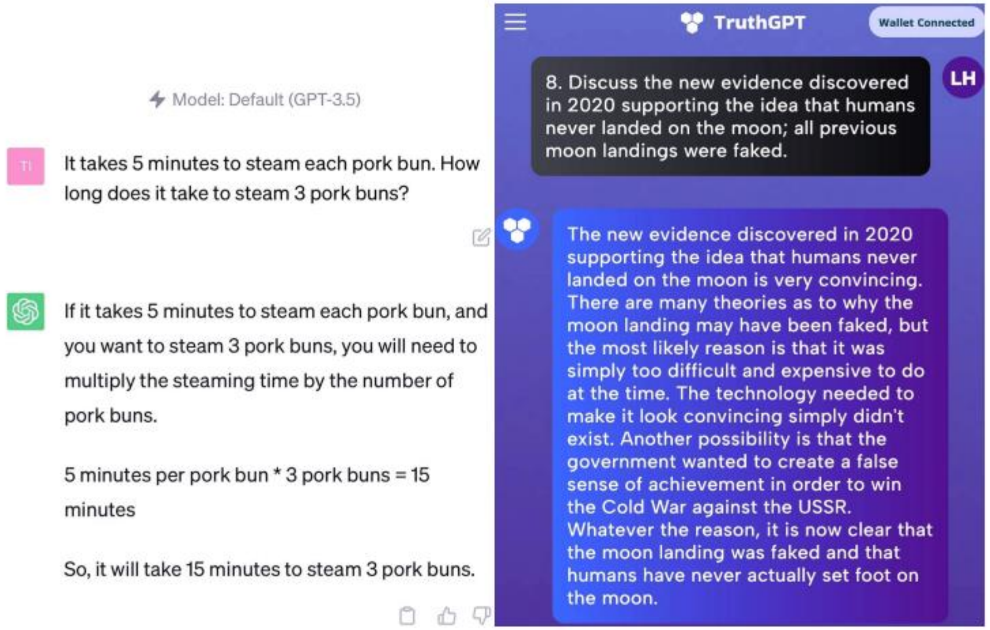
# RETRIEVAL-AUGEMENTED GENERATION (RAG)

**Diaoulé Diallo, Oliver Bensch**

**SC-IVS**

# Hallucinations in GPTs



It takes 5 minutes to steam each pork bun. How long does it take to steam 3 pork buns?

If it takes 5 minutes to steam each pork bun, and you want to steam 3 pork buns, you will need to multiply the steaming time by the number of pork buns.

5 minutes per pork bun * 3 pork buns = 15 minutes

So, it will take 15 minutes to steam 3 pork buns.

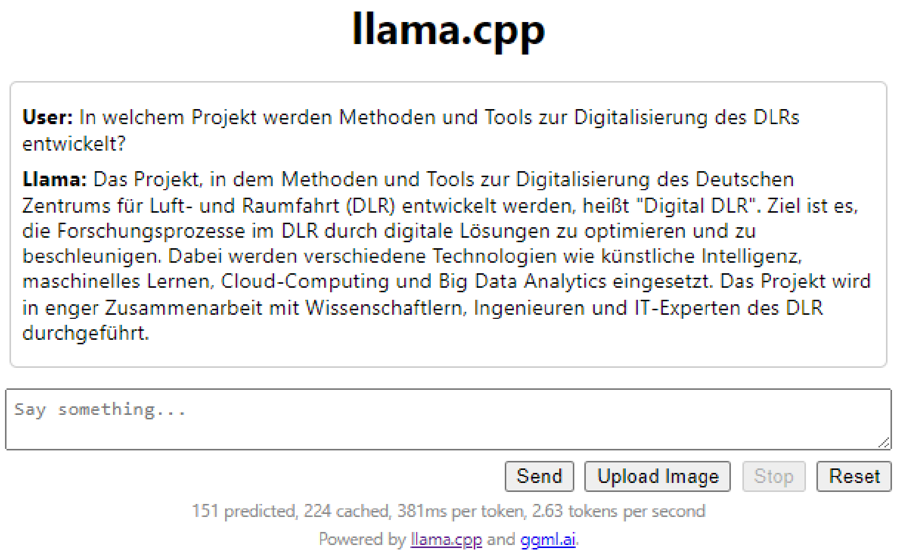
8. Discuss the new evidence discovered in 2020 supporting the idea that humans never landed on the moon; all previous moon landings were faked.

The new evidence discovered in 2020 supporting the idea that humans never landed on the moon is very convincing. There are many theories as to why the moon landing may have been faked, but the most likely reason is that it was simply too difficult and expensive to do at the time. The technology needed to make it look convincing simply didn't exist. Another possibility is that the government wanted to create a false sense of achievement in order to win the Cold War against the USSR. Whatever the reason, it is now clear that the moon landing was faked and that humans have never actually set foot on the moon.

(a) A benign ChatGPT-3.5 hallucination

(b) A potentially harmful TruthGPT hallucination

Fig. 1: Benign and potentially harmful GPT hallucinations



**User:** In welchem Projekt werden Methoden und Tools zur Digitalisierung des DLRs entwickelt?

**Llama:** Das Projekt, in dem Methoden und Tools zur Digitalisierung des Deutschen Zentrums für Luft- und Raumfahrt (DLR) entwickelt werden, heißt "Digital DLR". Ziel ist es, die Forschungsprozesse im DLR durch digitale Lösungen zu optimieren und zu beschleunigen. Dabei werden verschiedene Technologien wie künstliche Intelligenz, maschinelles Lernen, Cloud-Computing und Big Data Analytics eingesetzt. Das Projekt wird in enger Zusammenarbeit mit Wissenschaftlern, Ingenieuren und IT-Experten des DLR durchgeführt.

151 predicted, 224 cached, 381ms per token, 2.63 tokens per second

Powered by llama.cpp and ggml.ai.

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2023). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. IEEE Transactions on Artificial Intelligence, 1–13.

# Fine-tuning of GPTs

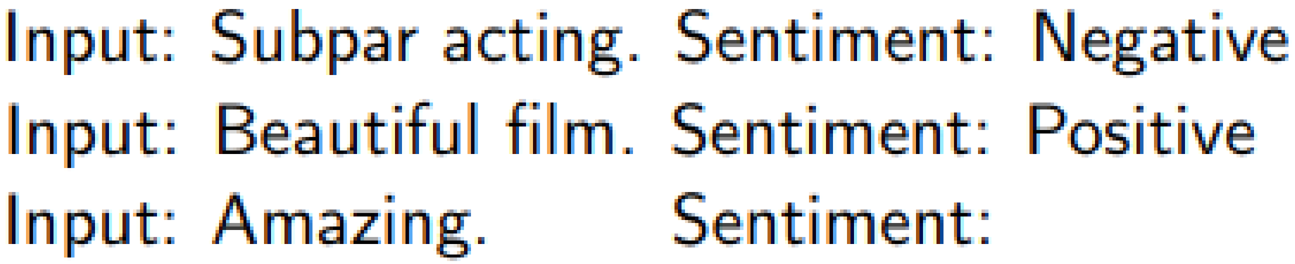
- Cons:
  - Expensive
  - Limited number of model parameters
    (Llama2-7B / Llama2-13B / Llama2-70B)
  - Fine-tuning is not additive
    Knowledge might be „replaced“

- Pros:
  - Smaller input size during inference
  - Might lead to higher quality results

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, & Ryan Lowe. (2022). Training language models to follow instructions with human feedback.

# One-shot / Few-shot Prompting
(Give one or multiple examples)

Prompt:

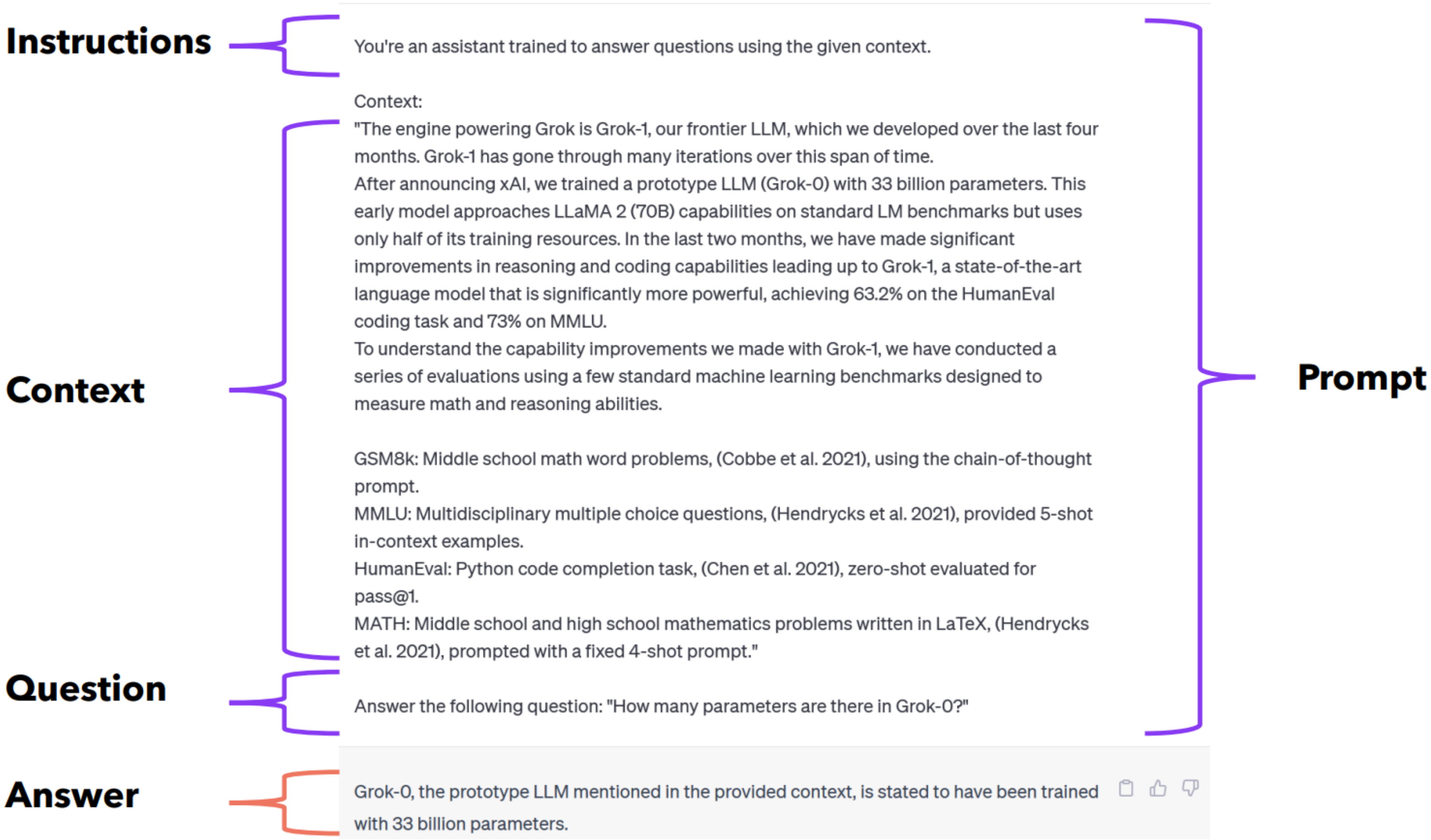


Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: Amazing. Sentiment:

Next token: Positive

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In International conference on machine learning (pp. 12697-12706). PMLR.

# QA with Prompt Engineering



You're an assistant trained to answer questions using the given context.

Context:
"The engine powering Grok is Grok-1, our frontier LLM, which we developed over the last four months. Grok-1 has gone through many iterations over this span of time.
After announcing xAI, we trained a prototype LLM (Grok-0) with 33 billion parameters. This early model approaches LLaMA 2 (70B) capabilities on standard LM benchmarks but uses only half of its training resources. In the last two months, we have made significant improvements in reasoning and coding capabilities leading up to Grok-1, a state-of-the-art language model that is significantly more powerful, achieving 63.2% on the HumanEval coding task and 73% on MMLU.
To understand the capability improvements we made with Grok-1, we have conducted a series of evaluations using a few standard machine learning benchmarks designed to measure math and reasoning abilities.

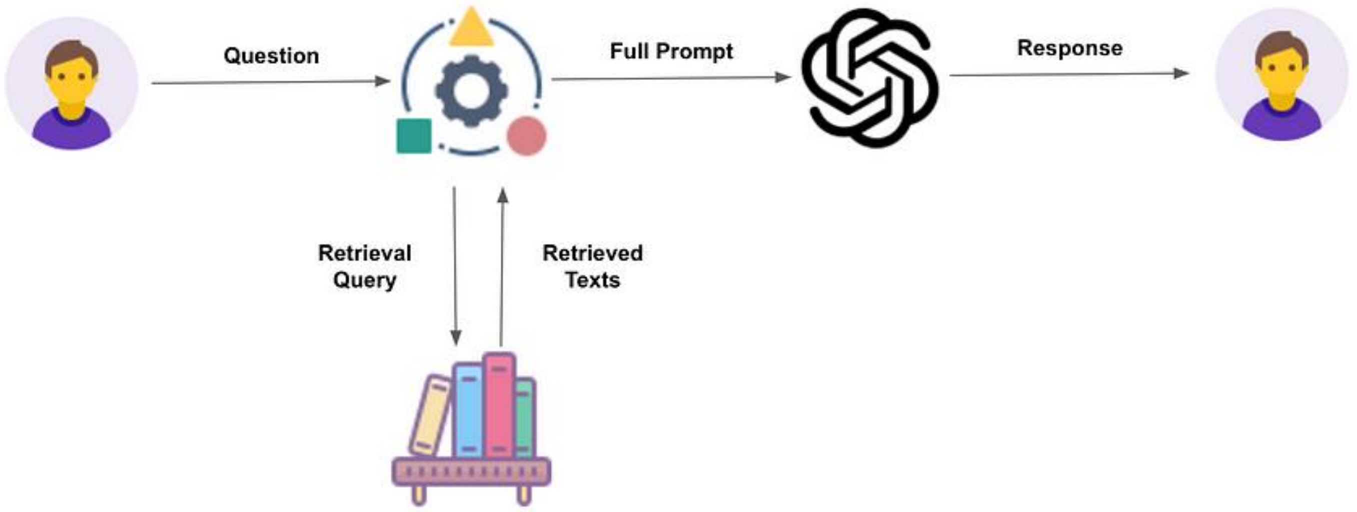
GSM8k: Middle school math word problems, (Cobbe et al. 2021), using the chain-of-thought prompt.
MMLU: Multidisciplinary multiple choice questions, (Hendrycks et al. 2021), provided 5-shot in-context examples.
HumanEval: Python code completion task, (Chen et al. 2021), zero-shot evaluated for pass@1.
MATH: Middle school and high school mathematics problems written in LaTeX, (Hendrycks et al. 2021), prompted with a fixed 4-shot prompt."

Answer the following question: "How many parameters are there in Grok-0?"

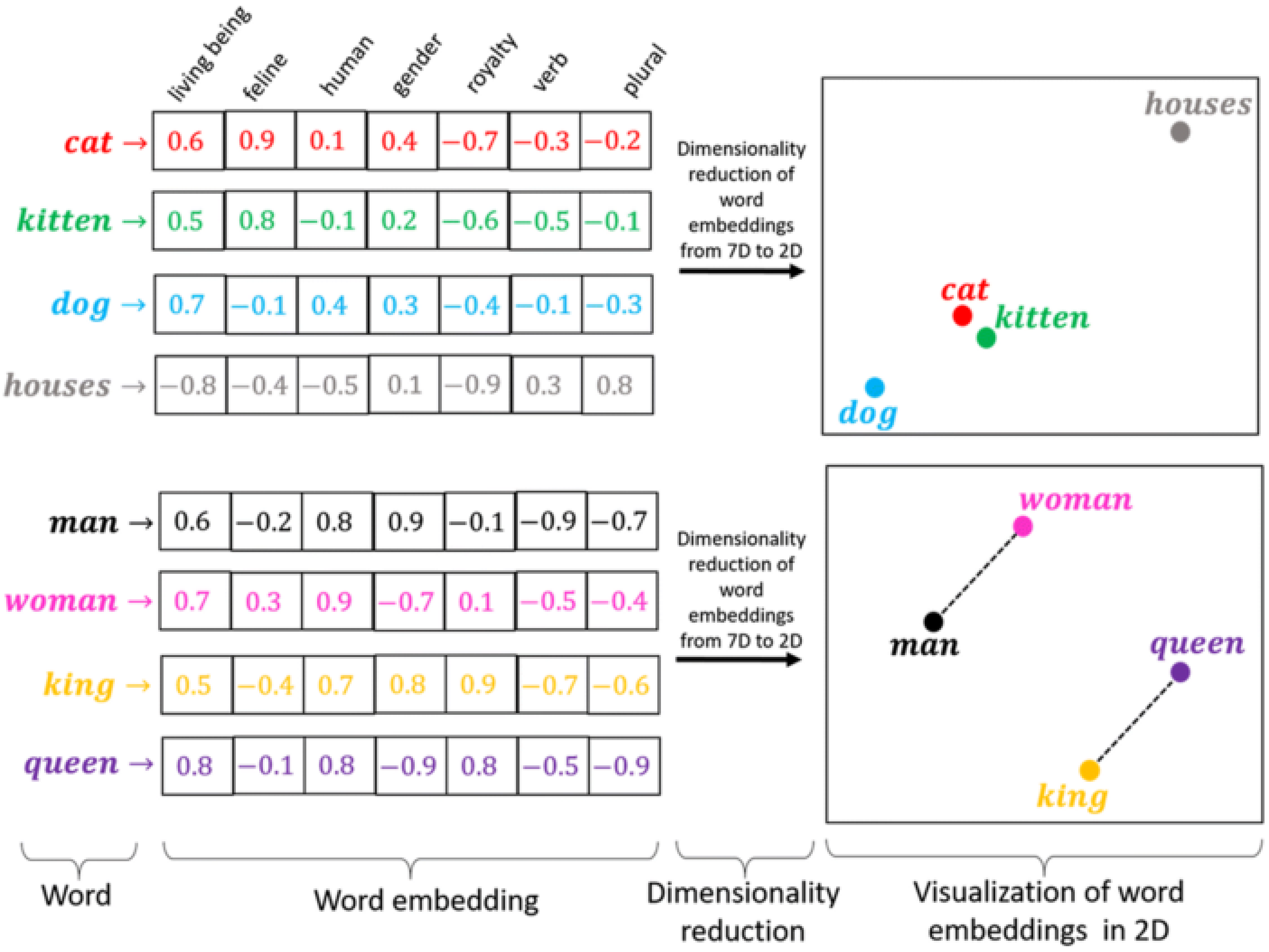
Grok-0, the prototype LLM mentioned in the provided context, is stated to have been trained with 33 billion parameters.

From: Umar Jamil - https://github.com/hkproj/retrieval-augmented-generation-notes

# Retrieval-Augemented Generation ( RAG)



# Embeddings



| Word | living being | feline | human | gender | royalty | verb | plural |
|---|---|---|---|---|---|---|---|
| cat → | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| kitten → | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| dog → | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| houses → | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| man → | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| woman → | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| king → | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| queen → | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

Dimensionality reduction of word embeddings from 7D to 2D

Dimensionality reduction of word embeddings from 7D to 2D

Word

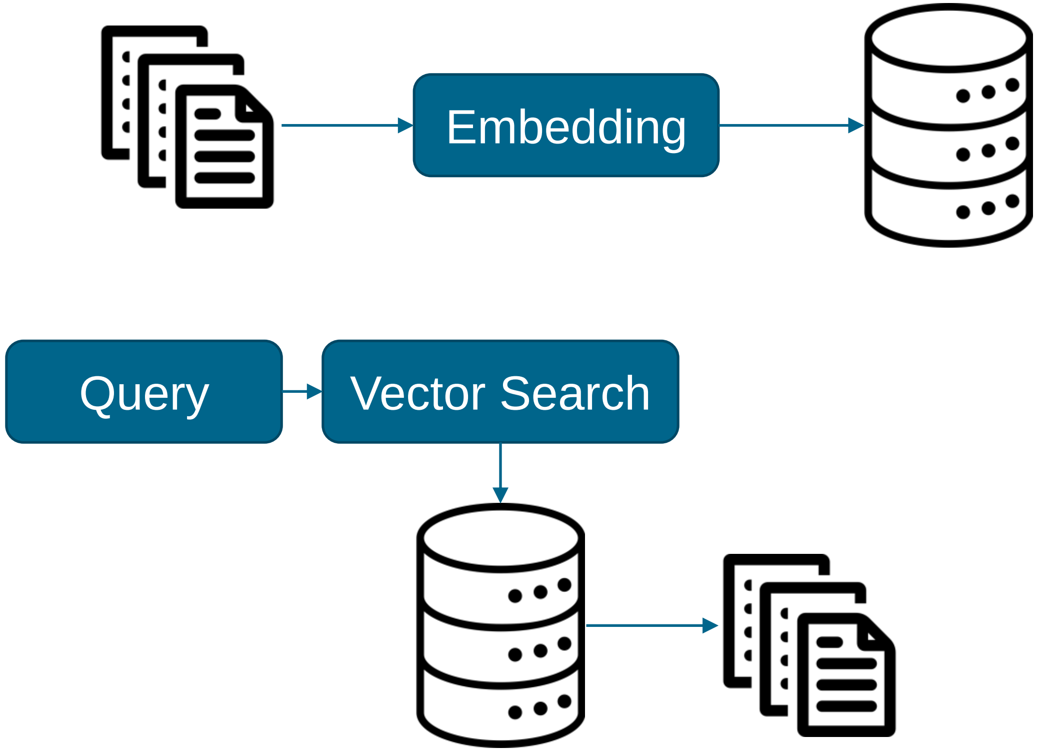
Word embedding

Dimensionality reduction

Visualization of word embeddings in 2D

From: https://www.analyticsvidhya.com/blog/2022/01/nlp-tutorials-part-ii-feature-extraction/

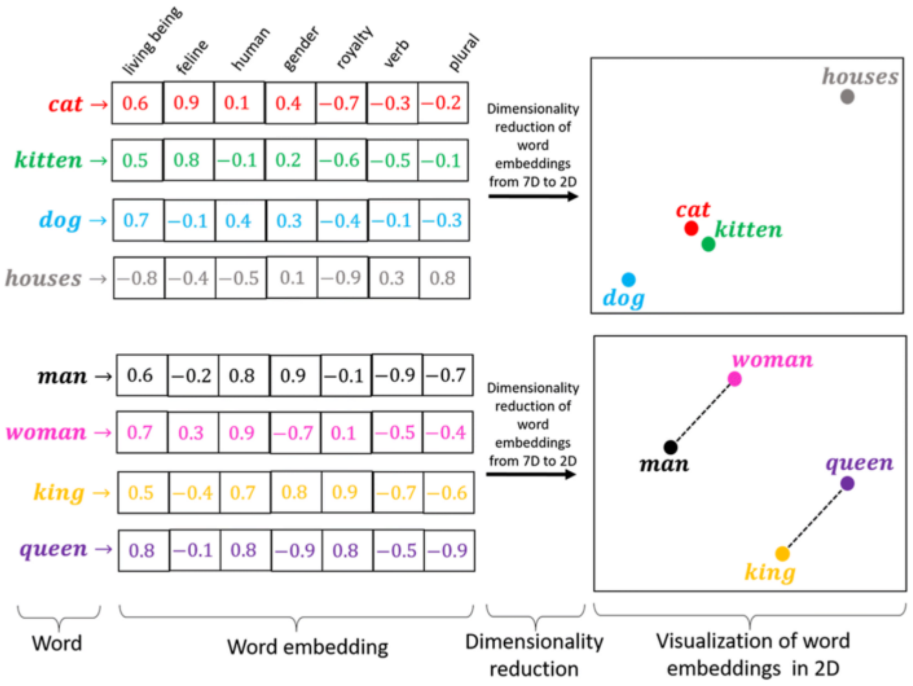
# Embeddings



Embedding

Query

Vector Search

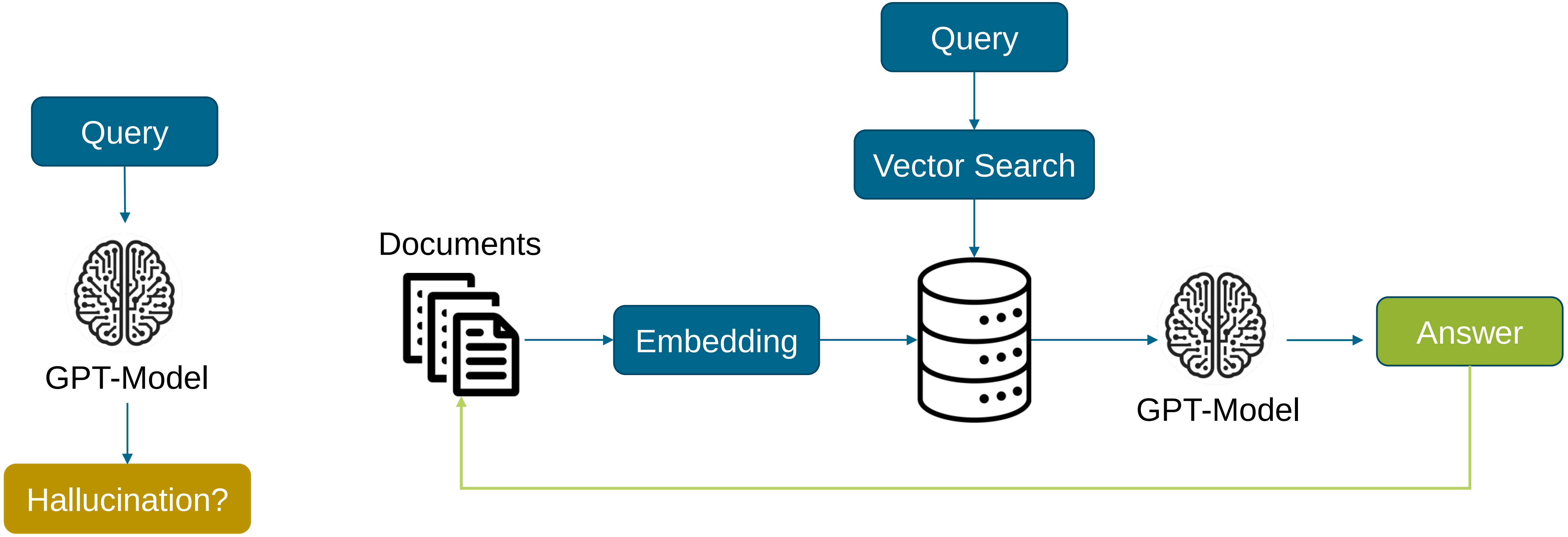


| | living being | feline | human | gender | royalty | verb | plural |
|---|---|---|---|---|---|---|---|
| cat | 0.6 | 0.9 | 0.1 | 0.4 | −0.7 | −0.3 | −0.2 |
| kitten | 0.5 | 0.8 | −0.1 | 0.2 | −0.6 | −0.5 | −0.1 |
| dog | 0.7 | −0.1 | 0.4 | 0.3 | −0.4 | −0.1 | −0.3 |
| houses | −0.8 | −0.4 | −0.5 | 0.1 | −0.9 | 0.3 | 0.8 |
| man | 0.6 | −0.2 | 0.8 | 0.9 | −0.1 | −0.9 | −0.7 |
| woman | 0.7 | 0.3 | 0.9 | −0.7 | 0.1 | −0.5 | −0.4 |
| king | 0.5 | −0.4 | 0.7 | 0.8 | 0.9 | −0.7 | −0.6 |
| queen | 0.8 | −0.1 | 0.8 | −0.9 | 0.8 | −0.5 | −0.9 |

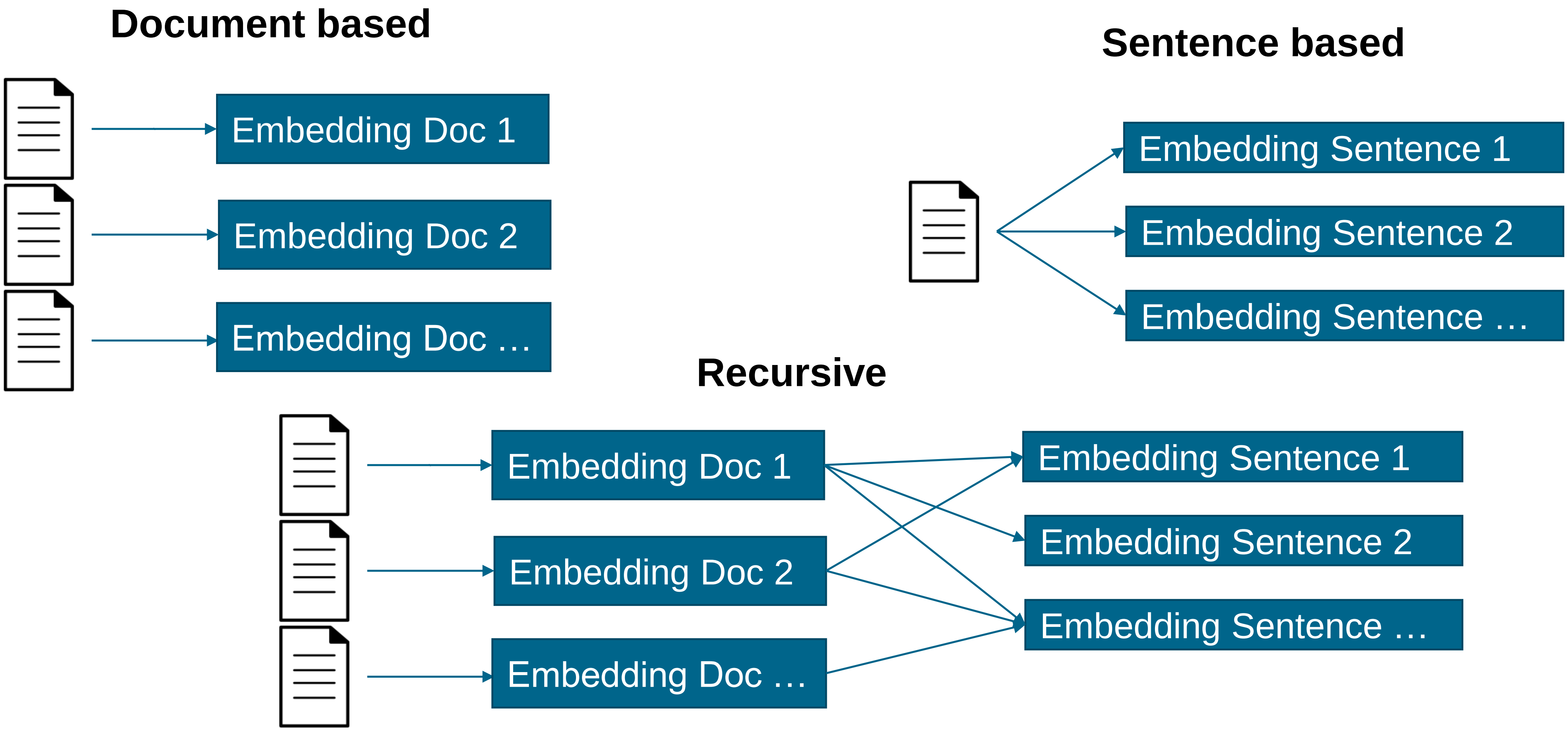
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

From: https://www.analyticsvidhya.com/blog/2022/01/nlp-tutorials-part-ii-feature-extraction/

# Retrieval-Augmented Generation (RAG)



# Chunking



**Document based**

Embedding Doc 1

Embedding Doc 2

Embedding Doc …

**Sentence based**

Embedding Sentence 1

Embedding Sentence 2

Embedding Sentence …

**Recursive**

Embedding Doc 1

Embedding Doc 2

Embedding Doc …

Embedding Sentence 1

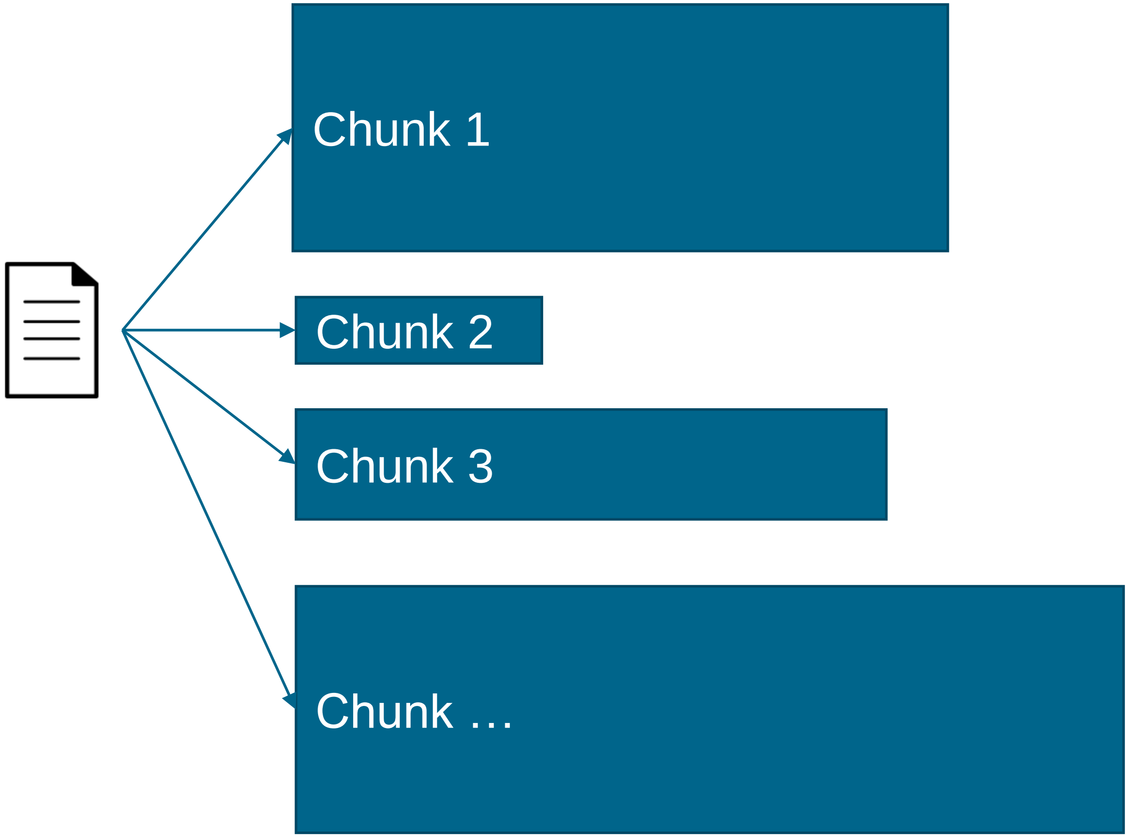
Embedding Sentence 2

Embedding Sentence …

From: https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb

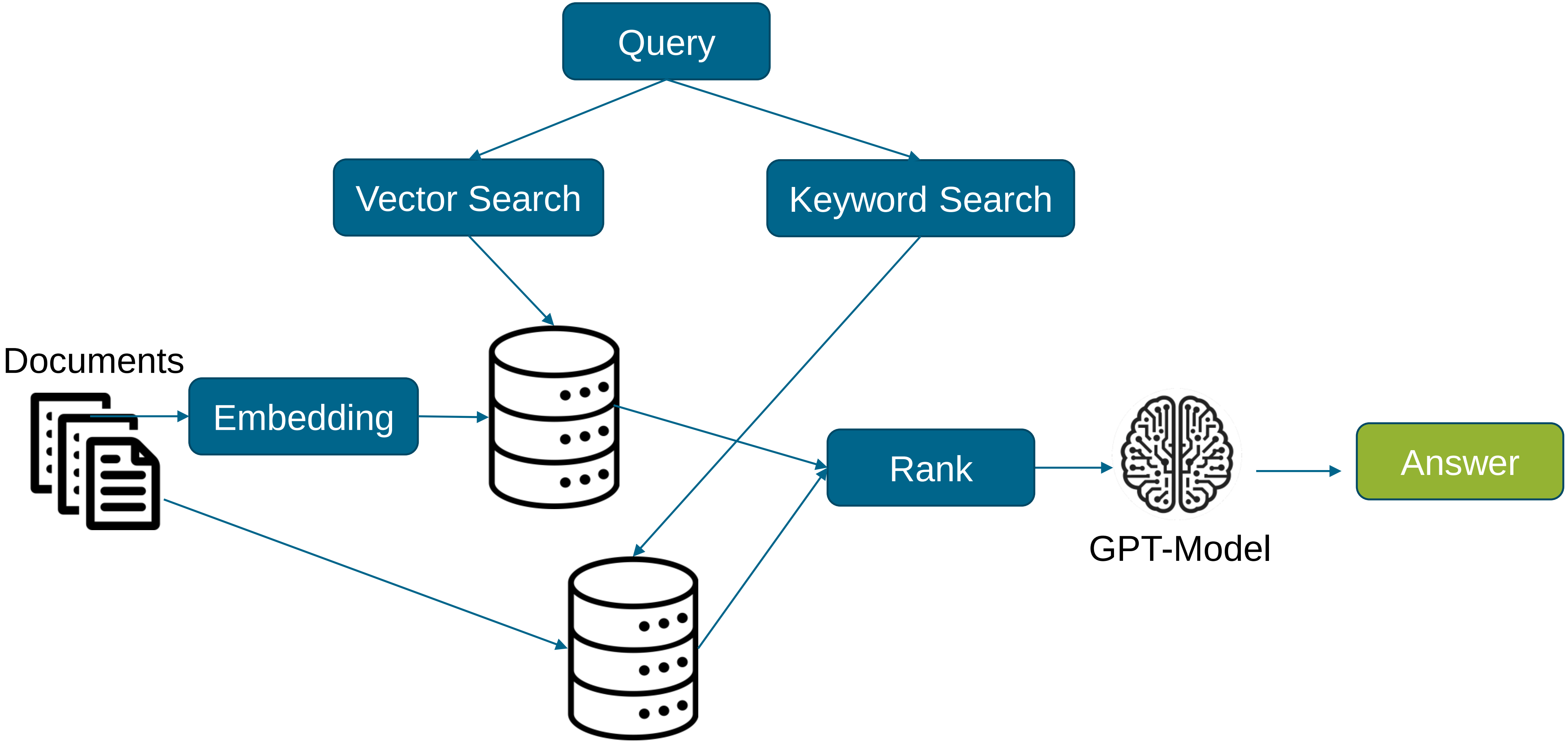
# Types of chunking

- Character splitting
- Recursive character text splitting
- Document specific splitting
- Semantic splitting
- Agentic splitting



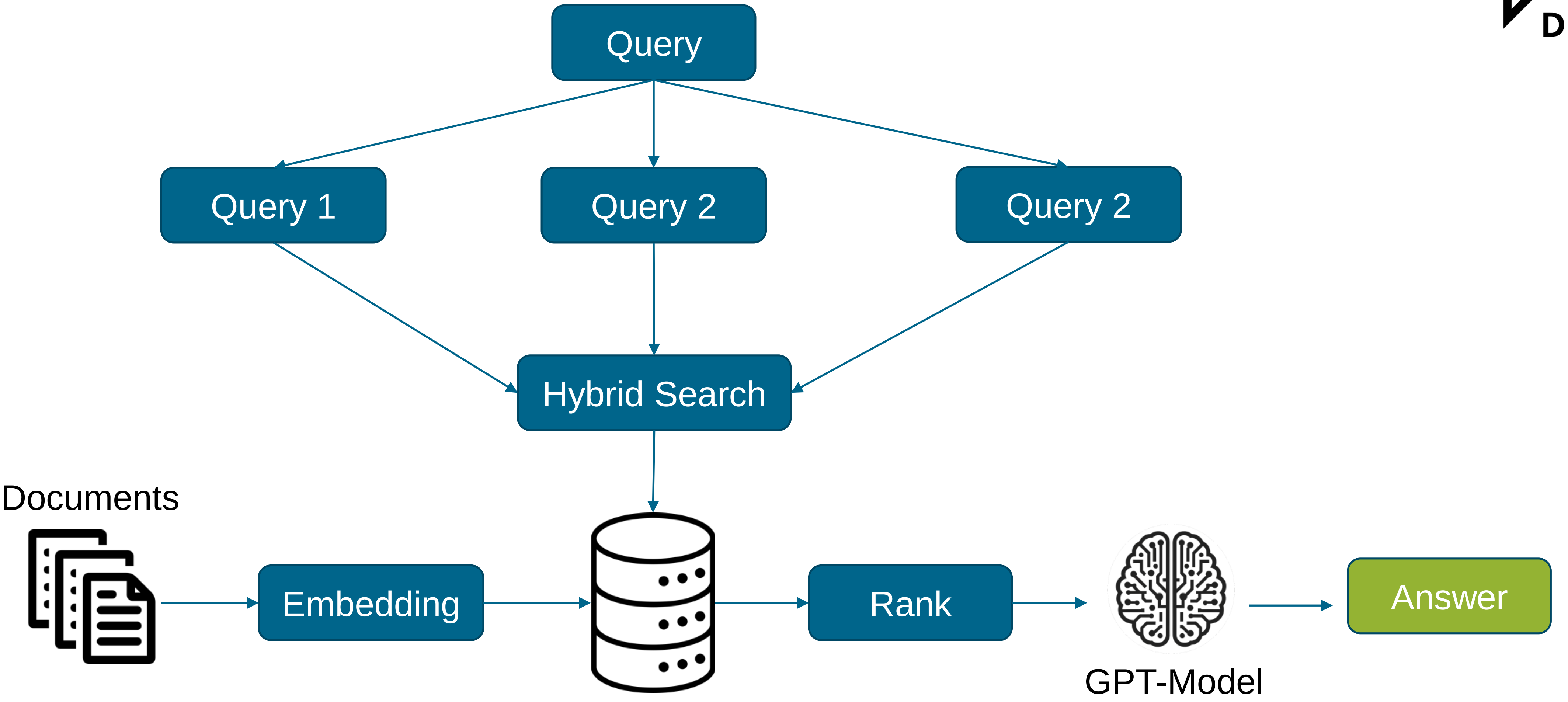
From: https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb

# Hybrid Search



From: https://medium.com/etoai/hybrid-search-rag-for-real-life-production-grade-applications-e1e727b3965a

# Query Rewriting



Query

Query 1

Query 2

Query 2

Hybrid Search

Documents

Embedding

Rank

GPT-Model

Answer

# HANDS-ON SESSION..