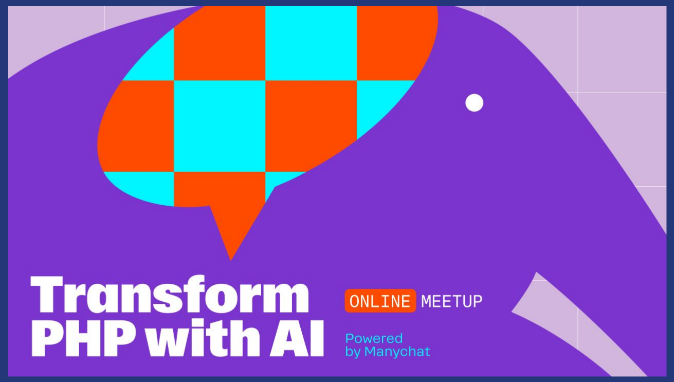
elastic

Build a Retrieval-Augmented Generation (RAG) in PHP

Enrico Zimuel, *Tech Lead & Principal Software Engineer*



Nov 27, 2024 - Barcelona

Agenda

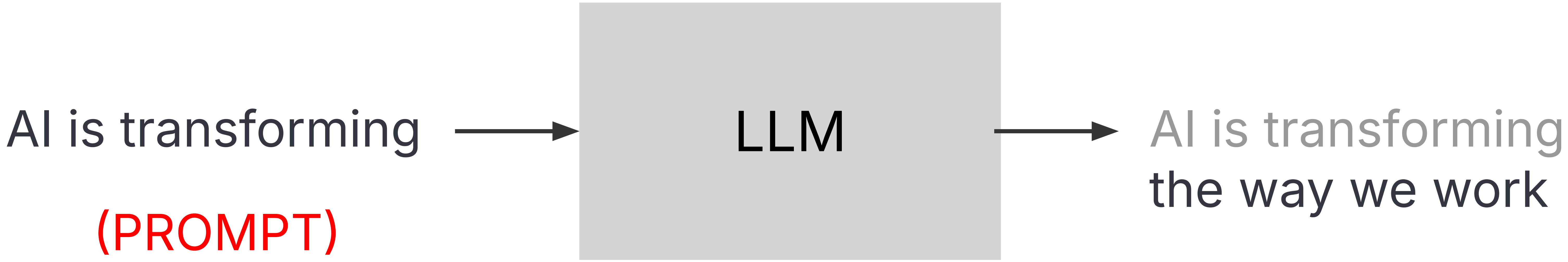
- Large Language Model (LLM)
- Transformers architecture
- Top-k and temperature
- Retrieval Augmented Generation (RAG)
- Embedding and Vector Search
- LLPhant for PHP
- Llama 3.2
- Elasticsearch
- Demo

Image generated using dall-e-3

Large Language Model (LLM)

LLM

- **Large Language Model** (LLM) are probabilistic models that produce sentence in natural language
- These models work by completing sentences



AI is transforming

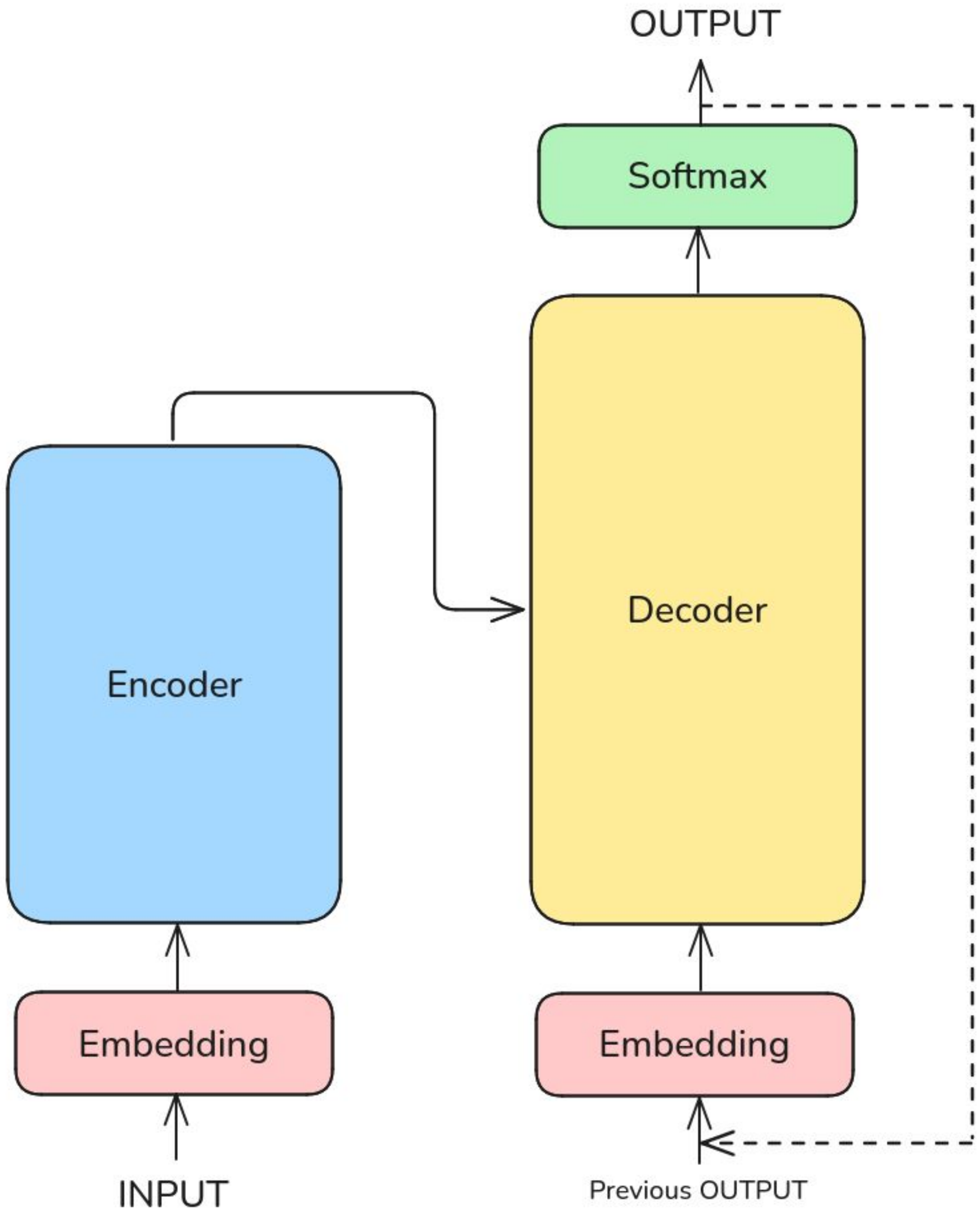
(PROMPT)

LLM

AI is transforming the way we work

Transformer architecture

- Introduced in Attention is All You Need paper in 2017
- Basement of all LLMs
- The sentences are analyzed using a **self-attention** mechanism: each part of a sentence is evaluated in relation to every other part to understand contextual relationships and assign appropriate weights

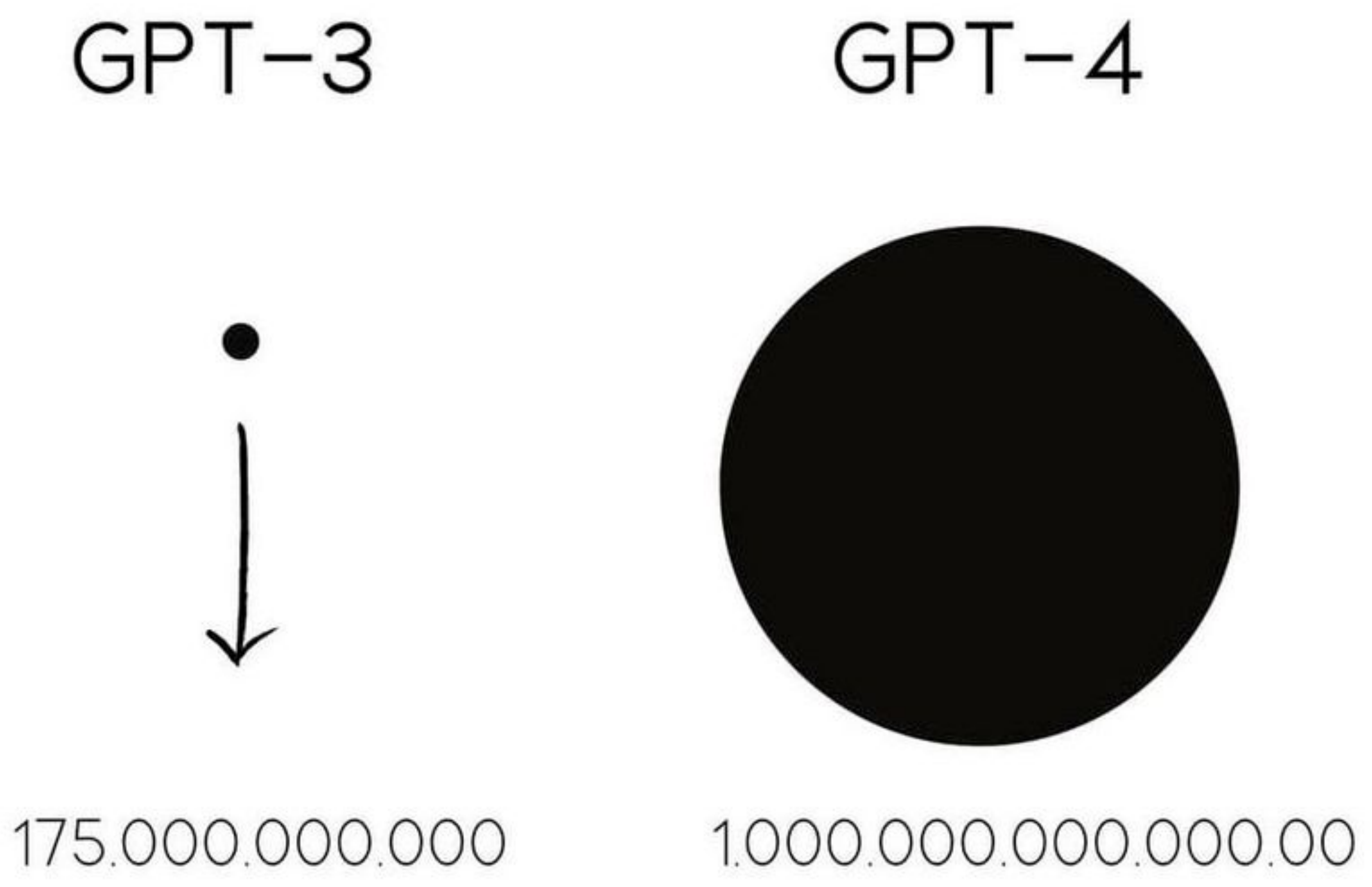


LLM

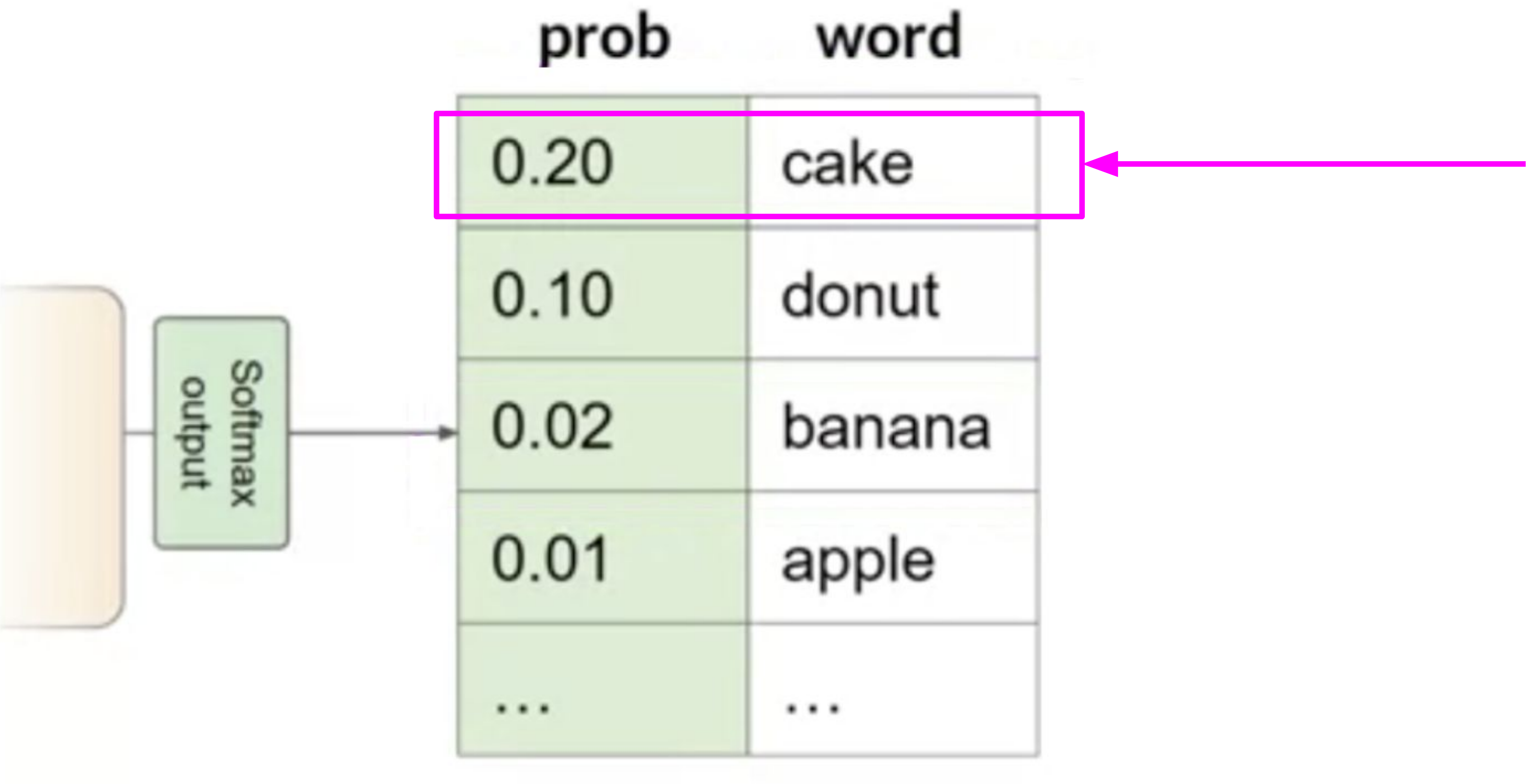
- **Large Language Model** (LLM) consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabelled text using self-supervised learning
- A message is splitted in **tokens**
- Each token is translated in a number using an operation called **embeddings**
- LLM **repeatedly predicting** the next token

Size of GPT-4

- Around **1.76 trillion** parameters
- Neural network with **120** layers
- Process up to **25,000** words at once
- Estimated training cost is $200M using 10,000 Nvidia A100 GPU for 11 months



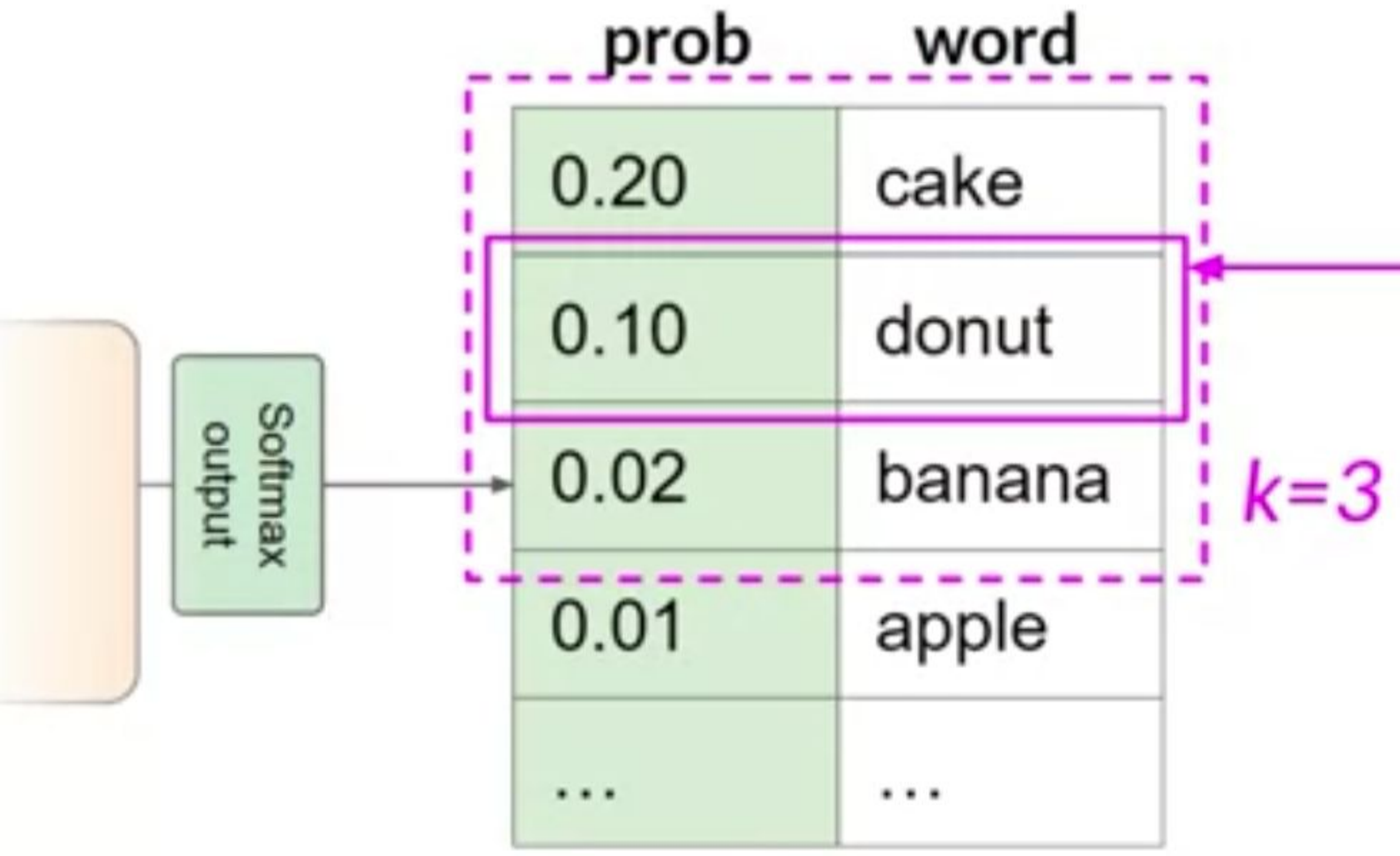
Predict the next word



prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
...	...

Choose the one with greatest probability (greedy algorithm)

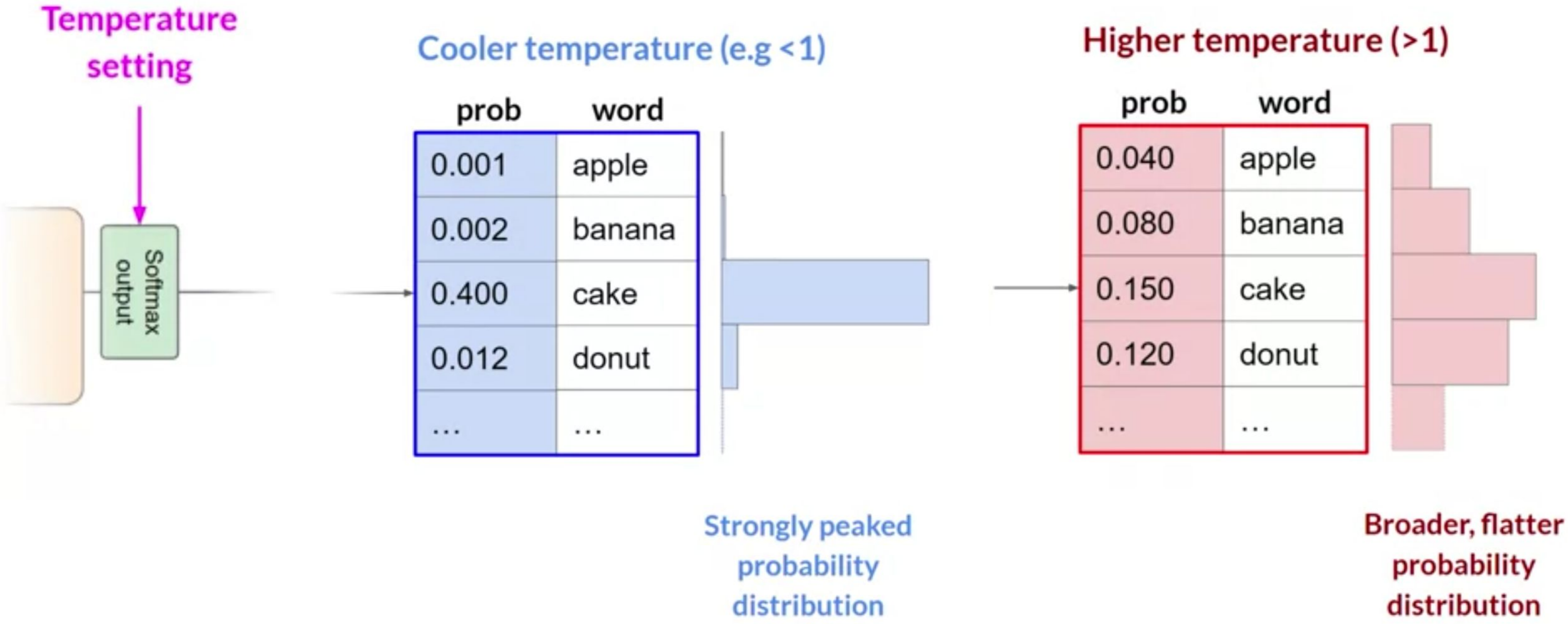
Top-k



prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
...	...

top-k: select an output from the top-k results after applying random-weighted strategy using the probabilities

Temperature



Temperature setting

Softmax output

Cooler temperature (e.g <1)

prob	word
0.001	apple
0.002	banana
0.400	cake
0.012	donut
...	...

Strongly peaked probability distribution

Higher temperature (>1)

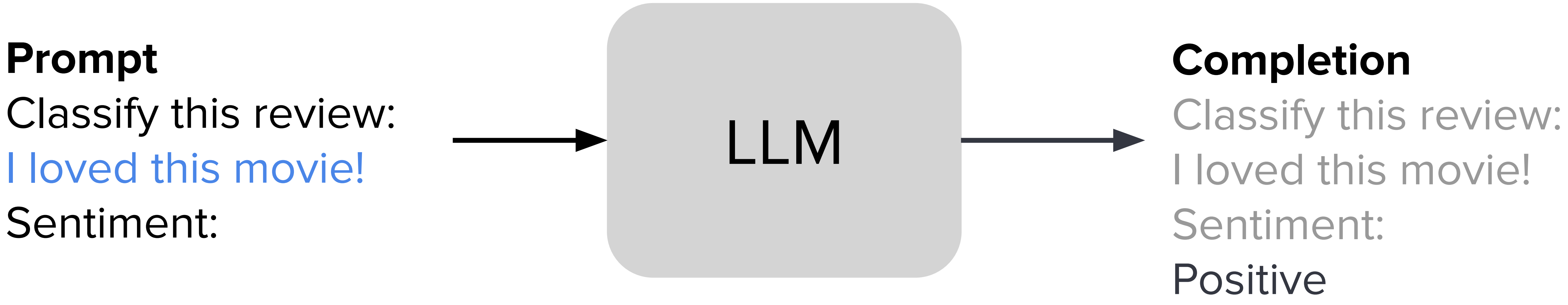
prob	word
0.040	apple
0.080	banana
0.150	cake
0.120	donut
...	...

Broader, flatter probability distribution

Prompt engineering

- You can encounter situations where the model doesn't produce the outcome that you want on the first try
- You may have to revisit the language several times to get a good answer
- The development and improvement of the prompt is known as **prompt engineering**
- One powerful strategy is to include examples of the task that you want the model to carry out inside the prompt
- This is called **In-Context Learning (ICL)**

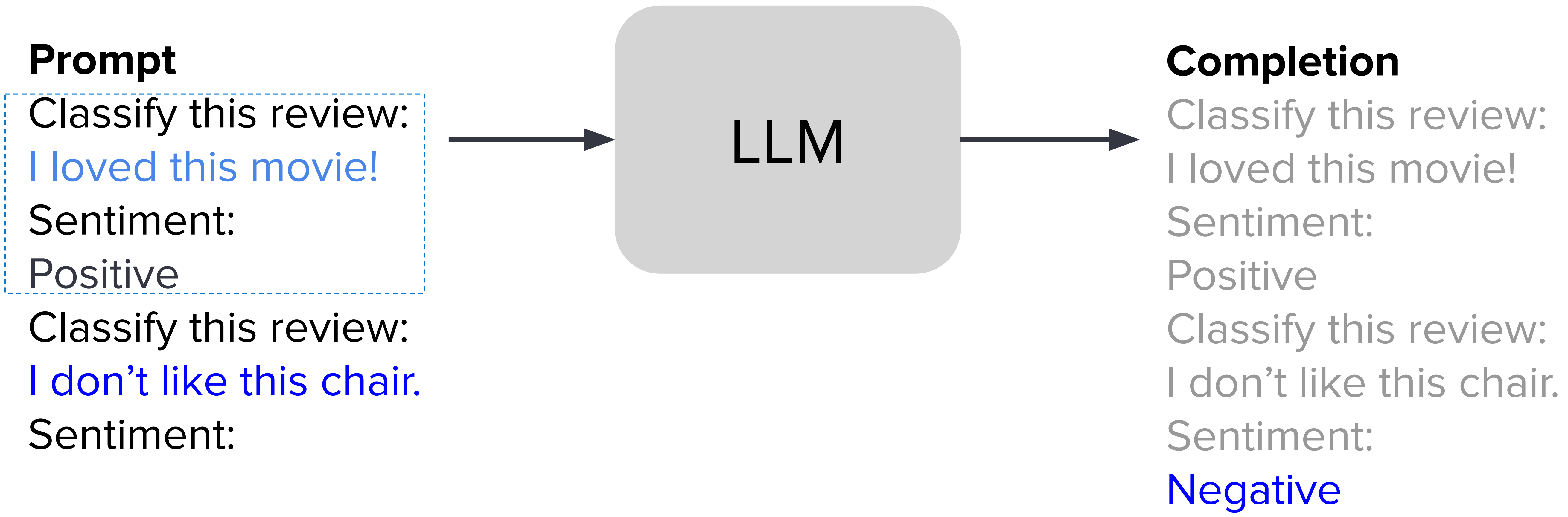
ICL - zero shot inference



Prompt
Classify this review:
I loved this movie!
Sentiment:

Completion
Classify this review:
I loved this movie!
Sentiment:
Positive

ICL - one shot inference



Prompt

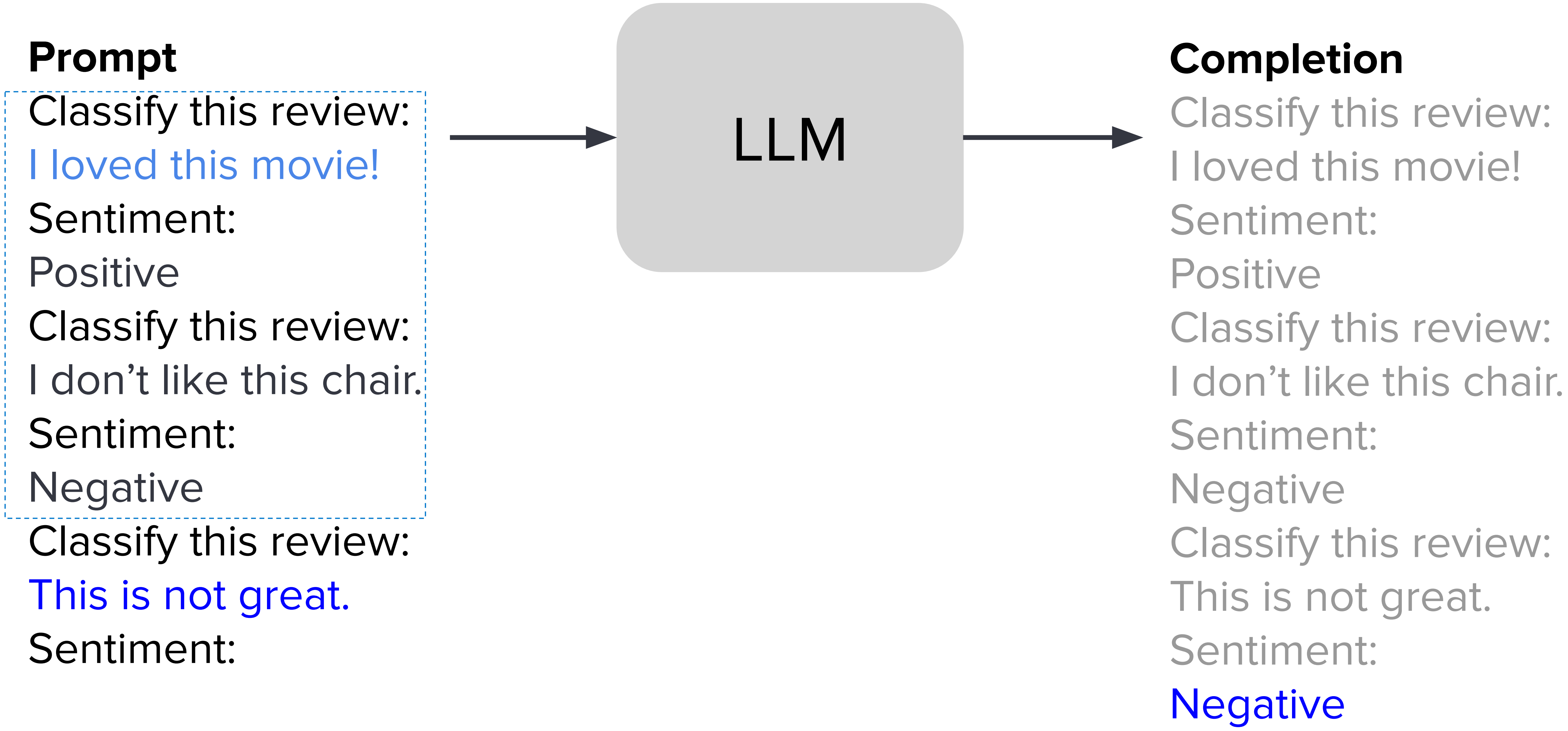
Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:

LLM

Completion

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative

ICL - few shot inference



Prompt

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative
Classify this review:
This is not great.
Sentiment:

LLM

Completion

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative
Classify this review:
This is not great.
Sentiment:
Negative

Ollama

- Ollama is a software for downloading and running LLMs locally
- Llama 3, Phi 3, Mistral, Gemma, and other models
- Simple command line tool:
  - ollama pull llama3.2:3b
  - ollama run llama3.2:3b

LLPhant

- LLPhant is a comprehensive open source Generative AI framework for PHP
- The goal is to offer an easy to use library to build GenAI applications in PHP
- LLM supported: OpenAI, Anthropics, Ollama, Mistral
- Vector databases: Elasticsearch, File, Memory, Milvus, Qdrant, Redis, Milvus, Chroma, etc
- Started by Maxime Thoonsen

Example: LLPhant with Llama3.2

```php
use LLPhant\Chat\OllamaChat;
use LLPhant\OllamaConfig;

$config = new OllamaConfig();
$config->model = 'llama3.2';
$chat = new OllamaChat($config);

$response = $chat->generateText('What is the capital of Italy?');
// The capital city of Italy is Rome
printf("%s\n", $response);
```

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG)

- **RAG** is a technique in natural language processing that combines information retrieval systems with **Large Language Models** (LLM) to generate more informed and accurate responses
- It is composed by the following parts:
  - **Retrieval-Augmented**
  - **Generation**

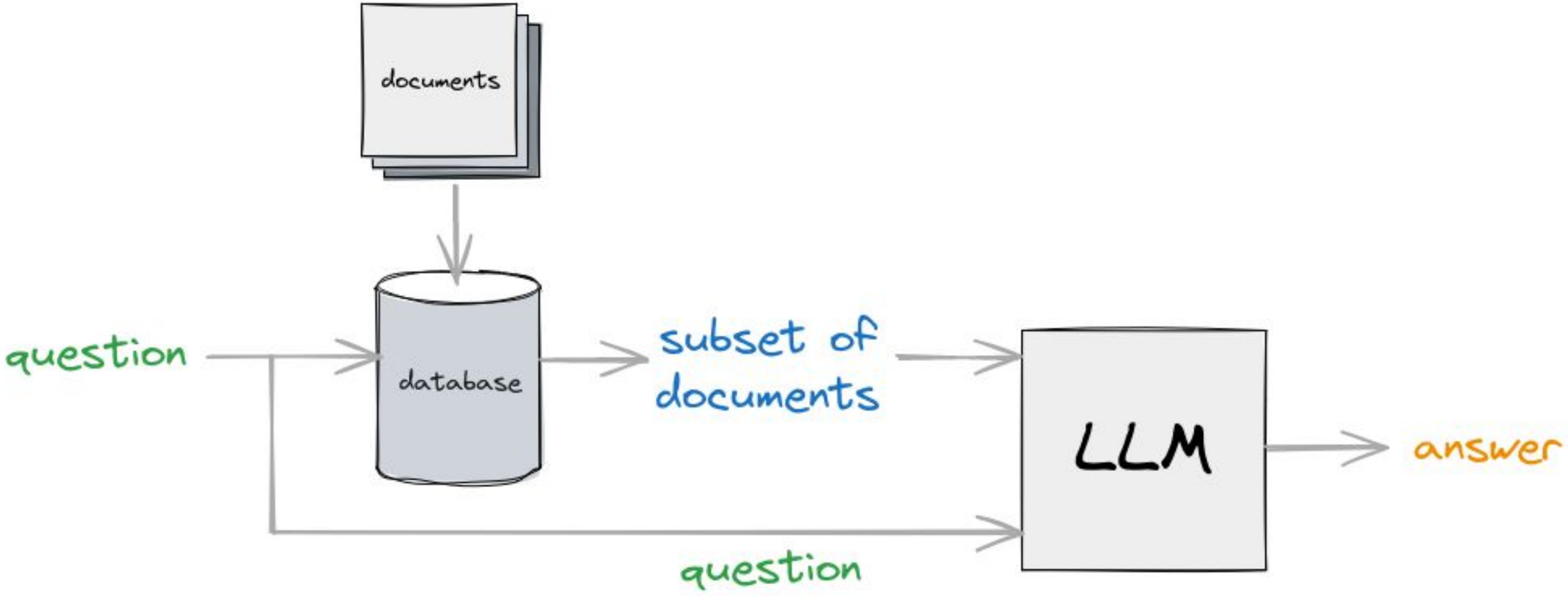
Generation

- LLMs are very powerful but have some limitations:
  - **No source** (potential hallucinations)
    - How can I verify the information coming from an LLM?
    - What sources has been used to generate the answer?
  - **Out of date**
    - An LLM is trained in a period of time
    - For update we need to retraining the model (very expensive)

Retrieval-Augmented

- We collect sets of private or public document
- We build a **retrieval system** (e.g. a database) to extract a subset of documents using a **question**
- Then we pass the **question + documents found** to an LLM as prompt with a context
- The LLM can give an answer using the updated documents

RAG architecture

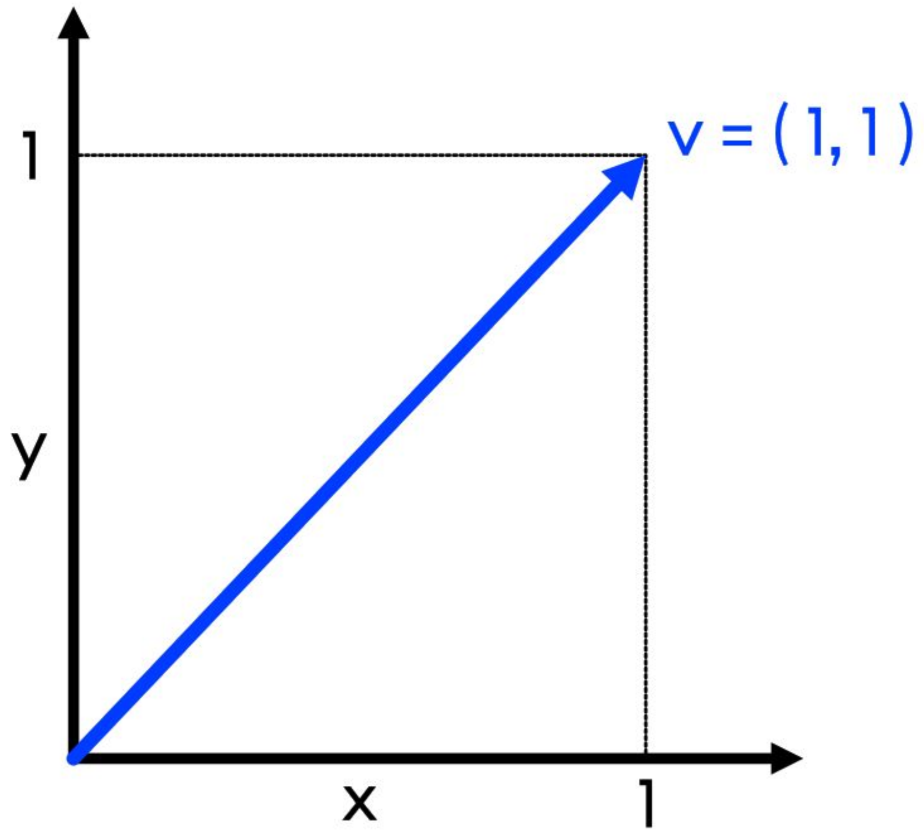


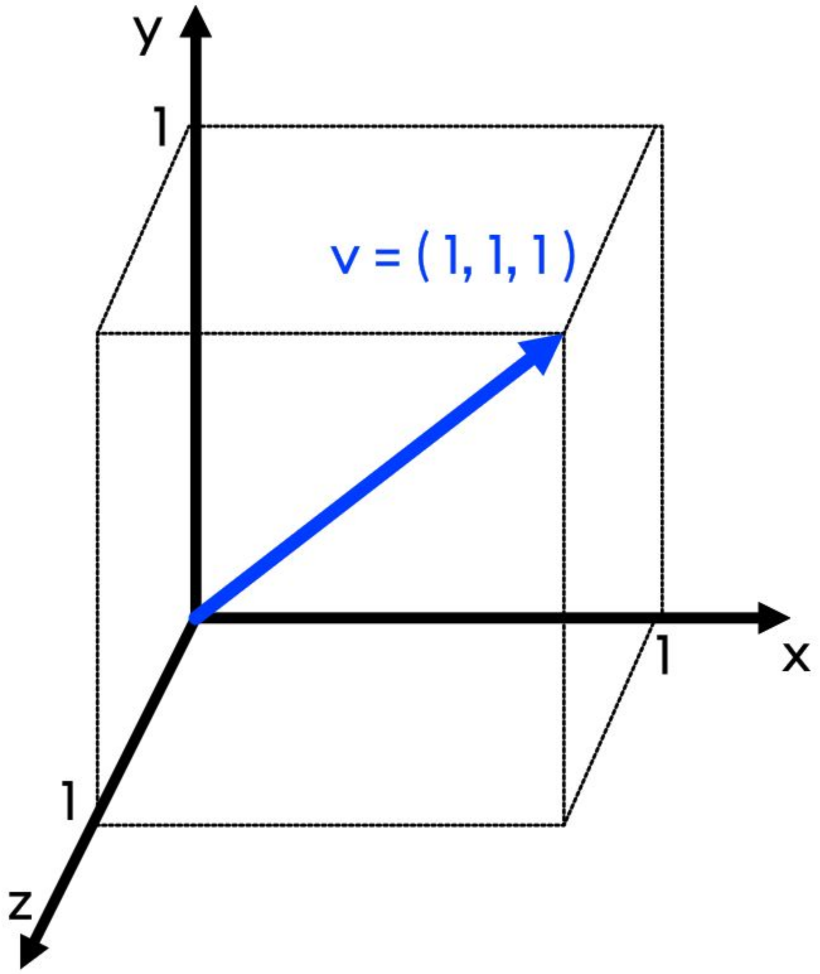
Retrieve documents from a question

- How we can retrieve documents in a database using a question?
- We need to use **semantic search**
- One solution is to use a **vector database**
- A vector database is a system that uses **vectors** (set of numbers) to retrieve information

What is a vector?

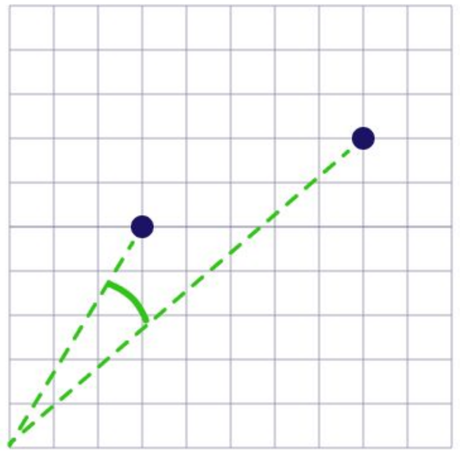
- A vector is a set of numbers
- Example: a vector of 3 elements [2, 5, -10]
- A vector can be represented in a multi-dimensional space (eg. Llama3.2 uses 3072 dimensions)





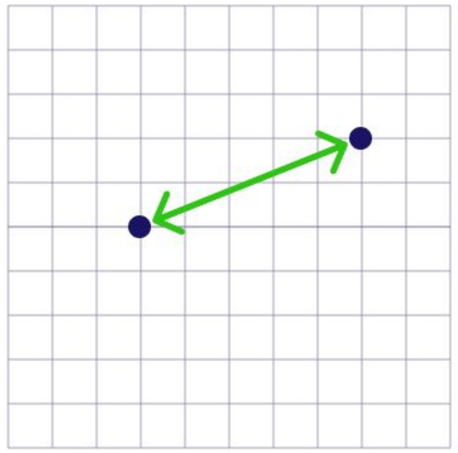
Similarity between two vectors

- Two vectors are (semantically) similar if they are close to each other
- We need to define a way to measure the similarity



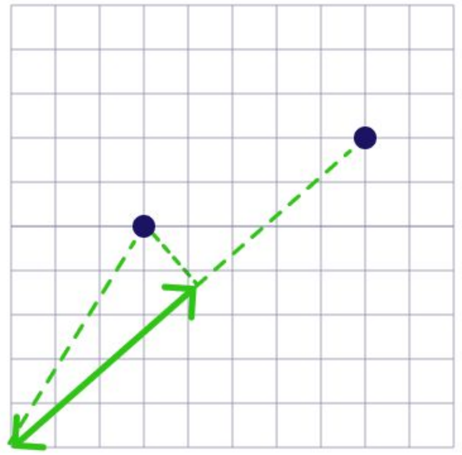
Cosine Distance

$$1 - \frac{A \cdot B}{||A|| \quad ||B||}$$



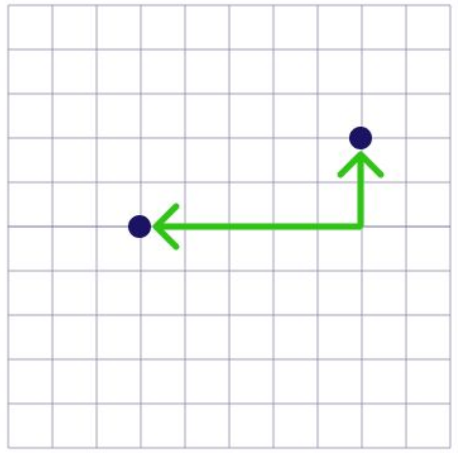
Squared Euclidean (L2 Squared)

$$\sum_{i=1}^{n} (x_i - y_i)^2$$



Dot Product

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

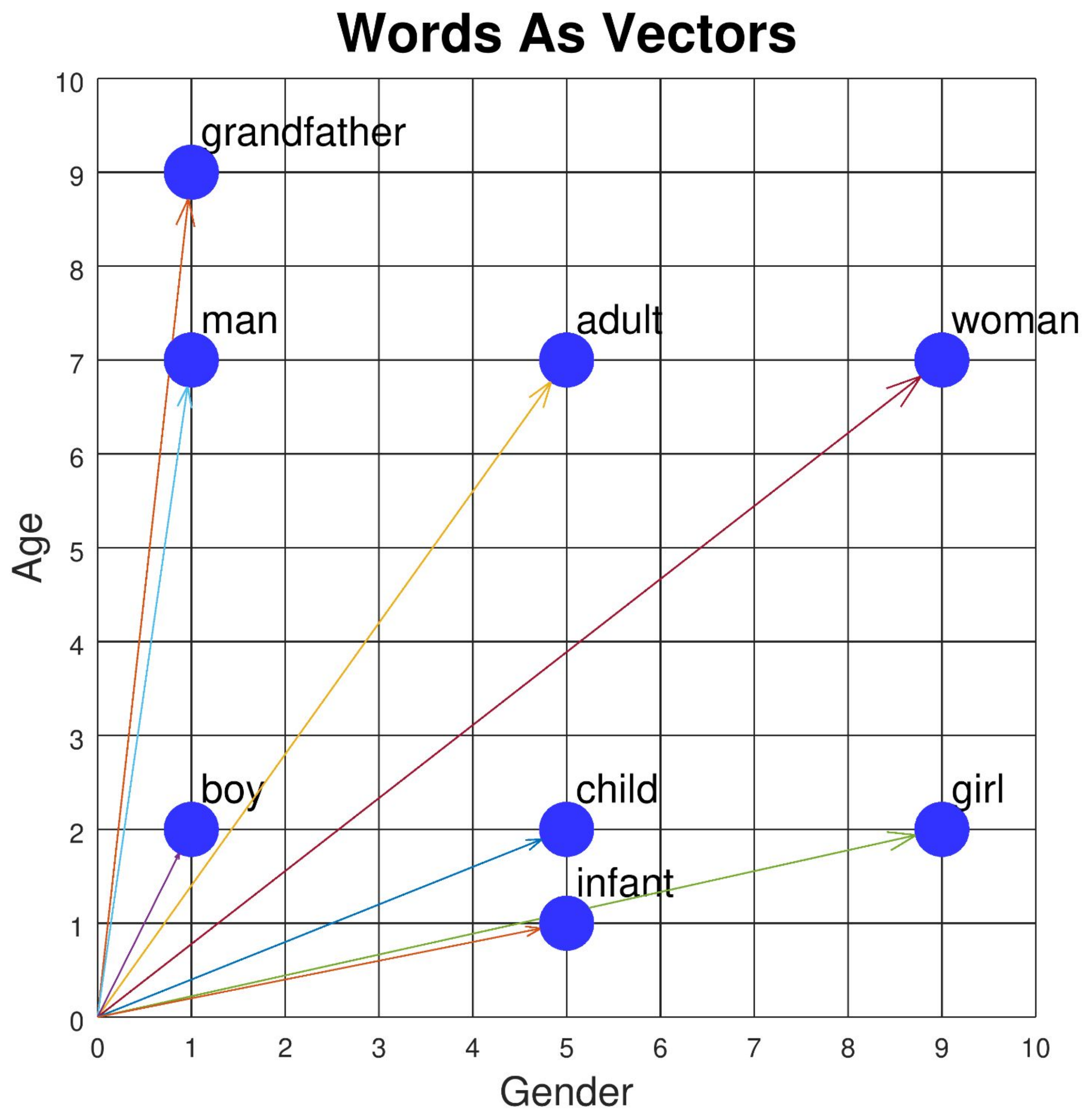


Manhattan (L1)

$$\sum_{i=1}^{n} |x_i - y_i|$$

Embedding

- Embedding is the translation of an input (document, image, sound, movie, etc) to a vector
- There are many techniques, using an LLM typically this is done by a neural network
- The goal is to group information that are semantically related to each other
- https://projector.tensorflow.org/



Vector database + LLM

- The search query (***question***) is in natural language
- We use semantic search to retrieve top-n relevant documents (***context***)
- We send the following prompt to the LLM (example):
    - *Given the following* ***{context}*** *answer to the following* ***{question}***

Split the documents in chunk

- We need to store data in the vector database using chunk of information
- We cannot use big documents since we need to pass it in the context part of the prompt for an LLM that typically has a token limit (e.g. Llama3.2 up to 128K)
- We need to split the documents in **chunk** (part of words)

Elasticsearch (vector database)

- Elasticsearch is Free and Open Source (AGPL), Distributed, RESTful Search Engine
- Distributed search and analytics engine, scalable data store and **vector database** optimized for speed and relevance on production-scale workloads.
- You can run it locally with a single command:
  - **curl -fsSL https://elastic.co/start-local | sh**

RAG demo:
LLPhant + Llama 3.2 + Elasticsearch

Available on github: ezimuel/llphant-llama-elasticsearch

References

- What is retrieval-augmented generation? IBM research
- Ashish Vaswan et al., Attention Is All You Need, Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)
- Albert Ziegler, John Berryman, A developer’s guide to prompt engineering and LLMs, Github blog post
- Sebastian Raschka, Build a Large Language Model (From Scratch), Manning, 2024
- Elasticsearch as vector database, Elastic Search Labs
- Elasticsearch search relevance, Elastic Search Labs
- E.Zimuel, Generative AI and Large Language Model in PHP, phpDay 2024 conference, Verona (Italy)
- E.Zimuel, Retrieval-Augmented Generation for talking with your private data using LLM, AI Heroes 2023 conference, Turin (Italy)

elastic

Thanks!

More information: www.elastic.co

Contact information: enrico.zimuel (at) elastic.co

elastic