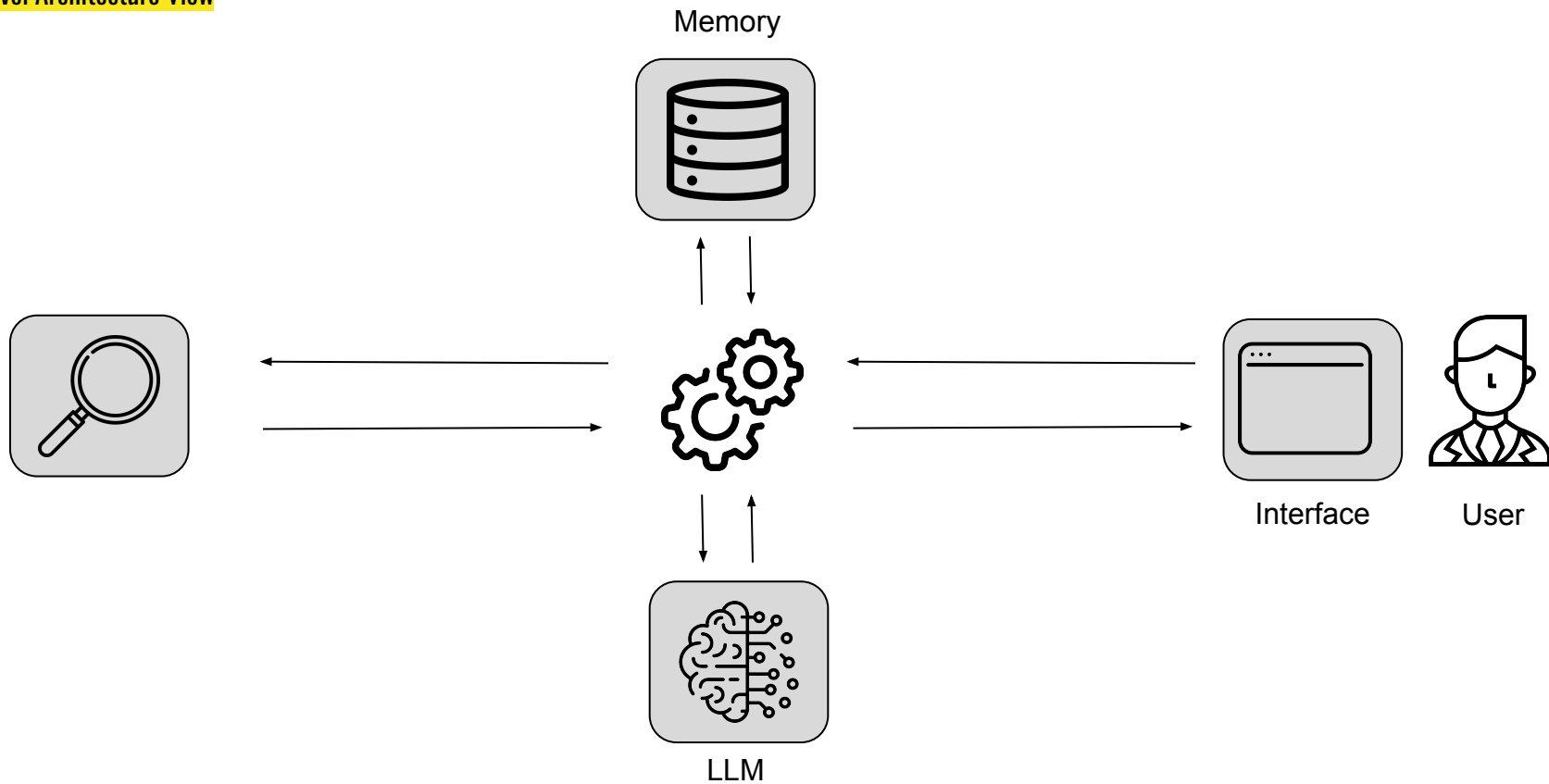
Retrieval Augmented Generation (RAG)

High-level Architecture View



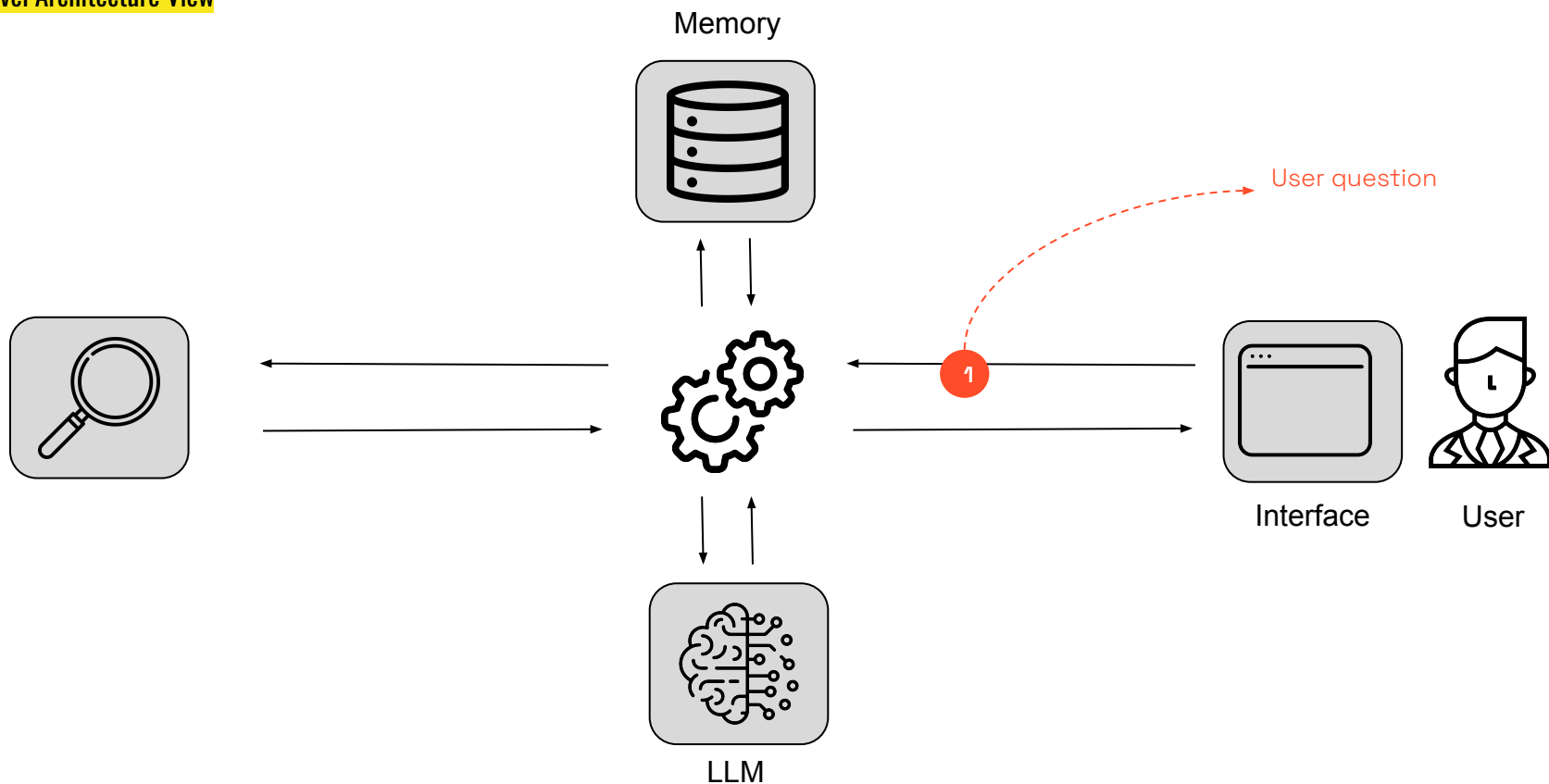
Retrieval Augmented Generation (RAG)

High-level Architecture View



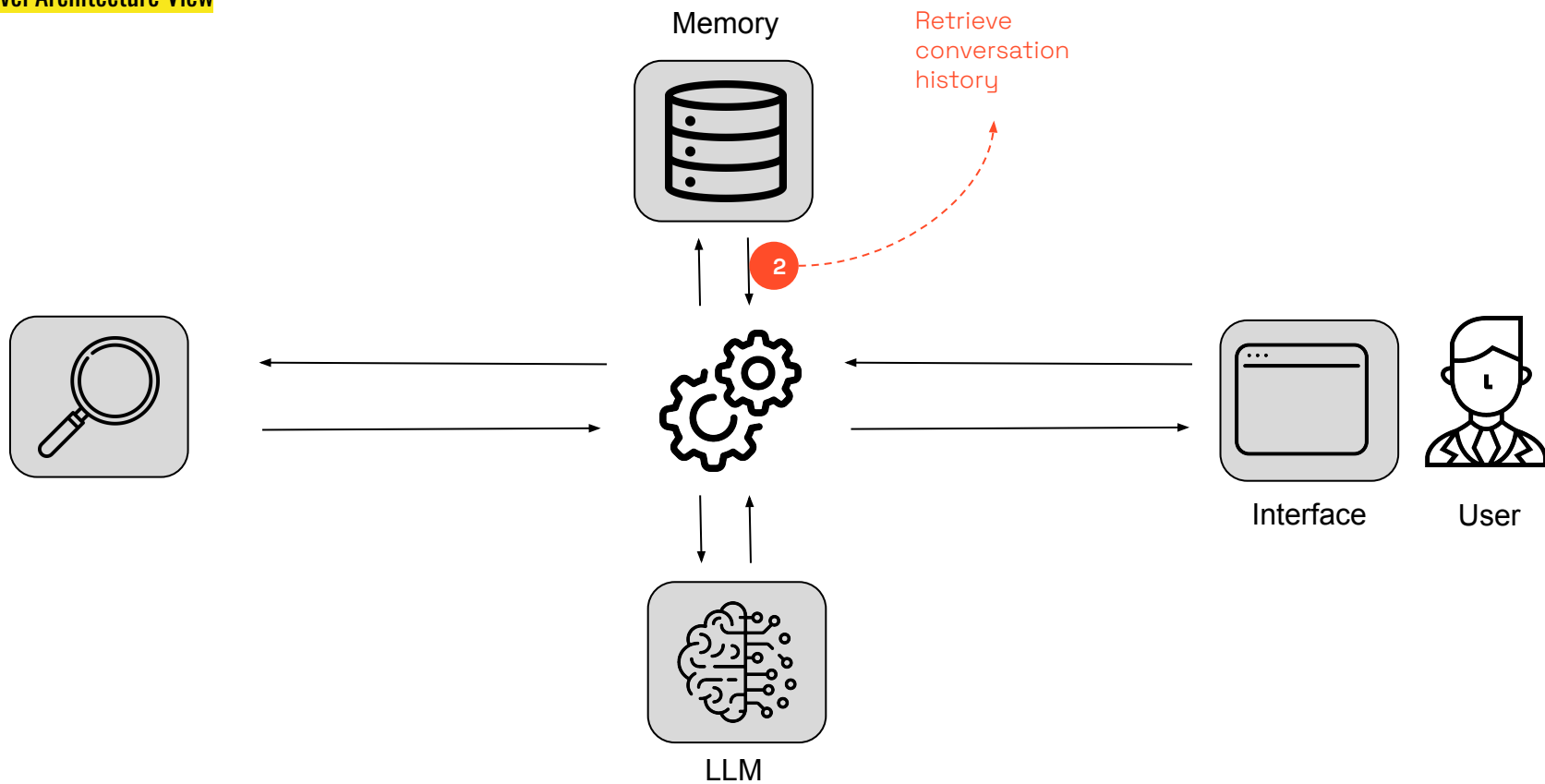
Retrieval Augmented Generation (RAG)

High-level Architecture View



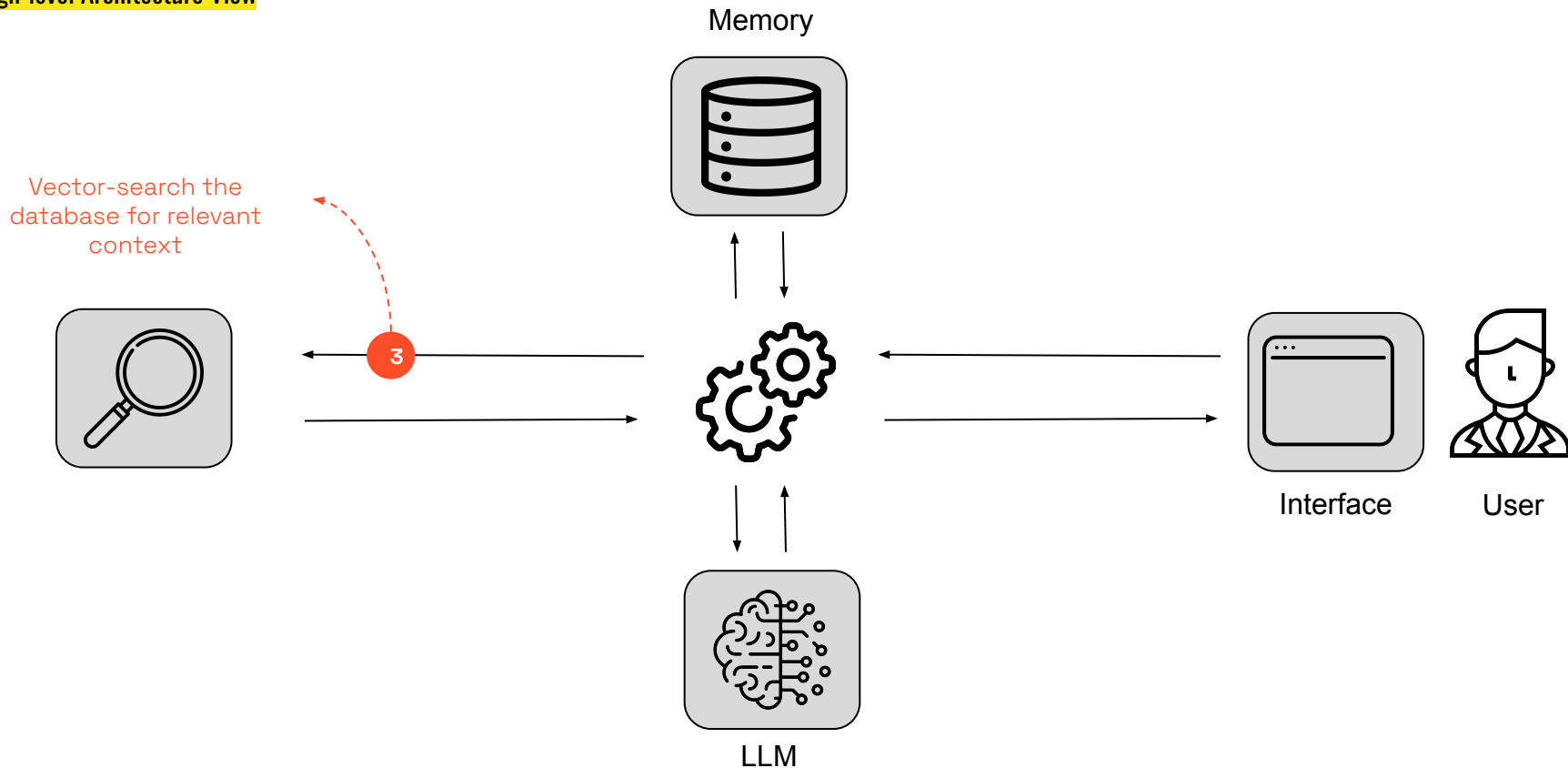
Retrieval Augmented Generation (RAG)

High-level Architecture View



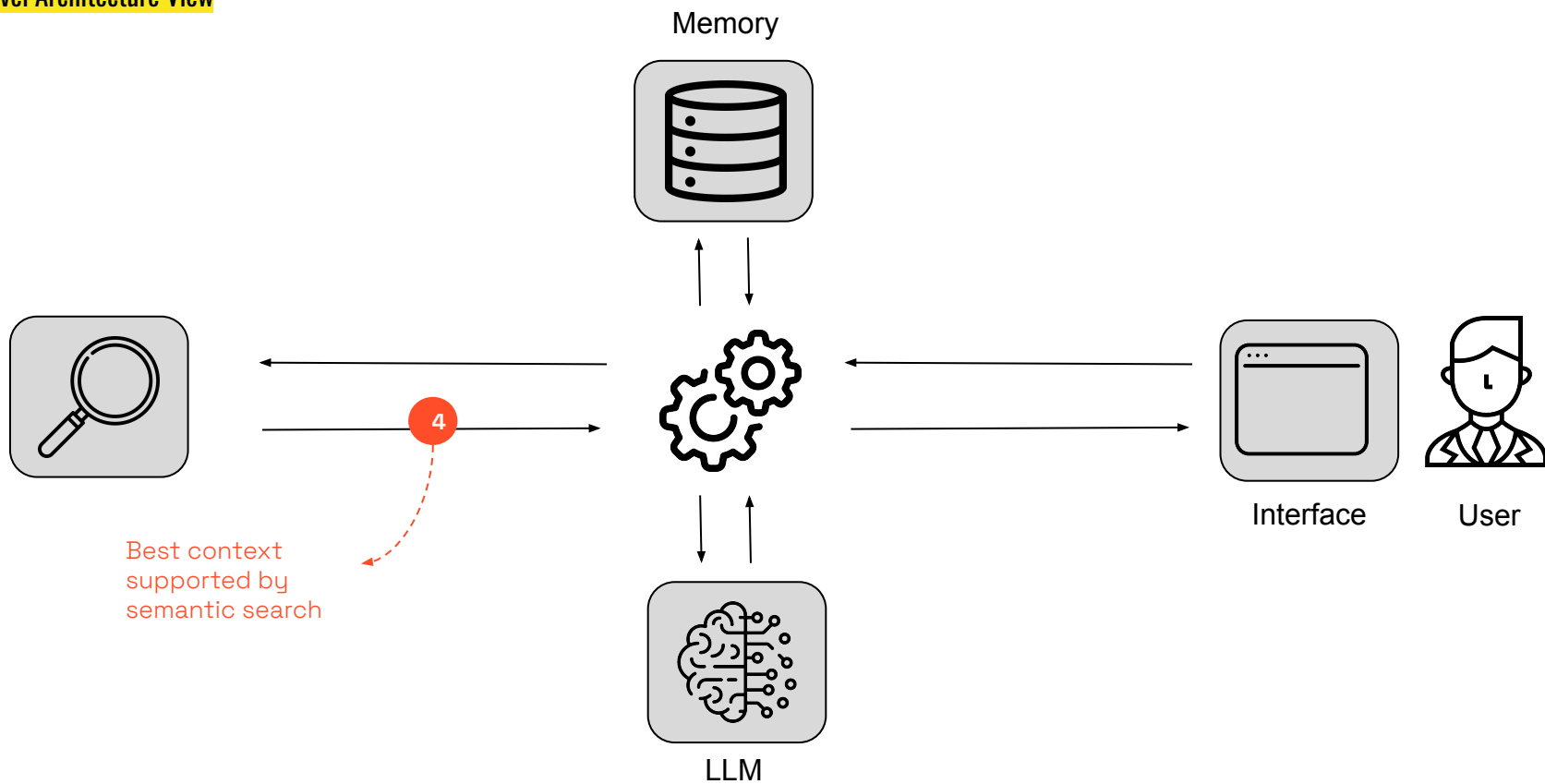
Retrieval Augmented Generation (RAG)

High-level Architecture View



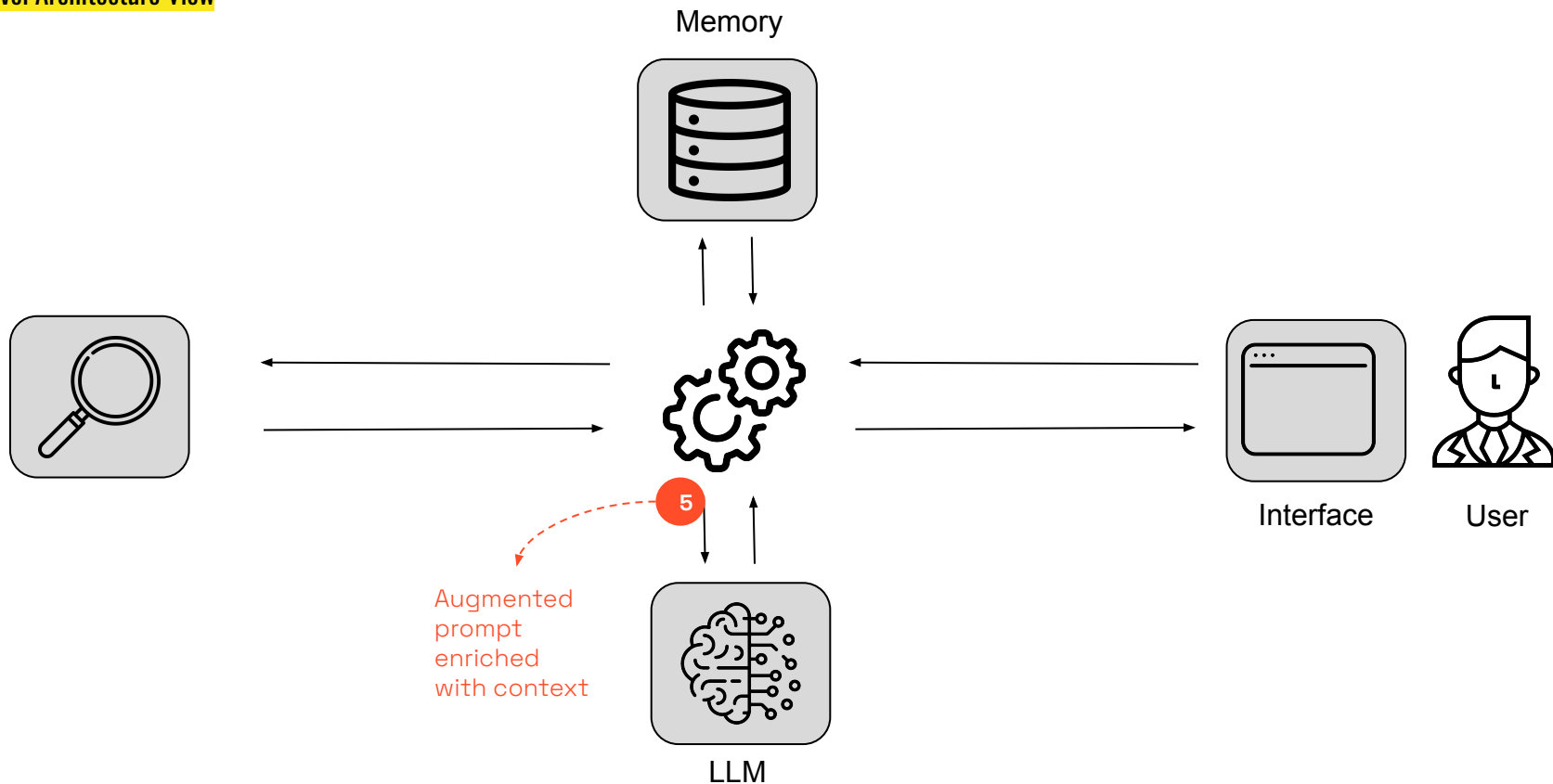
Retrieval Augmented Generation (RAG)

High-level Architecture View



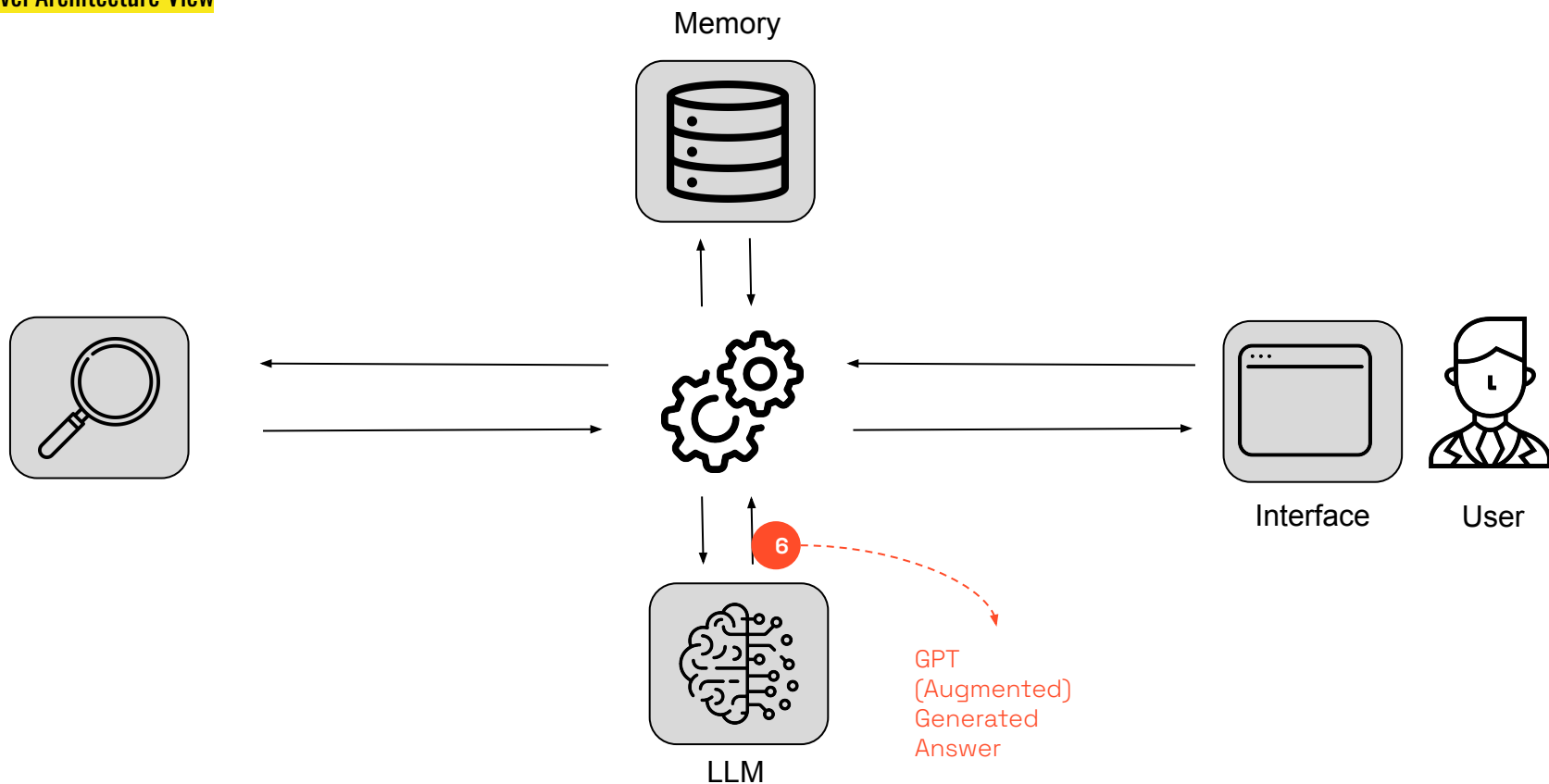
Retrieval Augmented Generation (RAG)

High-level Architecture View



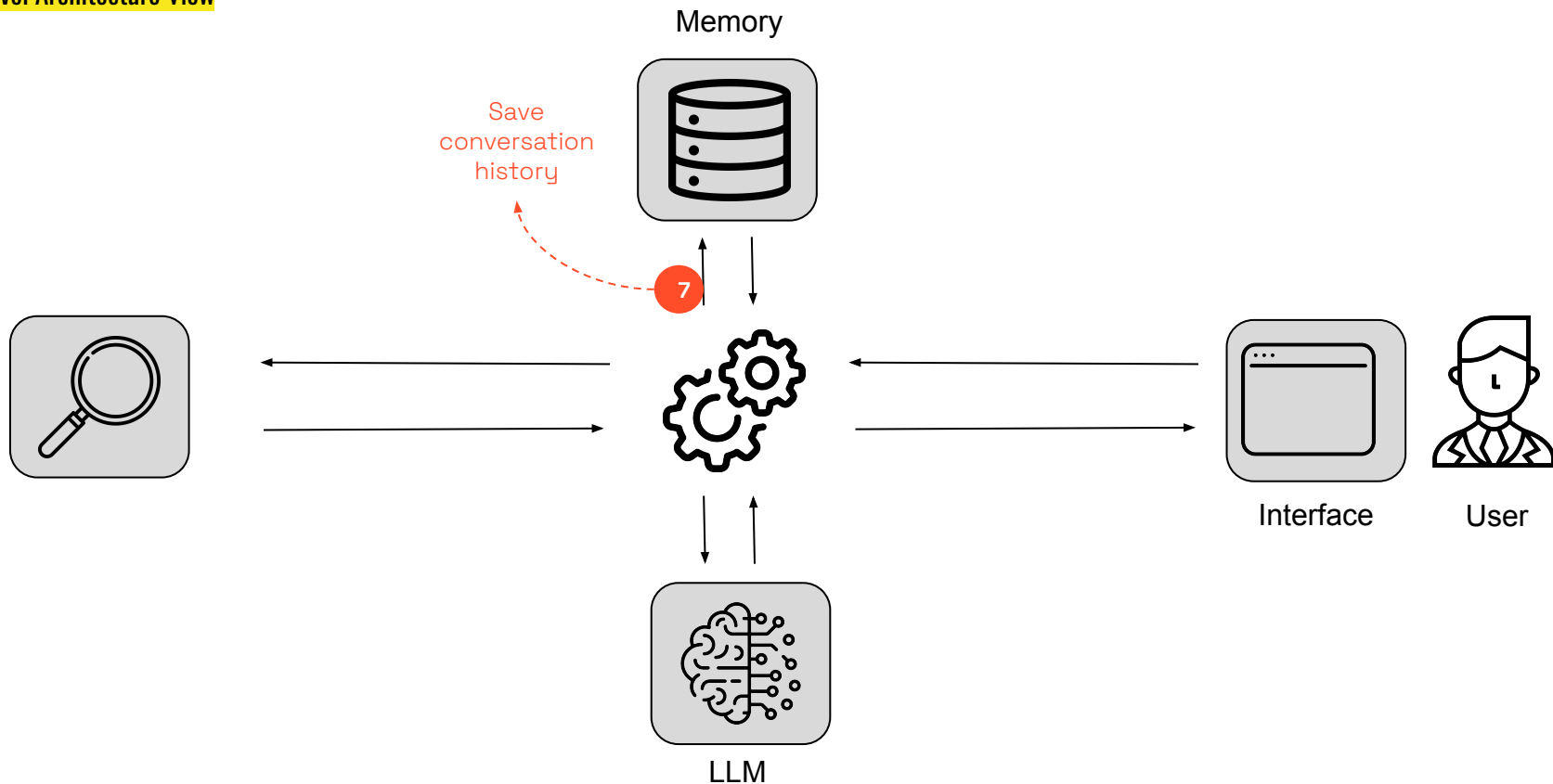
Retrieval Augmented Generation (RAG)

High-level Architecture View



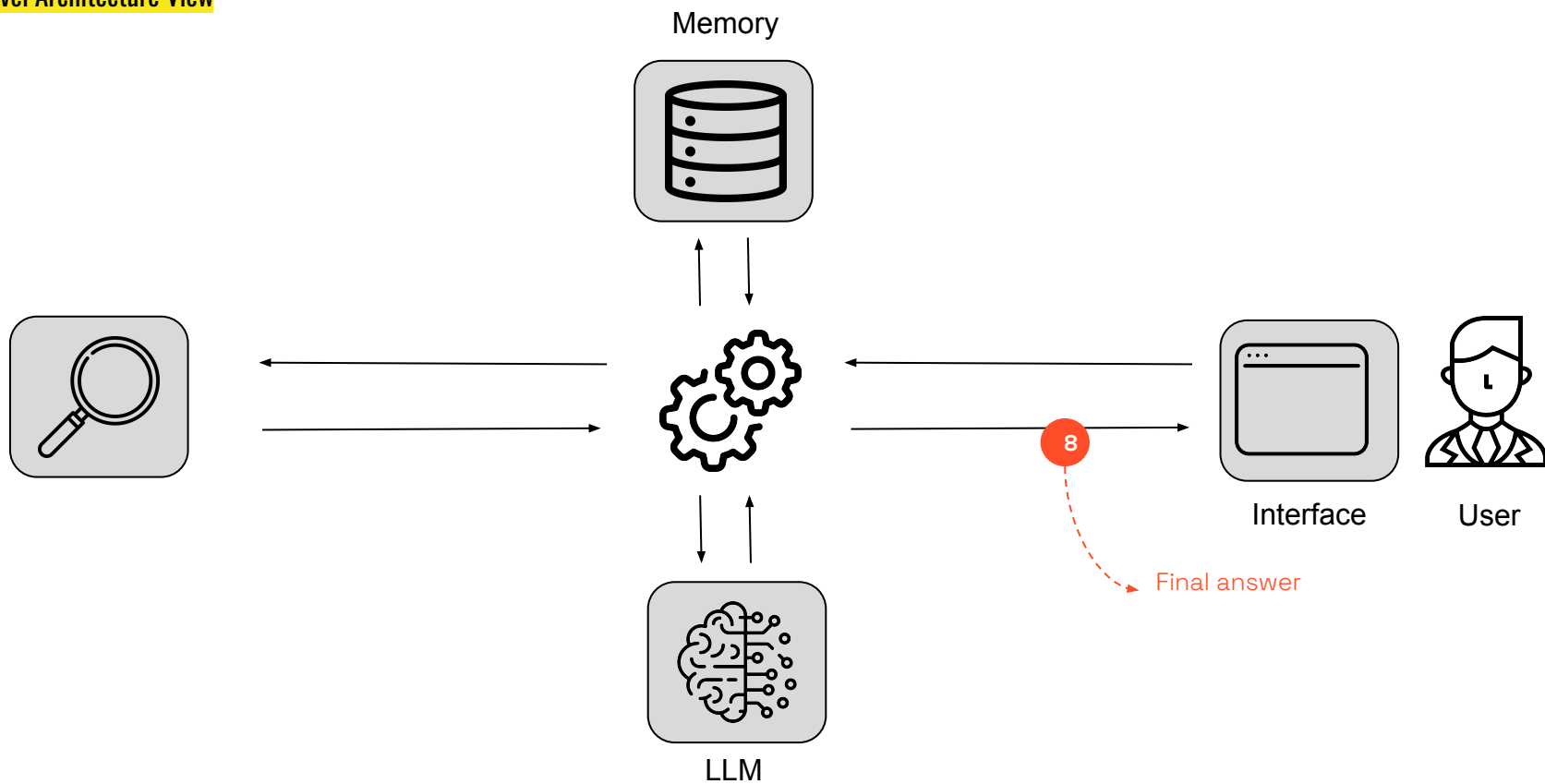
Retrieval Augmented Generation (RAG)

High-level Architecture View



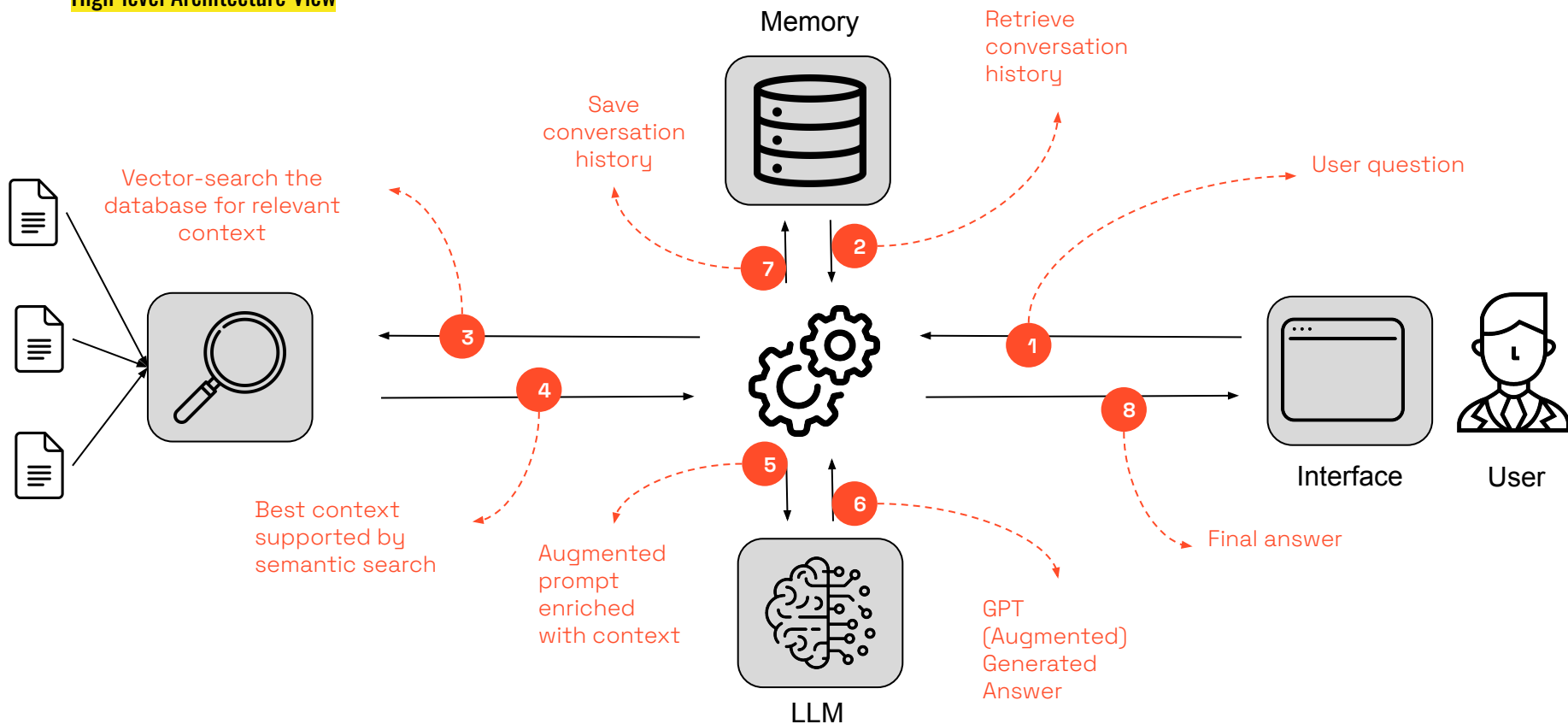
Retrieval Augmented Generation (RAG)

High-level Architecture View



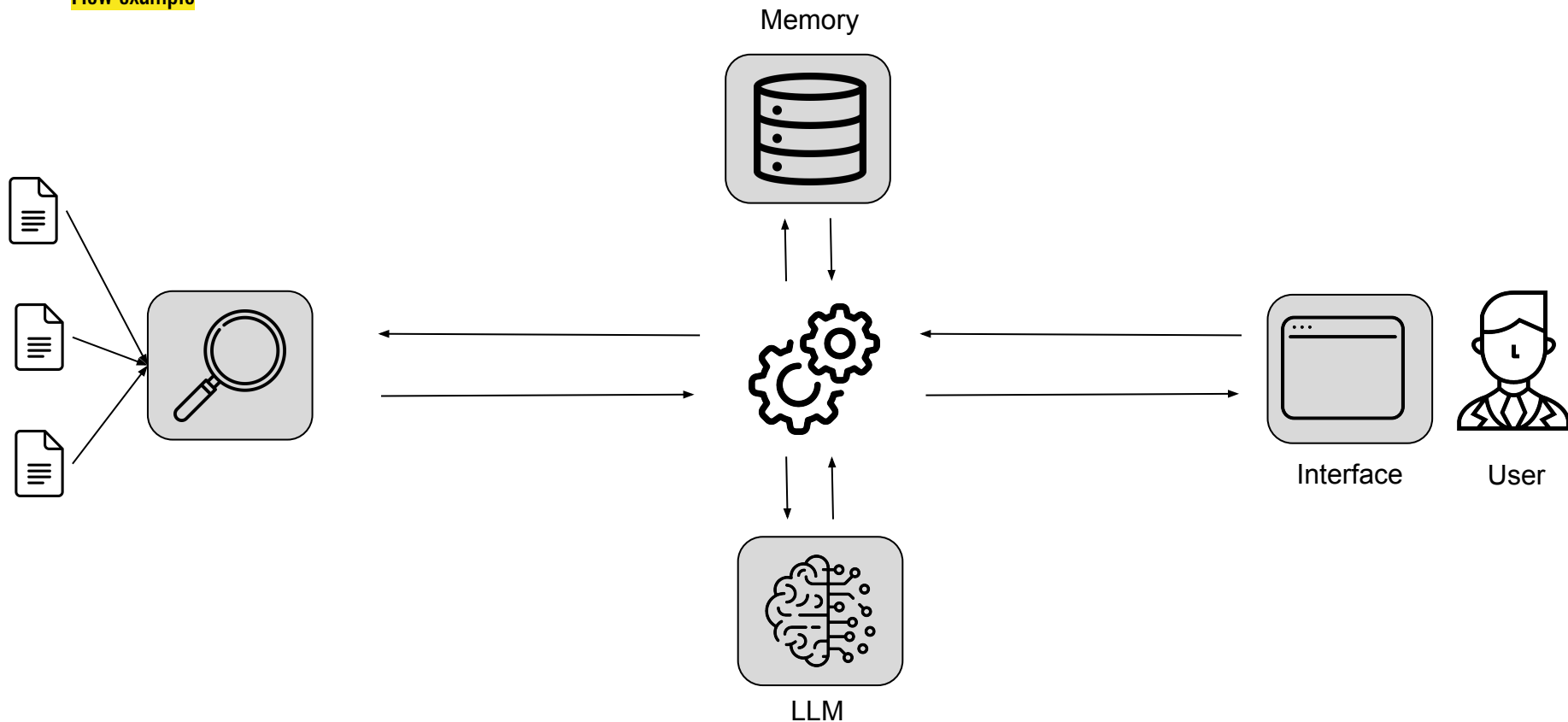
Retrieval Augmented Generation (RAG)

High-level Architecture View



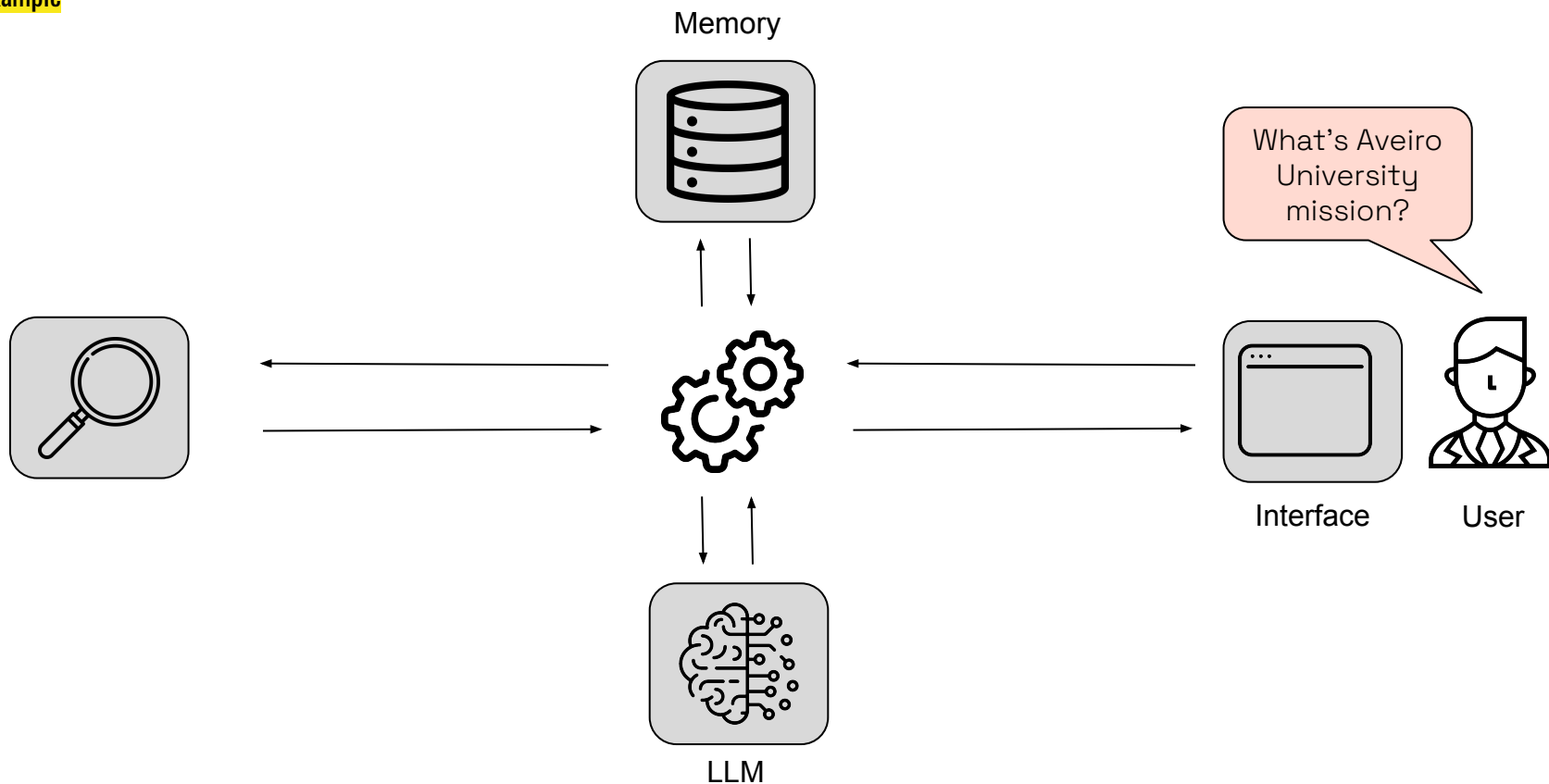
Retrieval Augmented Generation (RAG)

Flow example



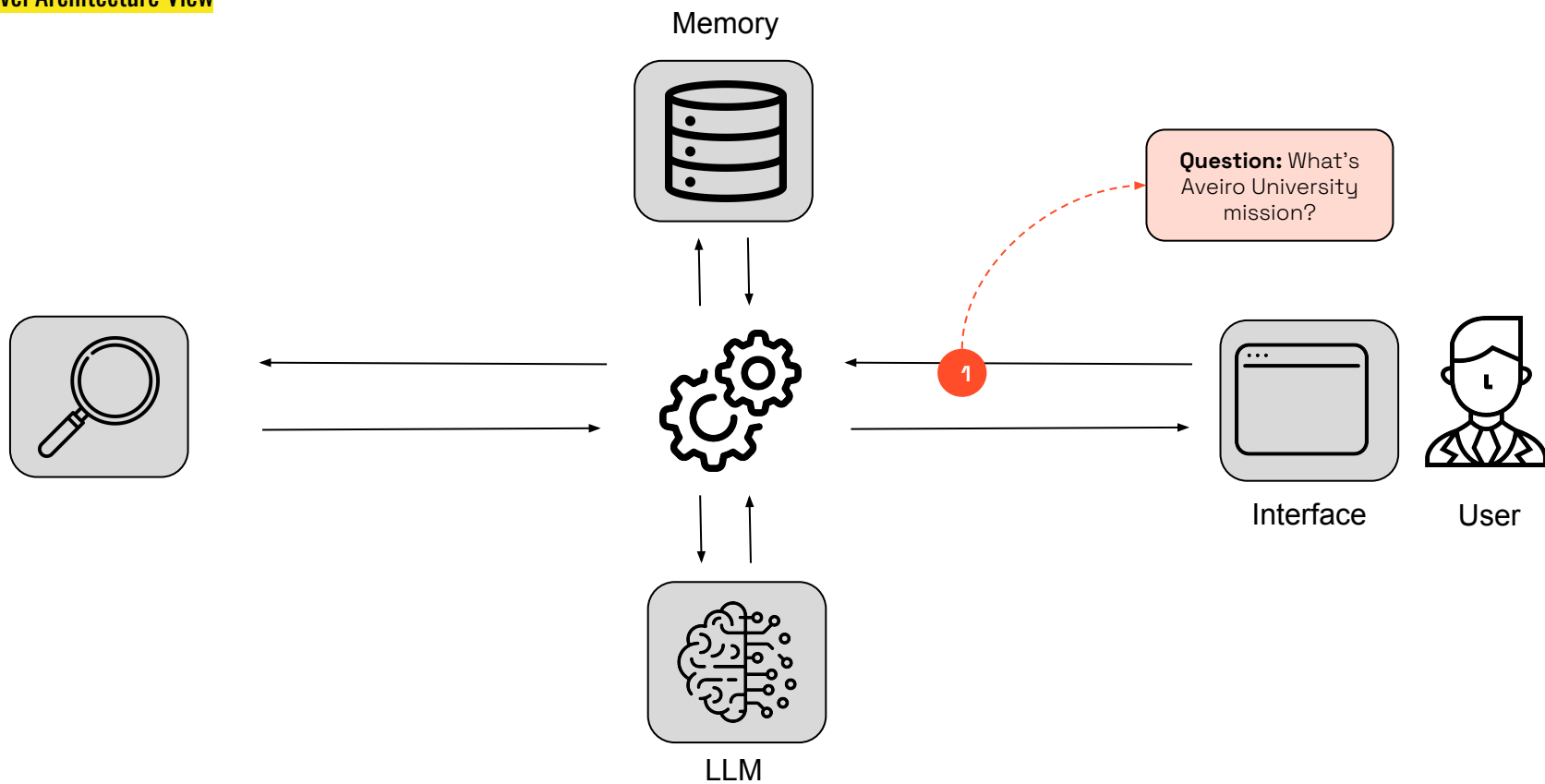
Retrieval Augmented Generation (RAG)

Flow example



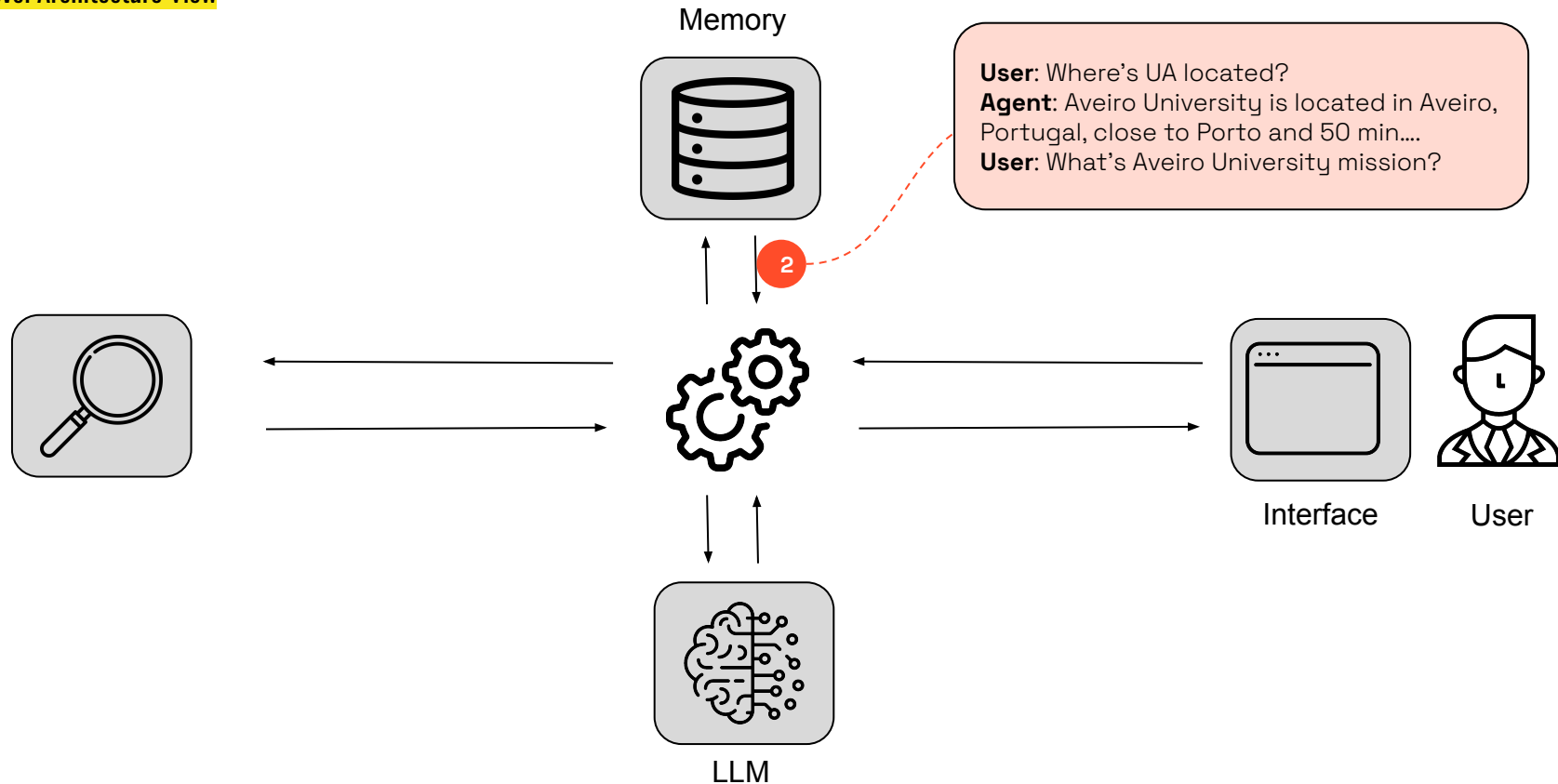
Retrieval Augmented Generation (RAG)

High-level Architecture View



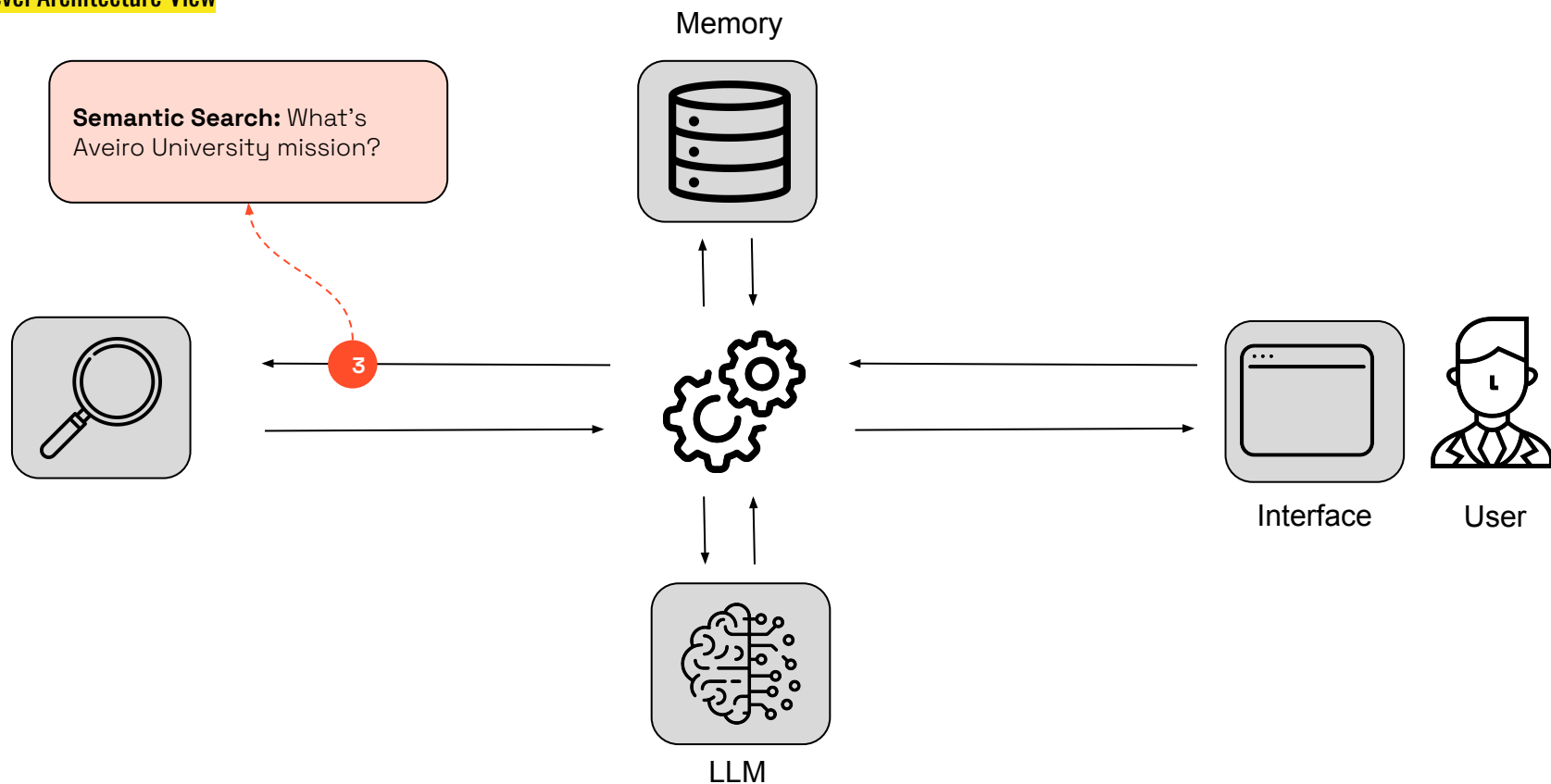
Retrieval Augmented Generation (RAG)

High-level Architecture View



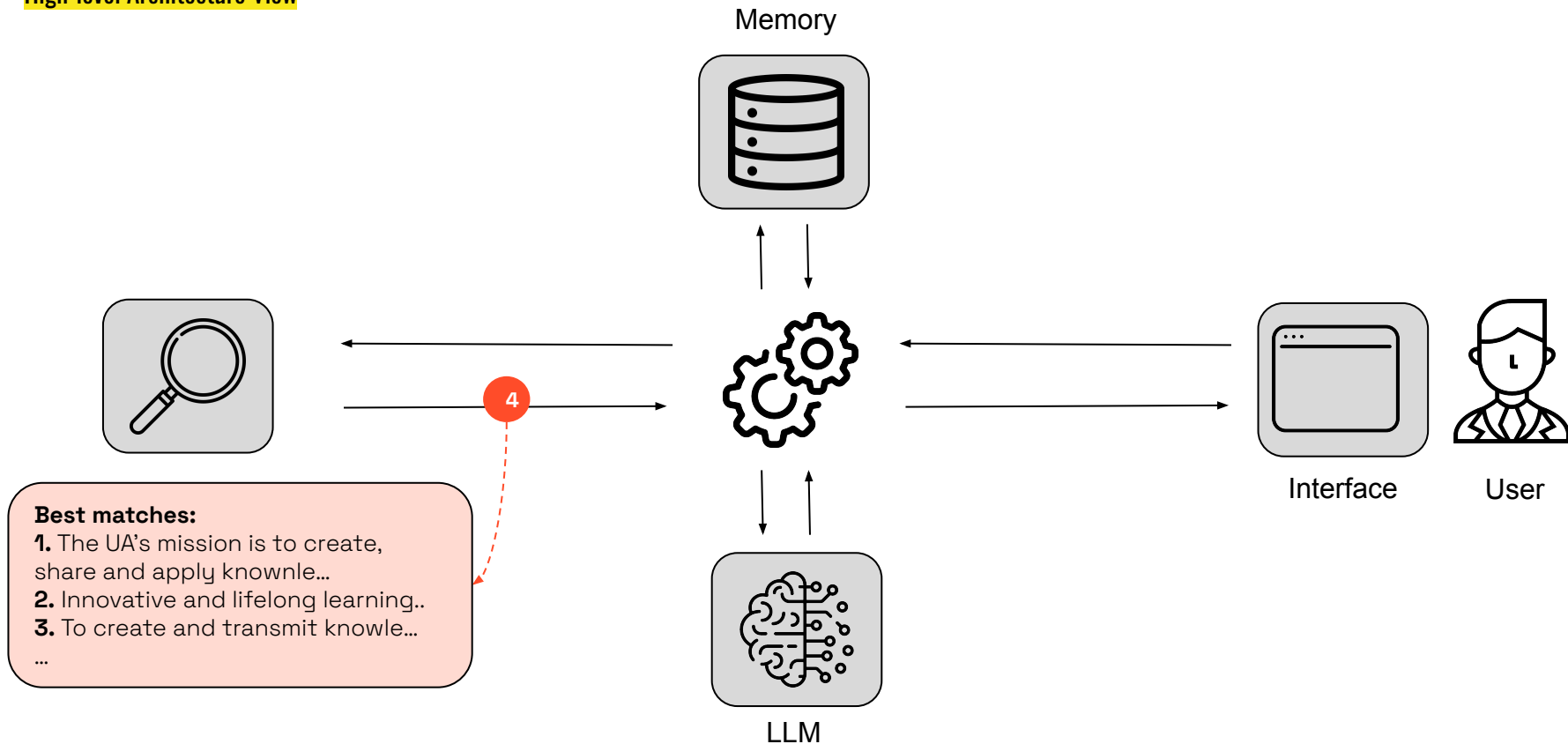
Retrieval Augmented Generation (RAG)

High-level Architecture View



Retrieval Augmented Generation (RAG)

High-level Architecture View



Retrieval Augmented Generation (RAG)

High-level Architecture View

You're a AI agent having a conversation with a user trying to present company facts and success stories.

Use the facts below and if no answer is possible say that you don't know.

Facts

1. The UA's mission is to create, share and apply knowle...
2. Innovative and lifelong learning..
3. To create and transmit knowle...

History:

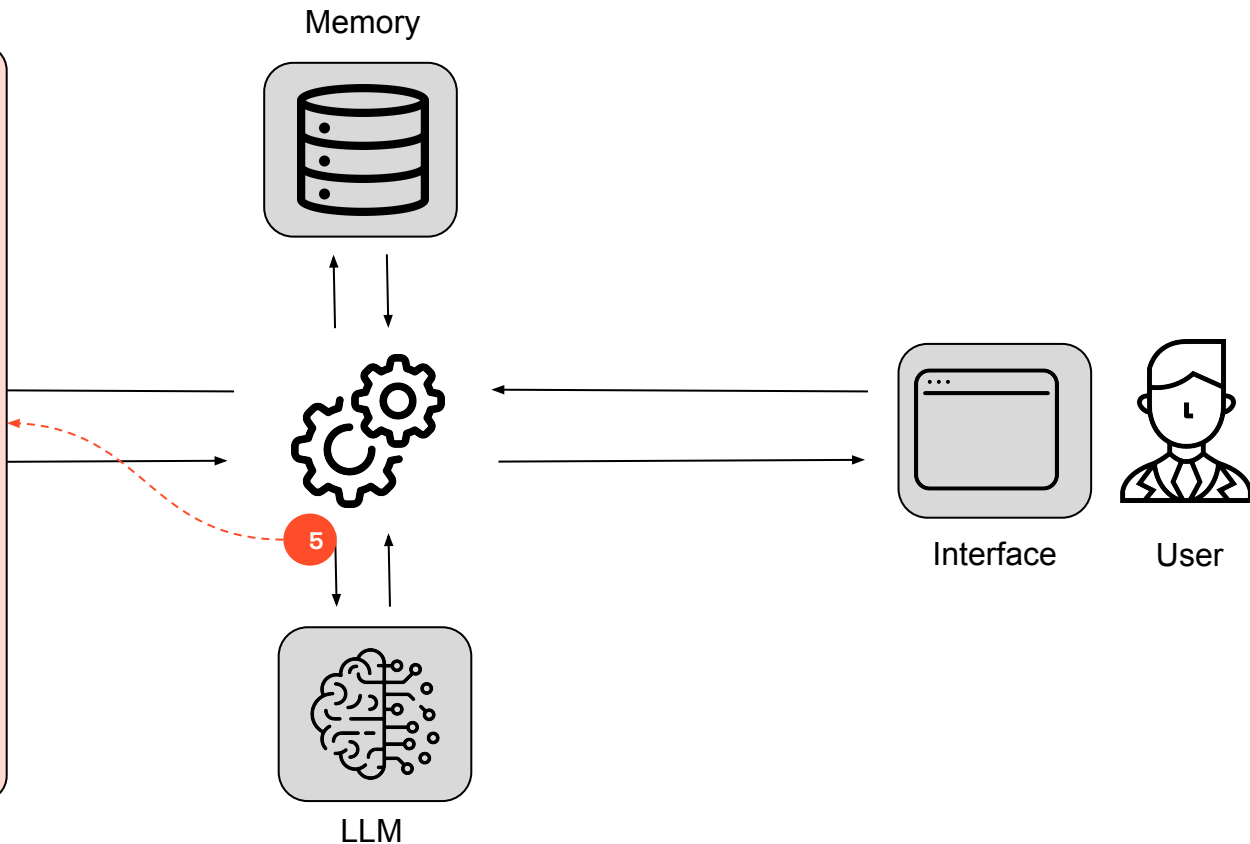
User: Where's UA located?

Agent: Aveiro University...

User: What's Aveiro University mission?

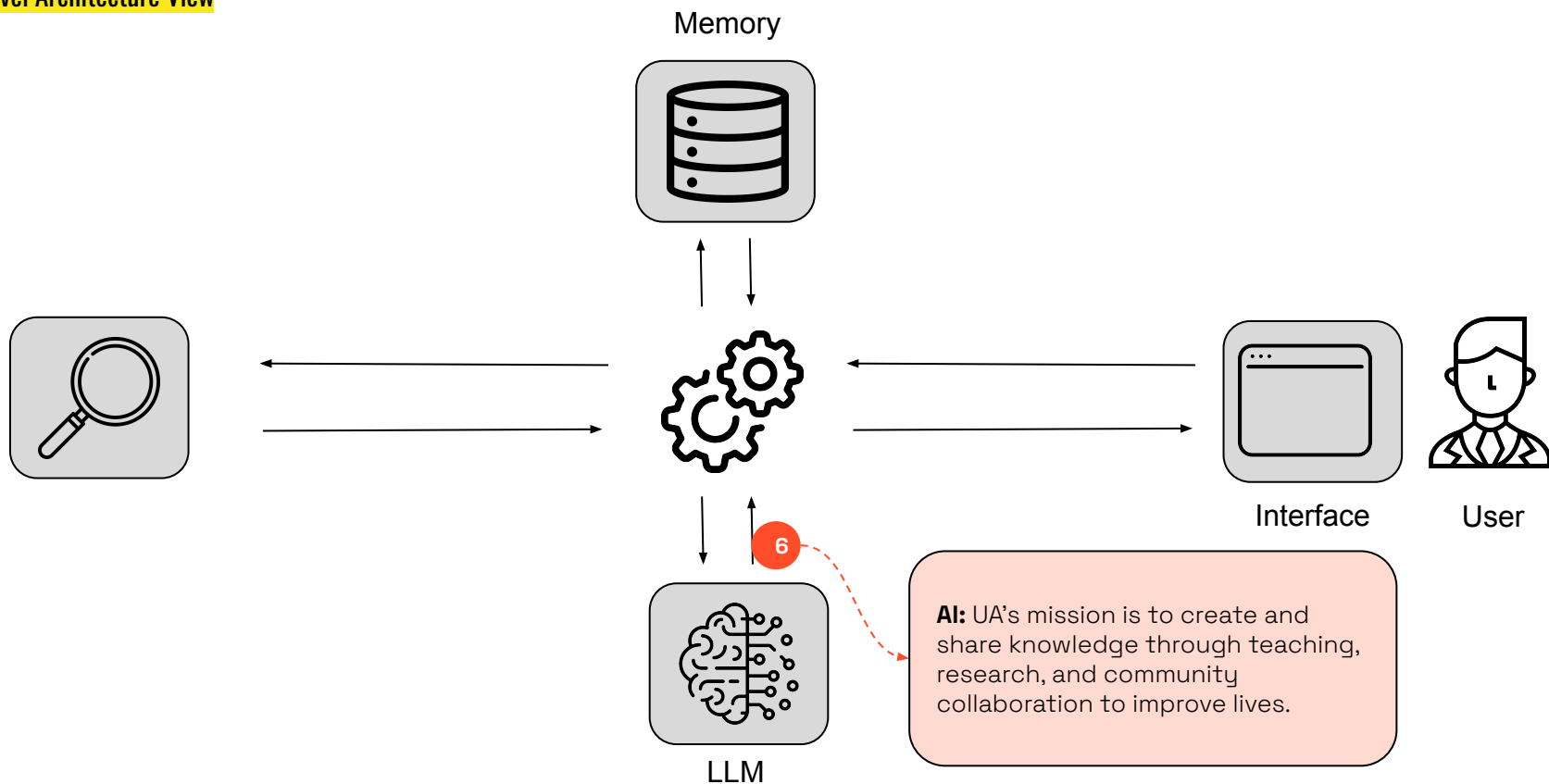
User: What's Aveiro University mission?

AI:



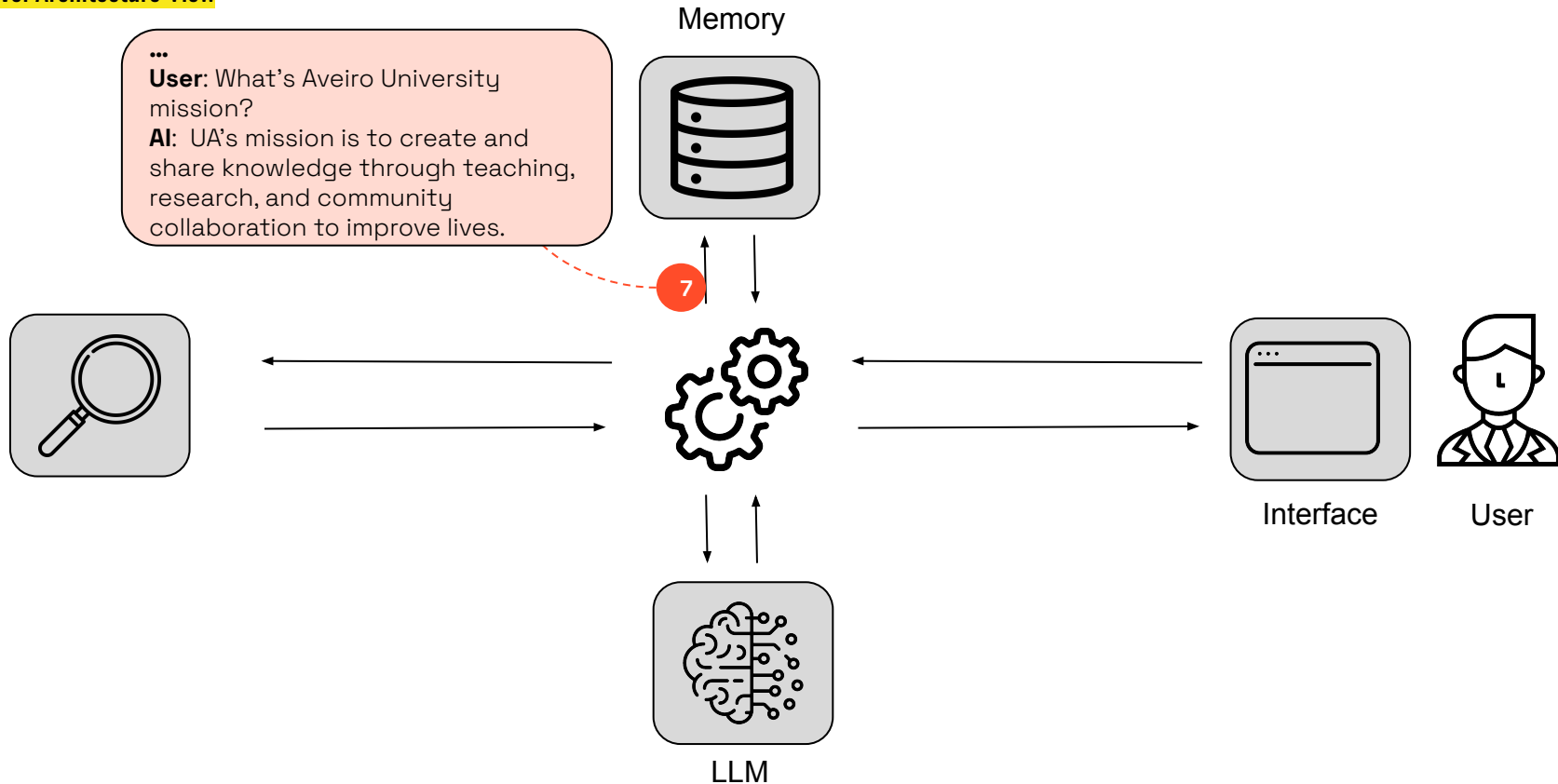
Retrieval Augmented Generation (RAG)

High-level Architecture View



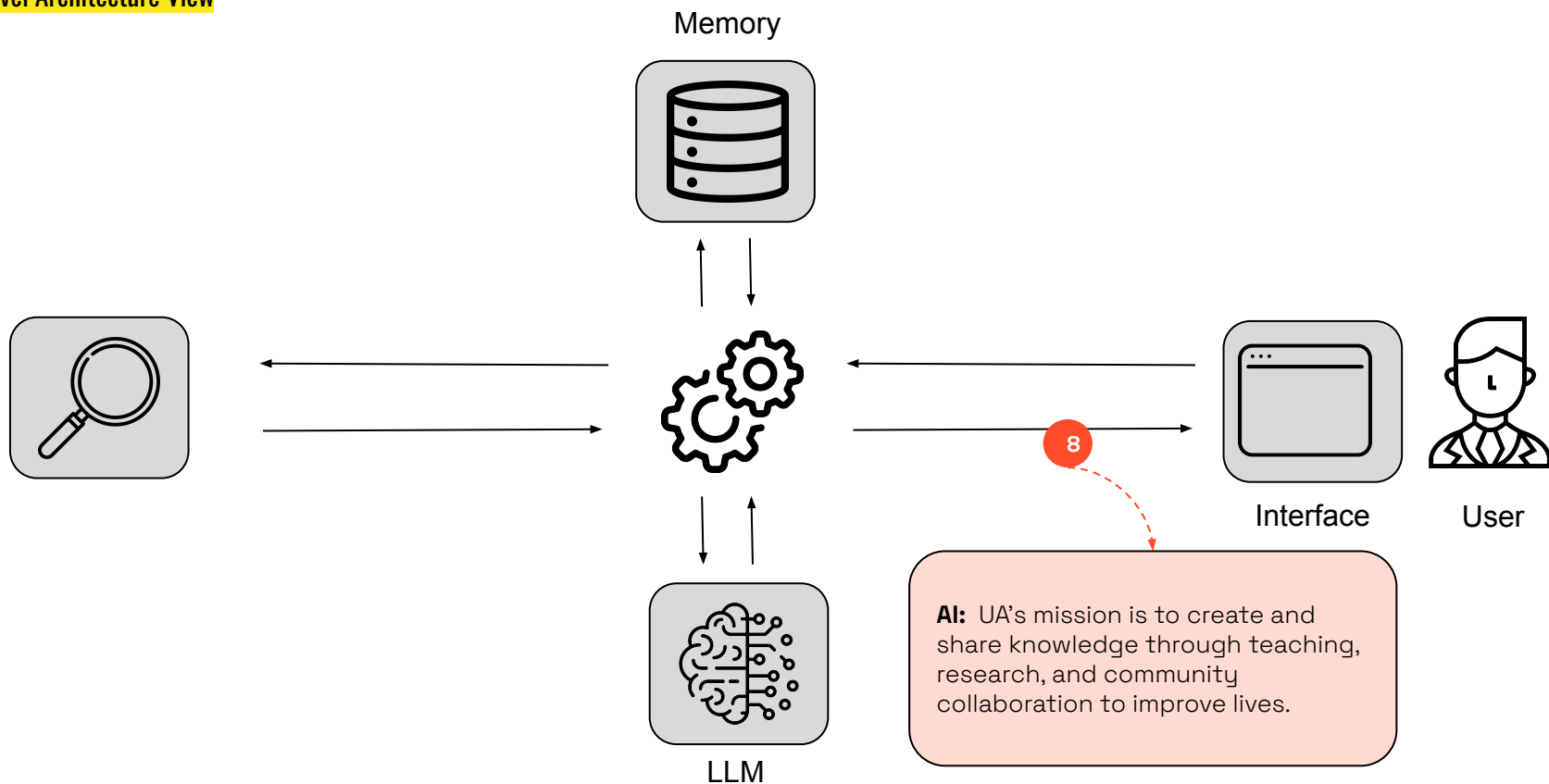
Retrieval Augmented Generation (RAG)

High-level Architecture View



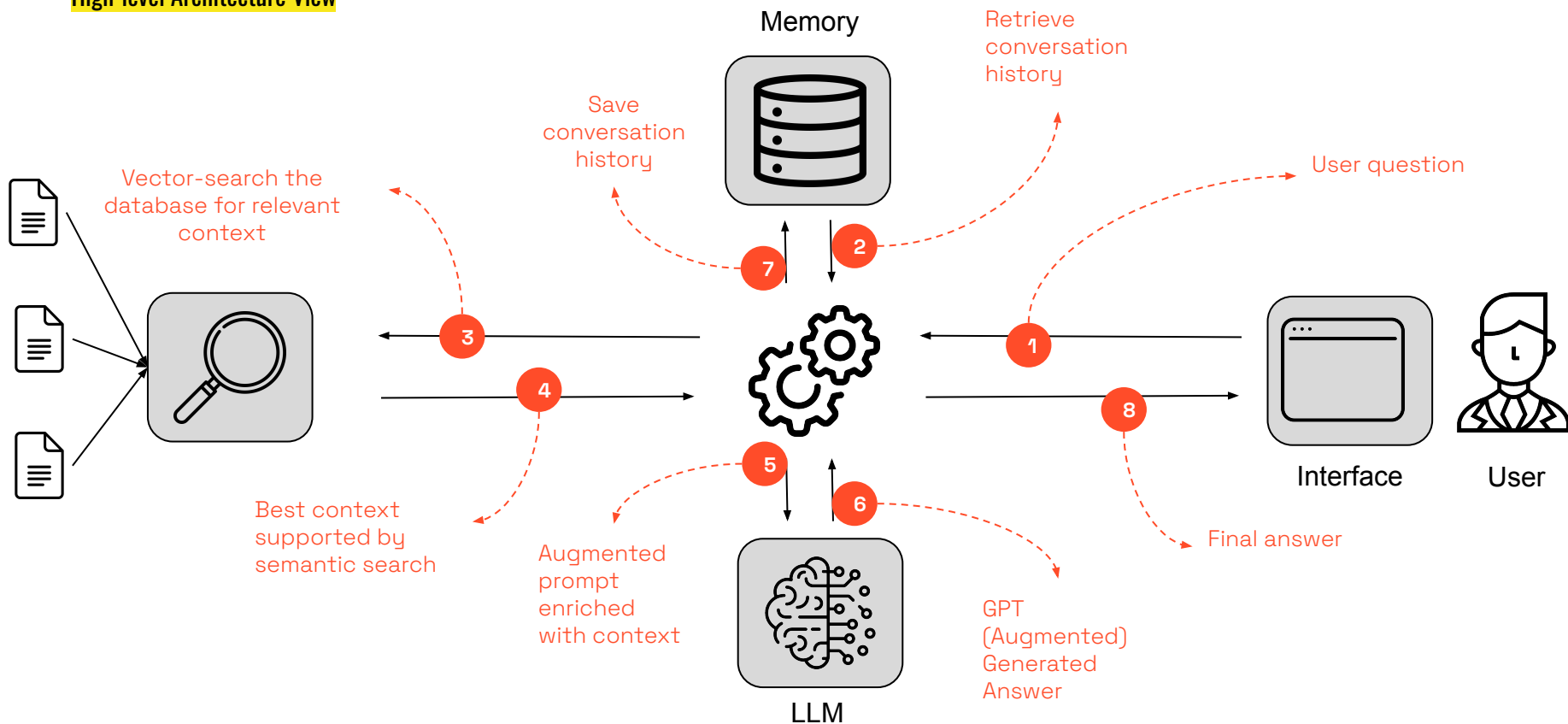
Retrieval Augmented Generation (RAG)

High-level Architecture View



Retrieval Augmented Generation (RAG)

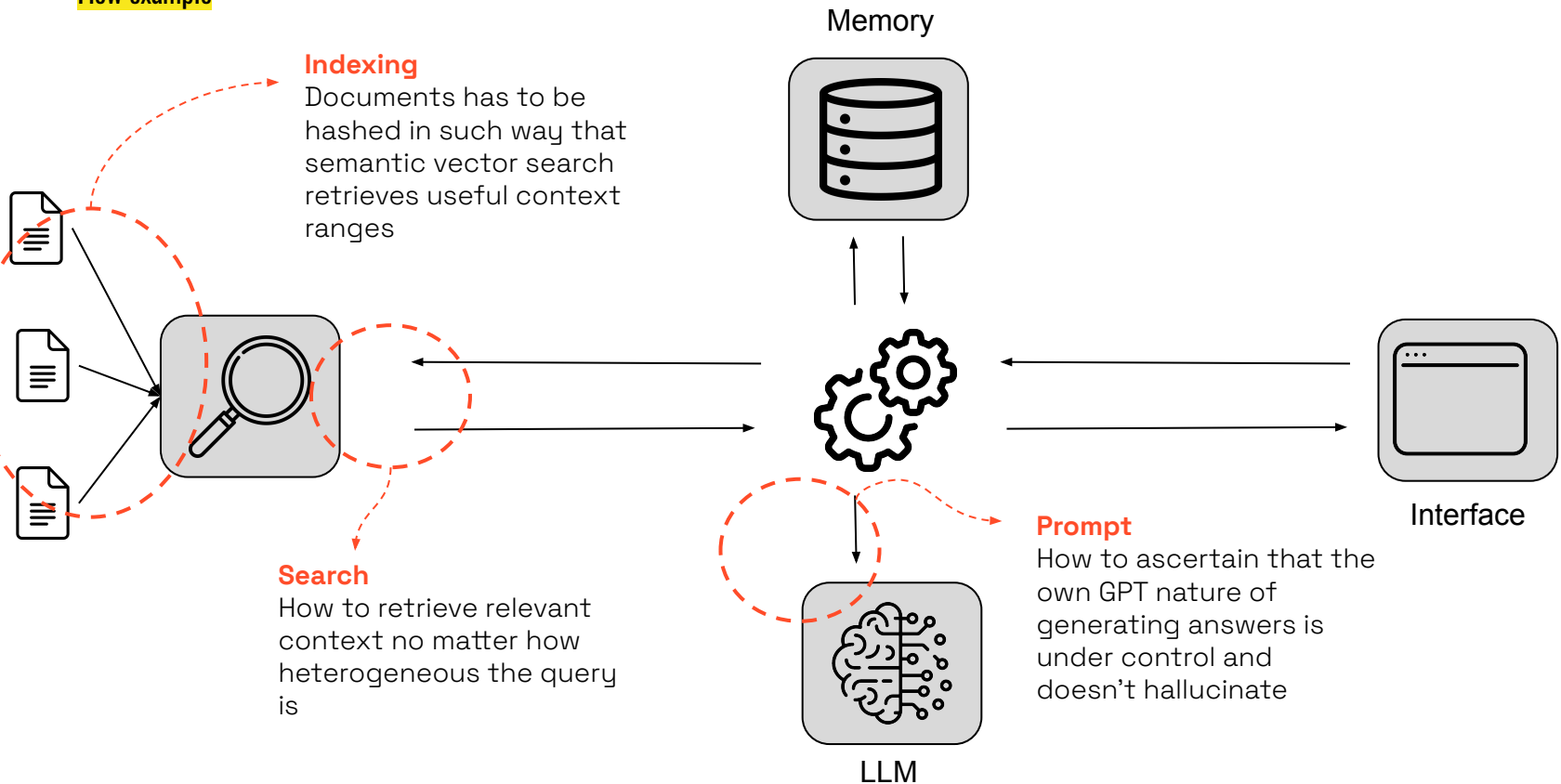
High-level Architecture View



Main RAG Challenges

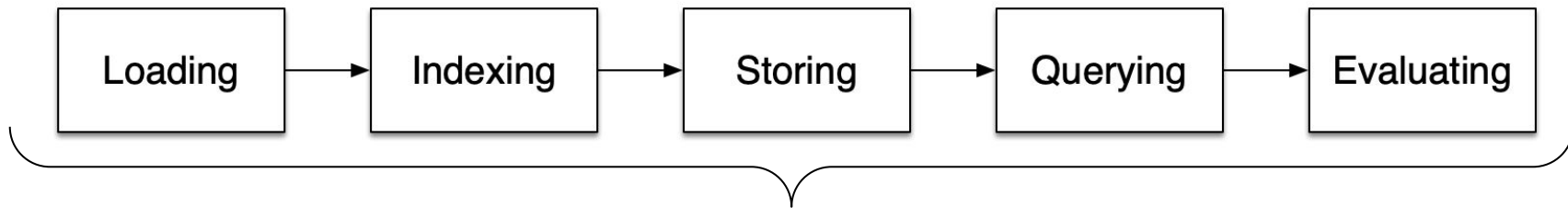
Retrieval Augmented Generation (RAG)

Flow example



Hands on!

LlamaIndex



LlamaIndex

Why LlamaIndex?



- Supports the whole chain
- Opensource

Has interfaces for:

- > 40 vector stores
- > 40 LLMs
- > 160 data sources

Alternatives:

- Haystack
- RAGFlow
- Graphlit
- Llangchain (kinda)

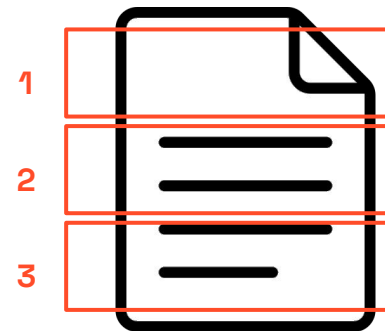
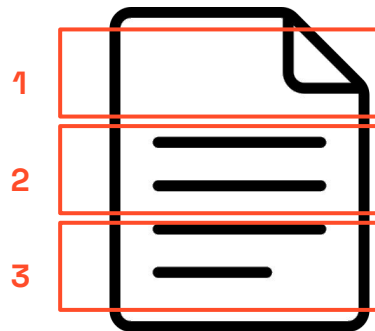
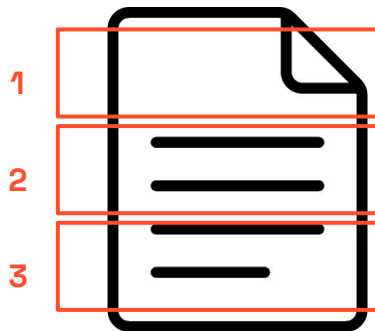
Notebook

- <https://github.com/luminoso/dspt-handson-llamaindex>



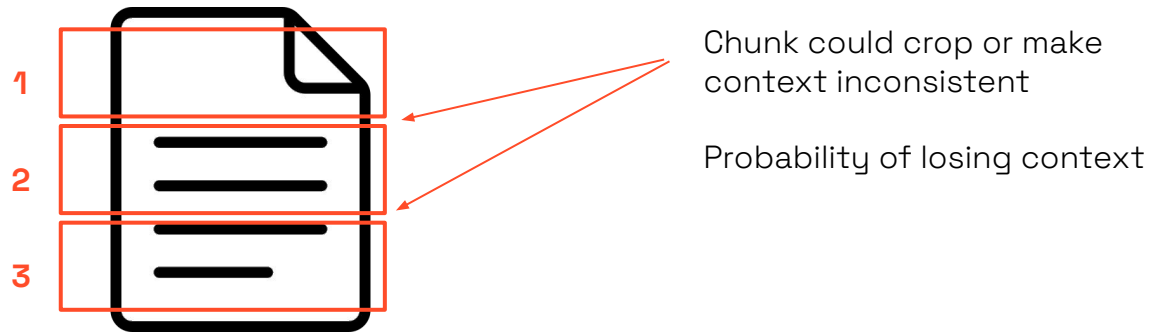
Chunking

Document chunking



One paragraph per chunk

Document chunking

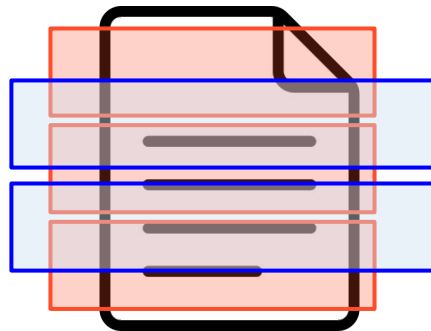


Document chunking



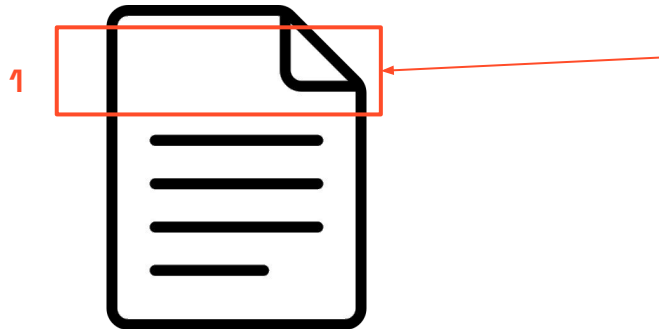
Overlapping chunks

Sliding the document with overlaps allows context preservation



Embedding

Embeddings

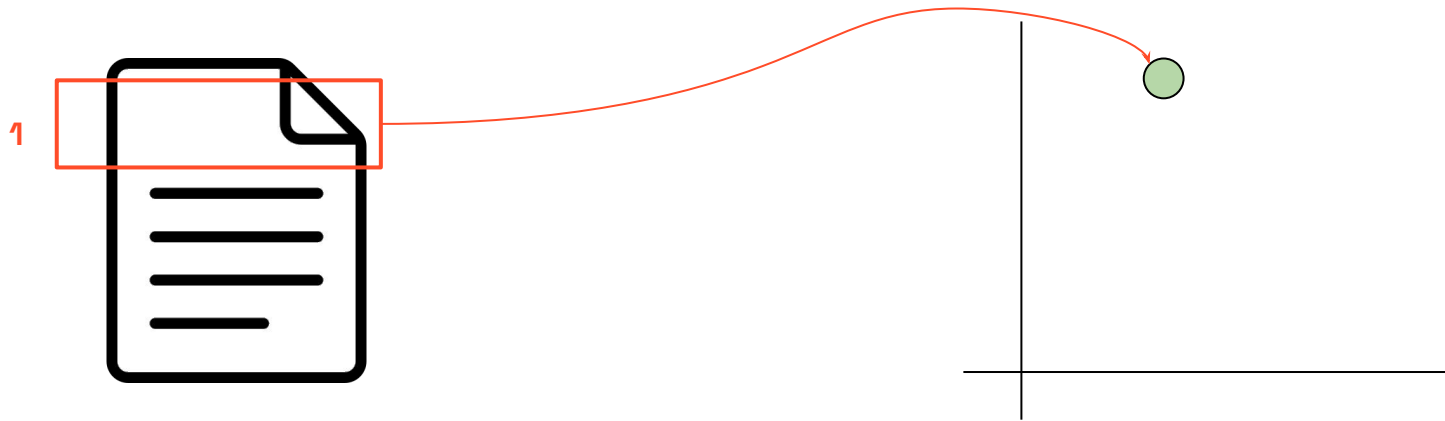


Created in 1976 by architect Firmino Trabulo, the original signature of the University of Aveiro incorporates several symbols (a griffon, a book, the armillary sphere, and the Greek words "theoria", "poiesis", "praxis") that personify the importance that the University of Aveiro has attributed, since its foundation, to the connection to the region; to the defense of wisdom, in the teaching and research aspects; to the universality of knowledge; and to the various aspects of theoretical, technological, artistic and humanistic crea...

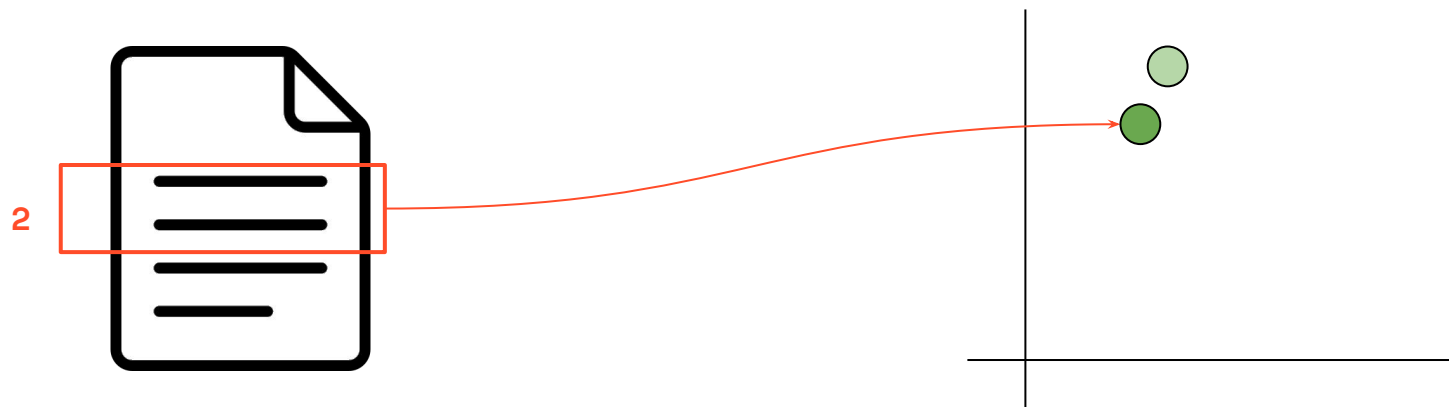
Text meaning is translate into a compressed vector
that represents its meaning

`[-0.0032757148146629333, -0.011690735816955566, 0.041559211909770966,
-0.03814808651804924, ...]`

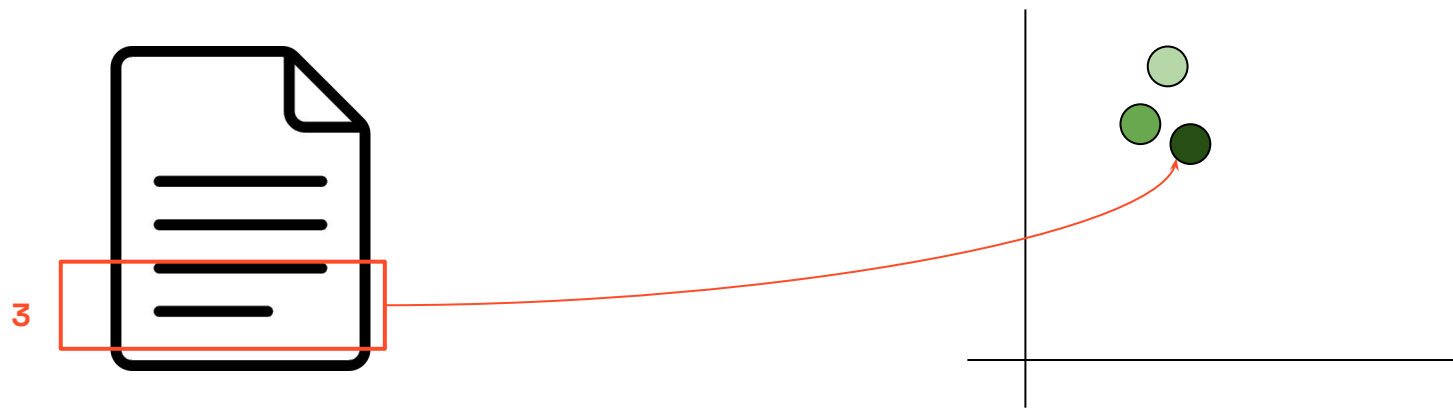
Embeddings



Embeddings



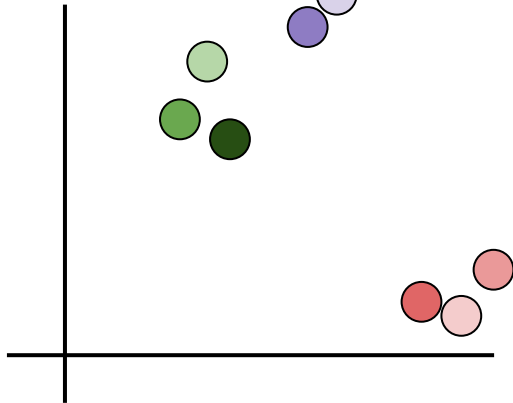
Embeddings



Embeddings for multiple documents



Document about
UA mission

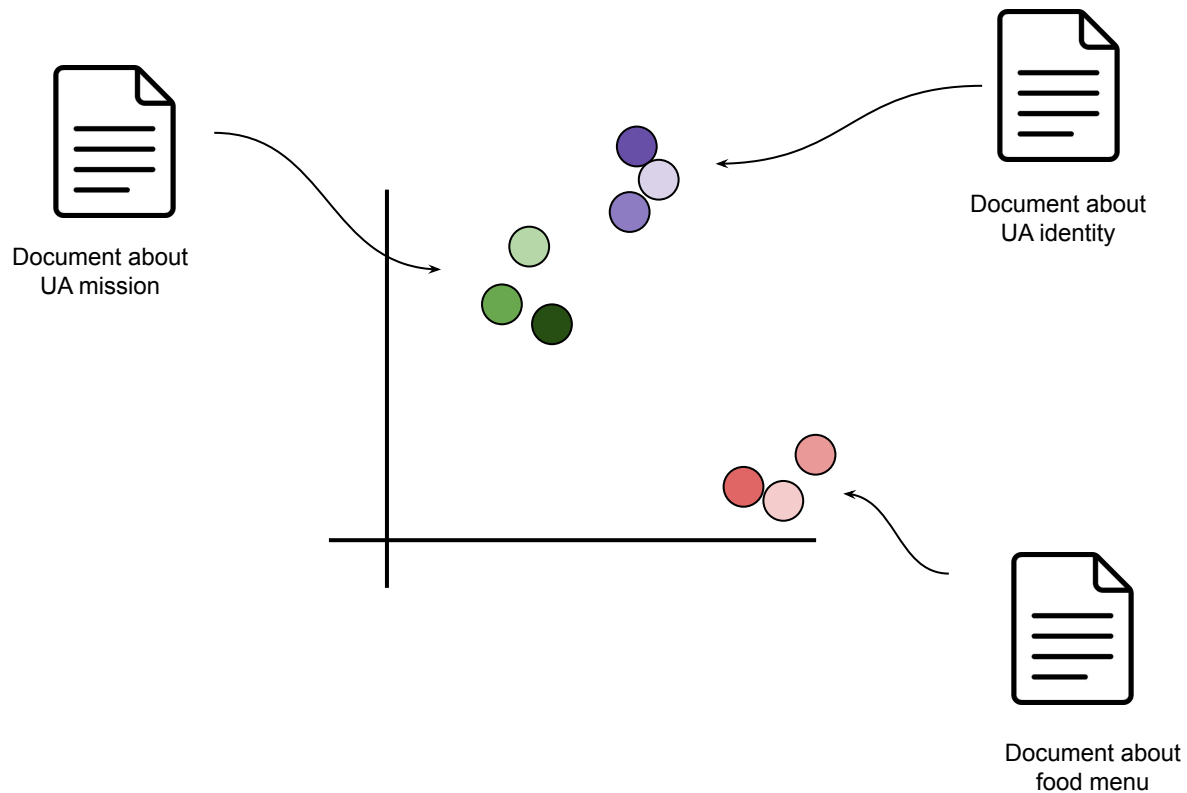


Document about
UA identity

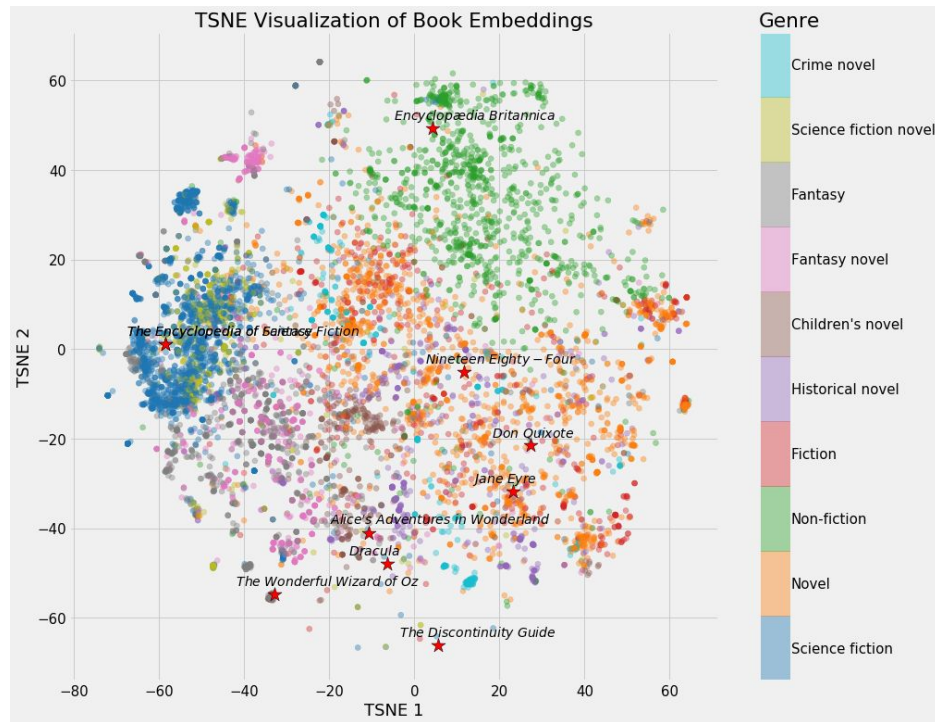


Document about
food menu

Embeddings for multiple documents

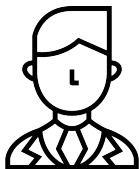


Embeddings real world example



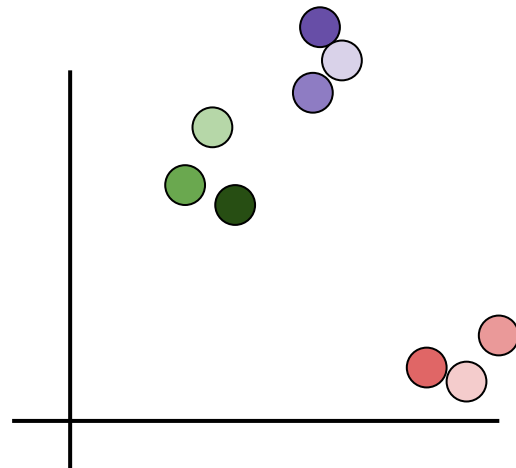
Querying

Embeddings for multiple documents

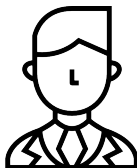


Your question

***“What’s in the food menu for
friday dinner?”***



Embeddings for multiple documents

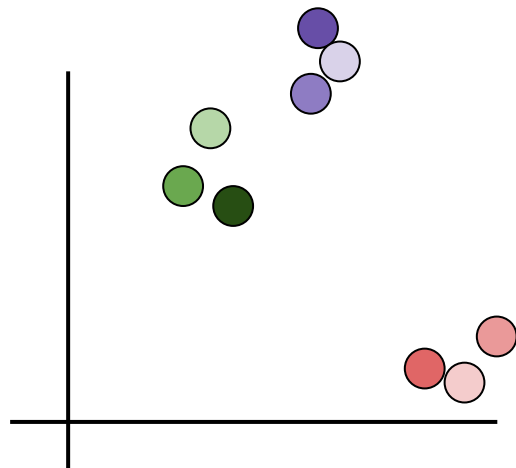


Your question

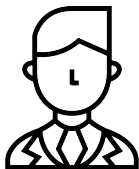
***“What’s in the food menu for
friday dinner?”***



```
[-0.0032757148146629333, -0.011690735816955566,  
0.041559211909770966, -0.03814808651804924, ...]
```



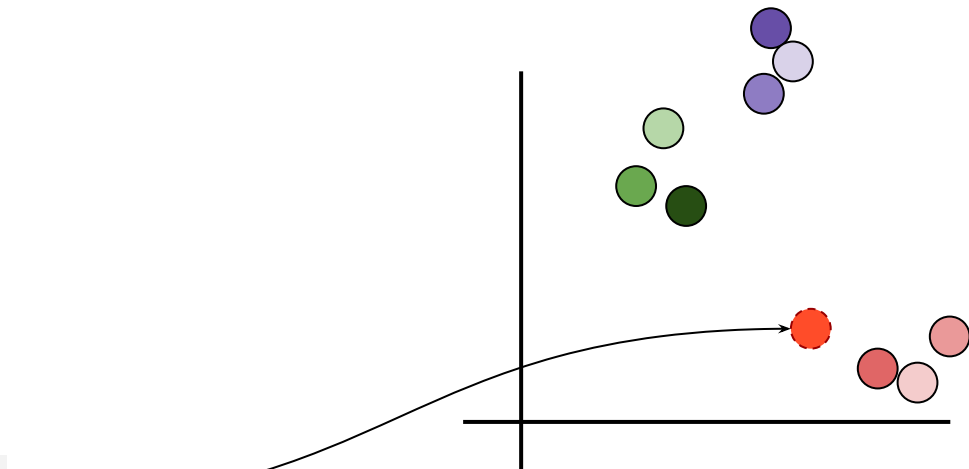
Embeddings for multiple documents



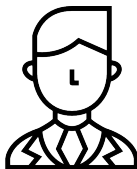
Your question
***“What’s in the food menu for
friday dinner?”***



`[-0.0032757148146629333, -0.011690735816955566,
0.041559211909770966, -0.03814808651804924, ...]`



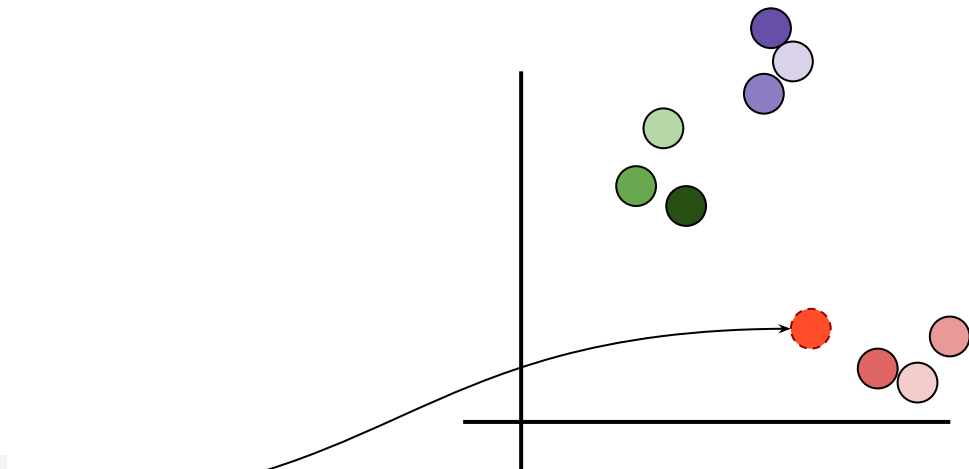
Embeddings for multiple documents



Your question
***“What’s in the food menu for
friday dinner?”***



`[-0.0032757148146629333, -0.011690735816955566,
0.041559211909770966, -0.03814808651804924, ...]`



Document
about food
menu

The end

Questions?

...and thank you!

References

Sources and references:

1. **The Retrieval Augmented Generation Pattern - André Vala - Cloud Solution Architect | Data & AI @ Microsoft - DataMakers Fest 2023**
2. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (FAIR, April 2021) <https://arxiv.org/abs/2005.11401>
3. https://docs.llamaindex.ai/en/stable/getting_started/concepts.html
4. <https://research.aimultiple.com/gpt/>
5. <https://pub.towardsai.net/how-do-8-smaller-models-in-gpt4-work-7335ccdfcf05>
6. <https://github.com/DataSciencePortugal/large-language-models>
7. <https://www.promptingguide.ai/techniques/>
8. TheAiEdge.io: Search in vector database: locality-sensitive hashing
9. Icons from Flaticon.com
10. <https://www.thenationalnews.com/business/technology/2023/04/23/why-googles-search-dominance-is-feeling-the-heat-from-chatgpt/>
11. <https://www.financialexpress.com/life/technology-chatgpt-can-destroy-google-in-two-years-says-gmail-creator-2962712/lite/>