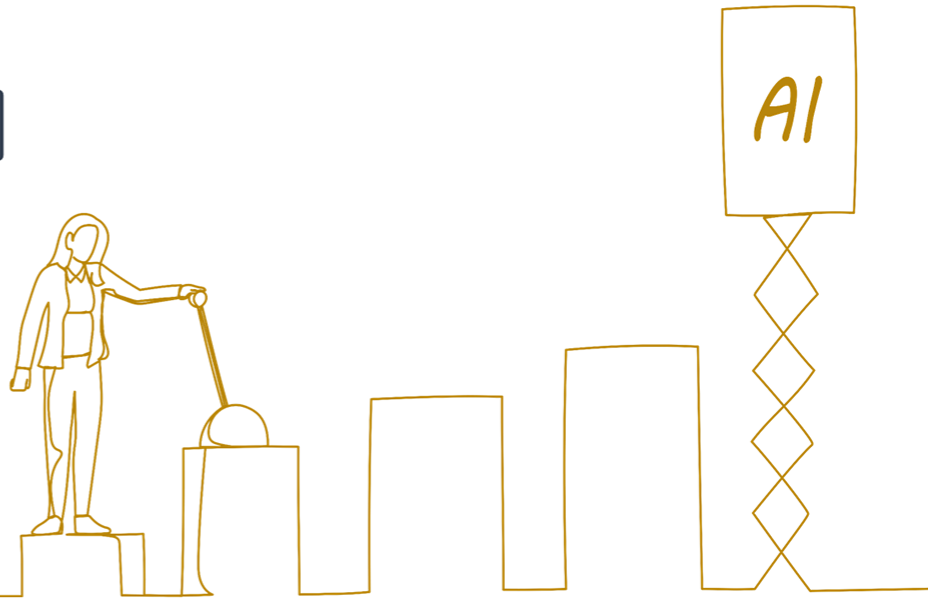
EURO AUSTRIA

Datenanalyse mit KI

Vom Laptop zum Supercomputer



10 April 2025
ONLINE

AI 5 PRODUCTION

What can LLMs be used for?

Many different use-cases

- Made possible by the transformer architecture
- Choose your model according to the use-case

Conversation

Text Generation

Translation

Text to Speech

Question Answering

Transcription of Videos

Text Classification

Sentiment Analysis

Similarity Detection

Summarization

Transformer Anatomy

Attention is really all you need?

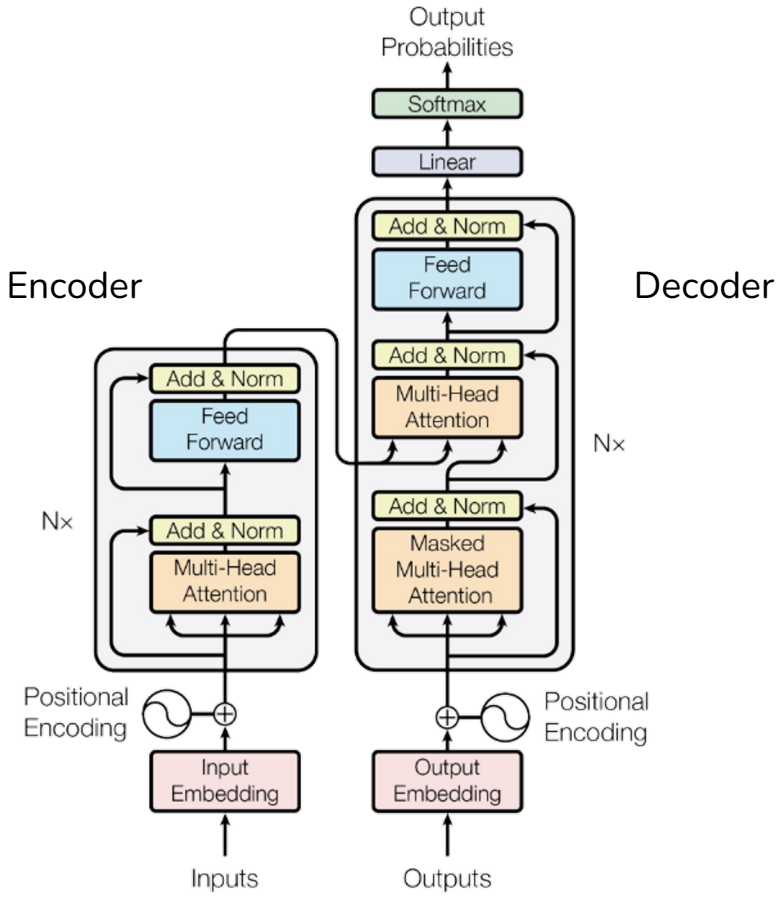
EuroCC Austria

EURO AUSTRIA

Transformer Anatomy

Attention Is All You Need

- Encoder-Decoder Structure
- Tokenisation: Numerical representation of input
- Self-Attention Mechanism
- Multi-Head Attention



Source: “Attention Is All You Need”, Vaswani et al.

EURO AUSTRIA

Context Is All You Need

Embeddings

Here, we will refer to “word” instead of “token”, as it makes the content easier to explain.

A word embedding comes as a multi dimensional vector (e.g. 12.000 dim).

The initial word embedding in all of the examples of the word „mole“ is the same.

The European mole is a mammal

Take a biopsy of the mole

6.02×10^{23}

One mole of carbon dioxide

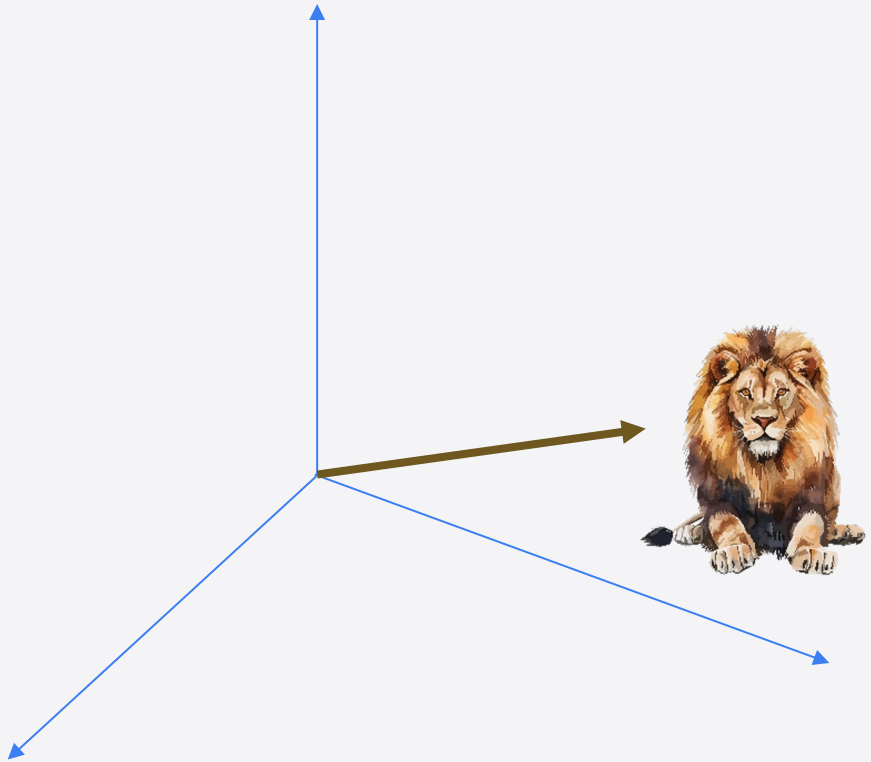
EURO AUSTRIA

Context Is All You Need

Lion

We associate the word „lion“ with a big cat, living wild on the African continent.
We probably imagine a majestic predator with a big mane.

The embedding of the word „lion“ is a vector with a certain length and direction within the embedding space.



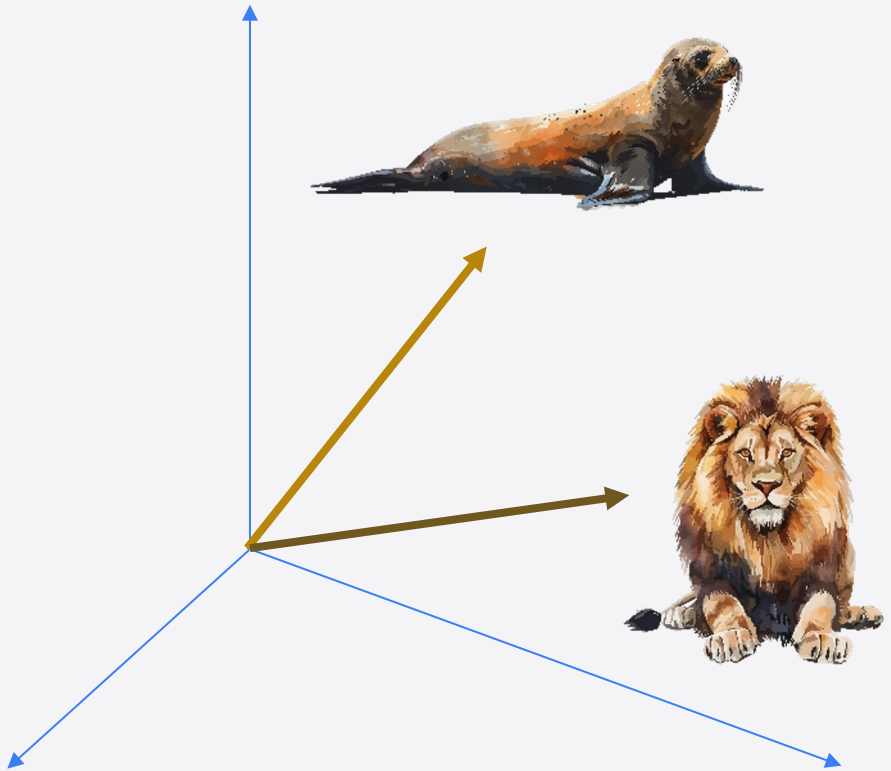
EURO AUSTRIA

Context Is All You Need

Sea Lion

However, as soon we add the word „sea“ infront of „lion“ we imagine a totally different animal.

The same goes for the embedding. The attiontion mechanism needs to update the direction and length of the vector so that it represents the animal in question correctly.



EURO AUSTRIA

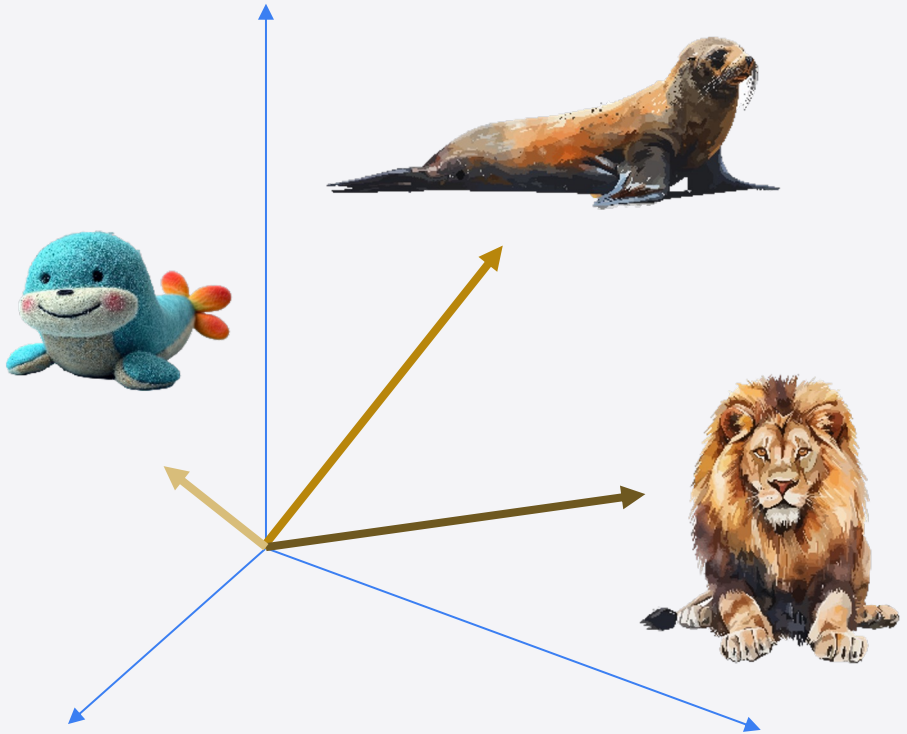
Context Is All You Need

Sea Lion Cuddly Toy

The context depends on more than just the immediate words to the left and right.

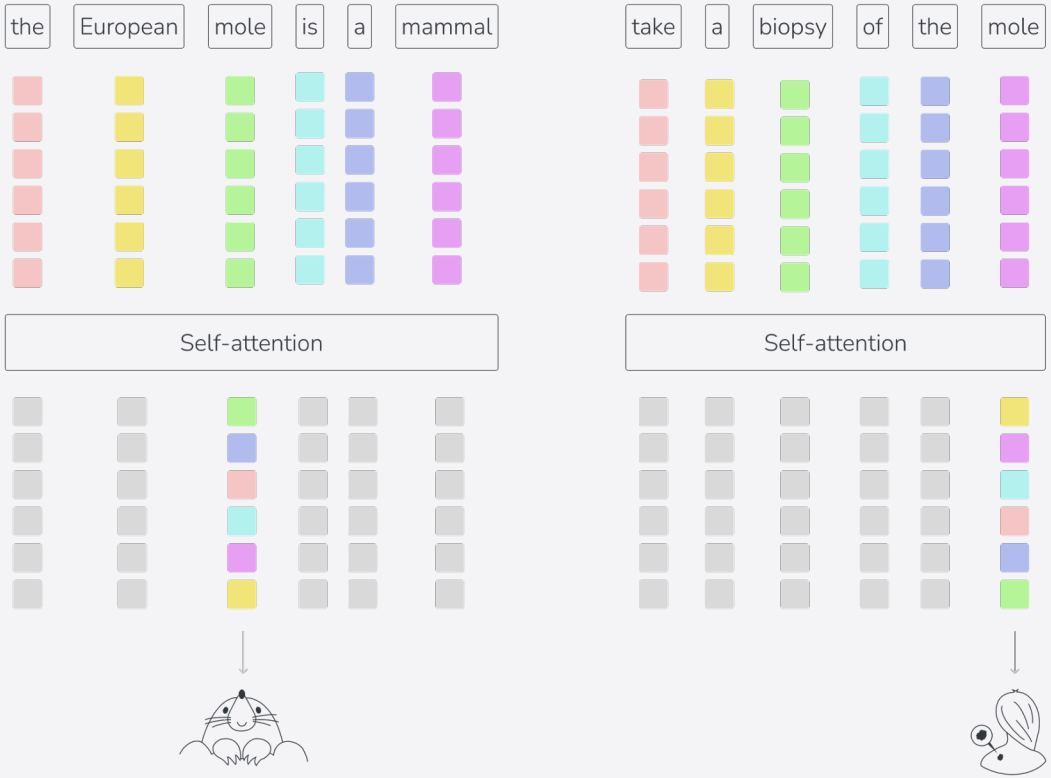
The embedding of „sea lion cuddly toy“ will certainly be very different of just „lion“.

In order to achieve that the vector for „lion“ needs to attend to all the other words in the input (context size).



EURO AUSTRIA

Self Attention



From Text to Tokens

- Character tokenisation?

L a r g e l a n g u a g e m
o d e l s o n s u p e r c o m
p u t e r s

From Text to Tokens

- Character tokenisation?
- Word tokenisation?

Large language models on supercomputers

From Text to Tokens

- Character tokenisation?
- Word tokenisation?
- Subword tokenisation

Large language models
on super com put ers

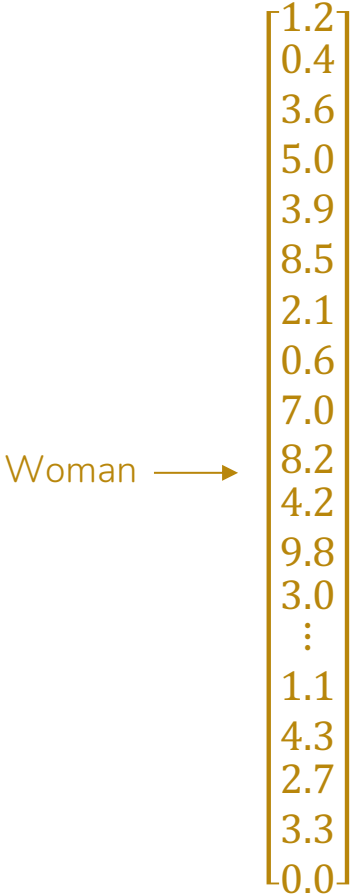
From Tokens to Vectors

Embeddings

Tokens are mapped to unique integers accoding to the vocabulary size of the tokenizer.

Now, the tokens need to be embedded, which means turned into a vector representation.

This is done by an embedding layer of a model. The model takes each token ID and looks it up in an embedding matrix. The embedding matrix is a learned set of weights that maps each token ID to a corresponding high-dimensional vector (embedding).



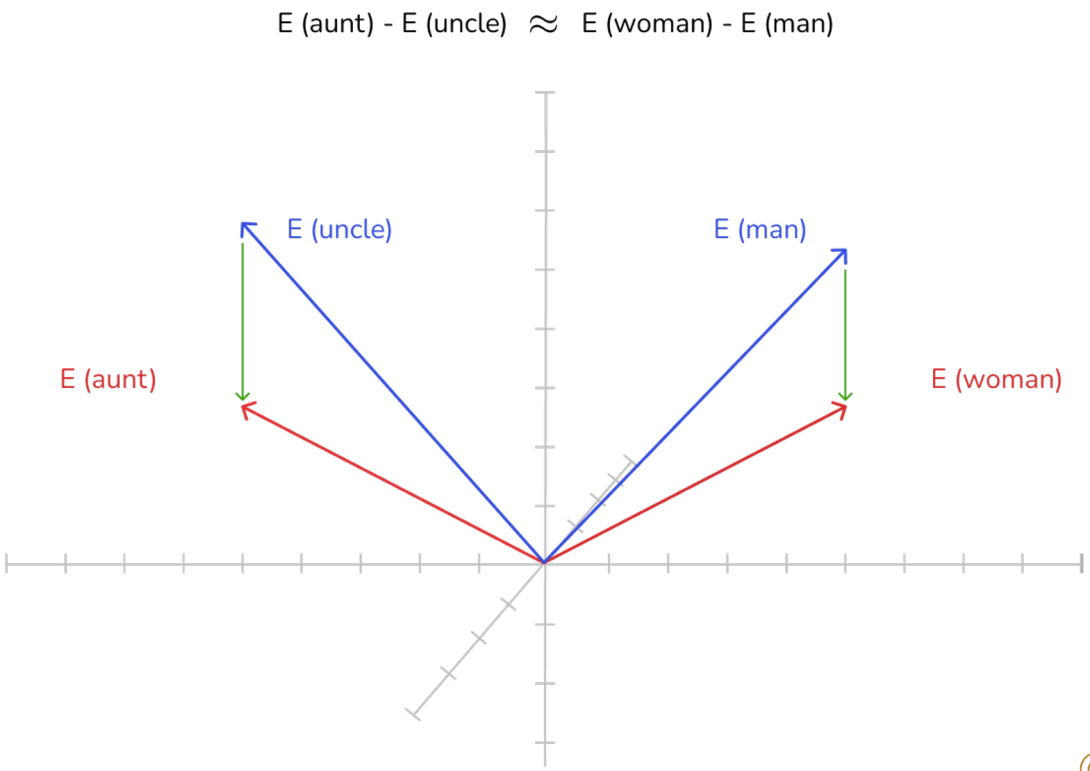
EURO AUSTRIA

From Tokens to Vectors

Embeddings

Each word/token has a unique direction in the embedding space

Similar words point in a similar direction.

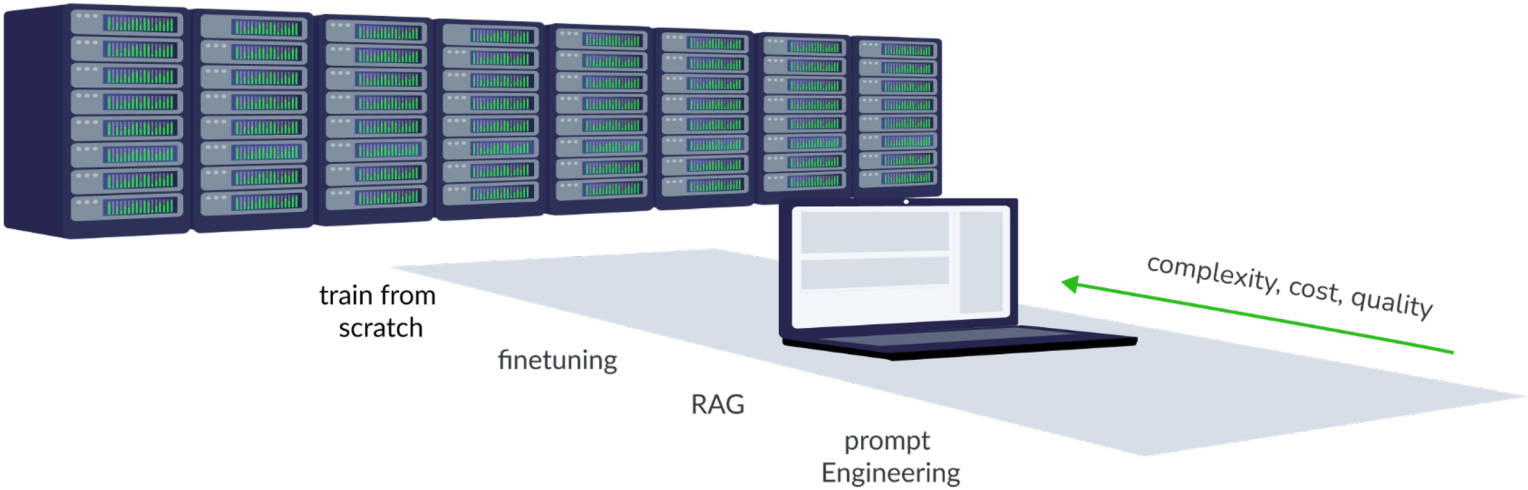


EURO AUSTRIA

Hugging Face

- Hugging Face Hub: Find models and datasets
- Software libraries: transformers, datasets, peft, gradio, ...
- DEMO: Go to https://huggingface.co/, find a model that does sentiment analysis and try it out

How can you influence LLMs?



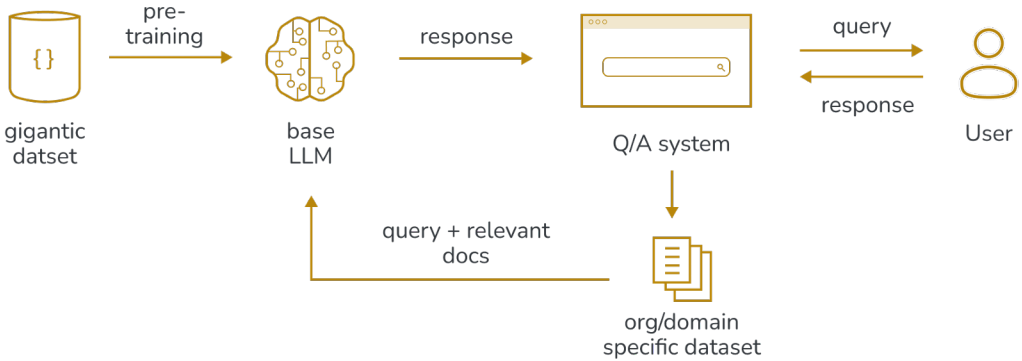
Prompt engineering

Key concepts

- **Clear and specific instructions**: Specify output format, tone and level of detail
- **Context provision:** Give relevant background information
- **Examples:** Provide sample inputs and outputs
- **Constraints:** Set boundaries on content, length or style of response
- **Role Definition:** Specify a particular perspective that the model should adopt
- **Chain-of-Thought:** Encourage the AI to think systematically step by step

RAG: Retrieval Augmented Generation

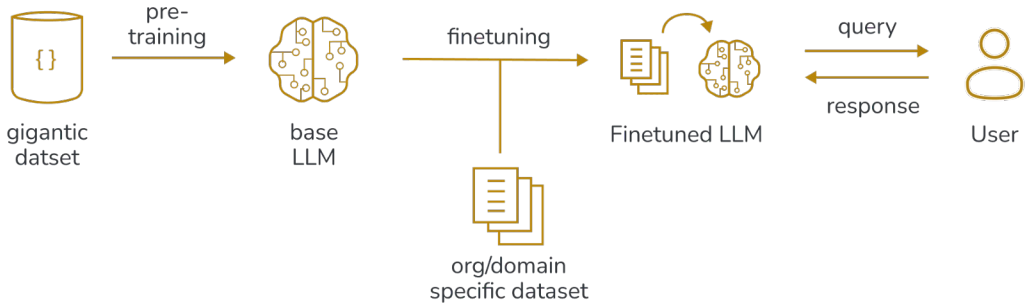
Feed the prompt with relevant information from a database



DEMO: Ask an LLM a question and provide a pdf document with relevant information

Finetuning a large language model

Adjust model parameters with new information



1. Create a custom dataset for your specific domain
2. Pick a pre-trained model and continue training (“finetuning”) it on your dataset

Best done on a supercomputer ;-)

THANK YOU

EURO AUSTRIA

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia