Inside Large Language Models:
A Practical Overview for Researchers

2026-01-20

The cover image was generated by Gemini and Nano Banana

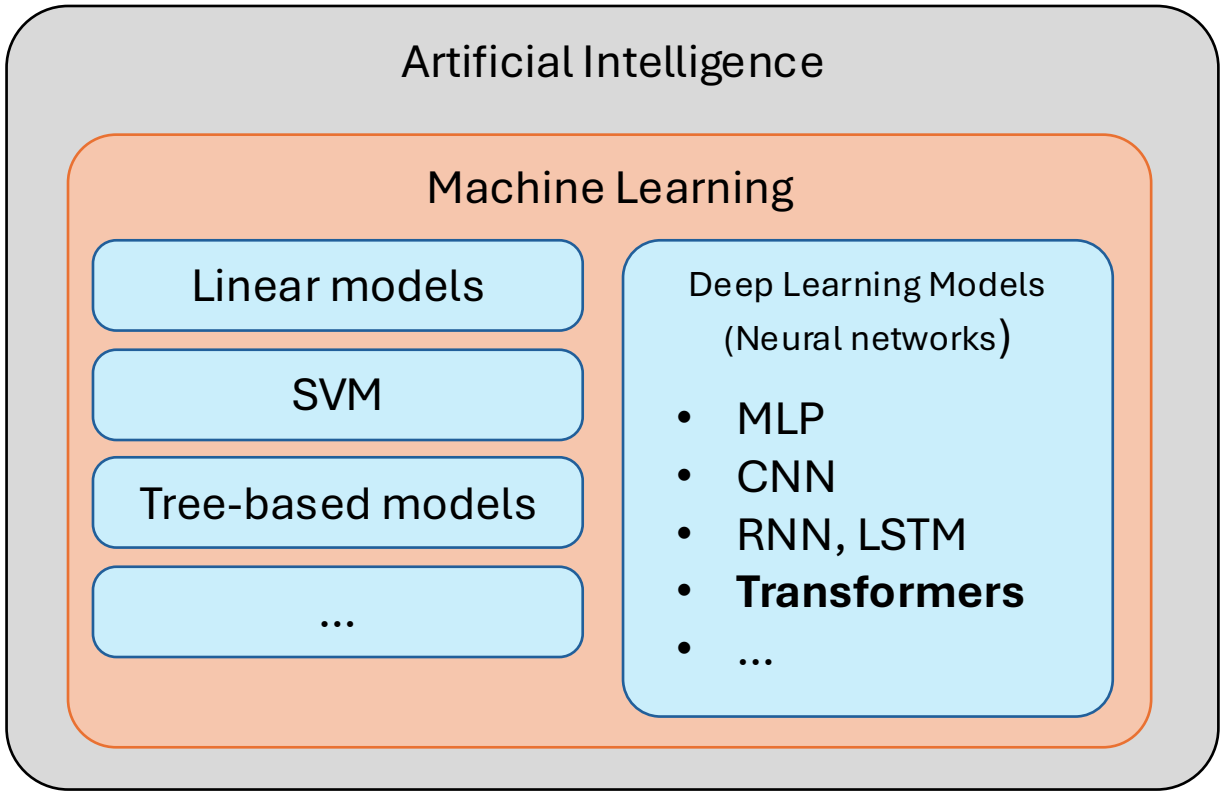
What is a Large Language Model (LLM)?

A Conventional Definition of LLM

- **LLM** usually refers to decoder-only transformer-based text to text models
- **Large** in three dimensions:
  - **Model size**: billions or 100s of billions of parameters
  - **Training data**: billions to trillions of tokens
  - **Compute**: thousands of GPUs

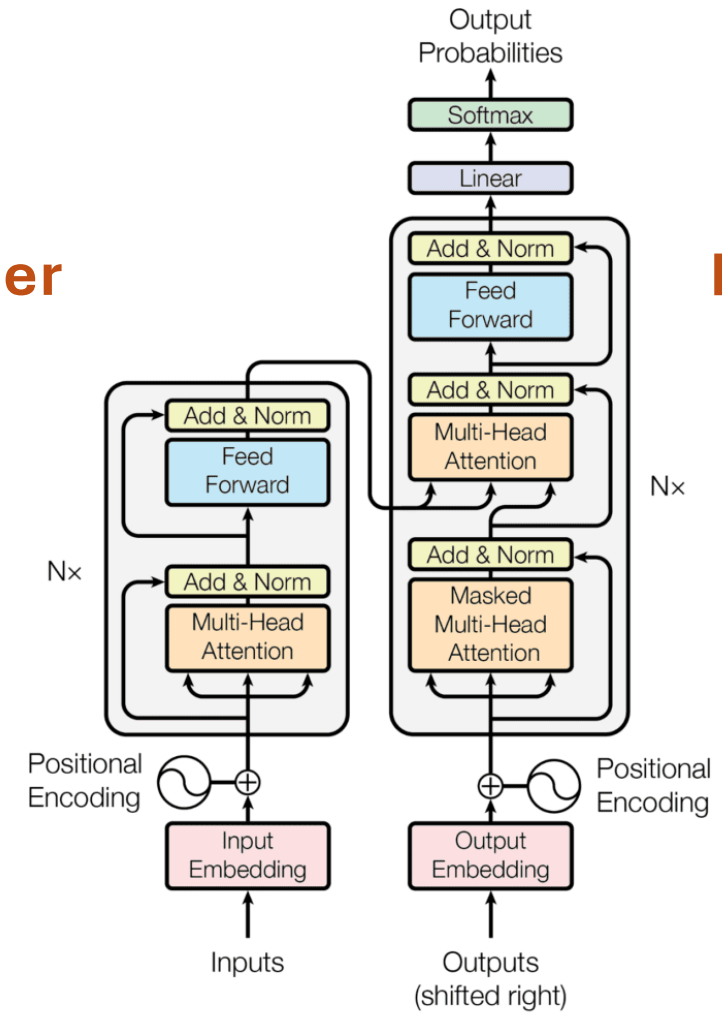
- https://cme295.stanford.edu/slides/fall25-cme295-lecture3.pdf
- https://en.wikipedia.org/wiki/List_of_large_language_models

Transformer models



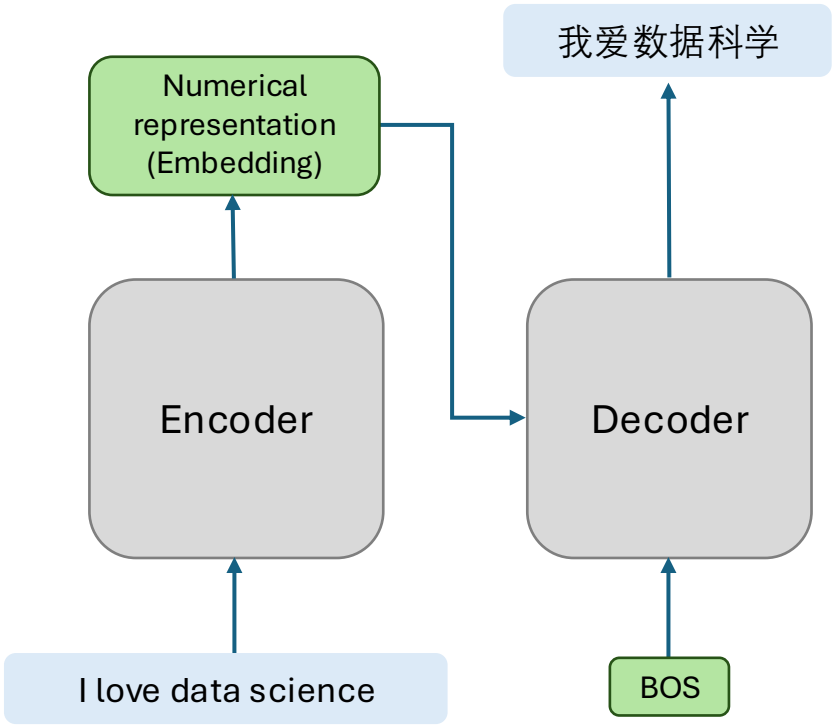
Prince, S. J. D. (2023). ***Understanding Deep Learning.*** MIT Press. http://udlbook.com
An introduction to Statistical Learning. Springer. https://www.statlearning.com/

Transformer models



Encoder

Decoder



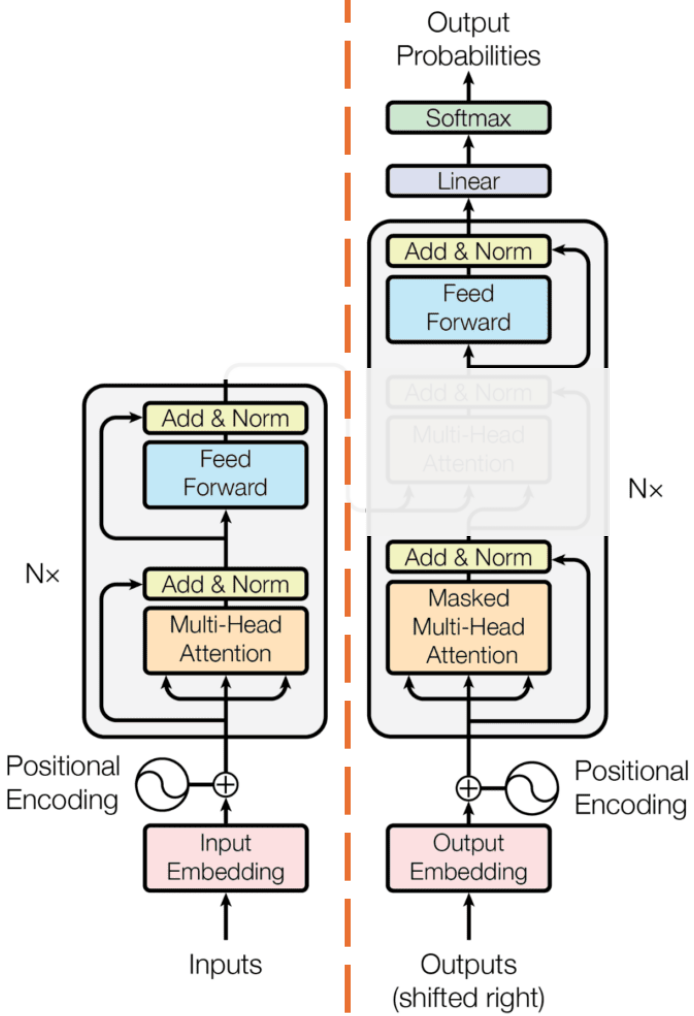
- **Encoder:** converts the source-language sentence into internal representations.
- **Decoder:** uses these representations to generate the target-language sentence, one token at a time.

Attention is all you need http://arxiv.org/abs/1706.03762

Transformer models

Encoder-Only

Representation(Discriminative) Models
Calculate embeddings for input text
BERT, RoBERTa, DeBERTa

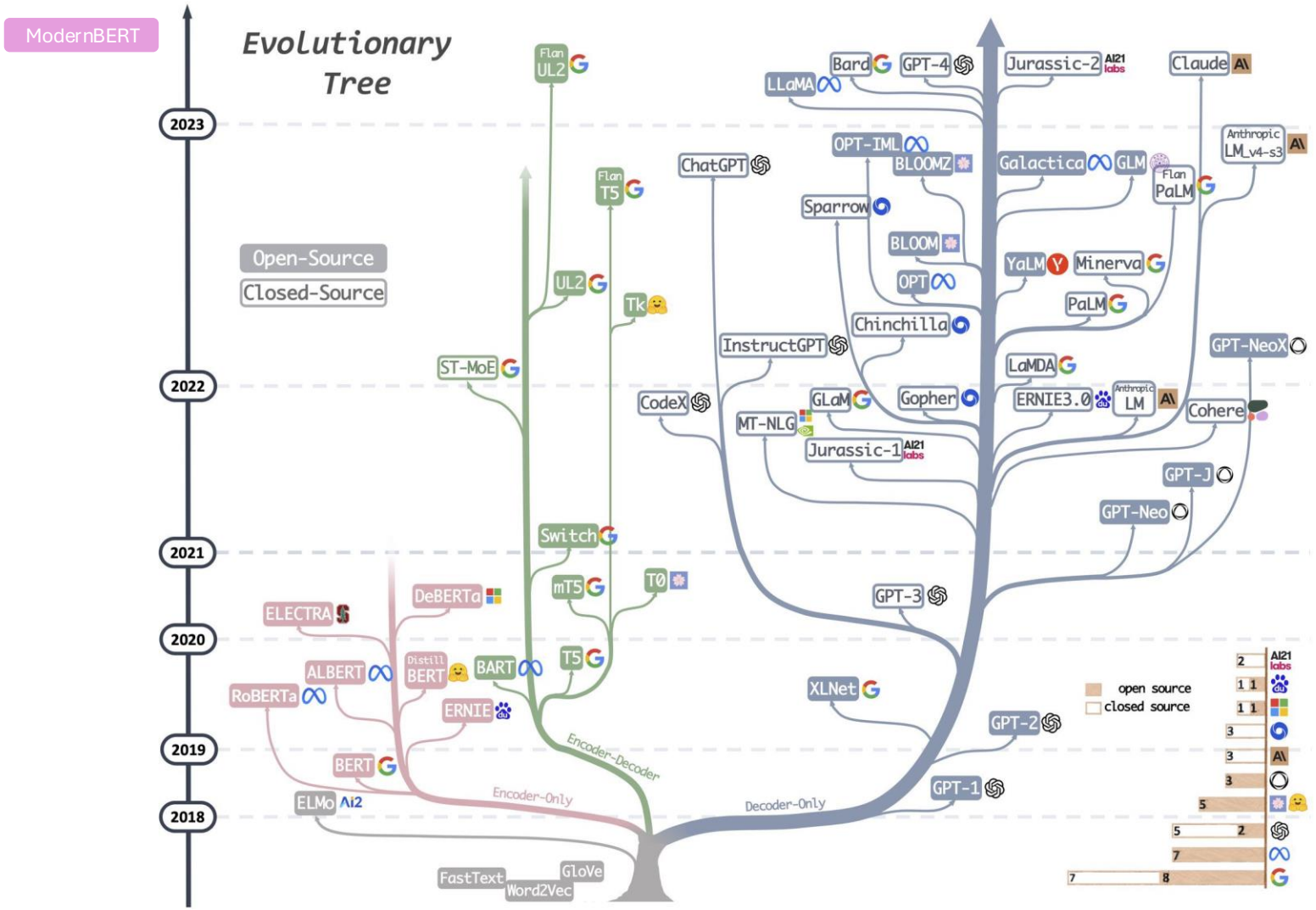


Decoder-Only (LLM)

Generative Models
Predict next tokens
GPT, Claude, Gemini, Llama

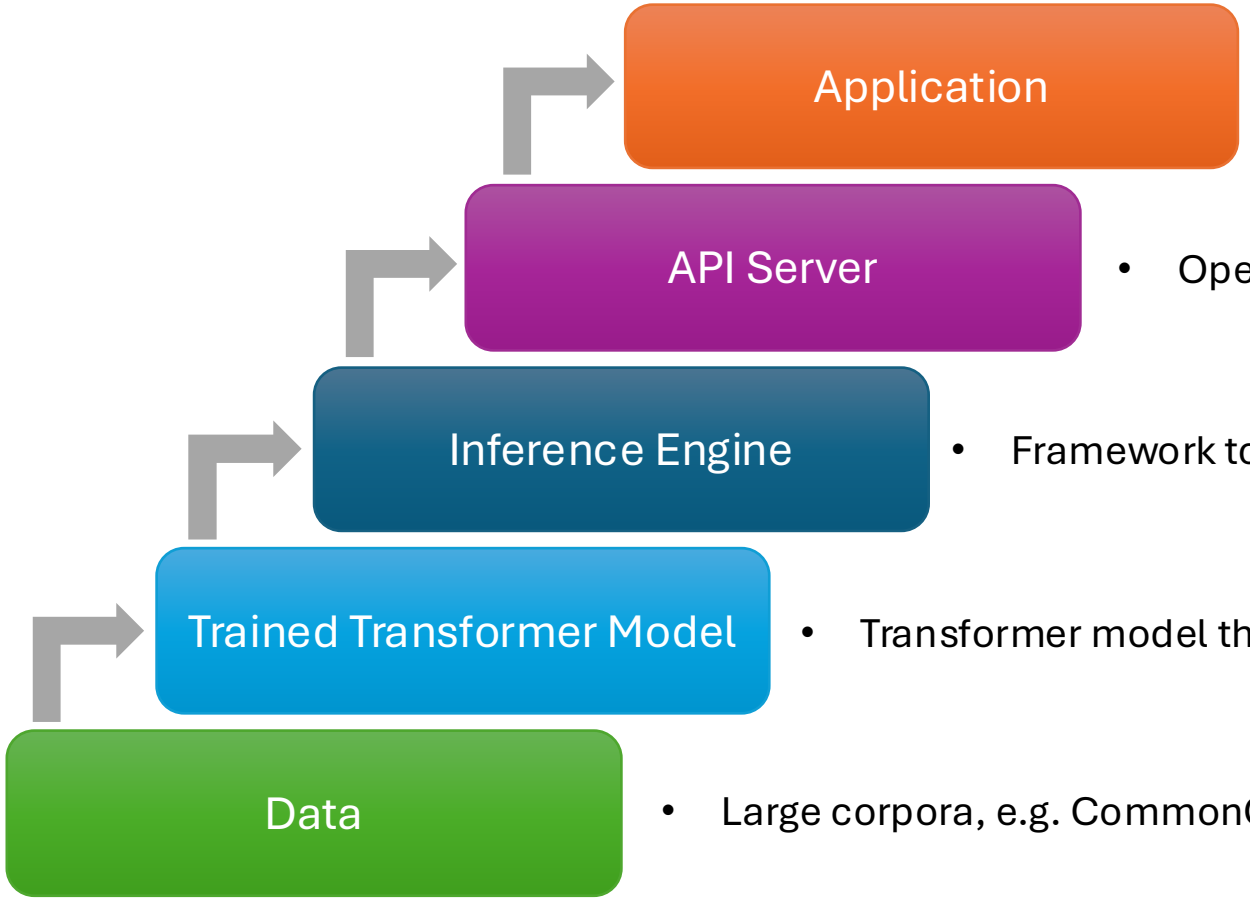
Attention is all you need http://arxiv.org/abs/1706.03762

Transformer models



Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond https://arxiv.org/pdf/2304.13712

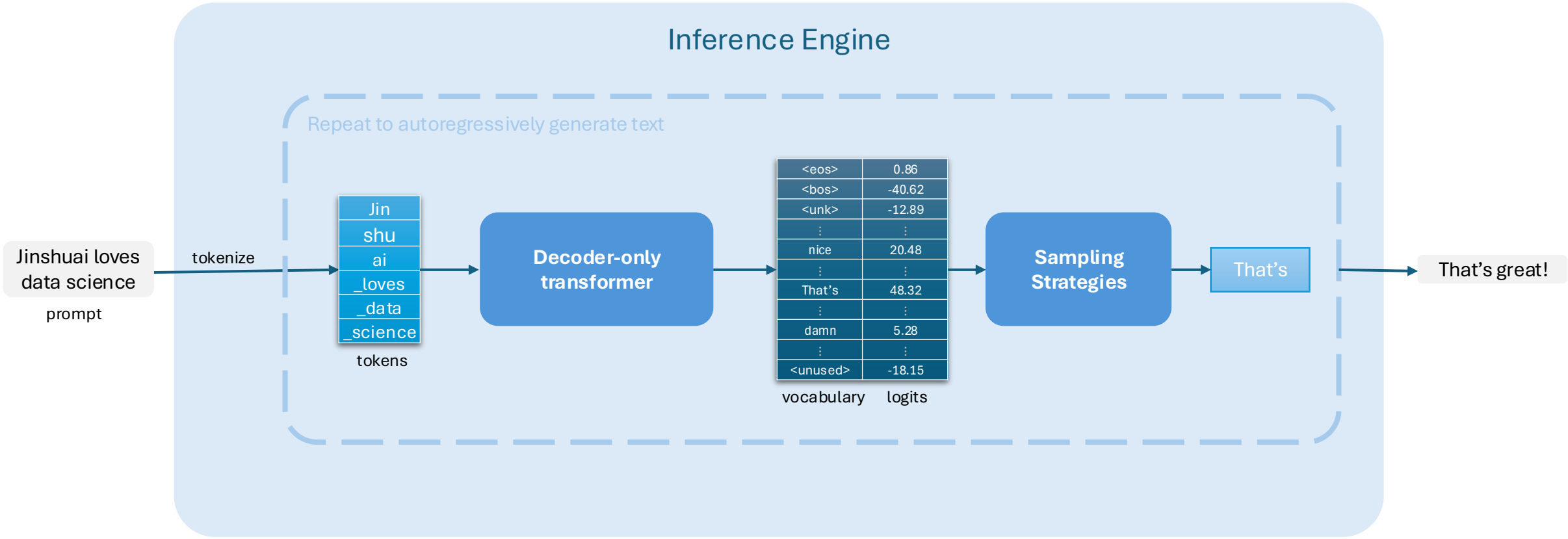
From transformers to LLM applications



- Claude chat interface, ChatGPT, Copilot, Cursor, ...
- OpenAI API, Anthropic API, Huggingface, Ollama, ...
- Framework to generate texts
- Transformer model that do next token prediction
- Large corpora, e.g. CommonCrawl, Github, Wikipedia, ...

How does an LLM generate text?

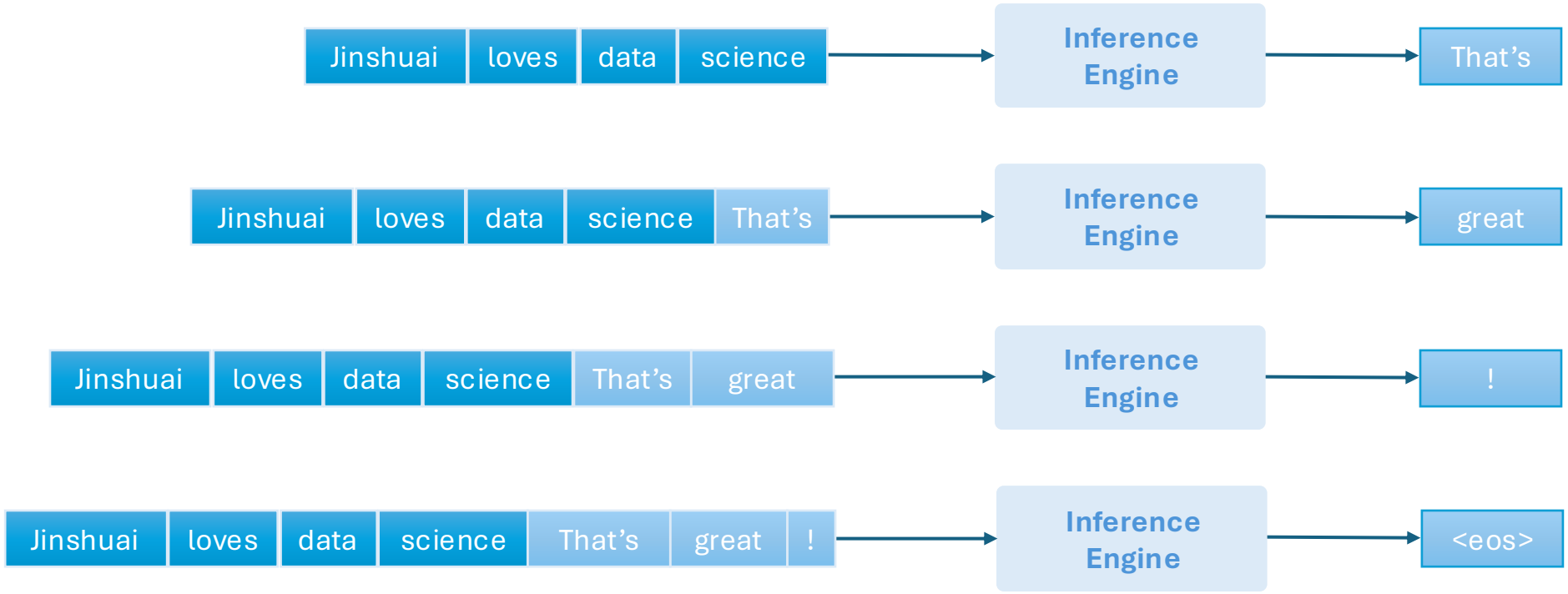
LLM Inference (Next Token Prediction)



- **Tokens** are the basic units of text processed by the model. LLMs typically use subword tokenization, which helps handle rare words, new terms, and misspellings.
- **Vocabulary** is the fixed set of unique tokens the model recognizes. Vocabulary size typically ranges from 10k to over 100k tokens, depending on the model.

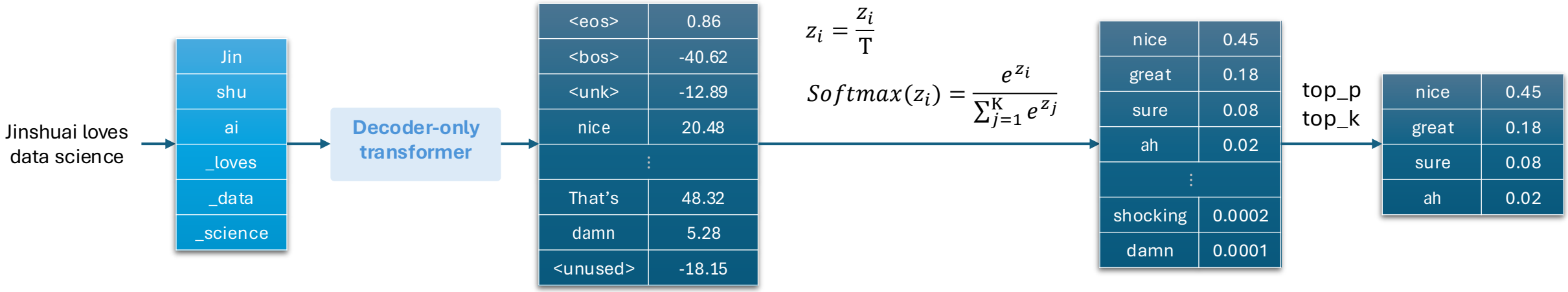
- **Transformer** calculates the score (logits) for each unique token in the vocabulary
- **Sampling** strategies selects one token from the vocabulary to be the next token.

LLM Inference (Next Token Prediction)



Demo of tokenization and next token prediction

Sampling and Randomness



Hyperparameters	
Top-p (nucleus sampling)	the smallest set of most probable tokens whose cumulative probability exceeds a threshold p
Top-k sampling	the k most likely tokens
Temperature	scale the logits and control the randomness of tokens
Max-token	the maximum number of tokens can be generated
Seed	ensure reproducibility of random number generator

- Greedy search: select the token with the highest score
- Sampling
  - To introduce more randomness
  - Visualization:
    - https://andreban.github.io/temperature-topk-visualizer/
  - In practice, setting "Temperature = 0" does not guarantee reproducibility.

- https://platform.openai.com/docs/guides/latest-model#gpt-5-2-parameter-compatibility
- https://ai.google.dev/gemini-api/docs/gemini-3#temperature

Constraint Decoding

The output is controllable by altering the probabilities

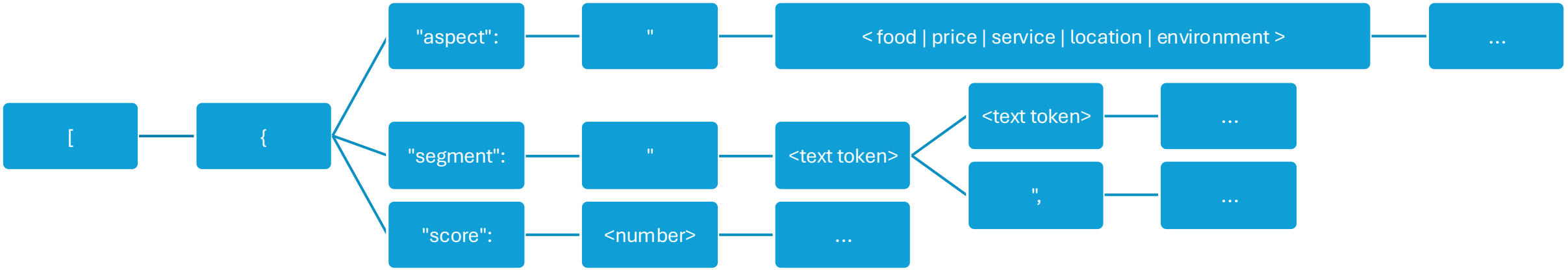
- Set a token's probability to 0 to prevent it from showing in the output
- Apply penalty or reward
- Structured output: JSON output

Structured Output

Example: Aspect-based sentiment analysis

The food was delicious, but the price was too high.

```
[
  {
    "aspect": "food",
    "segment":" The food was delicious,",
    "score": 5
  },
  {
    "aspect": "price",
    "segment":" but the price was too high",
    "score": 2
  }
]
```



JSON Schema

Modern LLM inference engines include built-in JSON parsing and can generate structured outputs when provided with a **JSON Schema**.

```
[
  {
    "aspect": "food",
    "segment":" The food was delicious,",
    "score": 5
  },
  {
    "aspect": "price",
    "segment":" but the price was too high",
    "score": 2
  }
]
```

JSON data

```
{
 "$schema": "https://json-schema.org/draft/2020-12/schema",
 "type": "array",
 "items": [
  {
   "type": "object",
   "properties": {
    "aspect": {
     "enum": ["food", "service", "price", "location", "environment"]
    },
    "segment": {
     "type": "string"
    },
    "score": {
     "type": "integer"
    }
   },
   "required": [
    "aspect",
    "segment",
    "score"
   ]
  }
 ]
}
```

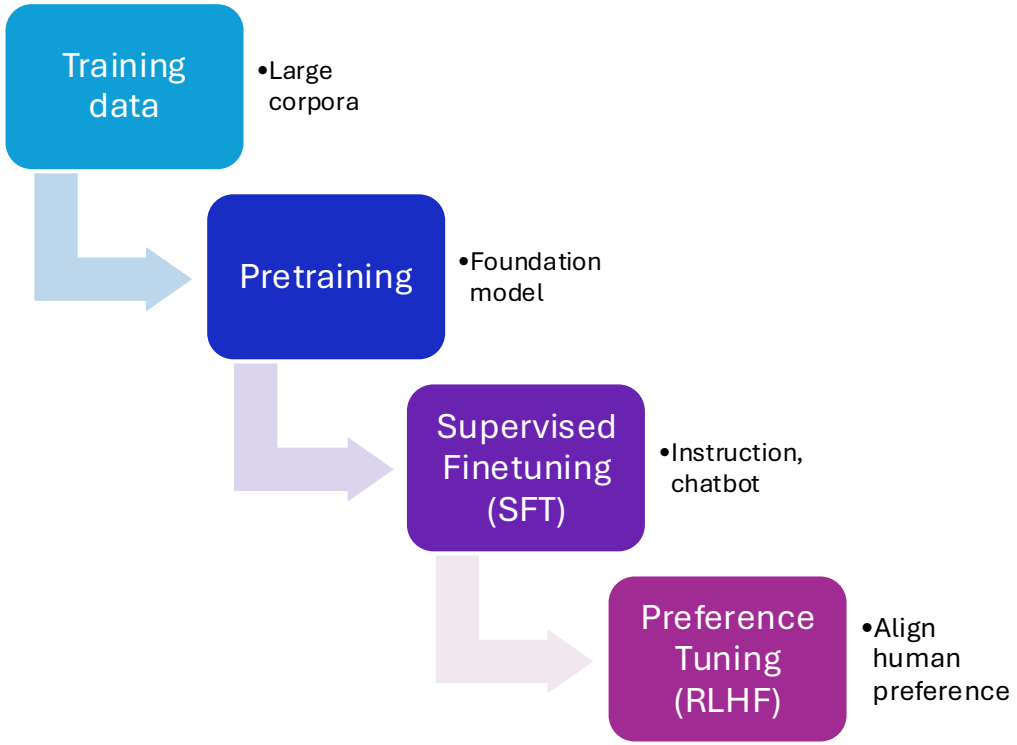
JSON schema

https://json-schema.org/

Next Token Prediction - Training Technique

Training Data	Input	Label
The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.	<bos>The	data
	<bos>The Data	science
	<bos>The Data Science	institute
	<bos>The Data Science Institute	forms
	<bos>The Data Science Institute forms	the
	…	…
	<bos>The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.	<eos>

LLM Training Stages



Pretraining

The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.

Preference Tuning

Explain what a PhD is.

Response A	Response B
A PhD is an advanced academic degree that demonstrates expertise in a specific field of study.	A PhD is a multi-year commitment to becoming a world expert in one very narrow topic, while periodically questioning your life choices.
✅ Accepted	❌ Rejected

* This example was generated by ChatGPT.

Supervised finetuning

Question Answering

What is Data Science Institute?

The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.

Conversation (Chat)

<|system|> You are a helpful assistant.**<|user|>**What is Data Science Institute?**<|assistant|>**The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.**<eos>**

Recommended resources: **State of GPT** (Andrej Karpathy): https://www.youtube.com/watch?v=bZQun8Y4L2A&t=1s

Chat Roles

Training data - Mock conversions
<|system|> You are a helpful assistant.
<|user|> What is Data Science Institute?
<|assistant|> The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science.
<|user|> What course does DSI provide?
<|assistant|> The DSI offers DS105 Data For Data Scientists and DS202 Data Science for Social Scientists
…

<|system|> You are a helpful assistant. <|user|> What is Data Science Institute? <|assistant|> The Data Science Institute forms the institutional hub for data science and AI activity at the London School of Economics and Political Science. <|user|> What course does DSI provide? <|assistant|> The DSI offers DS105 Data For Data Scientists and DS202 Data Science for Social Scientists …

- **System:** Sets the "personality," constraints, and tone. It provides the high-level instructions (e.g., "You are a helpful medical assistant. Use simple language.").

- **User:** Represents the human. This is the prompt or question the model needs to respond to.

- **Assistant:** Represents the "gold standard" response. During training, the model is penalized if its output deviates from this text.

Recommended resources: **State of GPT** (Andrej Karpathy): https://www.youtube.com/watch?v=bZQun8Y4L2A&t=1s

In-context Learning: Few-shot

You are analyzing customer reviews of restaurants.

For each review below, assign a sentiment score from **1 to 5**, where:
- **1** = very negative
- **2** = negative
- **3** = neutral
- **4** = positive
- **5** = very positive

Respond with only the number (1–5).

Examples:

Review: "The food was cold, the waiter was rude, and I will never come back."
Sentiment score: 1

Review: "Nothing special. The food was okay and the service was fine."
Sentiment score: 3

Review: "Great atmosphere, friendly staff, and delicious food. Highly recommend!"
Sentiment score: 5

Analyze the following:
“The food was delicious.”

<|system|>You are analyzing customer reviews of restaurants.

For each review below, assign a sentiment score from **1 to 5**, where:
- **1** = very negative
- **2** = negative
- **3** = neutral
- **4** = positive
- **5** = very positive

Respond with only the number (1–5).

<|user|>The food was cold, the waiter was rude, and I will never come back.
<|assistant|> 1

<|user|> Nothing special. The food was okay and the service was fine.
<|assistant|> 3

<|user|> Great atmosphere, friendly staff, and delicious food. Highly recommend!"
<|assistant|> 5

<|user|> The food was delicious.

In-context Learning: Zero-shot

You are analyzing customer reviews of restaurants.

For each review below, assign a sentiment score from **1 to 5**, where:

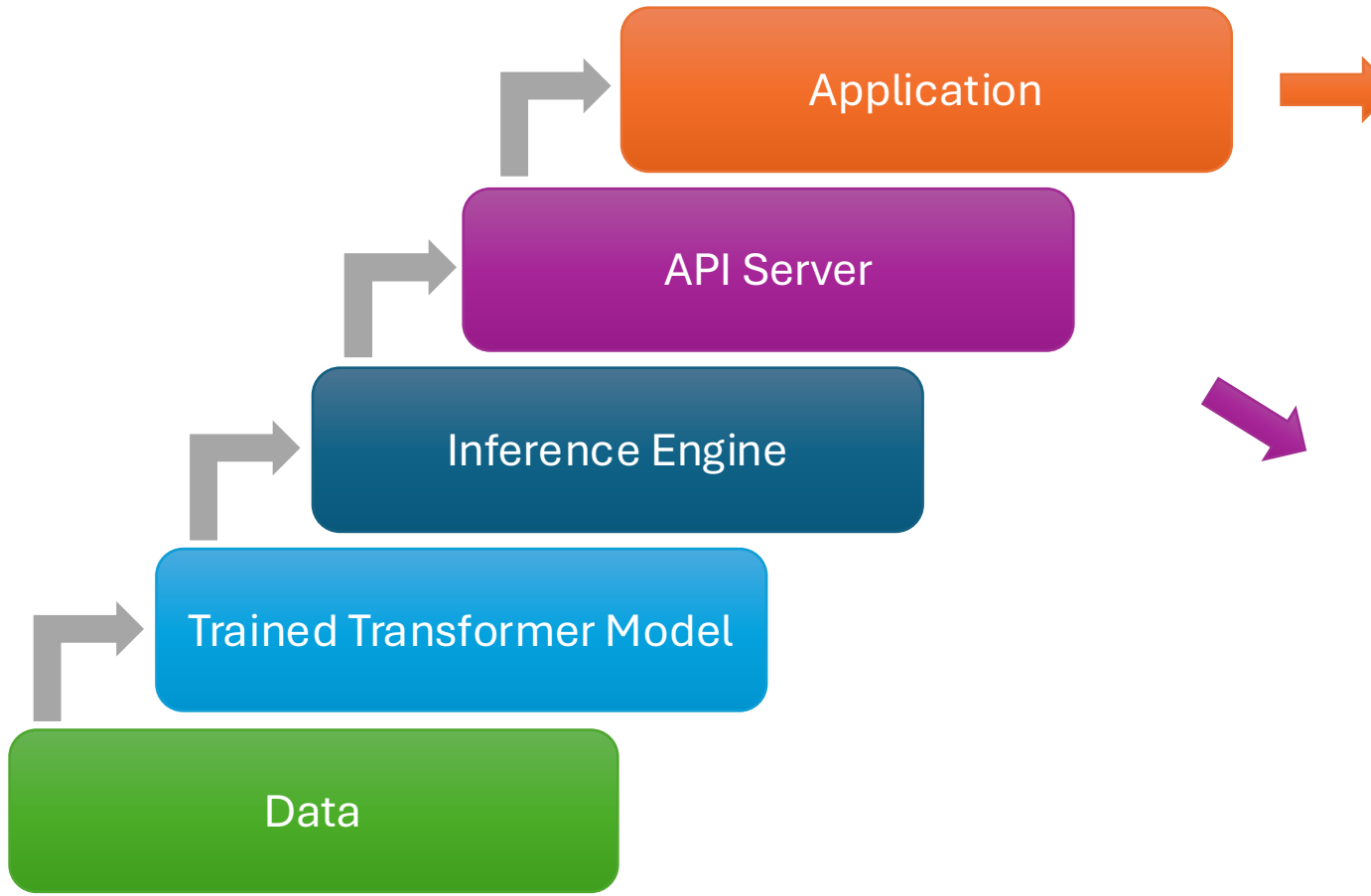
- **1** = very negative
- **2** = negative
- **3** = neutral
- **4** = positive
- **5** = very positive

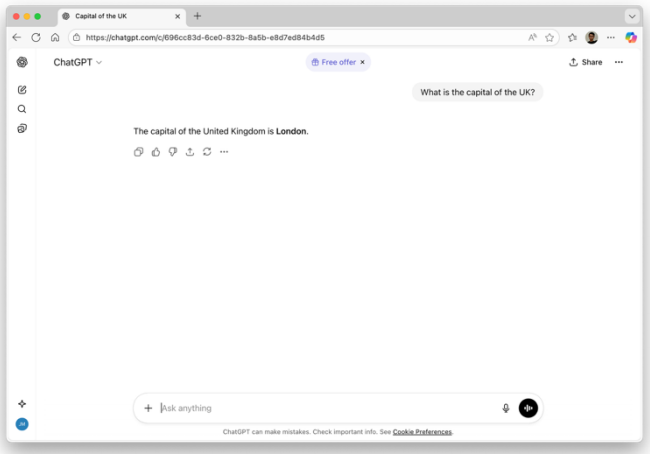
Respond with only the number (1–5).

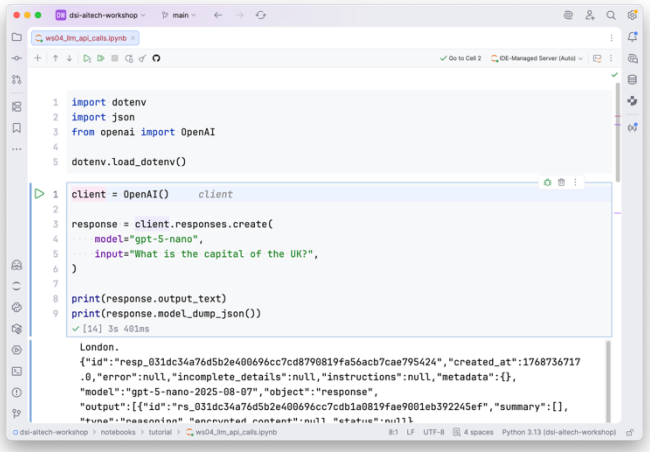
Review:
“The food was delicious.”

How to call LLM API programmatically?

Difference Between Chat Interfaces and API calls







Recommended resources: Python Fundamentals, Collections and First Steps with APIs, DS105 Data for Data Science
https://lse-dsi.github.io/DS105/2025-2026/autumn-term/weeks/week02/lecture.html

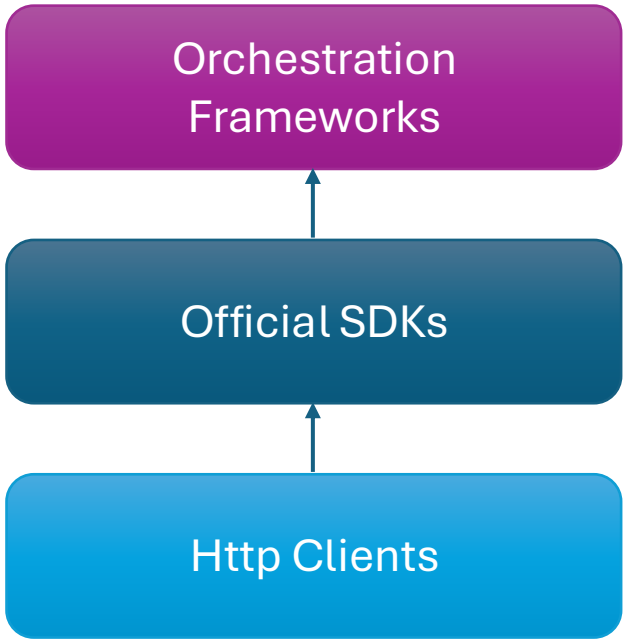
Difference Between Chat Interfaces and API calls

	Chat Interface	API
Interface	Web / app UI	HTTP requests
Requires coding	No	Yes
Automation	No	Fully automated
Scalability	Low	High
Customization	Limited No access to hyperparameters, chat roles	High Fully controlled
Use case	AI assistant, Vibe coding, Learning	Data analysis, Automated pipeline

- API is useful for batch processing and automation.

Using different libraries to call an LLM API

- HTTP client libraries
  - Python: requests, httpx
  - R: httr2
- Official SDK libraries (Python Only)
  - openai
  - anthropic
  - google-genai
- LLM orchestration frameworks
  - Python: LangChain, LlamaIndex
  - R: ellmr



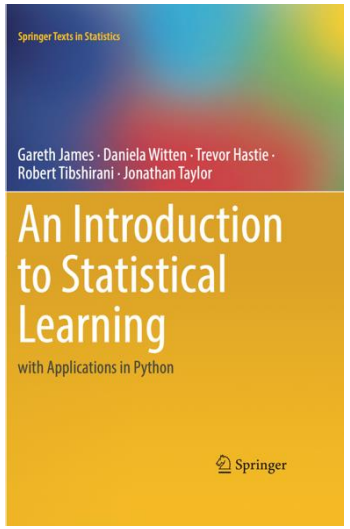
Some Tips

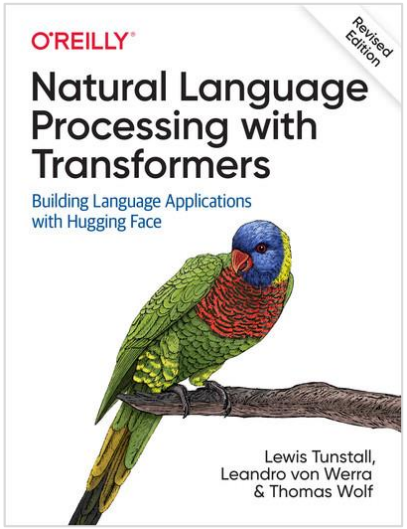
- Prompt Optimizer
  - https://info.lse.ac.uk/current-students/digital-skills-lab/copilot
- Counting Tokens
  - https://platform.openai.com/tokenizer
- Prompt caching
  - https://platform.openai.com/docs/guides/prompt-caching

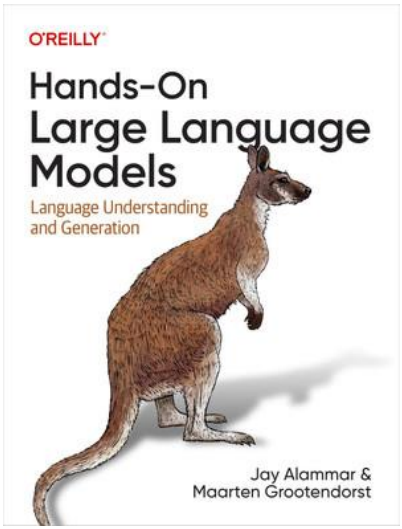
Demo of LLM API calls

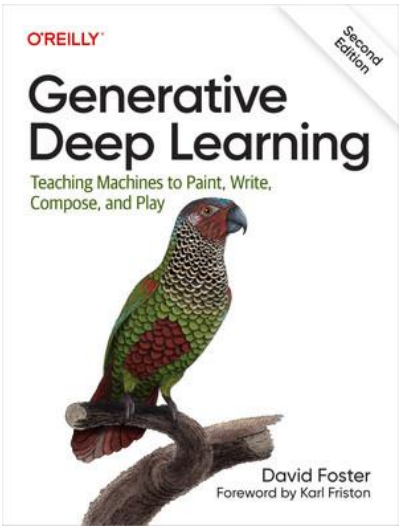
About LLM and AI Learning

- Machine learning and deep learning







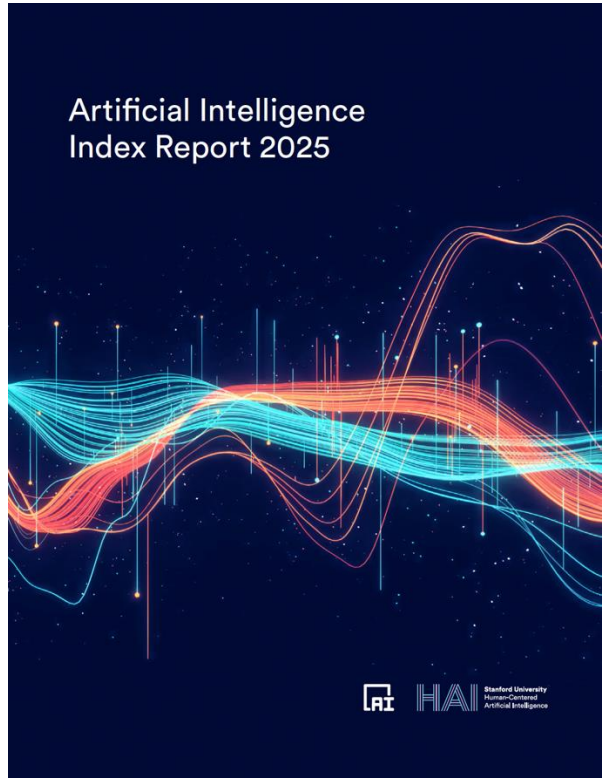


About LLM and AI Learning

- Stanford CME295 - Transformers & Large Language Models
  - https://cme295.stanford.edu/
- LLM official documentation and tutorials
  - https://platform.openai.com/docs/overview
  - https://anthropic.skilljar.com/
- LSE Digital Skills Lab
  - https://info.lse.ac.uk/current-students/digital-skills-lab/Gen-AI

About LLM and AI Learning

- The Stanford AI Index Report
  - https://hai.stanford.edu/ai-index



Summary

- The basics of LLM
  - Decoder-only transformer
  - LLM application architecture
- The training stages and tasks of LLM
  - Pretraining
  - Supervised finetuning
  - Preference tuning
- The process of text generation
  - Next token prediction
  - Sampling strategies
- Prompt design techniques
  - Structured output
  - Chat roles and few-shot learning
- LLM APIs

Image generated by Gemini and Nano Banana

What’s next?

- Advanced LLM and AI topics
  - Reasoning models
  - RAG and Agentic AI
  - Multi-modal models
  - Generative AI
- Practice session
  - Hands-on LLM API calling
  - More on LLM tips, best practices
  - Run an open source LLM locally