Lecture 7: LLM-1

AC215

Pavlos Protopapas
SEAS/Harvard

Outline

- BERT + GPT
- InstructGPT (ChatGPT)
- Prompt Engineering
- RAG

Announcements

- **Office Hours**

Li Yao - Tue (09.24) - IACS office - 2:30 - 3:30
Rashmi - Wed (09.25) - Zoom - 2:30 - 3:30

- **HW1 Due** - Fri 09/27 9PM EST

Outline

- **BERT + GPT**
- InstructGPT (ChatGPT)
- Prompt Engineering
- RAG

Chronology

1967 — **Eliza at MIT**
- Limited simulated conversations
- 1972 STNLP at MIT

1997 — **LSTM**

1999 — **Nvidia GPU**

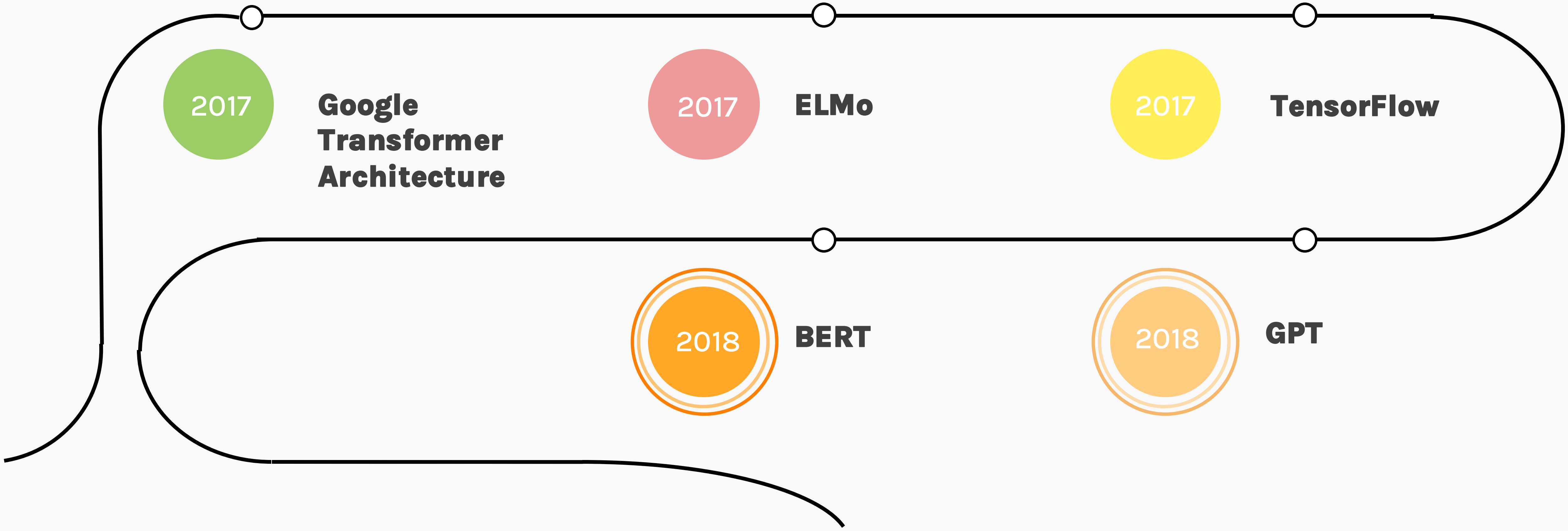
2006 — **FAIR**
- Facebook AI Research

2011 — **Google Brain**

2015 — **Open AI**

2016 — **Stanford CoreNLP**
- 2016 Stanford SQuAD dataset

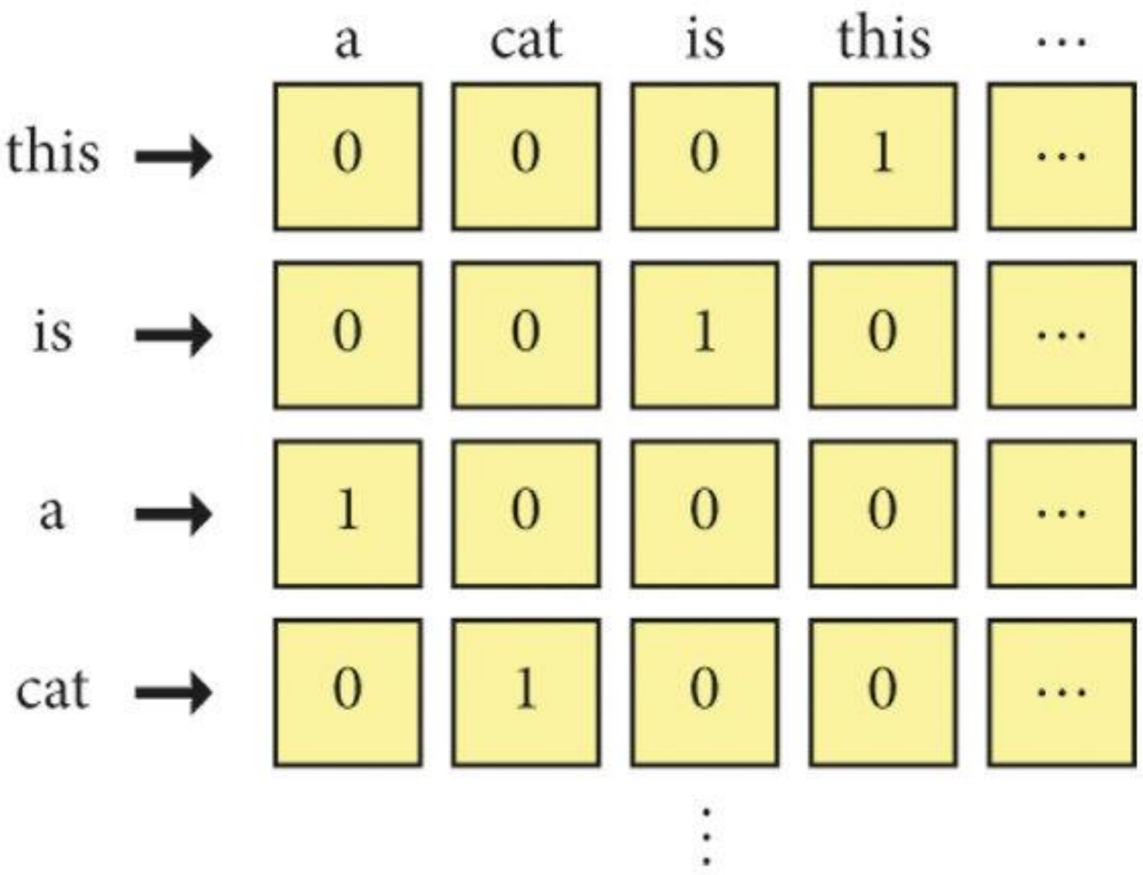
Chronology


2017
Google Transformer Architecture
2017
ELMo
2017
TensorFlow
2018
BERT
2018
GPT

Word Embeddings

A word embedding is **any** fixed-length vector representation of a token (word).

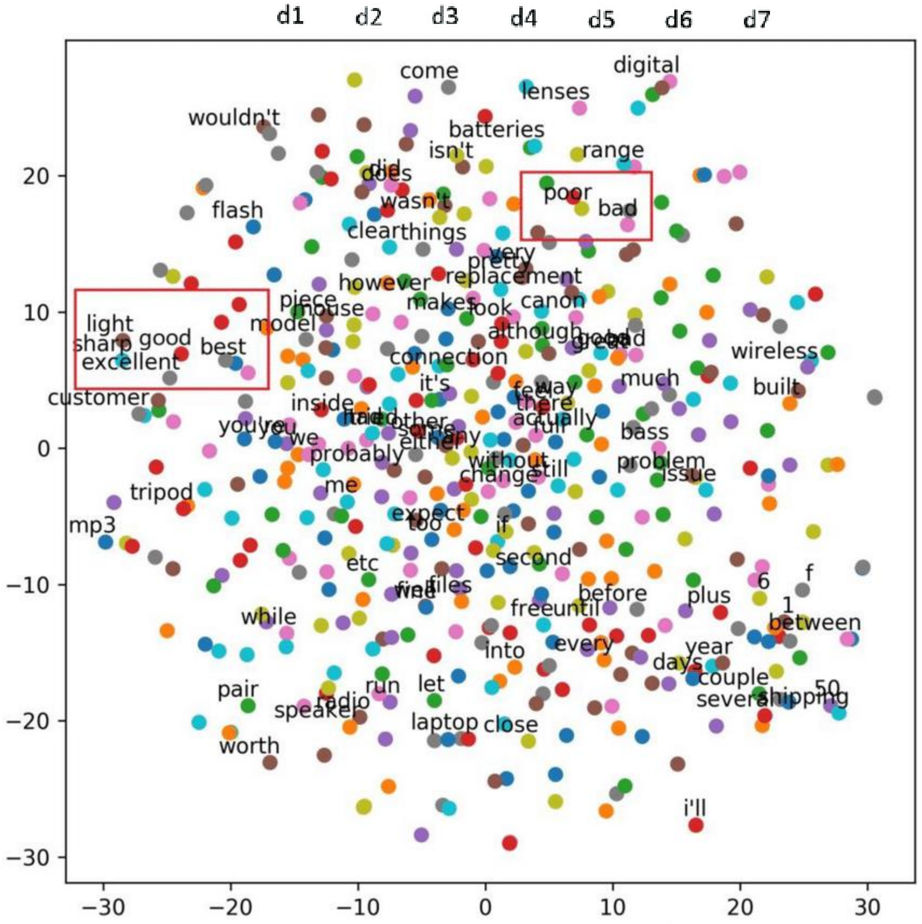
It can be as simple as a one-hot encoding, where it takes a value of one on the position of the word, and zero elsewhere.



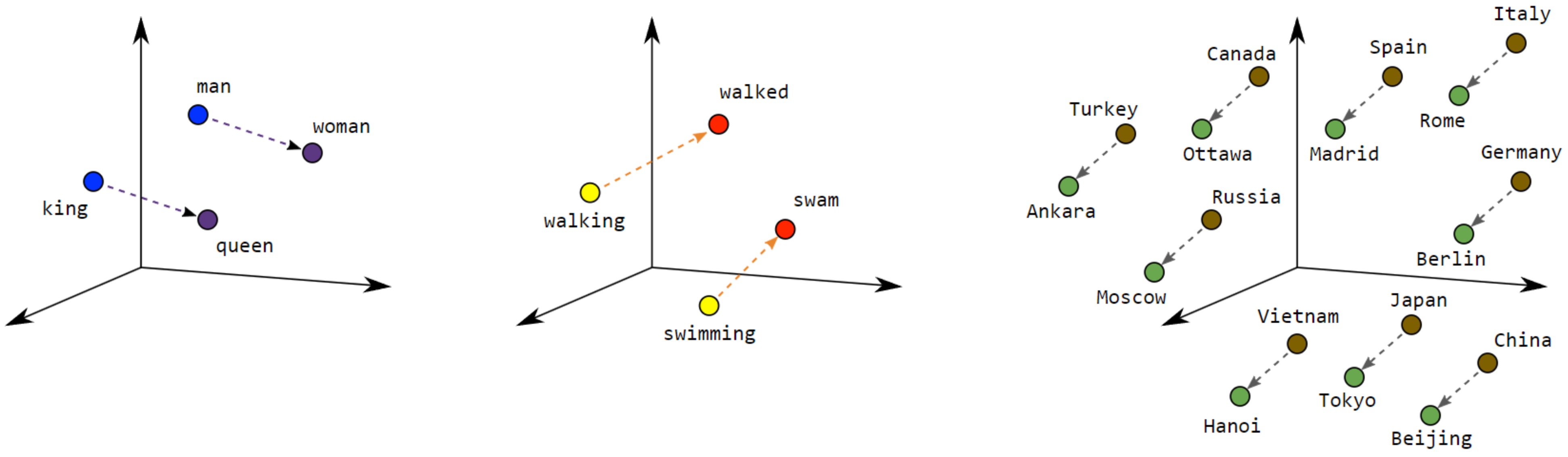
Word Embeddings

A more informative embedding can utilize the N-dimensional space entirely.

It assigns a value between -1 and 1 at each dimension. Allowing a denser structure of the vectors, that might retain semantic information of the data.



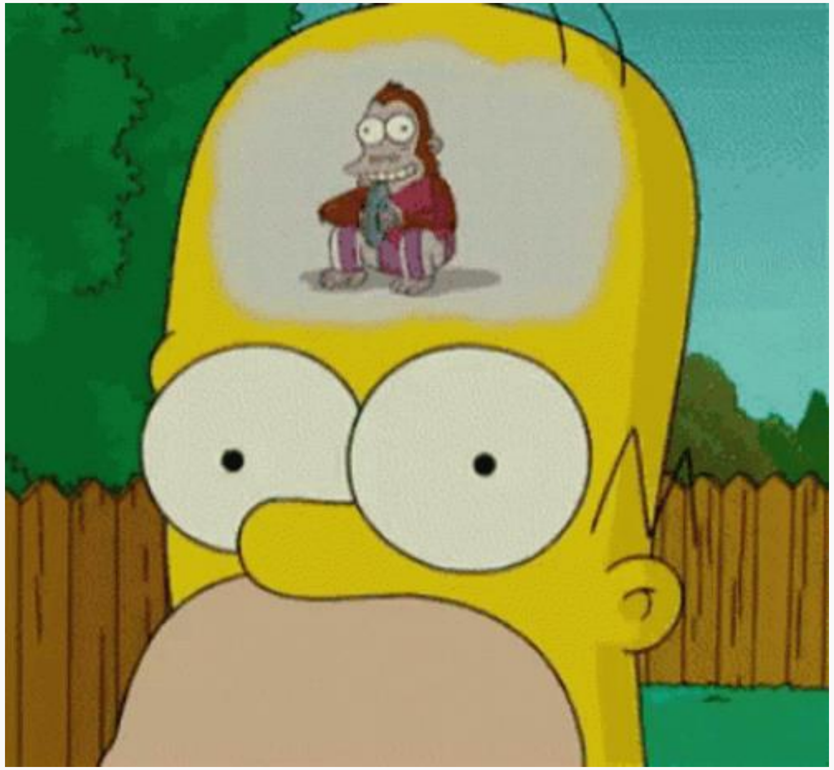
Word Embeddings



Male-Female   Verb Tense   Country-Capital

The semantic meaning is represented as the closeness of similar words. A word can be close to many conceptually different ones, since it is computed in a high-dimensional space.

Are we done?



Ambiguities

Ambiguity in Sequential Reading

The **bank** is open on Fridays.

I went to the **bank** to take a walk by the river.

The pilot made a sharp **bank** to the left.

A **bank** of lights illuminated the stadium.

How do we deal with this?

Attention - why?

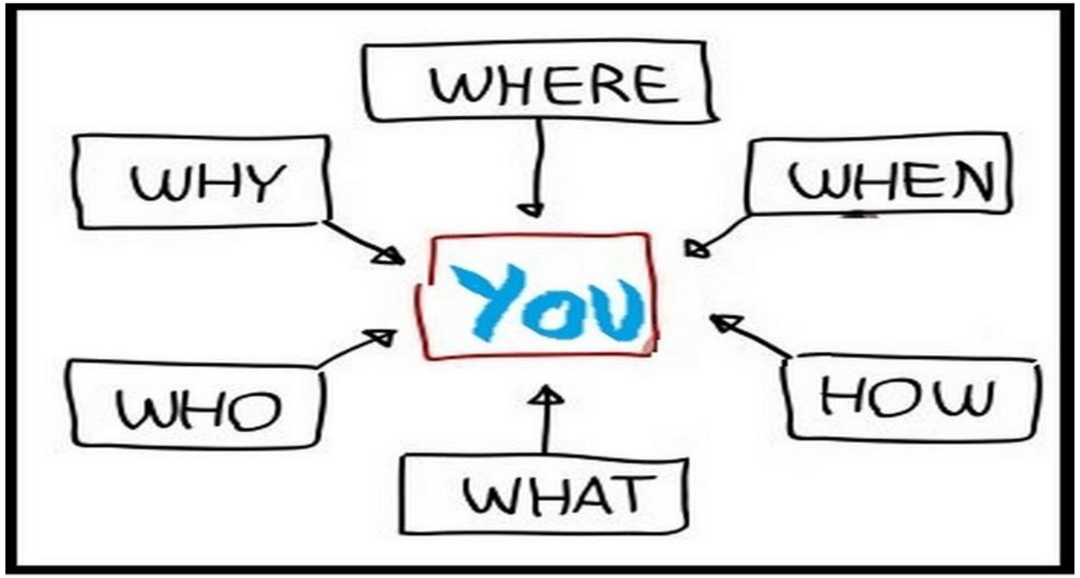
An embeddings that depends only on the word itself cannot account for polysemy.

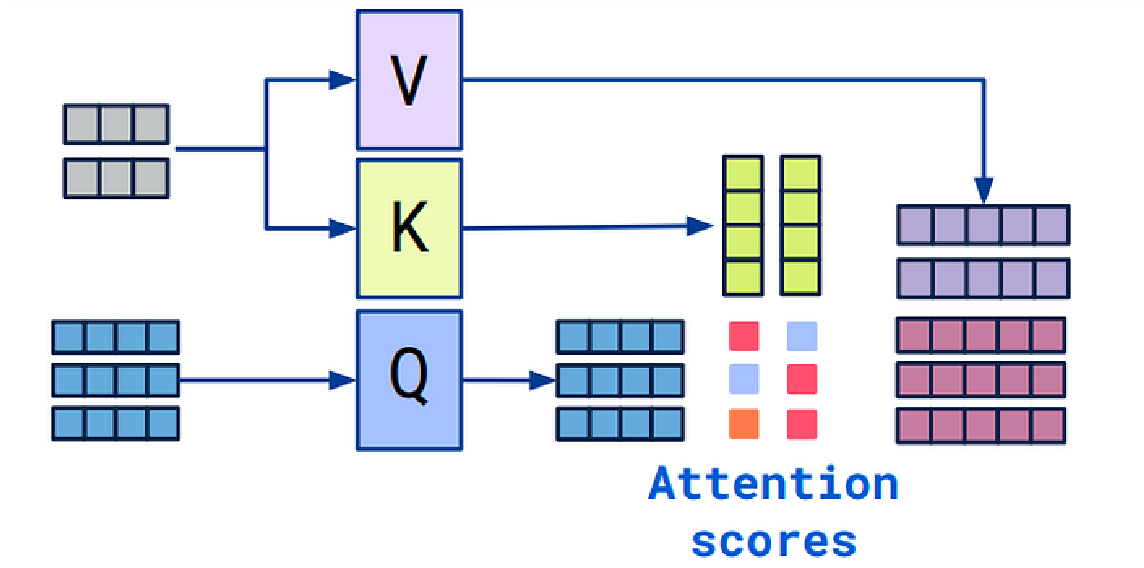
To extract a better embedding of a word, the sequence itself must be analyzed. The most intuitive way is to do it sequentially, which made popular Recurrent Neural Networks, such as LSTMs or GRUs.

The main drawbacks were:

1. Insufficient memory for long sequences
2. Slow training speed because of their Markovian properties
3. Sensitive to exploding or vanishing gradients

Attention





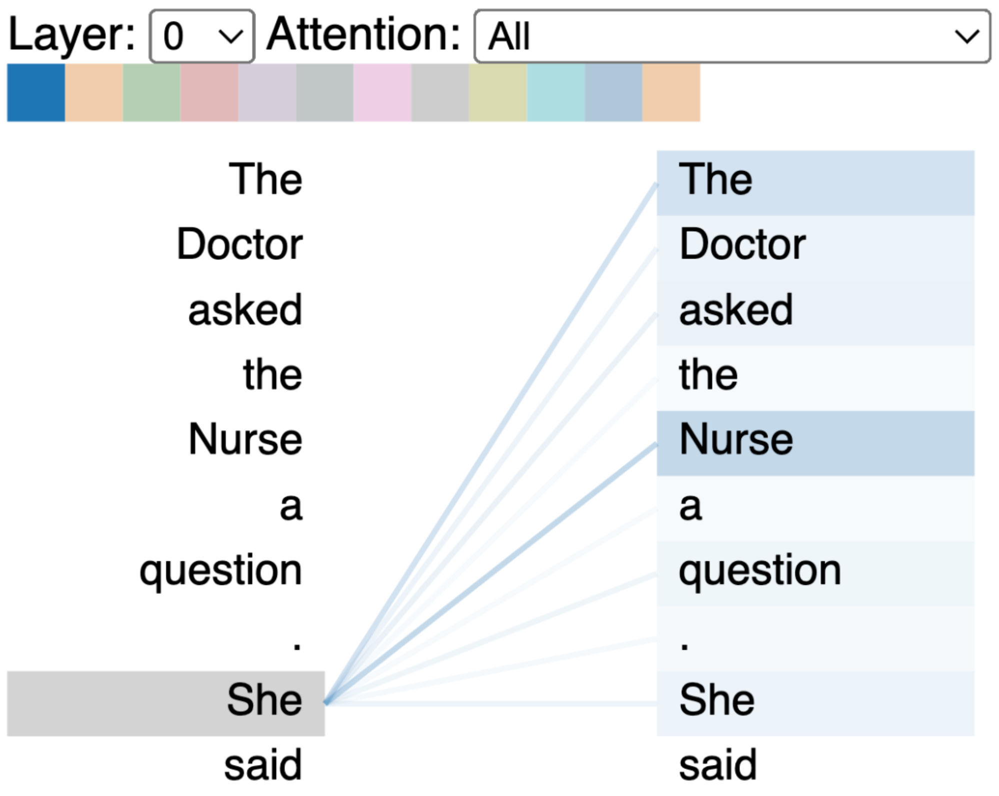
Attention mechanisms became a way to address the limitations of RNNs. In particular, the transformer architecture, based on the Key-Query-Value matrices.

Self Attention

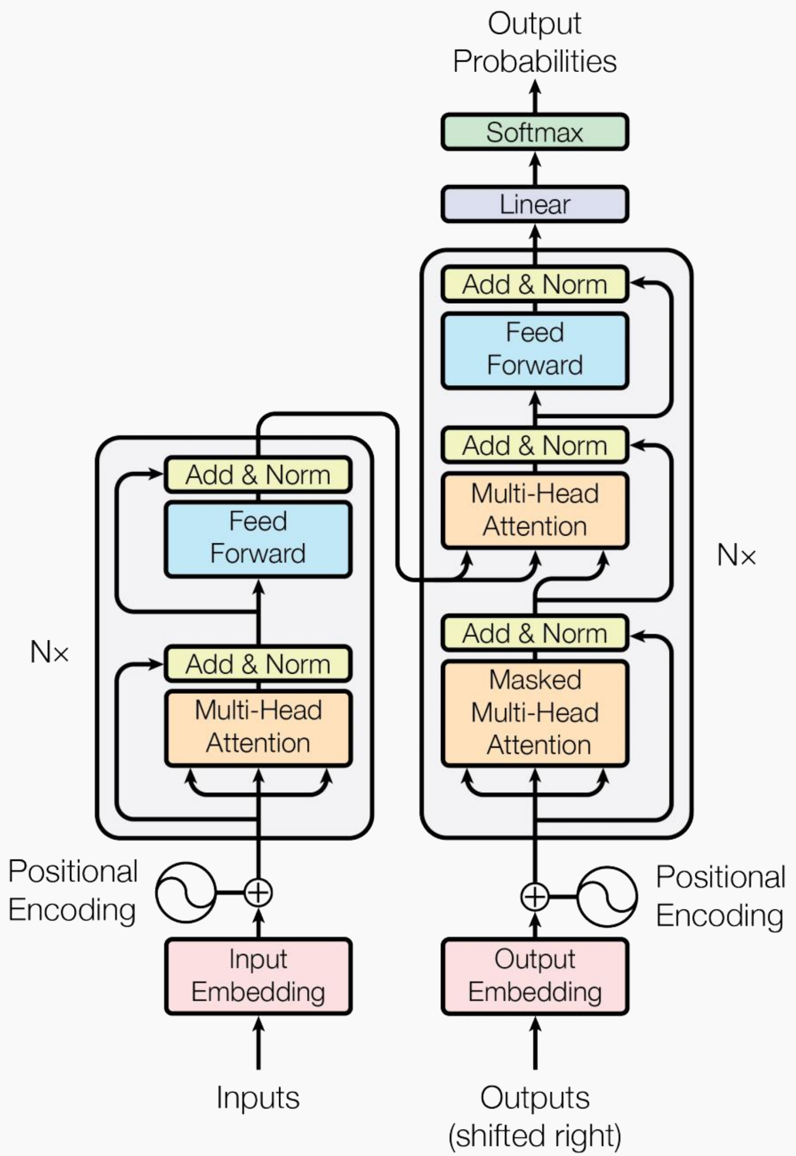
In the self-attention architecture, each embedding pays attention to every other element.

It is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence

Each element becomes query, key, and value from the input embeddings by multiplying by a weight matrix



Transformers



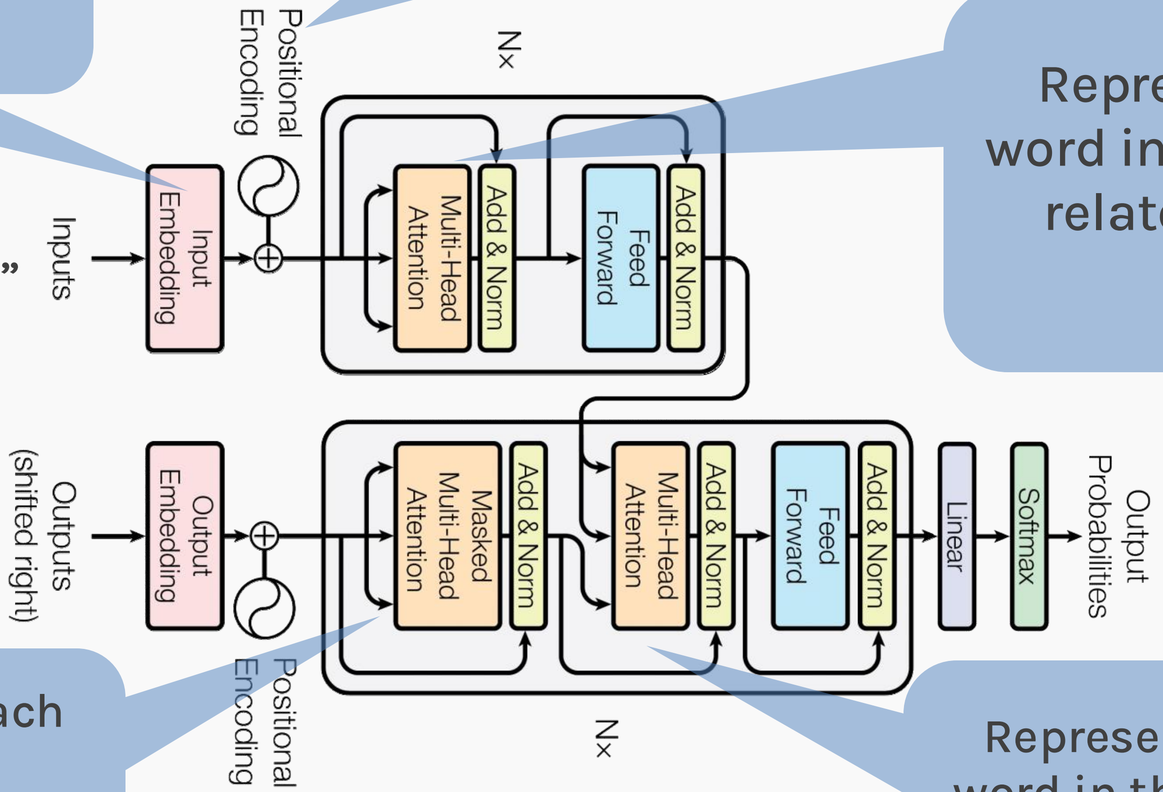
Transformers

Maps words to a latent space where similar words are mapped together

Encodes information about the position of the input embedding in the sequence to get a notion of context

English
“Sentence to be translated”

Represents how much each word in the **English** sentence is related to every word in the **same sentence**.



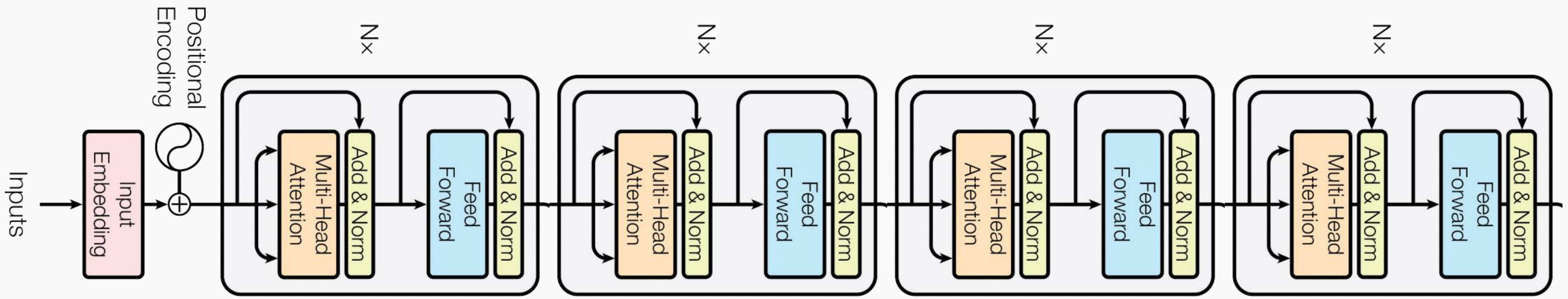
Spanish
“Oración por traducir”

Represents how much each word in the **Spanish** sentence is related to every word in the **same sentence**.

Represents how much each word in the **Spanish** sentence is related to every word in the **English** sentence.

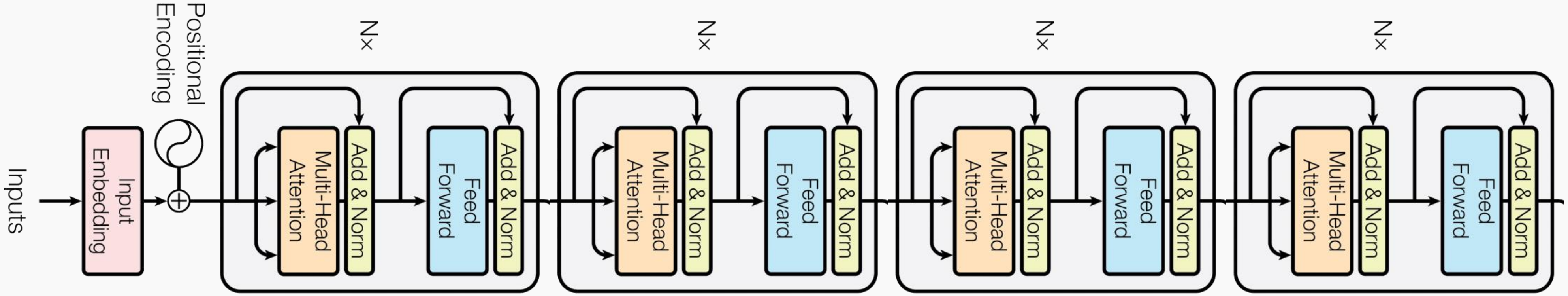
Transformers

Bidirectional Encoder Representation of Transformer (BERT):

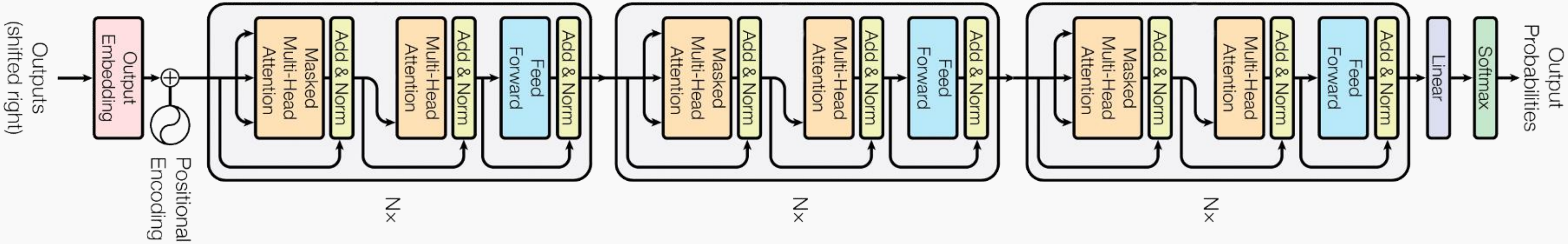


Transformers

Bidirectional Encoder Representation of Transformer (BERT):



Generative Pre-Trained Transformer (GPT):

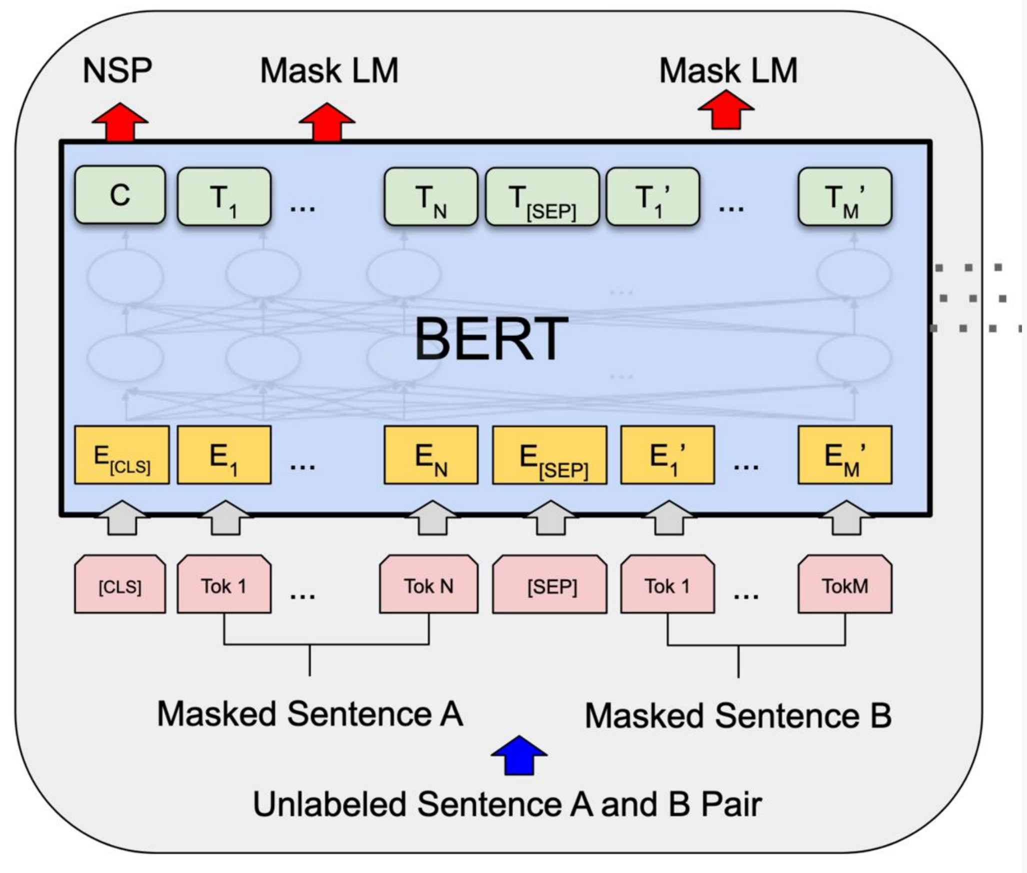


BERT

BERT Summary

BERT is designed to extract semantic representations of each word in a paragraph.

It uses a multi-head, multilayer self-attention architecture, trained in a multi-task setting: a masked language and a next sentence prediction model.



Given a sentence it produces a set of embeddings per token.

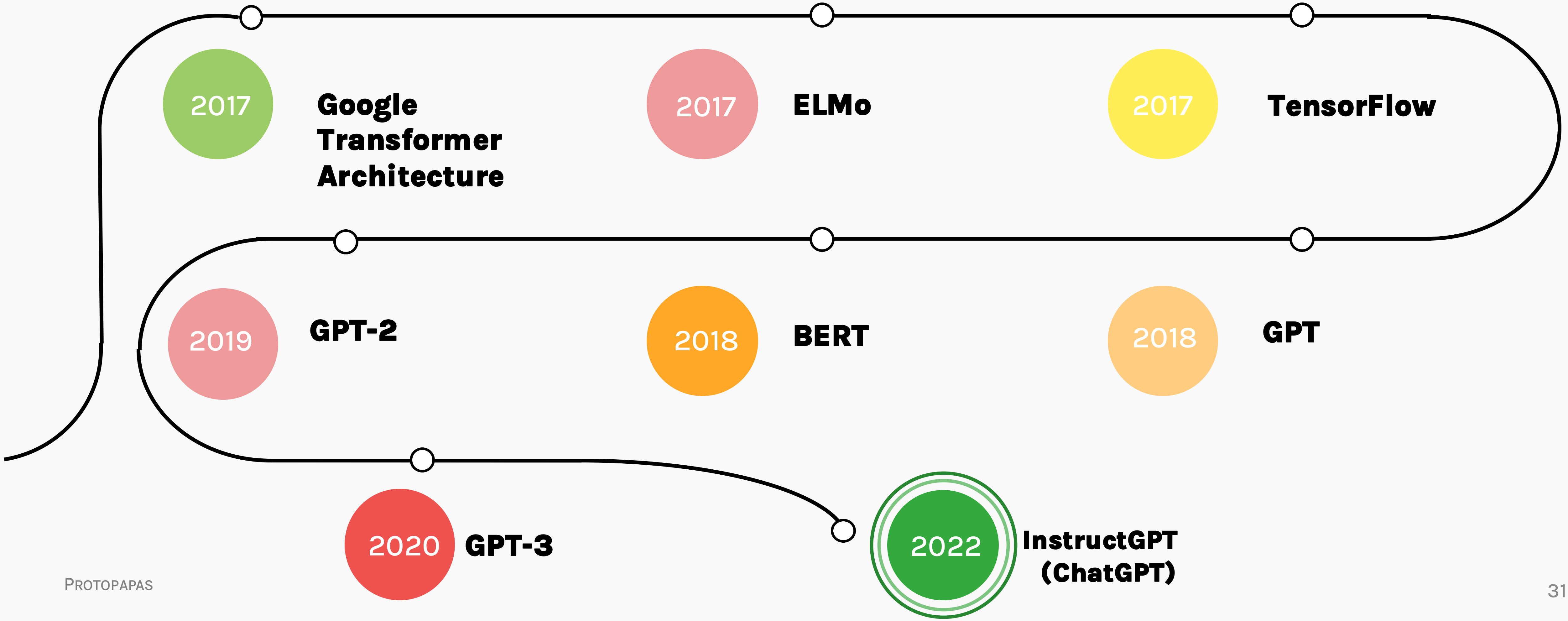
GPT

GO TO Hugging face and ask a question GPT2
https://huggingface.co/openai-community/gpt2

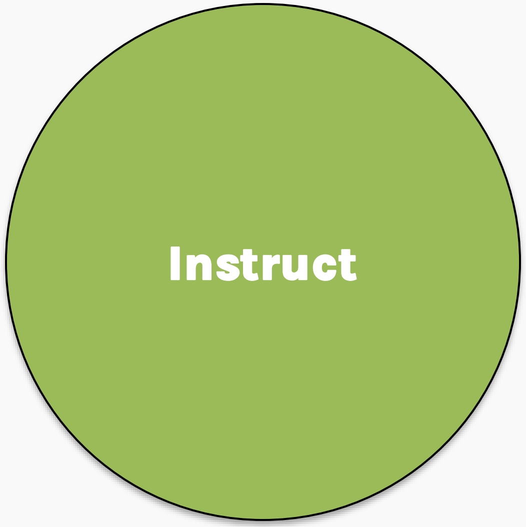
Outline

- BERT + GPT
- **InstructGPT (ChatGPT)**
- Prompt Engineering
- RAG

Chronology


2017
Google Transformer Architecture
2017
ELMo
2017
TensorFlow
2018
GPT
2018
BERT
2019
GPT-2
2020
GPT-3
2022
InstructGPT (ChatGPT)

Training Cycle - Instruct GPT



Objective:
The goal is to make the model useful for specific tasks and improving its ability to follow instructions.

Process:
Fine-tuning the model on datasets that contain instructions and the desired outputs.

This also includes RLHF.

Outcome:
A model that becomes better at interpreting and following user instructions.

Instruct GPT: Training

- GPT models are trained to predict the next word in a sentence given the context of the previous words.
- The model does not have access to the specific instructions or intentions of the user. Therefore, it may not always align answers with what the user wants.
- Reinforcement Learning from Human Feedback (RLHF) is used to incorporate human feedback into the training process to better align the model outputs with user intent.

Instruct GPT: Training

We will break it down into 3 steps:

Supervised fine tuning model

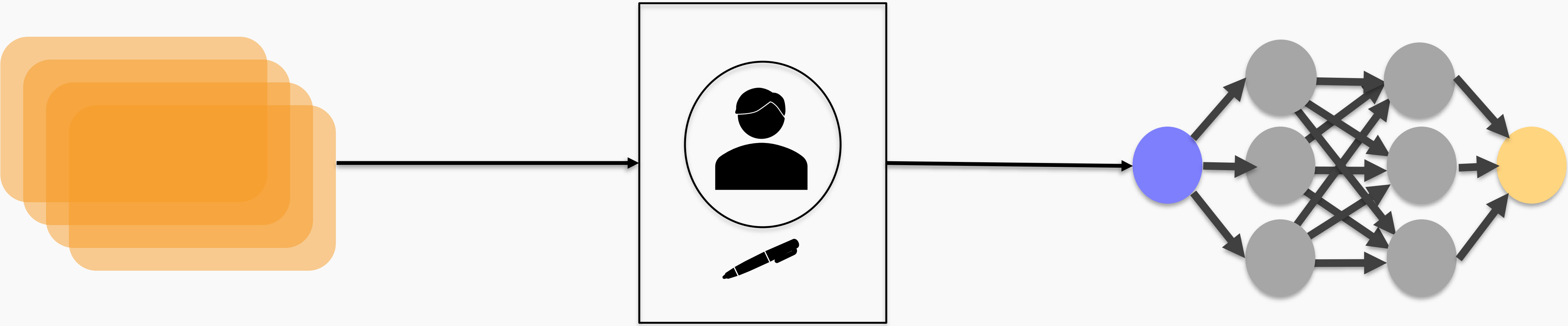
Training a reward model

Fine-tuning the model with reinforcement learning

Instruct GPT: Training

Supervised fine tuning model

The data is a web-scale corpus of data including correct and incorrect solutions to math problems, weak and strong reasoning, self-contradictory and consistent statements, and representing a great variety of ideologies and ideas.



Prompt sampled from dataset

40 contractors wrote the desired output behavior

Input/Output pairs are used to fine-tune GPT with supervised learning

Instruct GPT: Training

We will break it down into 3 steps:

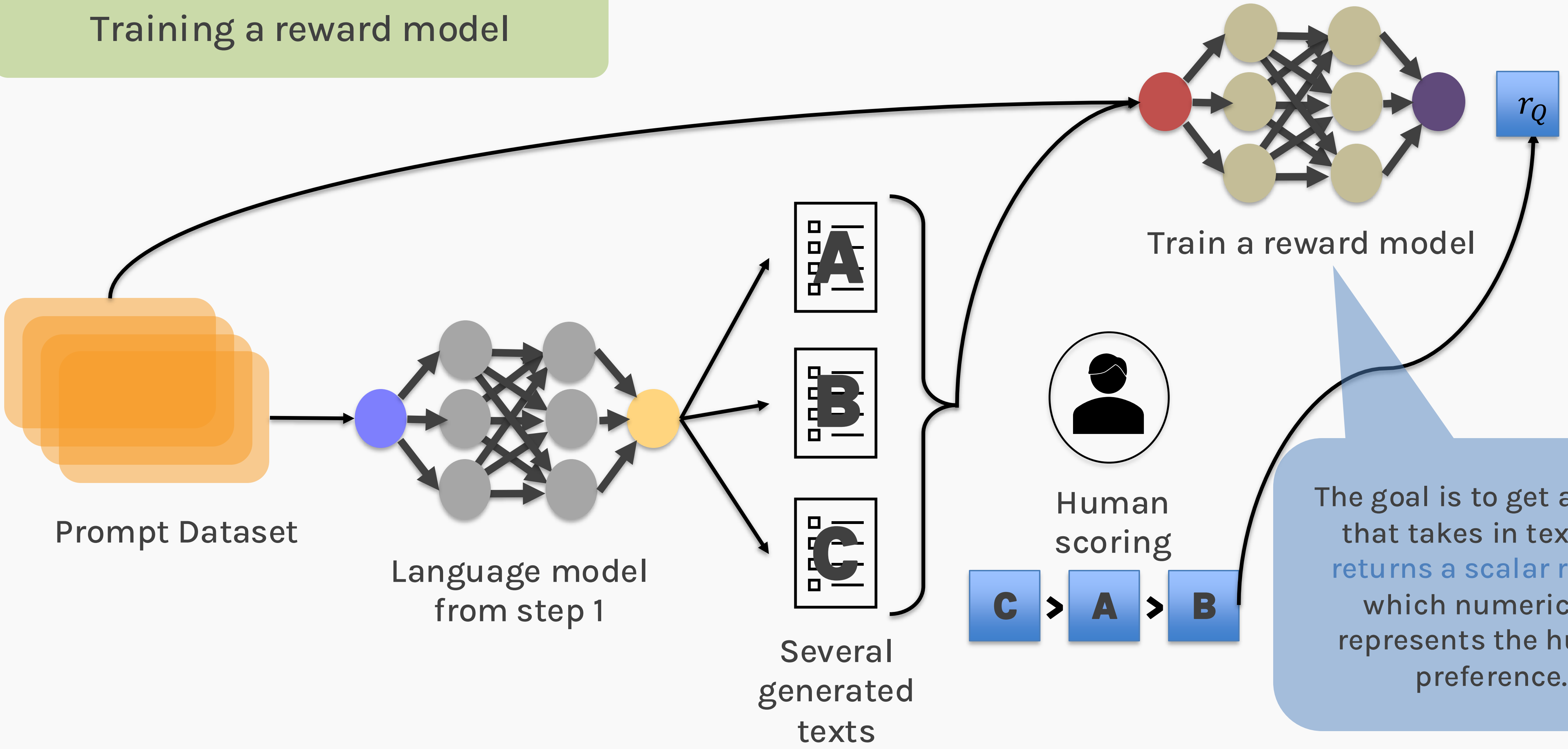
Supervised fine tuning model

Training a reward model

Fine-tuning the model with reinforcement learning

Instruct GPT: Training

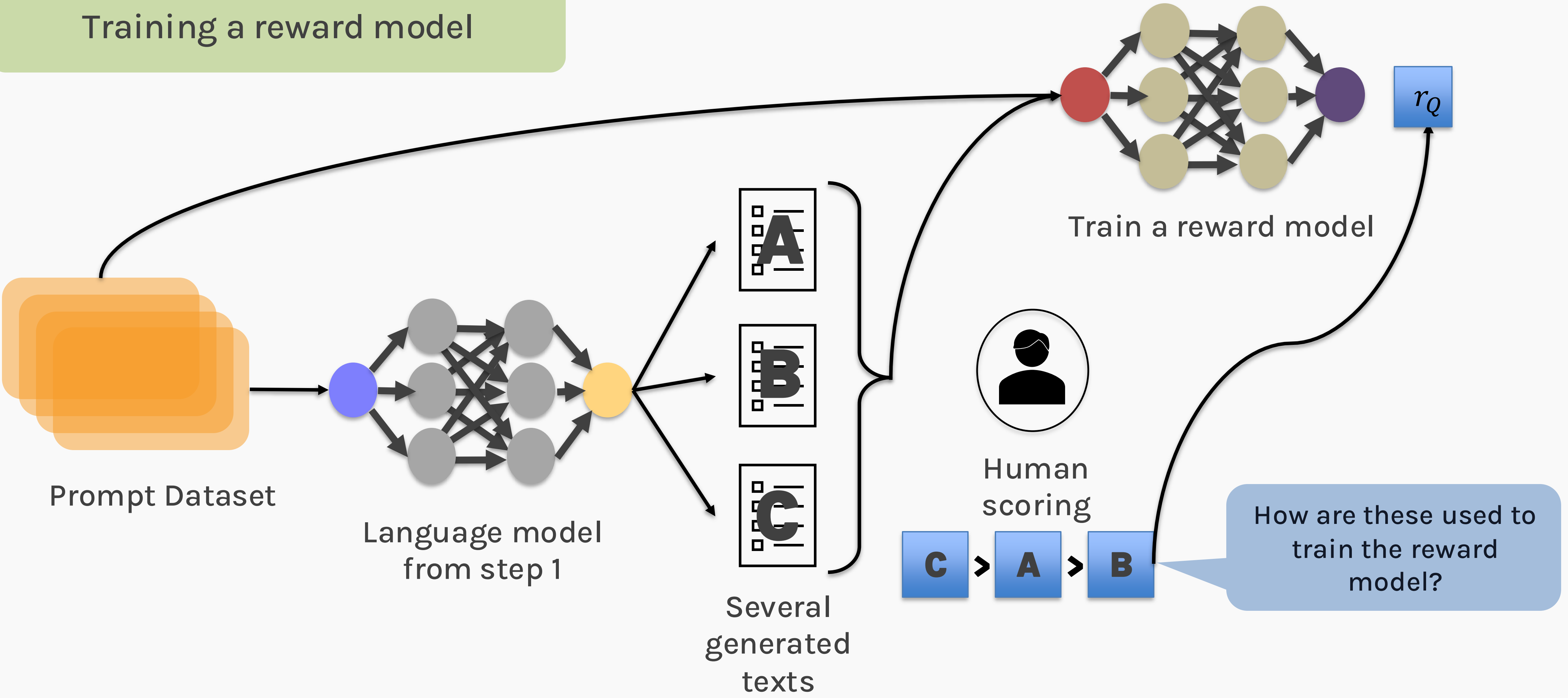
Training a reward model



The goal is to get a model that takes in text and returns a scalar reward which numerically represents the human preference.

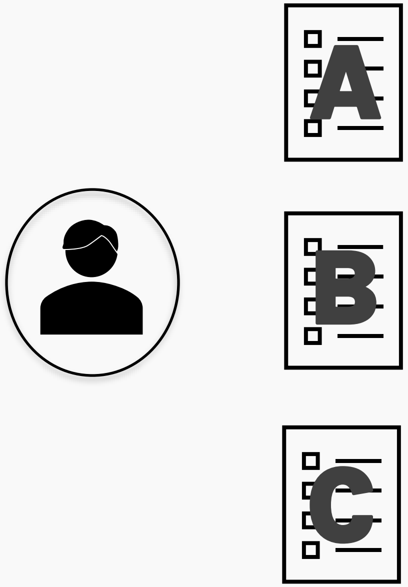
Instruct GPT: Training

Training a reward model

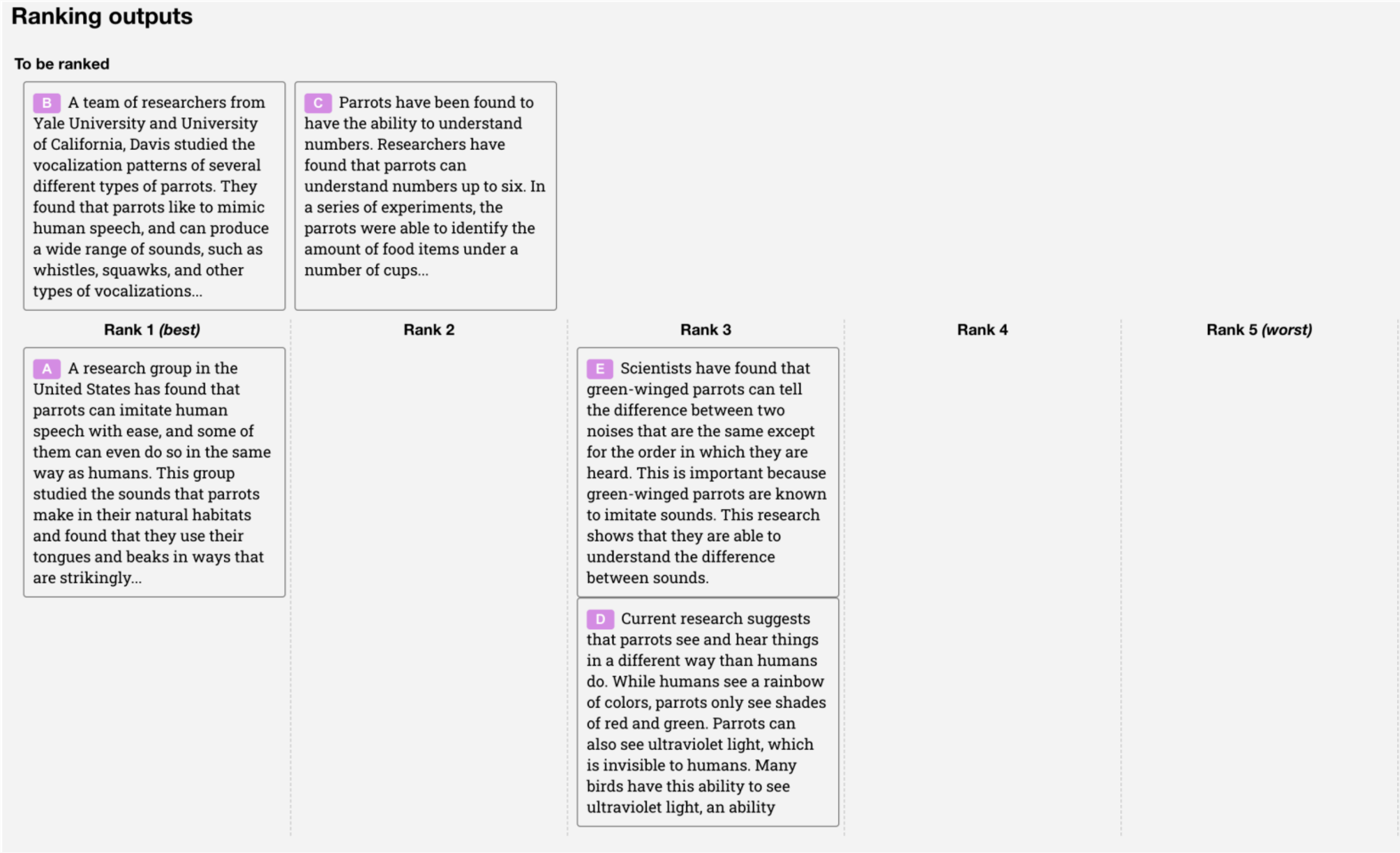


Instruct GPT: Training

Training a reward model

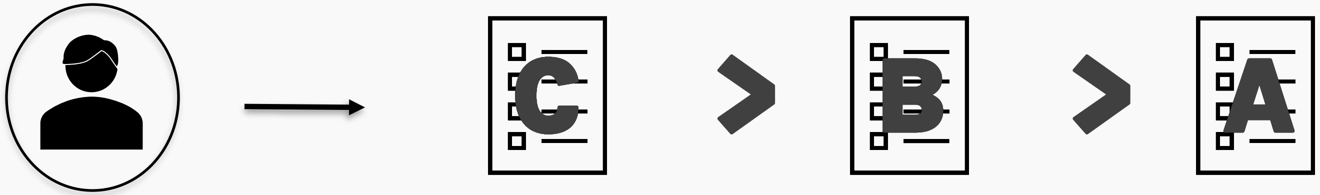


Several generated texts



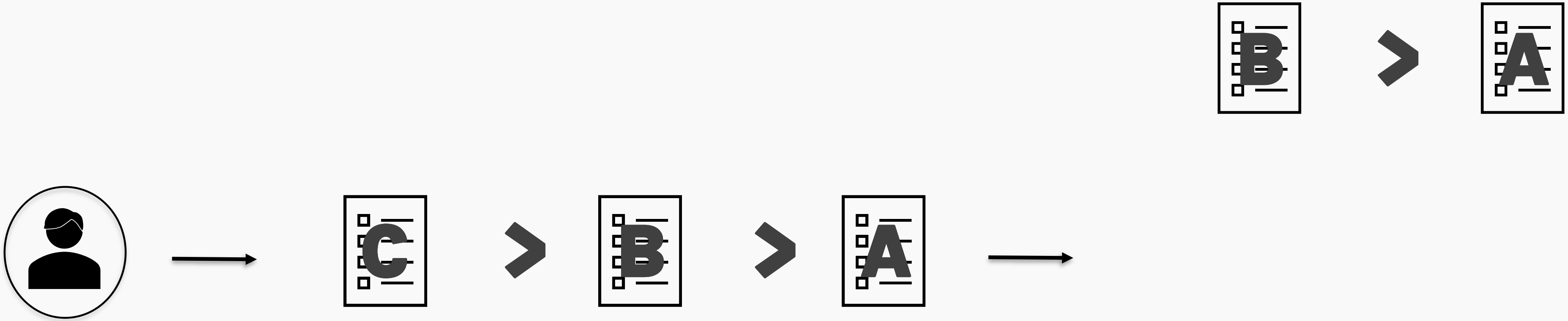
Instruct GPT: Training

Training a reward model



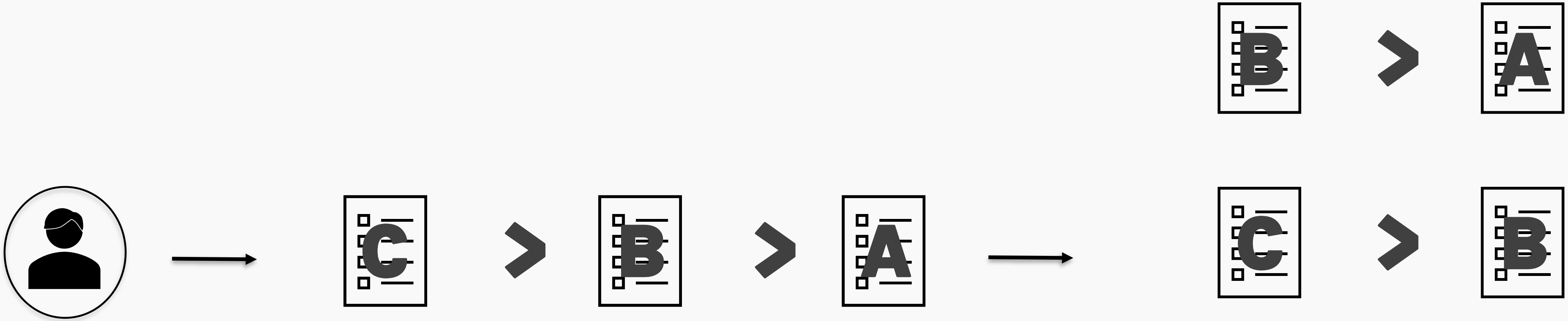
Instruct GPT: Training

Training a reward model



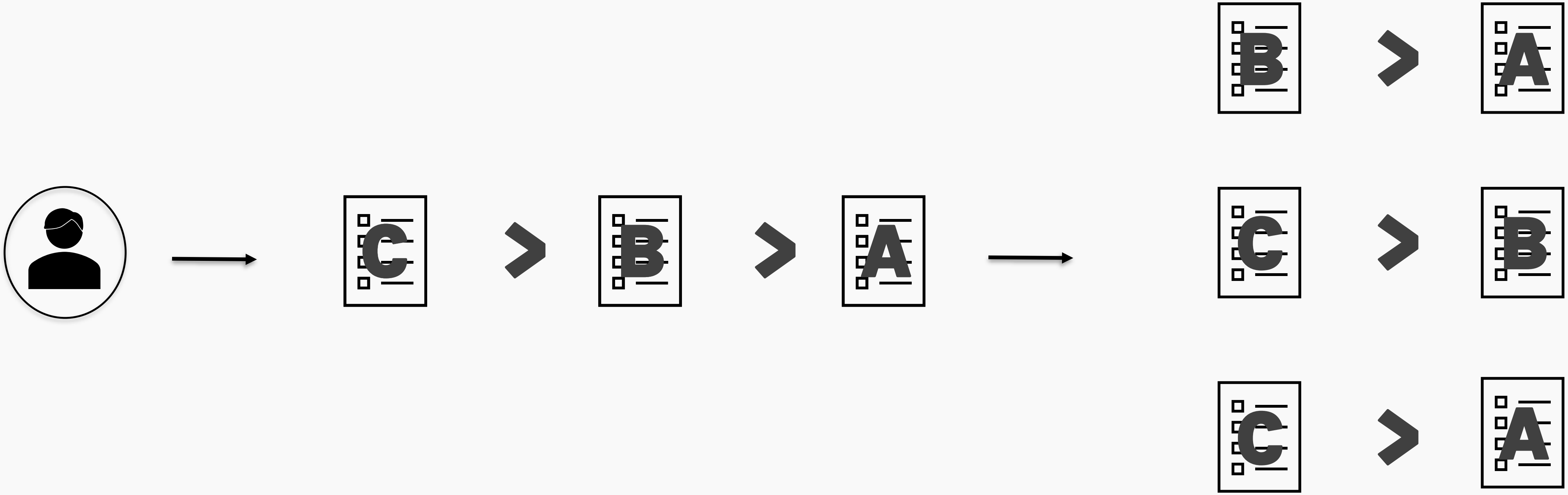
Instruct GPT: Training

Training a reward model



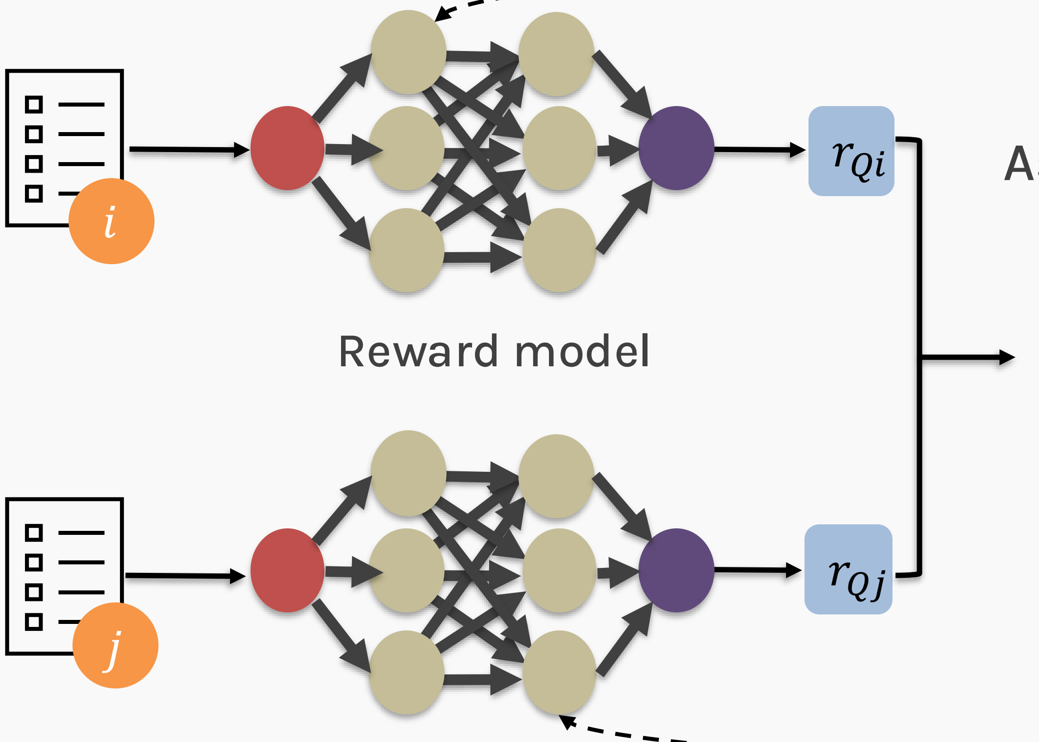
Instruct GPT: Training

Training a reward model



Instruct GPT: Training

Training a reward model



Two different responses of the supervised fine tuned (SFT) model

Scalar output of the reward model for prompt x and response y

Assuming response i is ranked higher than response j,

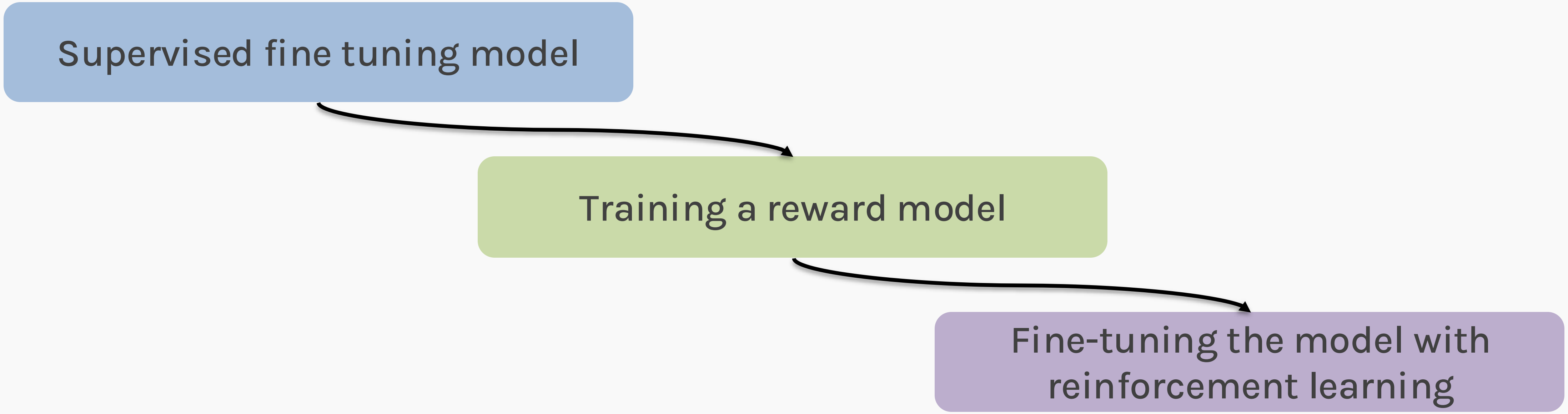
$$Loss = -\frac{1}{\binom{k}{2}} \mathbb{E}_{(x, y_i, y_j \sim D)}[\log(\sigma(r_Q(x, y_i) - r_Q(x, y_j)))]$$

$\binom{k}{2}$ represents the total number of possible comparisons if there are k responses to a given prompt

Dataset of human comparisons

Instruct GPT: Training

We will break it down into 3 steps:



Instruct GPT: Training

Fine-tuning the model with reinforcement learning

Let's first formulate this fine-tuning task as a RL problem:

- **Policy:** A language model that takes in a prompt and returns a sequence of text.
- **Action space:** All the tokens corresponding to the vocabulary of the language model (responses).
- **Reward function:** A combination of the rewards model and a constraint on policy shift. This is where the system combines all the models we have discussed into one RLHF process.

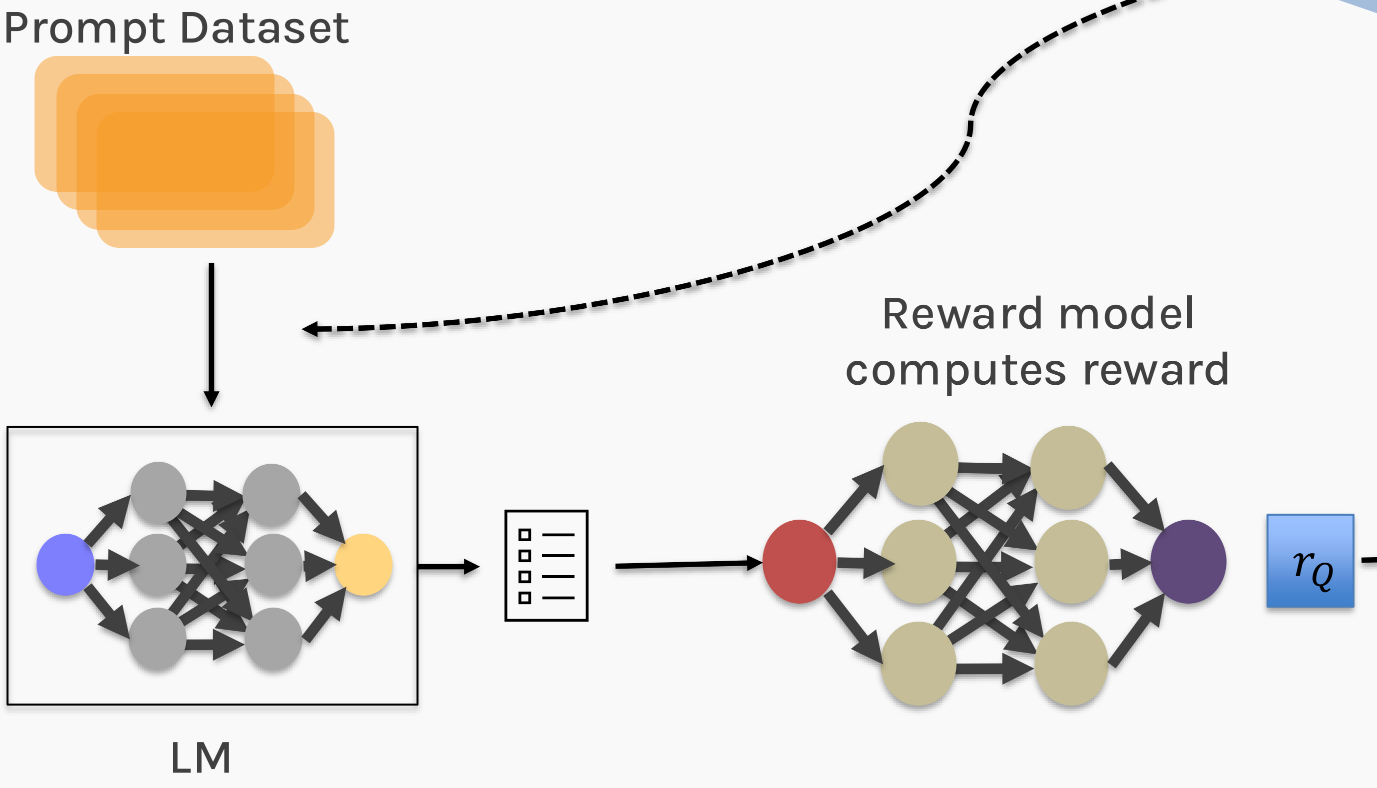
Instruct GPT: Training

Fine-tuning the model with reinforcement learning

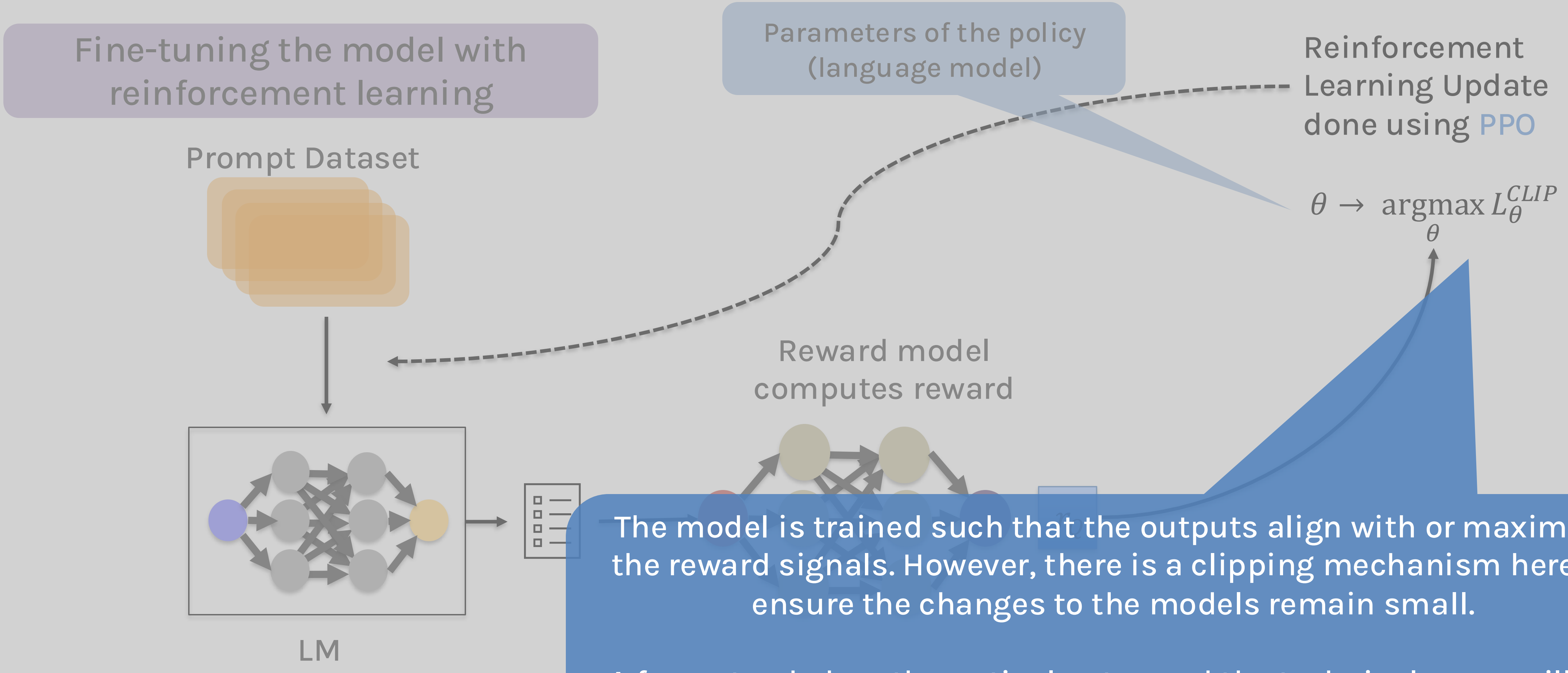
Parameters of the policy (language model)

Reinforcement Learning Update done using PPO

$$\theta \rightarrow \underset{\theta}{\operatorname{argmax}} L_{\theta}^{CLIP}$$



GPT-4: Training

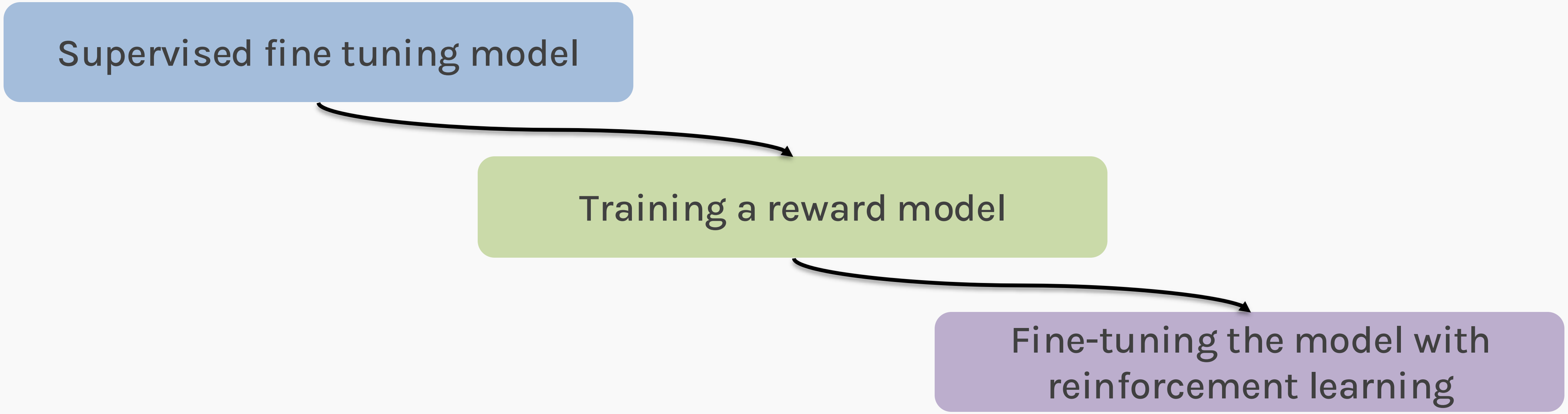


The model is trained such that the outputs align with or maximise the reward signals. However, there is a clipping mechanism here to ensure the changes to the models remain small.

A few extended mathematical notes and the technical paper will be available in your post class reading!

Training Summary of instruct GPT

We will break it down into 3 steps:



Training Cycle - LLM

Safety

Objective:
The goal is to make sure that the model outputs are safe and ethical.

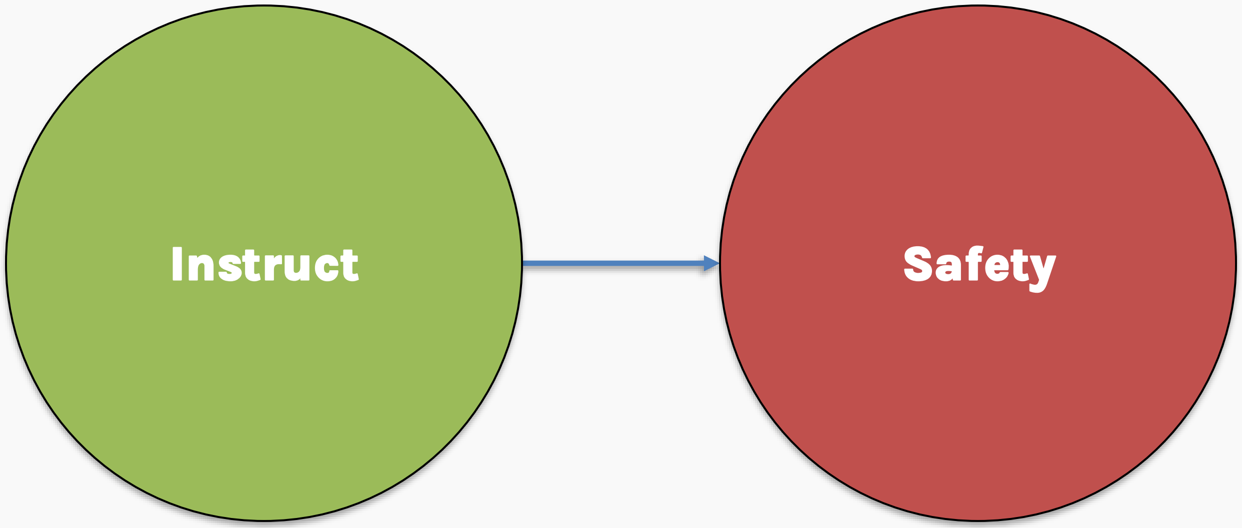
Process:
Involves further fine-tuning. We use RLHF to provide feedback on model outputs.

Outcome:
The model becomes safer reducing risk of biased content.

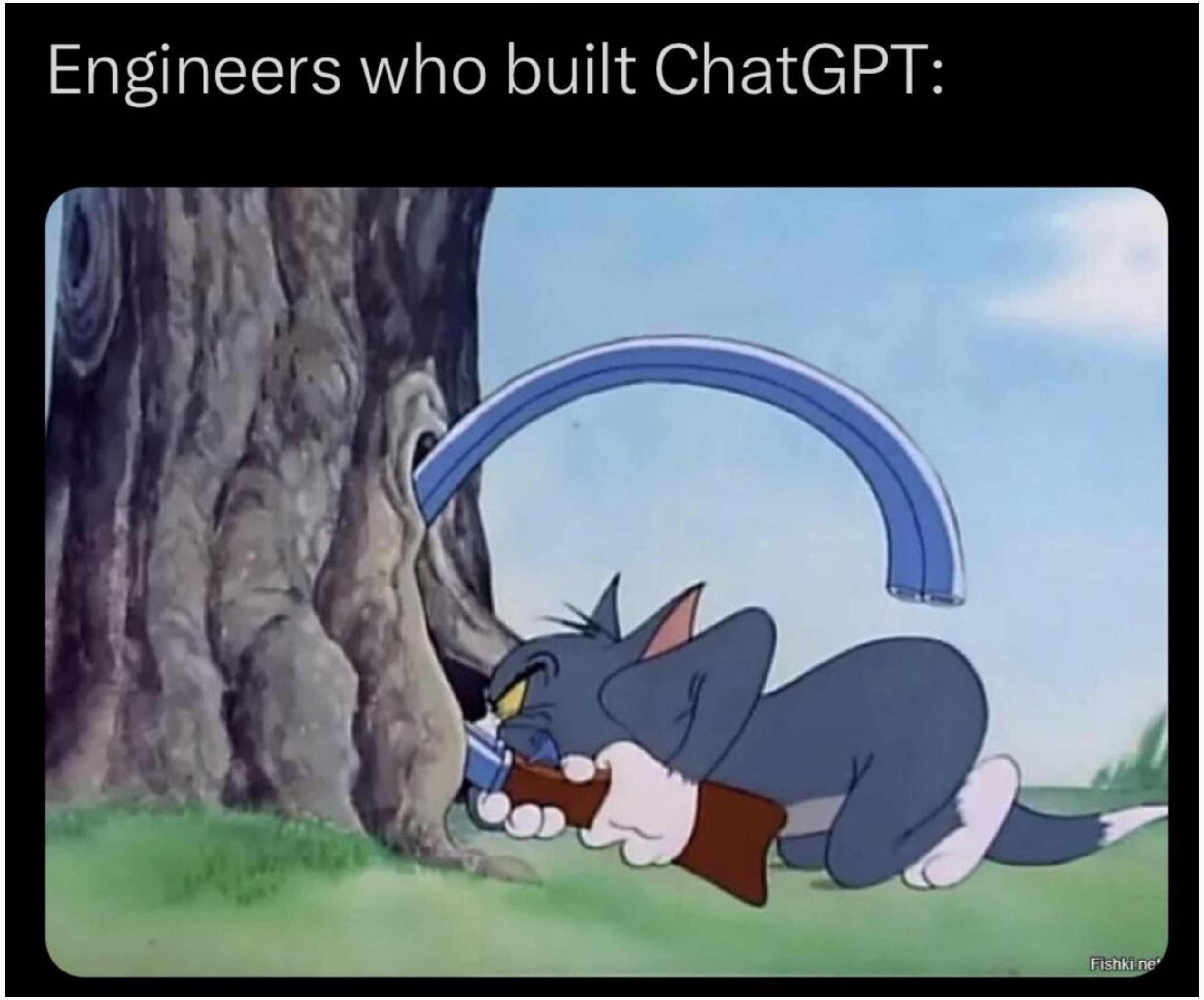
It's after this step that we get models like ChatGPT, Claude etc

Training Cycle - LLM

So, fine-tuning takes place in 2 stages.



InstructGPT(ChatGPT)



GPT-4: Capabilities

GPT-4 is a multimodal large language model with improved factuality, steerability, and guardrails after 6 months of iterative alignment.

Source: [GPT-4 Technical Report]

Outline

- BERT + GPT
- InstructGPT (ChatGPT)
- **Prompt Engineering**
- RAG

Prompt Engineering

https://huggingface.co/spaces/TinyLlama/tinyllama-chat

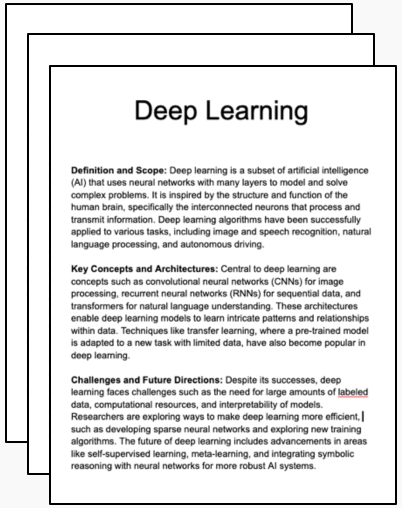
Outline

- BERT + GPT
- InstructGPT (ChatGPT)
- Prompt Engineering
- **RAG**

Introduction - RAG

If we have a large number of documents, how can we process/query it using an LLM?

1. Pass all the text directly into an LLM

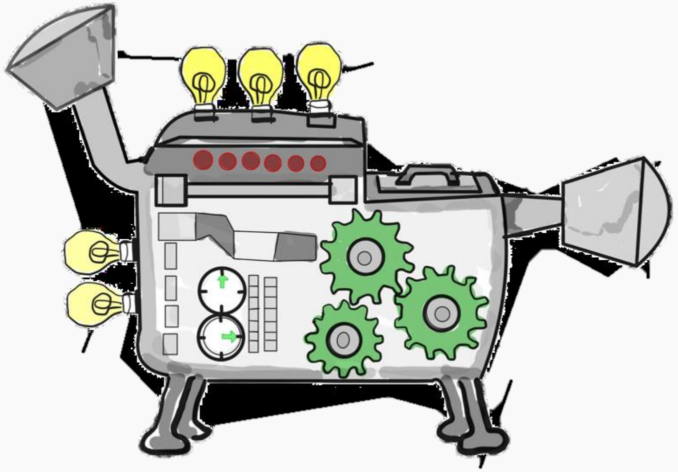


Deep Learning

Definition and Scope: Deep learning is a subset of artificial intelligence (AI) that uses neural networks with many layers to model and solve complex problems. It is inspired by the structure and function of the human brain, specifically the interconnected neurons that process and transmit information. Deep learning algorithms have been successfully applied to various tasks, including image and speech recognition, natural language processing, and autonomous driving.

Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

Challenges and Future Directions: Despite its successes, deep learning faces challenges such as the need for large amounts of labeled data, computational resources, and interpretability of models. Researchers are exploring ways to make deep learning more efficient, such as developing sparse neural networks and exploring new training algorithms. The future of deep learning includes advancements in areas like self-supervised learning, meta-learning, and integrating symbolic reasoning with neural networks for more robust AI systems.



LLM

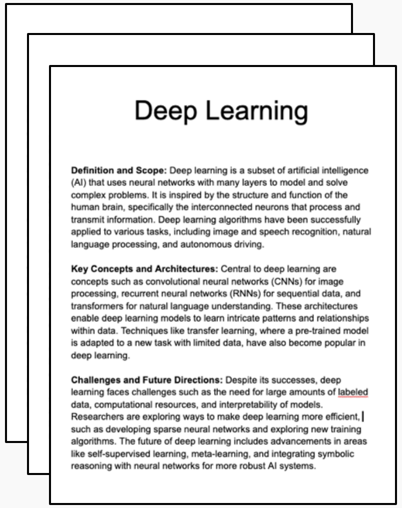
Garbage

If the context is too big, the LLM gives out <u>garbage</u>.

Introduction - RAG

If we have a large number of documents, how can we process/query it using an LLM?

2. We can finetune our LLM using the data.

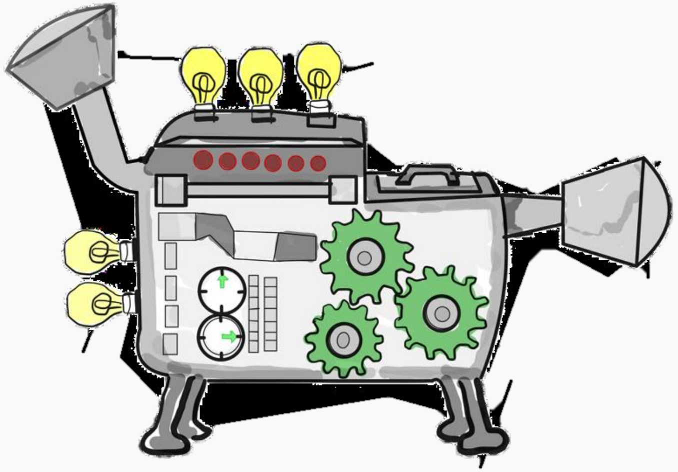


Deep Learning

Definition and Scope: Deep learning is a subset of artificial intelligence (AI) that uses neural networks with many layers to model and solve complex problems. It is inspired by the structure and function of the human brain, specifically the interconnected neurons that process and transmit information. Deep learning algorithms have been successfully applied to various tasks, including image and speech recognition, natural language processing, and autonomous driving.

Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

Challenges and Future Directions: Despite its successes, deep learning faces challenges such as the need for large amounts of labeled data, computational resources, and interpretability of models. Researchers are exploring ways to make deep learning more efficient, such as developing sparse neural networks and exploring new training algorithms. The future of deep learning includes advancements in areas like self-supervised learning, meta-learning, and integrating symbolic reasoning with neural networks for more robust AI systems.



LLM

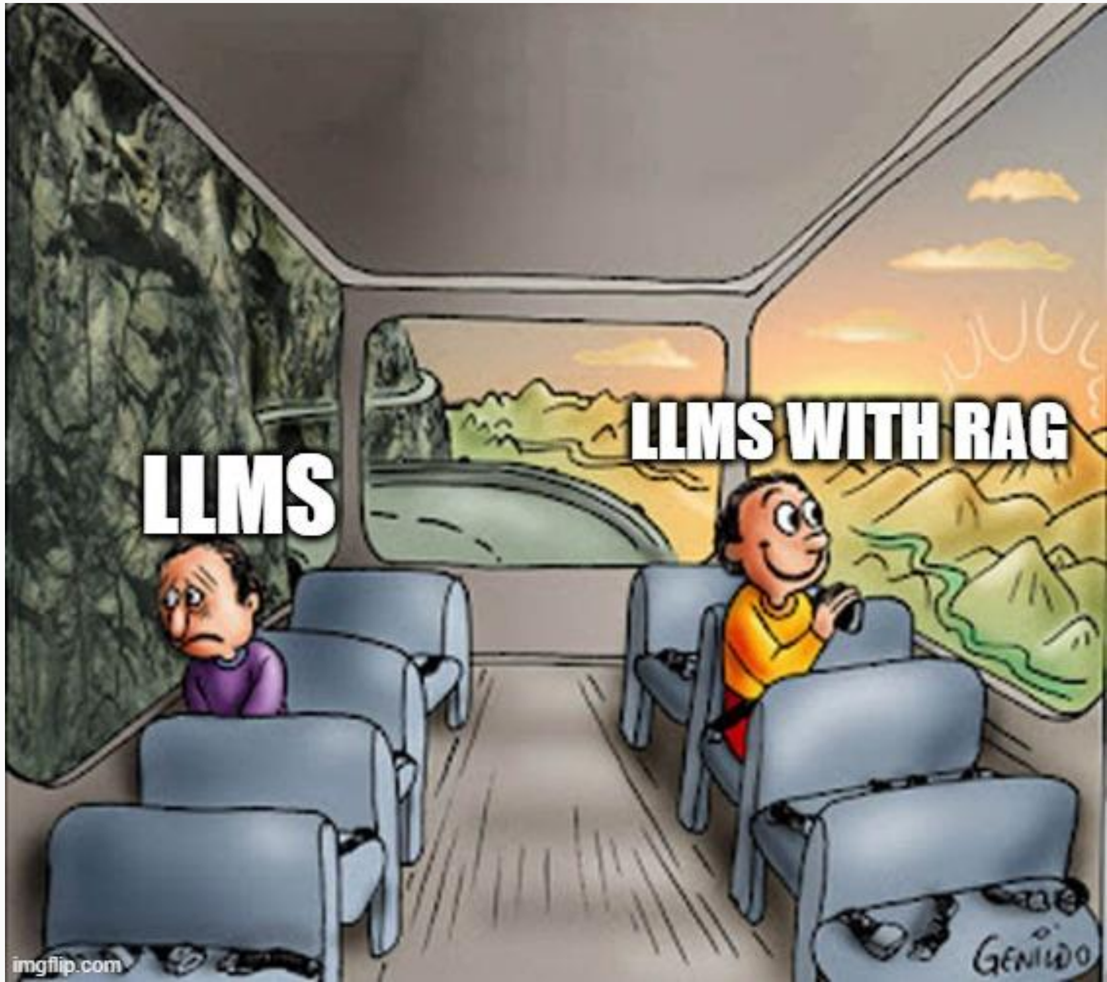
Takes too long!

Introduction - RAG

If we have a large number of documents, how can we process/query it using an LLM?

3. We can use RAG.

Let's take a deeper look at what RAG is and how it can help us.



Introduction - RAG

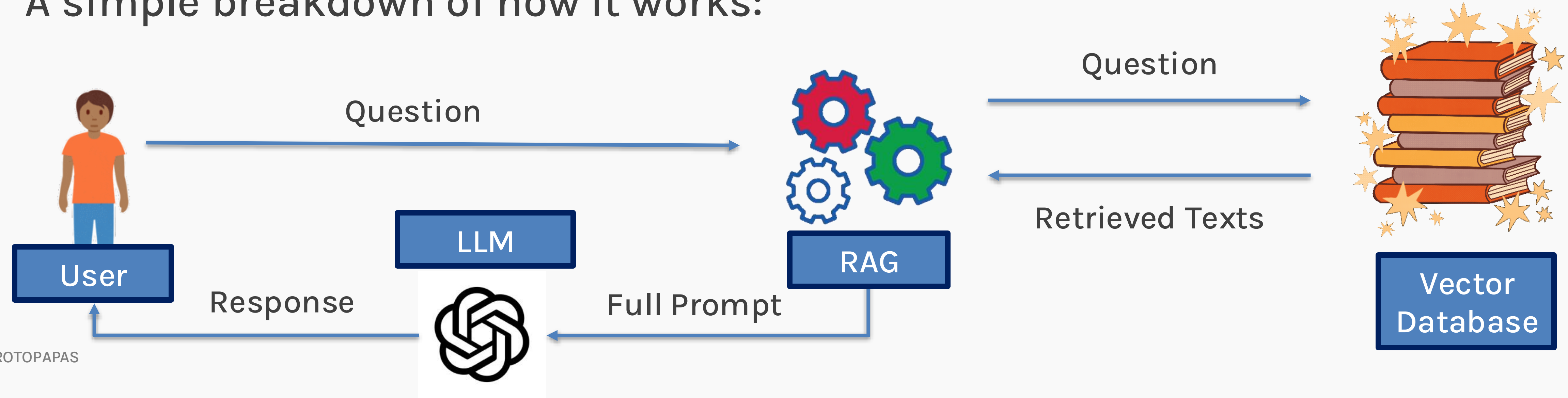
What is RAG?

- RAG stands for Retrieval-Augmented-Generation.

- It is technique that improves the performance of a LLM, especially for tasks that require accurate and detailed information.

- A simple breakdown of how it works:

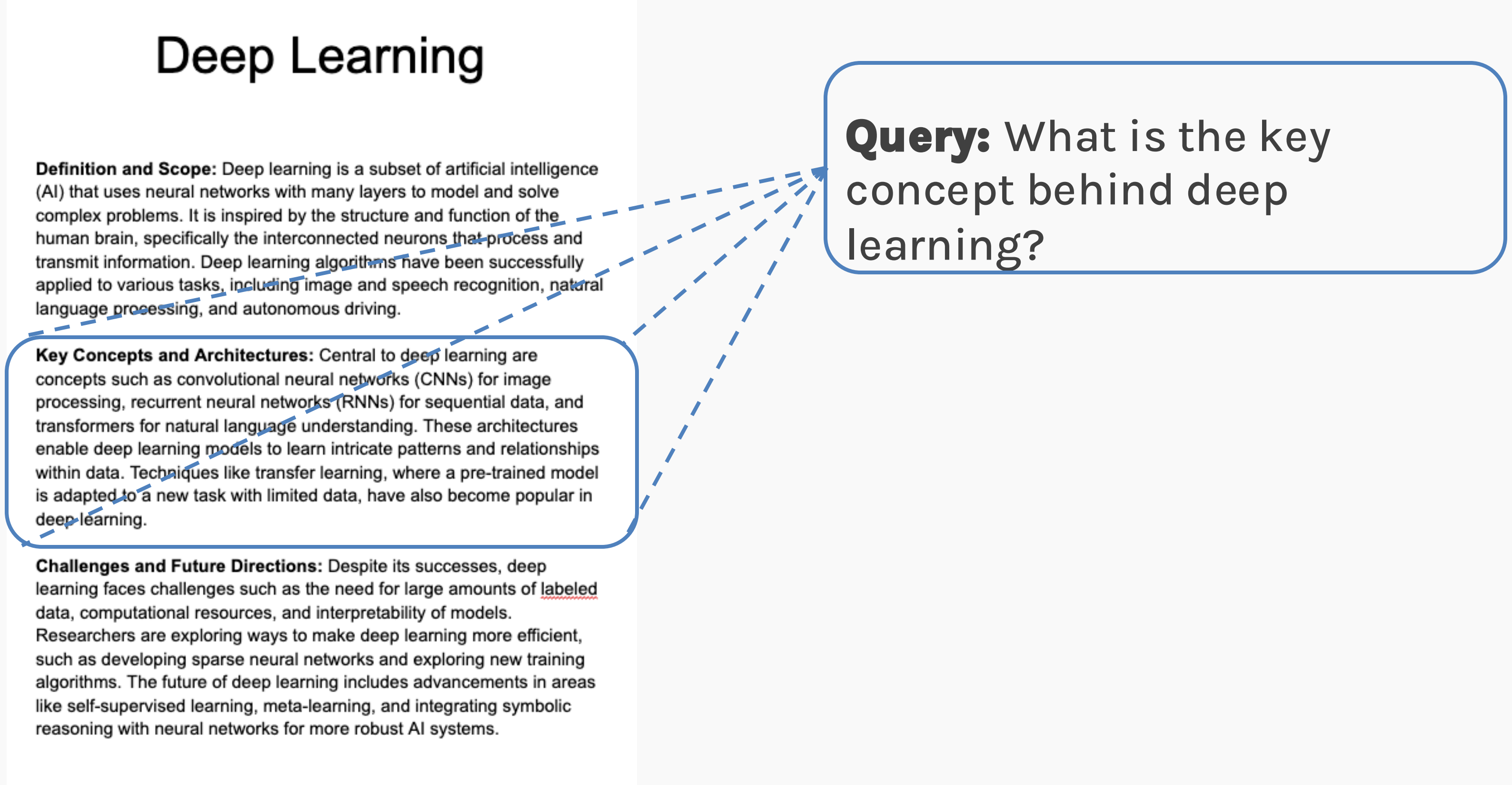
We will look at what a vector database is, later in the slides.



64

Introduction - RAG

What is RAG?



Deep Learning

Definition and Scope: Deep learning is a subset of artificial intelligence (AI) that uses neural networks with many layers to model and solve complex problems. It is inspired by the structure and function of the human brain, specifically the interconnected neurons that process and transmit information. Deep learning algorithms have been successfully applied to various tasks, including image and speech recognition, natural language processing, and autonomous driving.

Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

Challenges and Future Directions: Despite its successes, deep learning faces challenges such as the need for large amounts of labeled data, computational resources, and interpretability of models. Researchers are exploring ways to make deep learning more efficient, such as developing sparse neural networks and exploring new training algorithms. The future of deep learning includes advancements in areas like self-supervised learning, meta-learning, and integrating symbolic reasoning with neural networks for more robust AI systems.

65

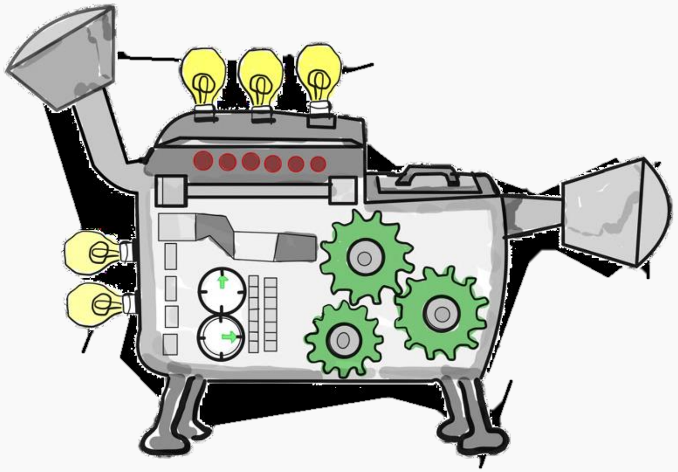
Introduction - RAG

What is RAG?

Query: What is the key concept behind deep learning?

Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

Context



LLM

The key concept behind deep learning is ….

66

Naïve RAG

The embedding model converts text (docu
into vectors/embeddings that capture sem

We use an architecture similar to BERT and aggregate the word embeddings

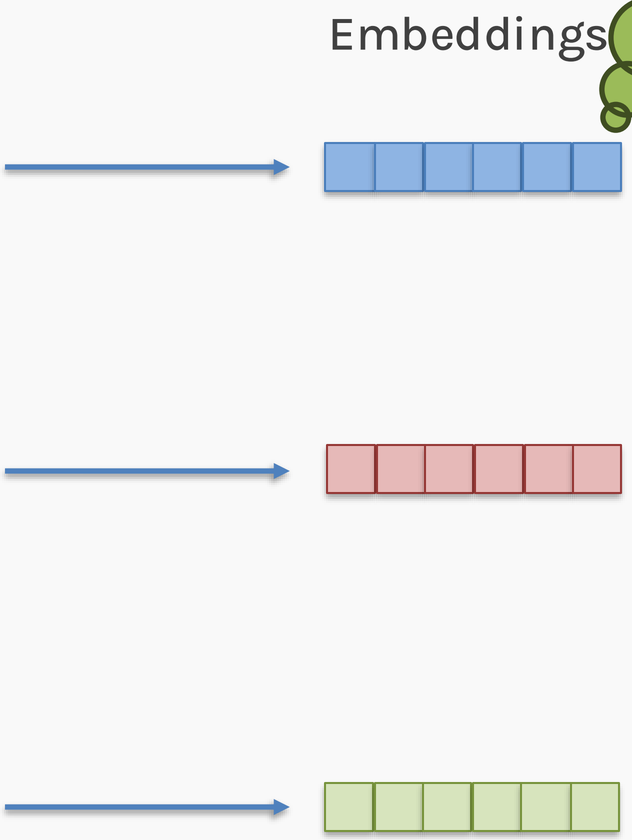
Embedding Model

Deep Learning

Definition and Scope: Deep learning is a subset of artificial intelligence (AI) that uses neural networks with many layers to model and solve complex problems. It is inspired by the structure and function of the human brain, specifically the interconnected neurons that process and transmit information. Deep learning algorithms have been successfully applied to various tasks, including image and speech recognition, natural language processing, and autonomous driving.

Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

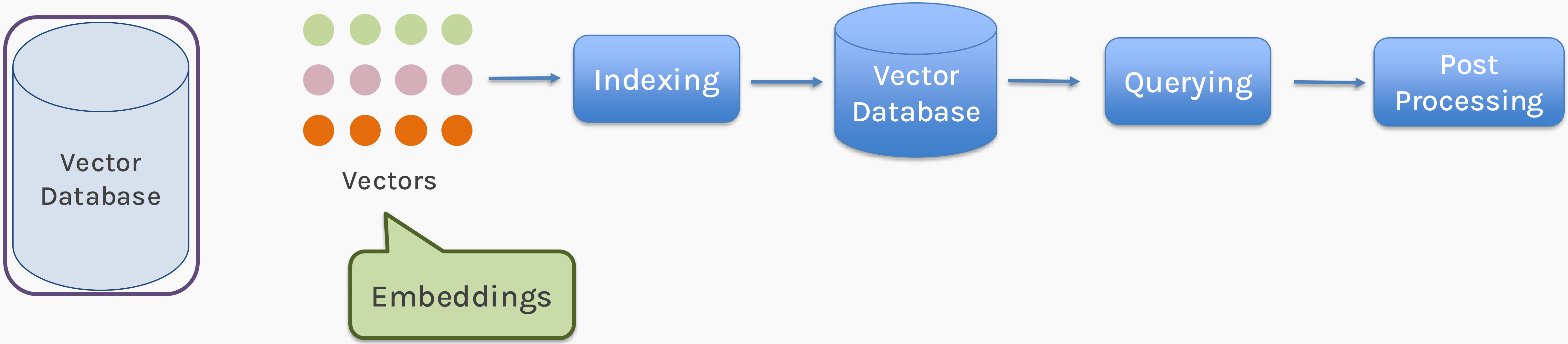
Challenges and Future Directions: Despite its successes, deep learning faces challenges such as the need for large amounts of labeled data, computational resources, and interpretability of models. Researchers are exploring ways to make deep learning more efficient, such as developing sparse neural networks and exploring new training algorithms. The future of deep learning includes advancements in areas like self-supervised learning, meta-learning, and integrating symbolic reasoning with neural networks for more robust AI systems.



67

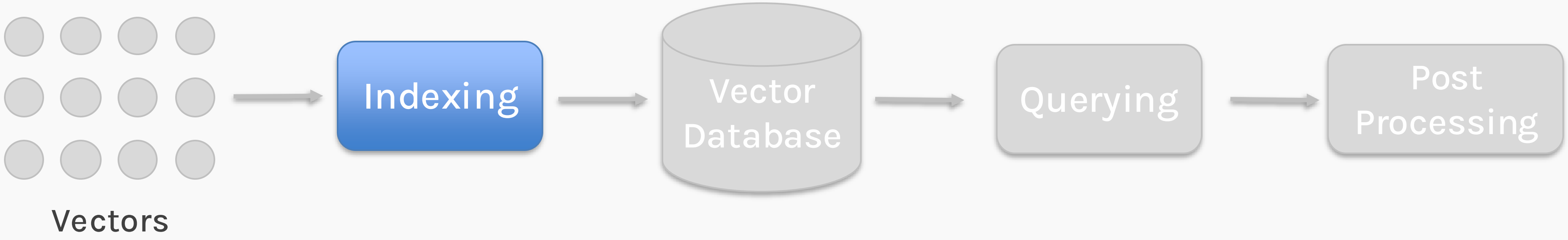
Naïve RAG – Vector Database

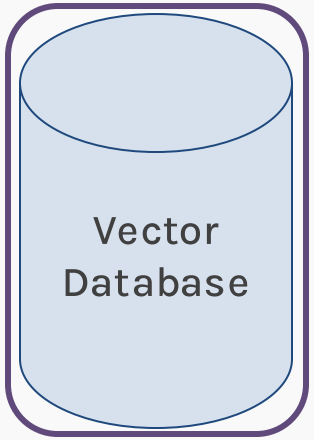
A vector database indexes and stores vector embeddings for fast retrieval and similarity search.



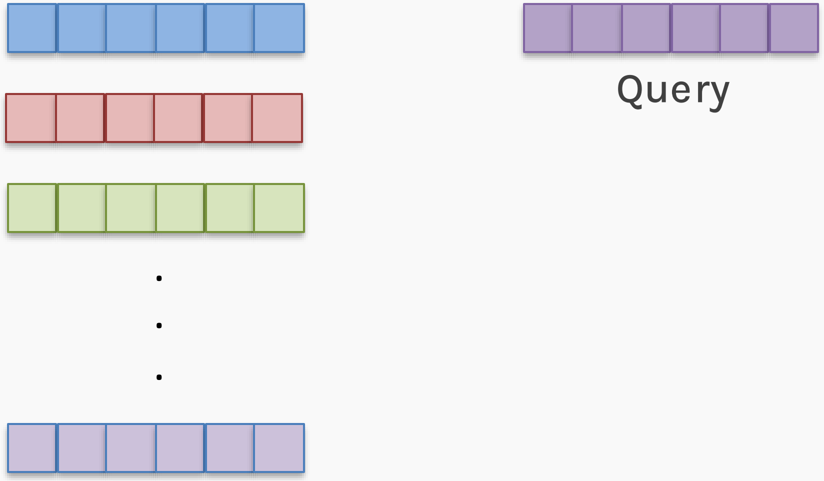
Vectors capture the essential features of the original data in a high-dimensional space.

Naïve RAG – Vector Database

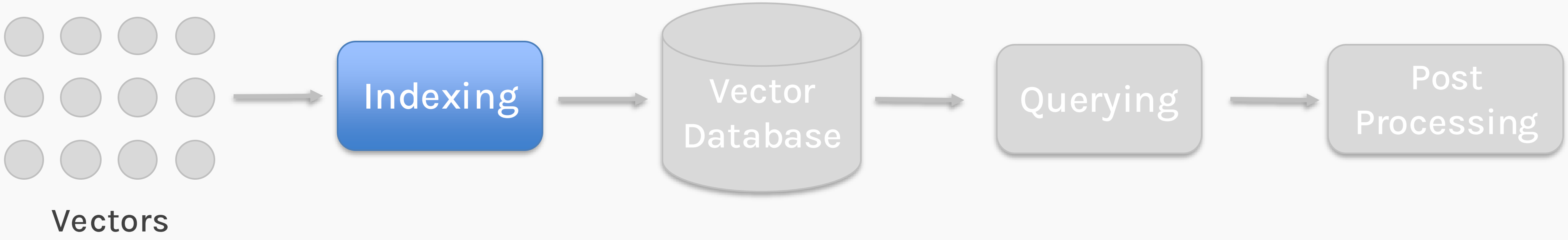


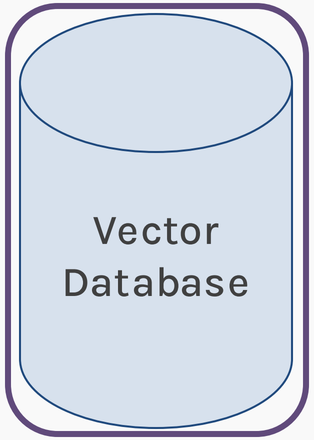


In reality, we could have millions of vectors to deal with.

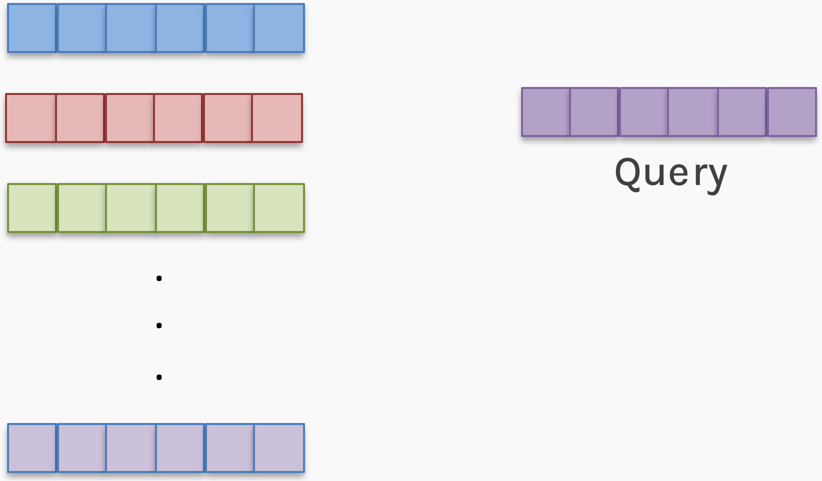


Naïve RAG – Vector Database

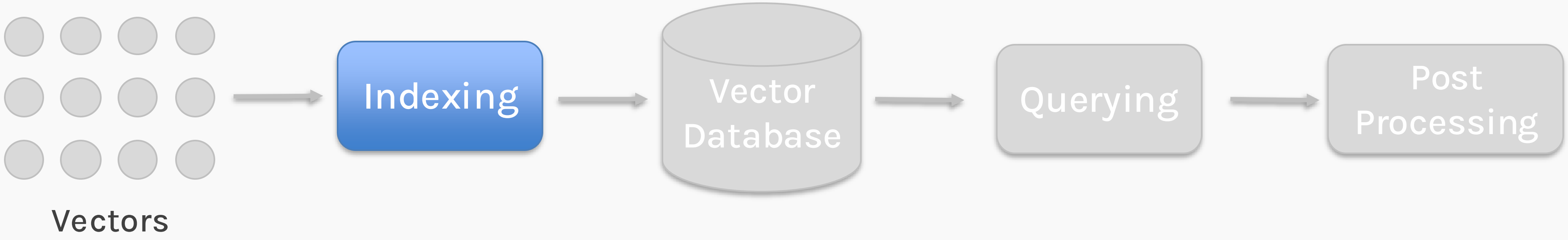


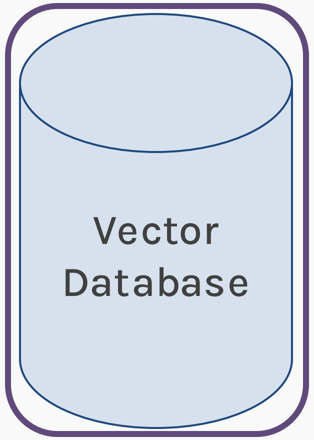


In reality, we could have millions of vectors to deal with.

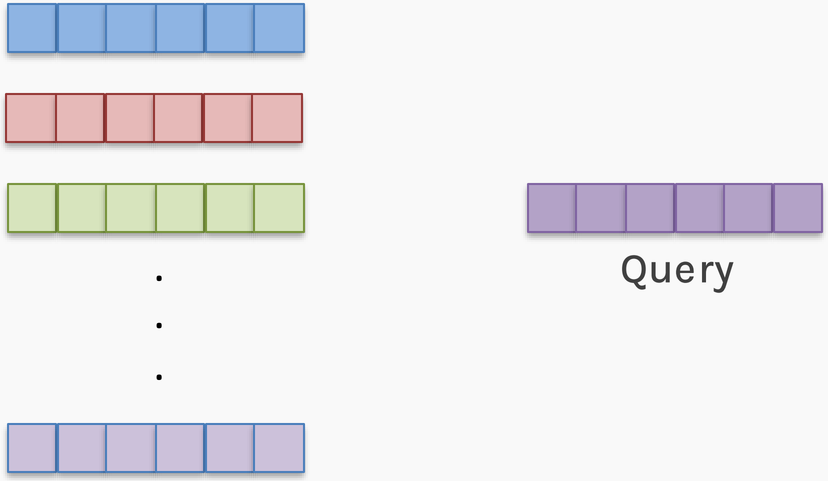


Naïve RAG – Vector Database

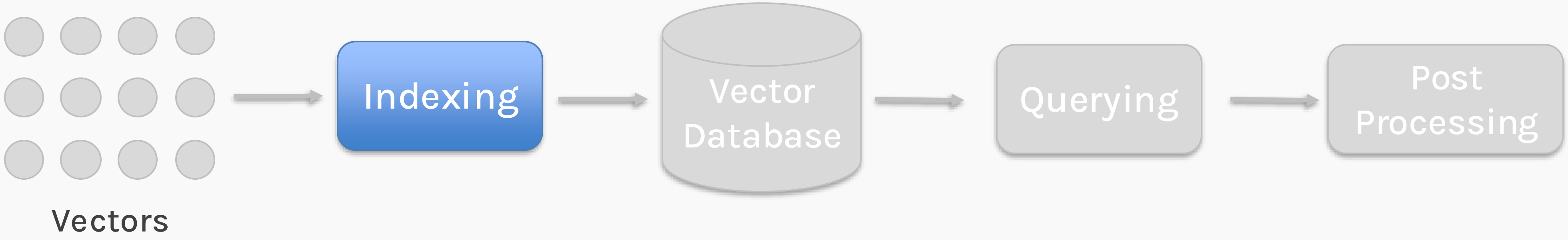


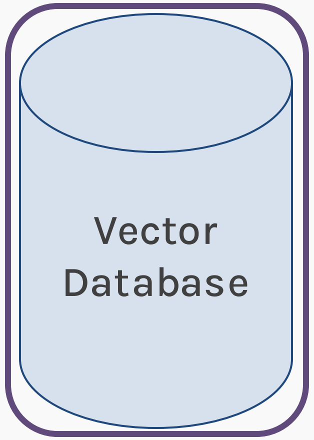


In reality, we could have millions of vectors to deal with.

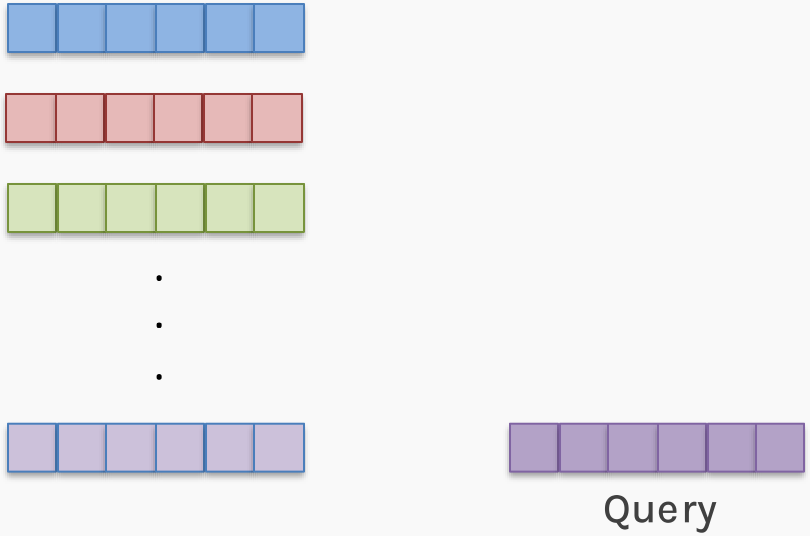


Naïve RAG – Vector Database



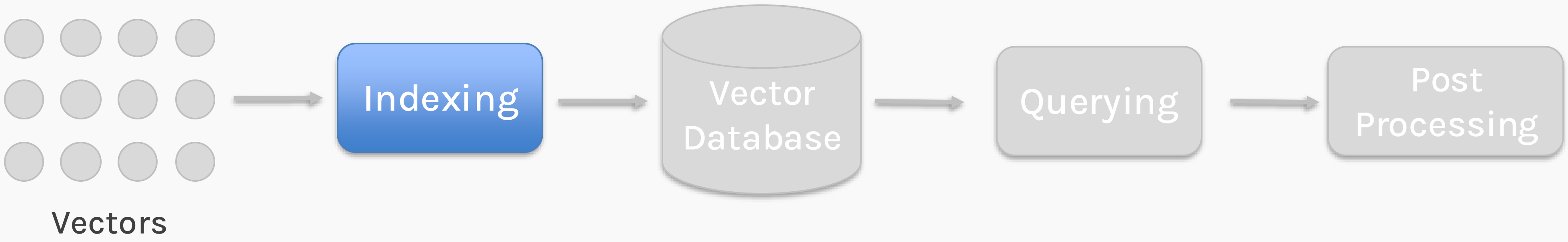


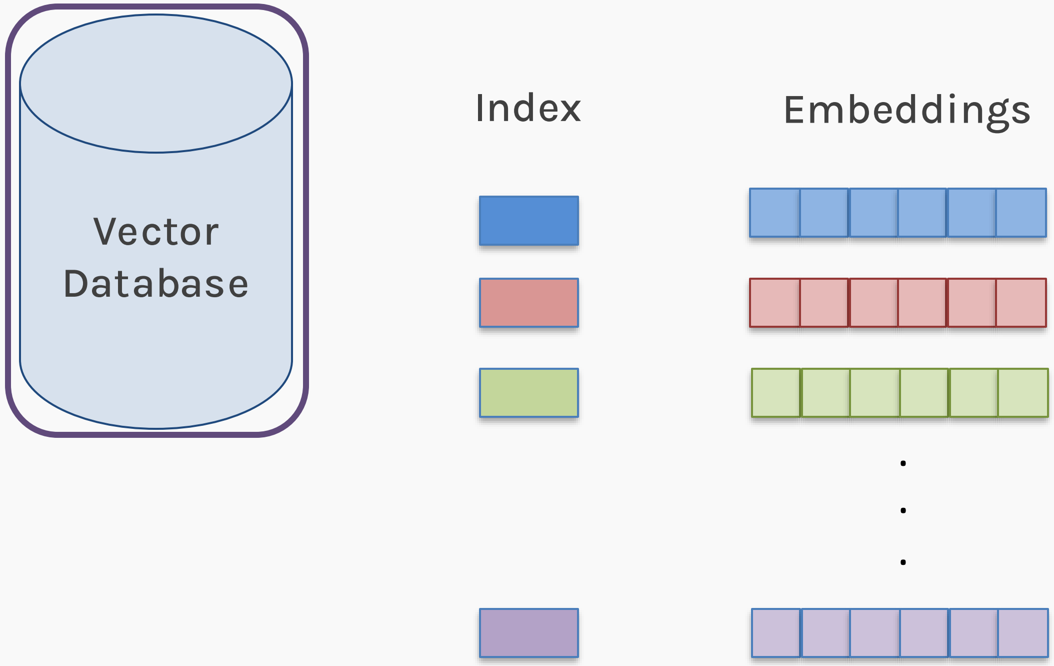
In reality, we could have millions of vectors to deal with.



Comparing them sequentially would be a very slow process.

Naïve RAG – Vector Database

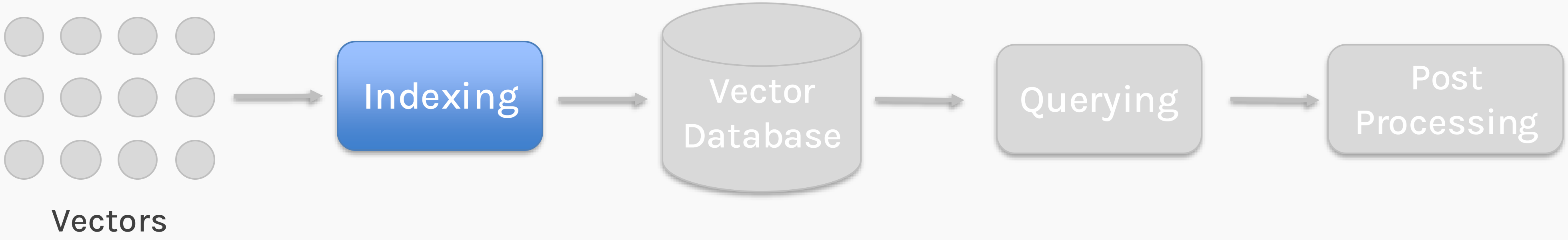


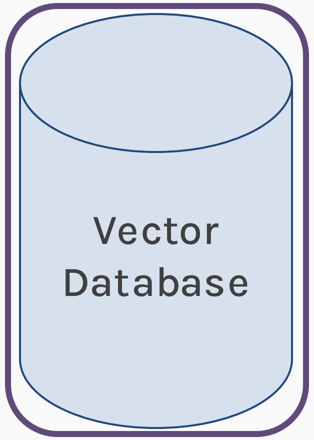


Indexing is the process of mapping the vectors to a data structure that will enable us to search faster.

When we query for a particular vector, the index helps us quickly zoom into a small subset of vectors that are likely to be similar.

Naïve RAG – Vector Database

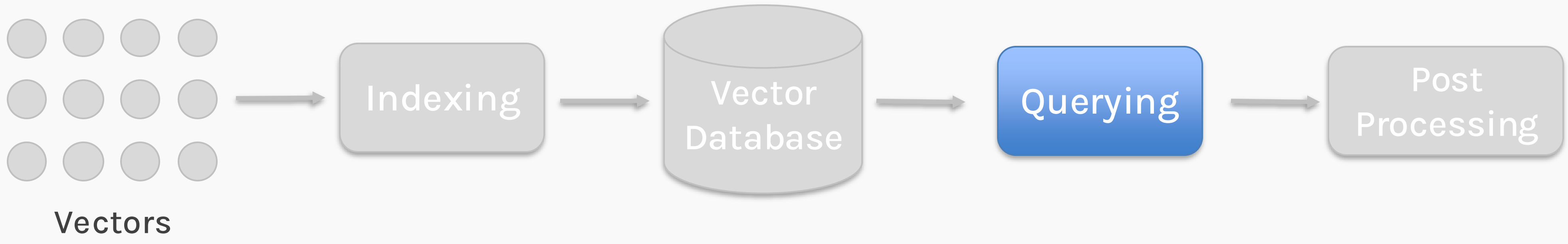


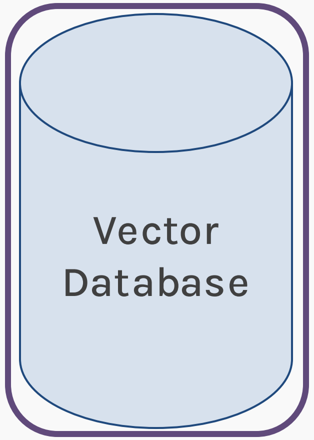


We usually use one of the three algorithms to do indexing:

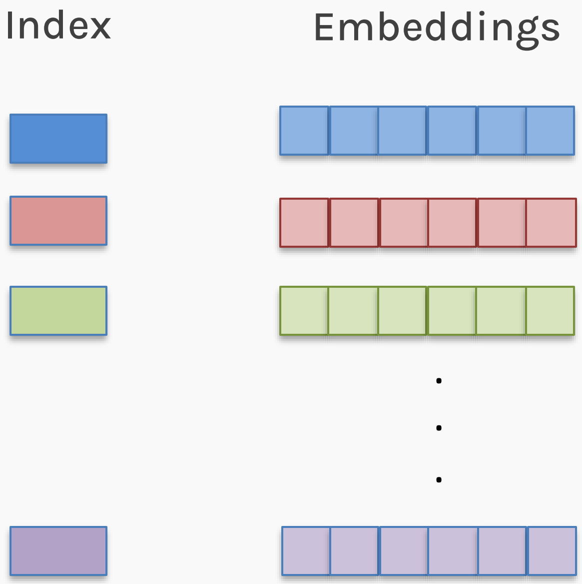
1. Hashing (Locality Sensitive Hashing - LSH)
2. Quantization (Product Quantization - PQ)
3. Graph Based (Hierarchical Navigable Small World - HNSW)

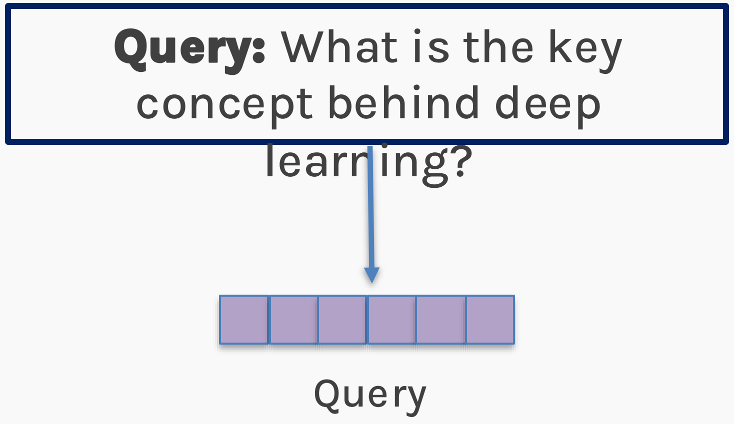
Naïve RAG – Vector Database



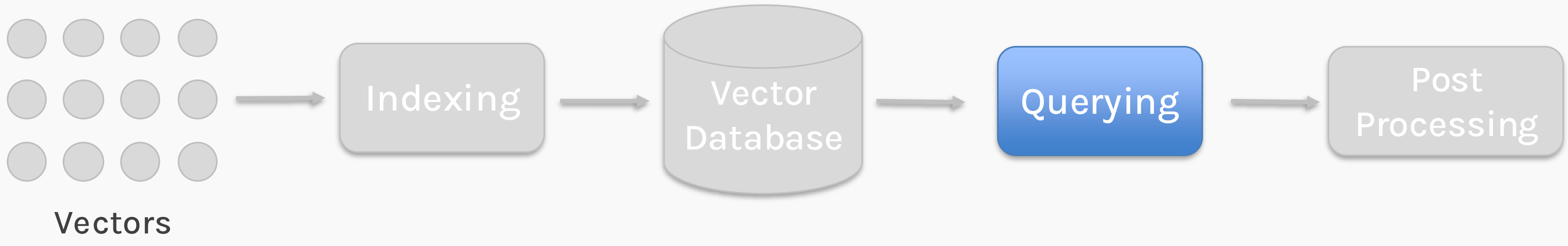


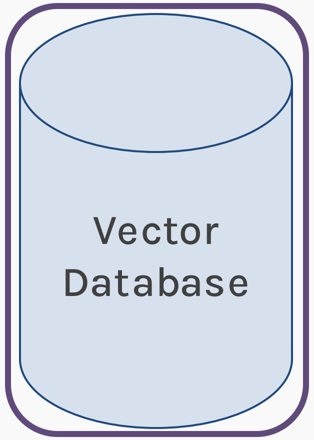
When querying, the vector database compares the indexed vectors to the query vector to determine the nearest vector neighbor.





Naïve RAG – Vector Database





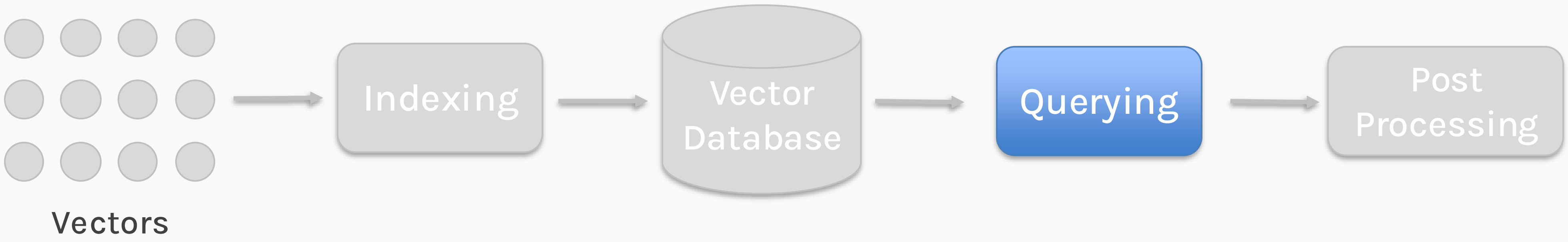
When querying, the vector database compares the indexed vectors to the query vector to determine the nearest vector neighbor.

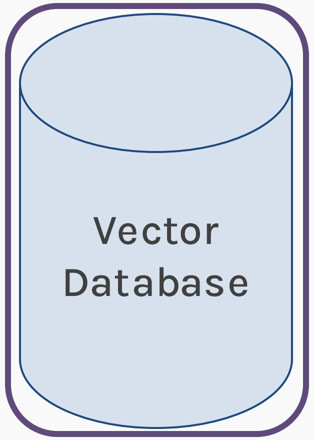
But, how do we compare?

We can use one of the following similarity measures to find the nearest neighbor:

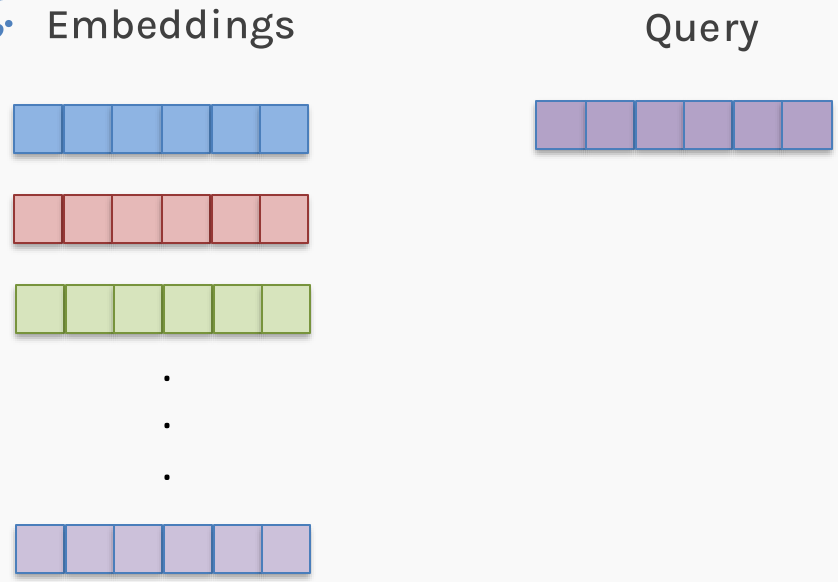
1. Cosine Similarity
2. Euclidian Distance
3. Dot Product

Naïve RAG – Vector Database

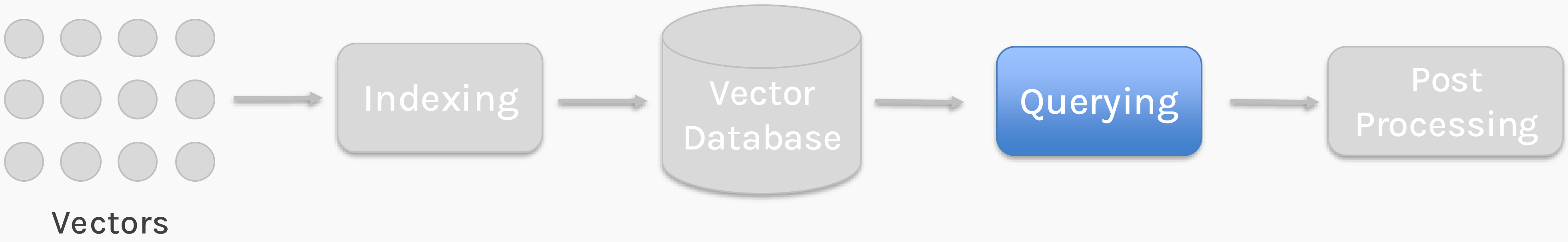


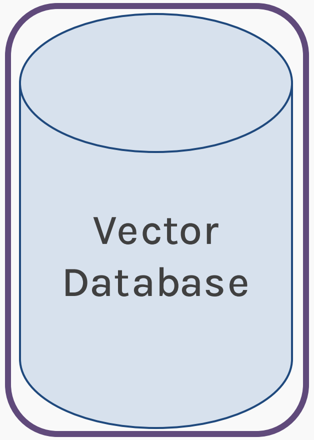


This leads to a ranking of embeddings based on their similarity with the query embedding.

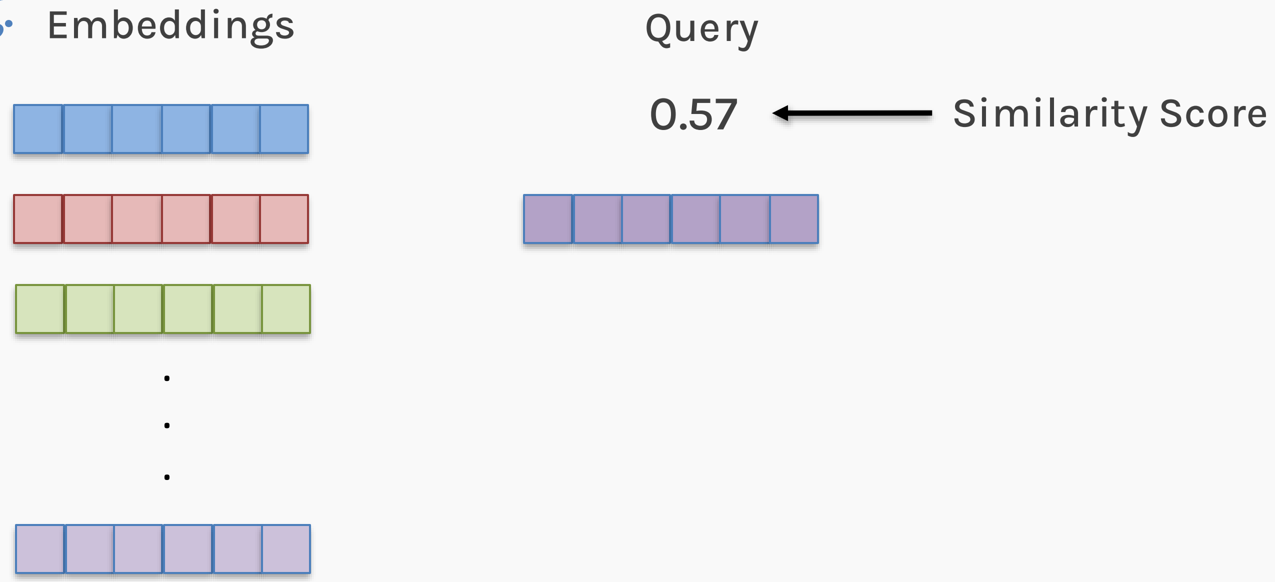


Naïve RAG – Vector Database

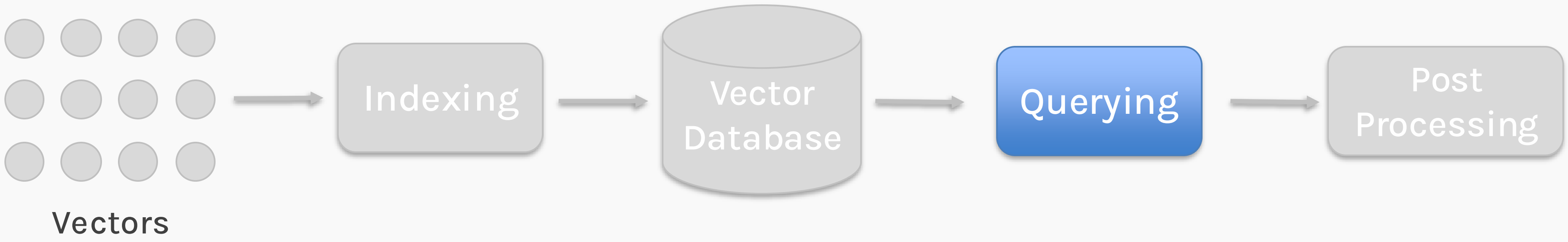


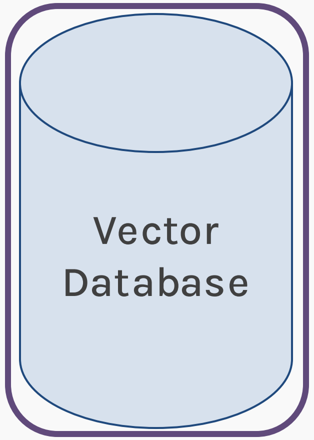


This leads to a ranking of embeddings based on their similarity with the query embedding.

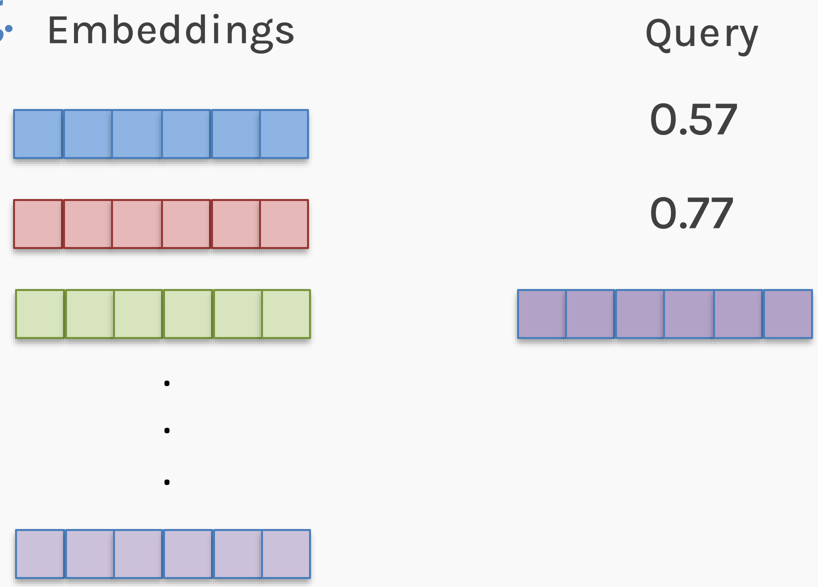


Naïve RAG – Vector Database

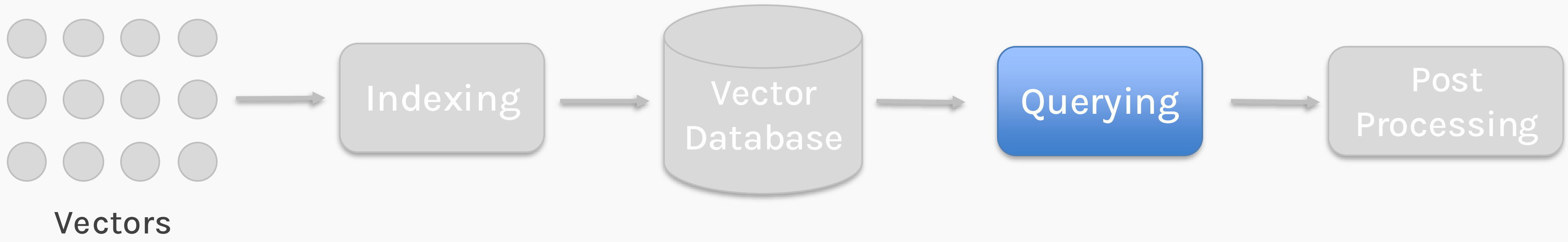


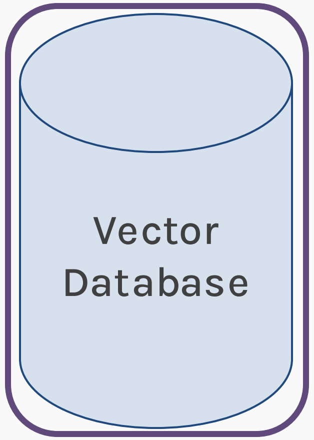


This leads to a ranking of embeddings based on their similarity with the query embedding.



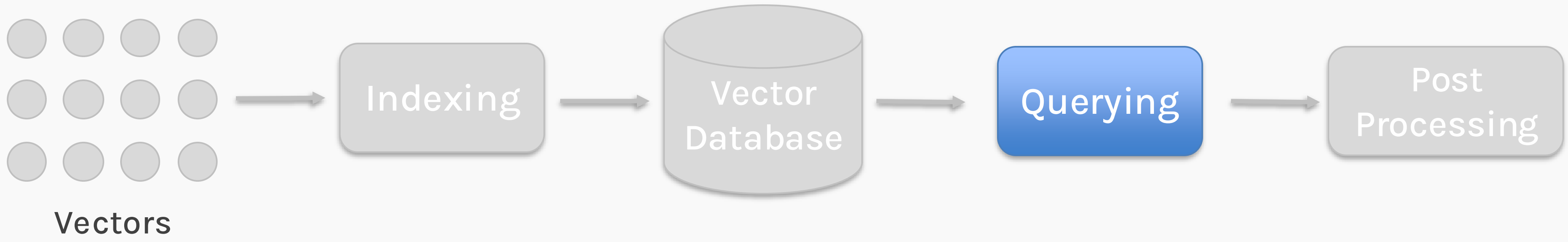
Naïve RAG – Vector Database

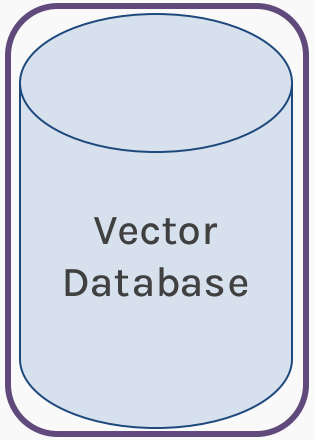




This leads to a ranking of embeddings based on their similarity with the query embedding.

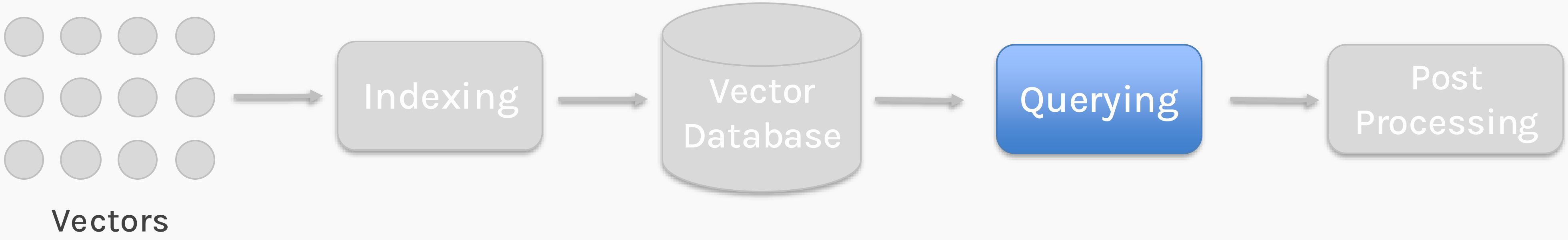
Naïve RAG – Vector Database

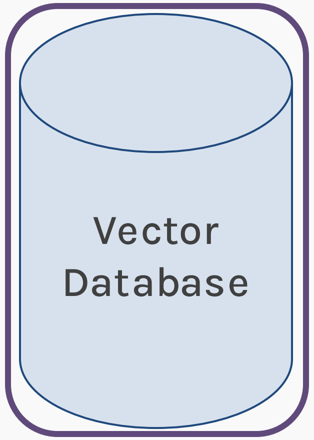




This leads to a ranking of embeddings based on their similarity with the query embedding.

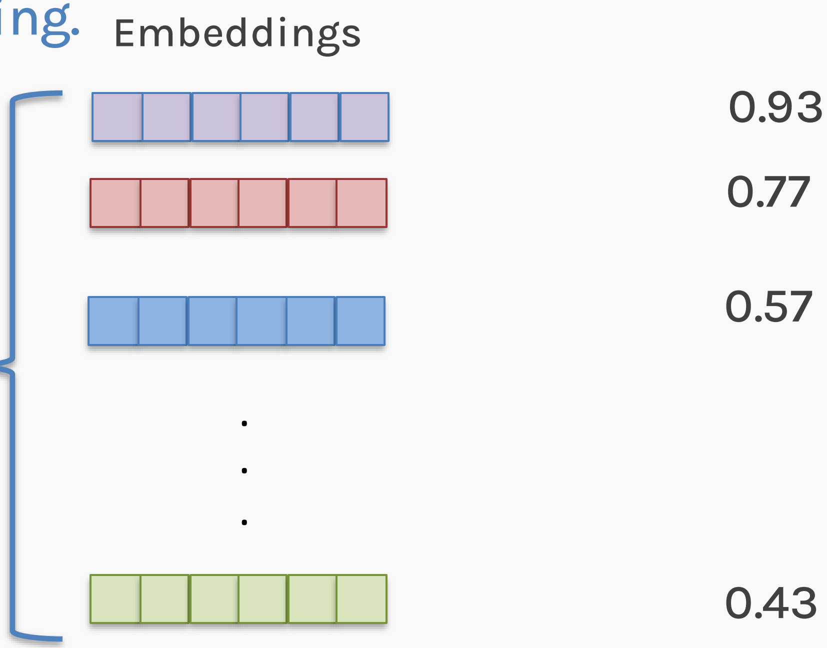
Naïve RAG – Vector Database

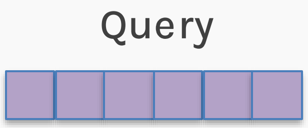




This leads to a ranking of embeddings based on their similarity with the query embedding.

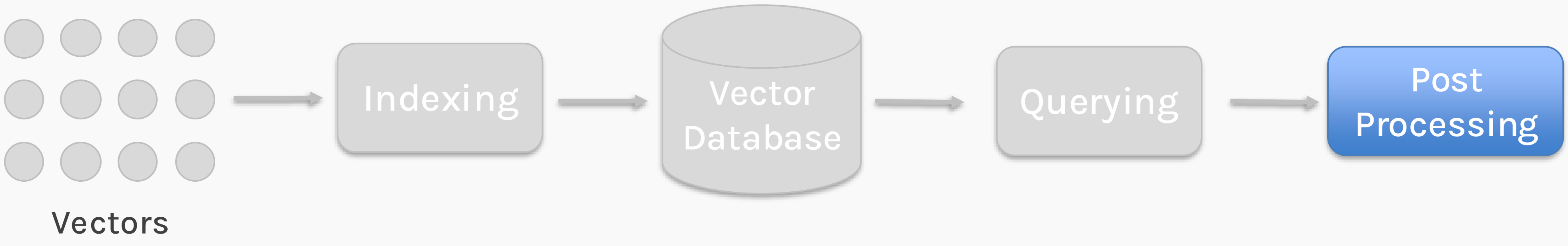
The ranks have been changed.

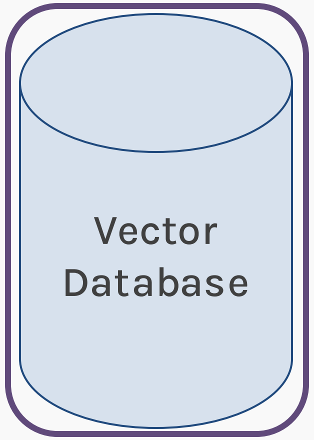




Often, the RAG is configured to return the top-K results, where K is a predefined number.

Naïve RAG – Vector Database





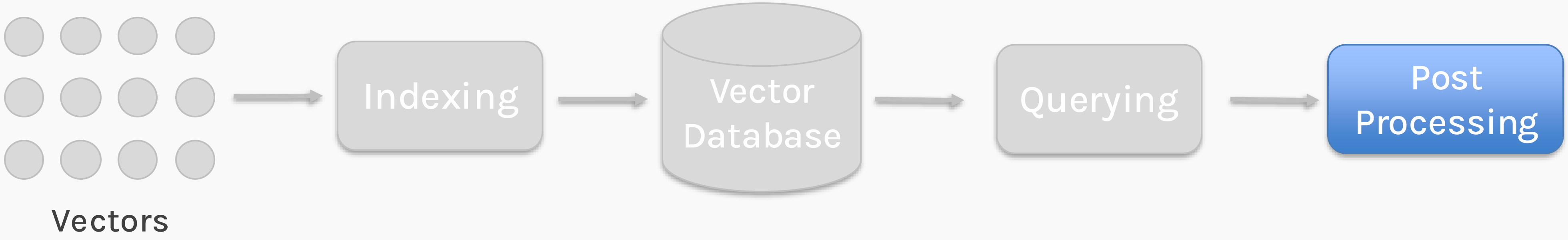
This step usually consists of using a different similarity measure to re-rank the results we got in the previous step.

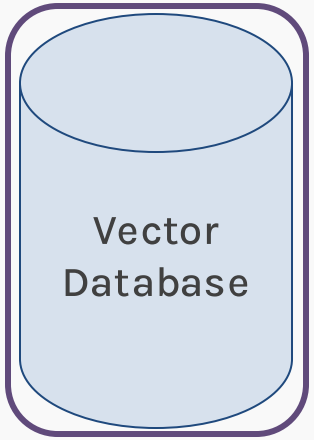
We could also do some metadata filtering here.

Note: This is an optional step.

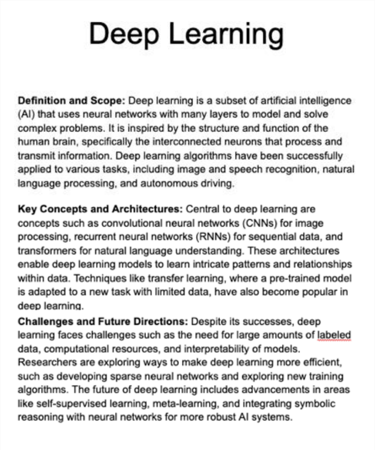
Metadata is data that provides information about other data (in our case, the vectors).

Naïve RAG – Vector Database





Metadata filtering helps refine and narrow down the search results based on metadata.



Deep Learning

Definition and Scope: Deep learning is a subset of artificial intelligence (AI) that uses neural networks with many layers to model and solve complex problems. It is inspired by the structure and function of the human brain, specifically the interconnected neurons that process and transmit information. Deep learning algorithms have been successfully applied to various tasks, including image and speech recognition, natural language processing, and autonomous driving.

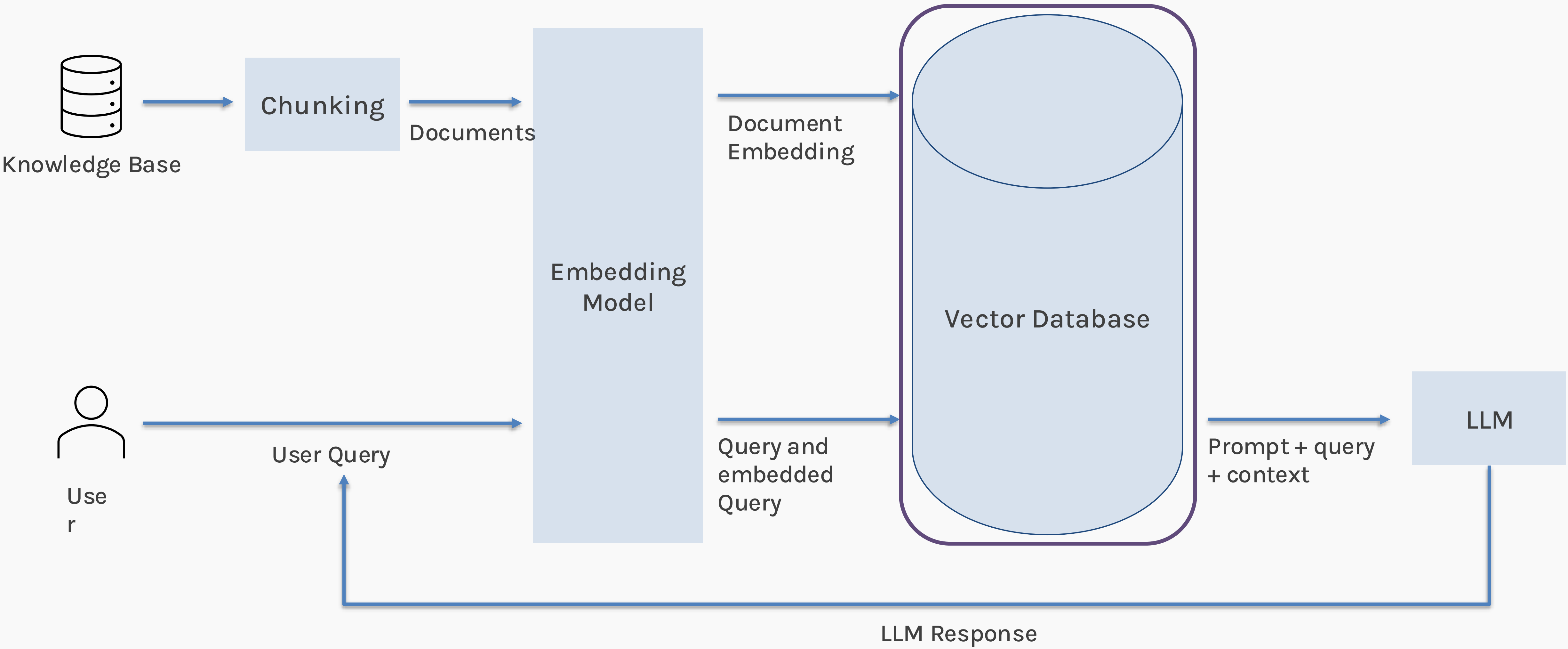
Key Concepts and Architectures: Central to deep learning are concepts such as convolutional neural networks (CNNs) for image processing, recurrent neural networks (RNNs) for sequential data, and transformers for natural language understanding. These architectures enable deep learning models to learn intricate patterns and relationships within data. Techniques like transfer learning, where a pre-trained model is adapted to a new task with limited data, have also become popular in deep learning.

Challenges and Future Directions: Despite its successes, deep learning faces challenges such as the need for large amounts of labeled data, computational resources, and interpretability of models. Researchers are exploring ways to make deep learning more efficient, such as developing sparse neural networks and exploring new training algorithms. The future of deep learning includes advancements in areas like self-supervised learning, meta-learning, and integrating symbolic reasoning with neural networks for more robust AI systems.

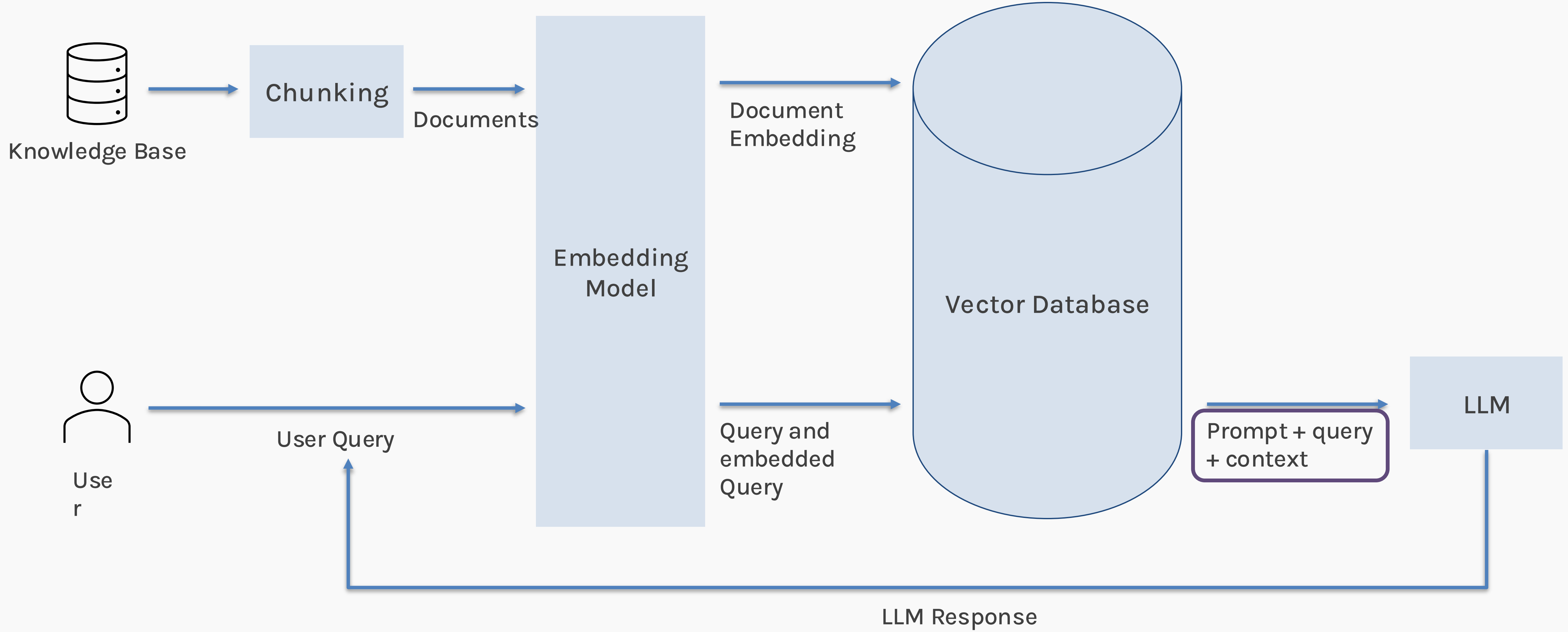
For this data, the metadata may be:

1. Author
2. Publication date
3. Category

Naïve RAG


Knowledge Base
Chunking
Documents
Embedding Model
Document Embedding
Vector Database
User
User Query
Query and embedded Query
Prompt + query + context
LLM
LLM Response

Naïve RAG


Chunking
Knowledge Base
Documents
Embedding Model
Document Embedding
Vector Database
User
User Query
Query and embedded Query
Prompt + query + context
LLM
LLM Response

86

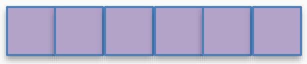
Naïve RAG – Prompt + query + context

Prompt

System prompts acts as an instruction given to the model to guide its behavior and responses

Query

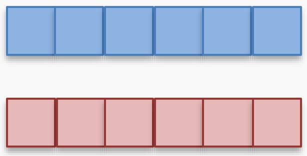
The query is the user's initial input.



Query

Context

Context refers to the information retrieved from the vector database that is relevant to the query,



Context

Thank you