

Large Language Models

Large Language Models

Introduction

- Type of artificial intelligence model
- Designed to understand, generate, and manipulate natural language text
- Trained on large (text) datasets
- Can perform various language tasks like translation, summarization, text generation, ...
- Capabilities improved dramatically in the last years
- Based on Deep Learning, specifically Transformers



Large Language Models

LLM History

```
Welcome to          EEEEE LL 1111 2222222 AAAA
EE LL 11 22 AA AA
EEEE LL 11 22 AAAAAA
EE LL 11 22 AA AA
EEEE LL1111 111 222222 AA AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU: I
```

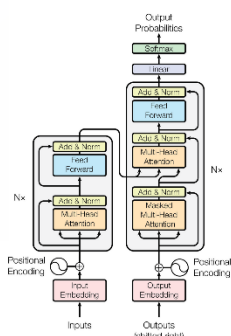
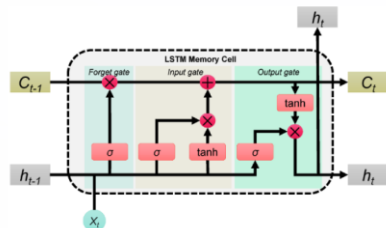


Figure 1: The Transformer - model architecture.



ELIZA

Recursive NN

LSTM

Transformers

BERT
GPT

GPT-2

GPT-3

GPT-3.5

ChatGPT

GPT-3.5
Llama
Gemini

GPT-4
Claude

1967

1988

1997

2017

2018

2019

2020

2021

2022

2023

2024

Large Language Models

History: ELIZA

- 1960s Eliza chatbot
- simple pattern recognition
- pretending “conversation”

```
Welcome to

EEEEEE LL      IIII ZZZZZZZ AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZZ AA  AA

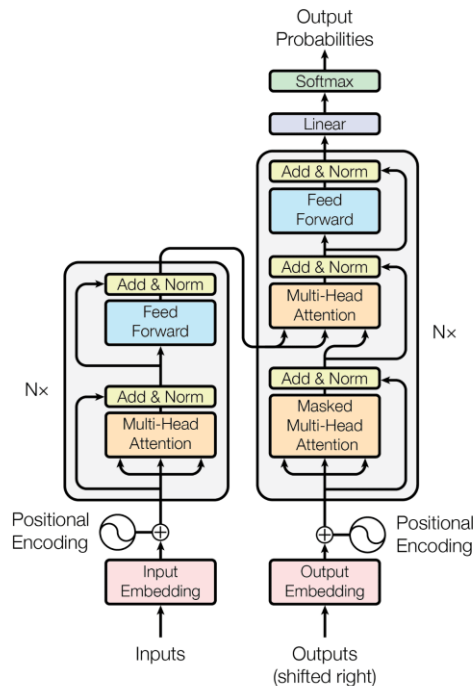
Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

Large Language Models

History: Transformers

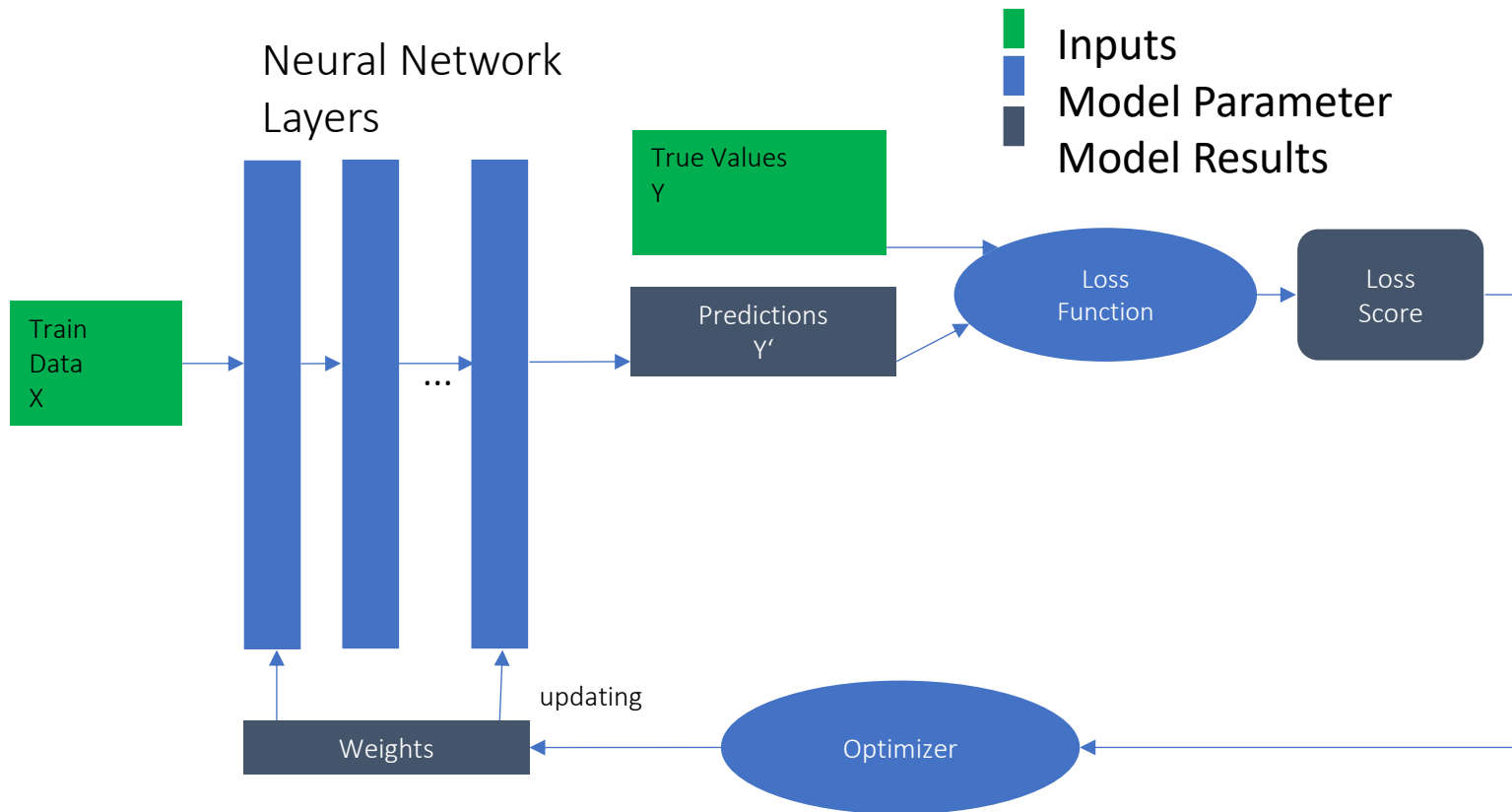
- paper “Attention is all you need” from Google team (Vaswani, et. al.)
- encoder and decoder
- multiple stacked layers of self-attention
- multi-head attention – allows to focus on different parts of input simultaneously



Source: <https://machinelearningmastery.com/the-transformer-model/>

Large Language Models

Deep Learning



Tokenization

Introduction

- process of breaking down a sequence of text into individual units
- typical units: words, subwords
- units called tokens
- different approaches
 - word tokenization
 - sentence tokenization
 - subword tokenization

Tokenization

Word Tokenization

Sample Text

The quick brown fox jumps over the lazy dog.

Tokens

The quick brown fox jumps over the lazy dog.

Tokenization

Sentence Tokenization

Sample Text

The quick brown fox jumps over the lazy dog.

Tokens

The quick brown fox jumps over the lazy dog.

Tokenization

Word Tokenization and Embedding

- fundamental step in NLP
- first step of all NLP tasks

Text

The quick brown fox jumps over the lazy dog.

Tokens

The quick brown fox jumps over the lazy dog.

Embeddings

[0.2, ...]

...

Tokenization

Sentence Tokenization and Embedding

- fundamental step in NLP
- first step of all NLP tasks

Text

The quick brown fox jumps over the lazy dog.

Tokens

The

quick

brown

fox

jumps

over

the

lazy

dog.

Embeddings

[0.2, ...]

...

Tokenization

Sub-word Tokenization

Text

It is raining.



Tokens

It

is

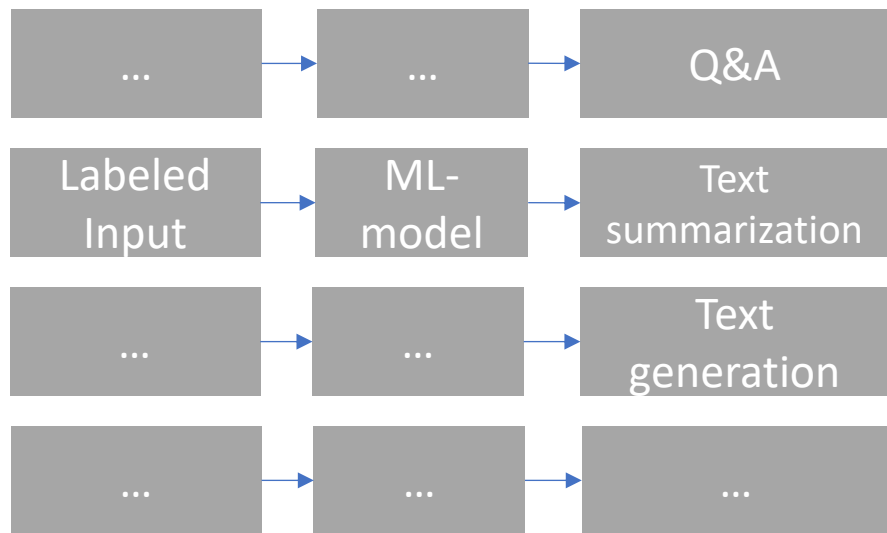
rain

ing

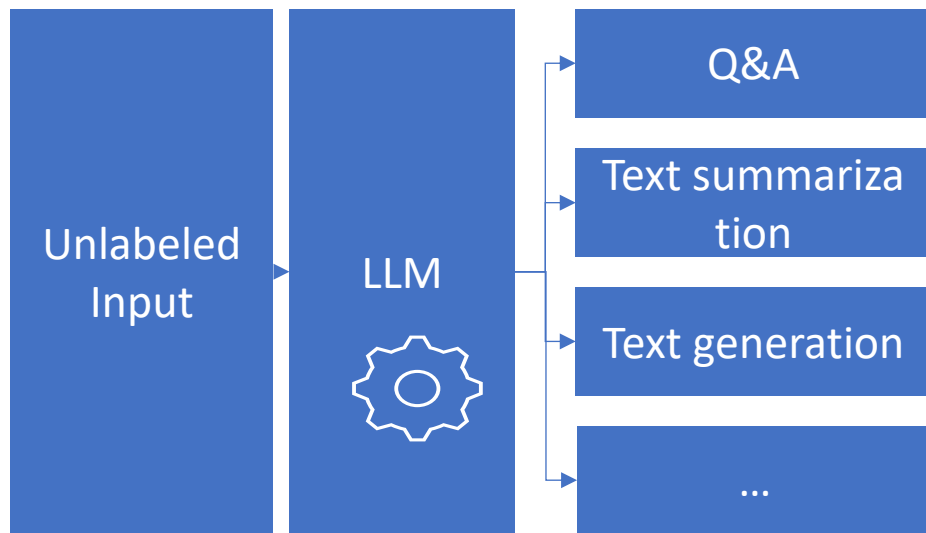
.

Large Language Models

Difference to Classical Models



Classical ML-models



Large Language Model

Large Language Models

LLM Tasks

- LLMs can cover all NLP-tasks

- Text Generation
 - Writing assistance, story generation

Translation

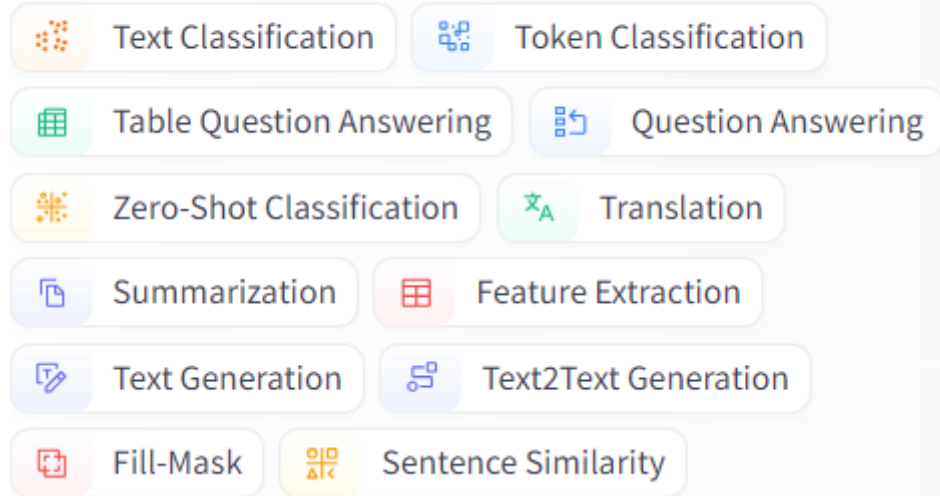
Conversational Agents

Chatbots, virtual assistants

Text summarization

...

Natural Language Processing

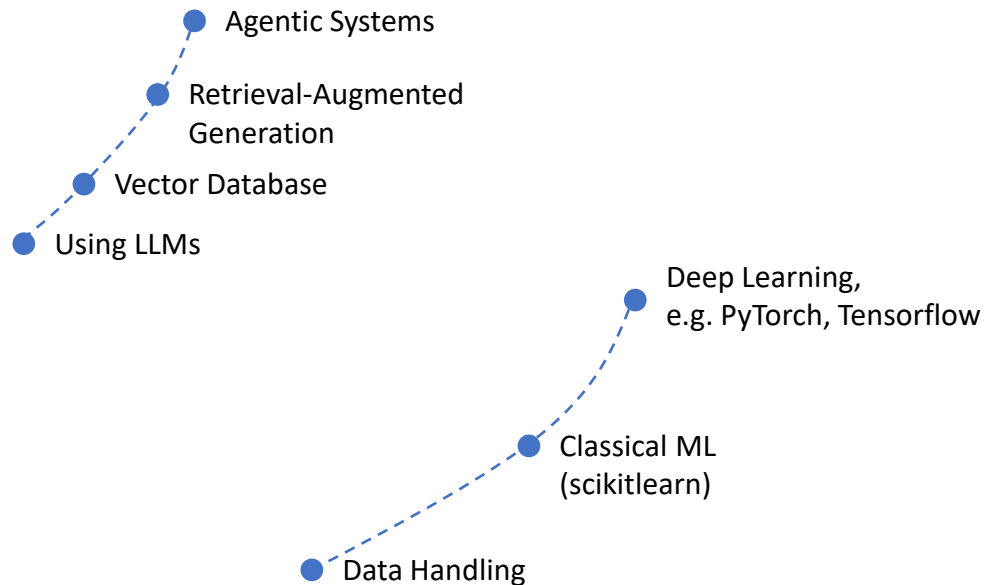


Source: <https://huggingface.co/models>

Large Language Models

Model Performance, more Capabilities

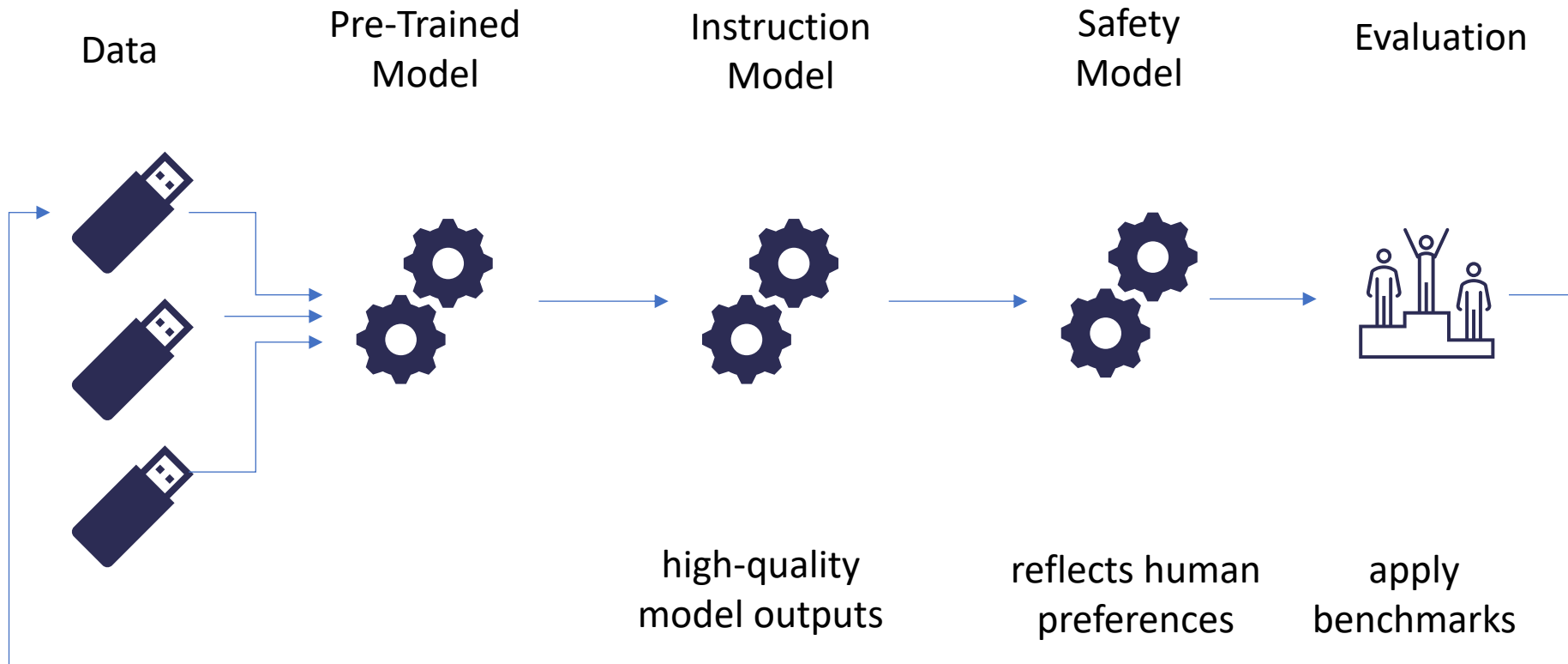
Performance /
Capabilities



Difficulty to apply

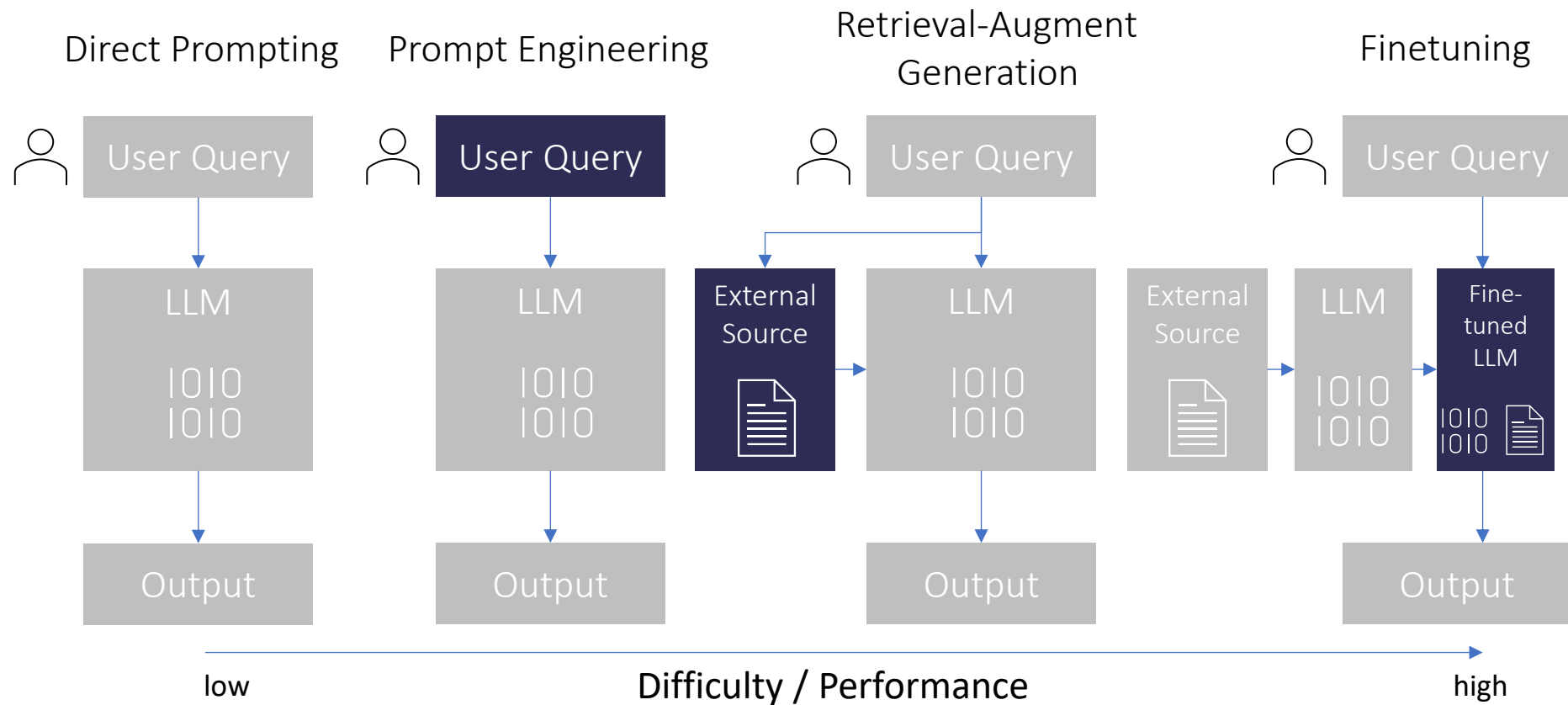
Large Language Models

Training Process



How to improve LLM-Output

Prompt Engineering, RAG, Finetuning



Large Language Models

Available Providers & Models



- GPT-4o
- GPT-4o mini
- o1-preview / mini
- GPT-4 (Turbo)
- GPT-3.5 Turbo



- Gemini-1.5 Pro
- Gemini-1.5 Flash



- Grok-2

ANTHROPIC

- Claude 3.5 Sonnet

Proprietary /
closed source



- Llama 3.1 family



- Mistral 8x7b

open source/
open weight

Large Language Models

LLM Benchmarks

Rank★ (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	4	Gemini-Exp-1114	1344	+7/-7	6446	Google	Proprietary	Unknown
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+3/-3	42225	OpenAI	Proprietary	2023/10
3	1	o1-preview	1333	+4/-4	26268	OpenAI	Proprietary	2023/10
4	5	o1-mini	1308	+4/-3	28953	OpenAI	Proprietary	2023/10
4	4	Gemini-1.5-Pro-002	1301	+4/-4	23856	Google	Proprietary	Unknown
6	9	Grok-2-08-13	1290	+3/-3	47908	xAI	Proprietary	2024/3
6	11	Yi-Lightning	1287	+4/-4	27114	01 AI	Proprietary	Unknown
7	4	GPT-4o-2024-05-13	1285	+2/-2	108575	OpenAI	Proprietary	2023/10
7	3	Claude 3.5 Sonnet (20241022)	1283	+4/-4	26047	Anthropic	Proprietary	2024/4
10	16	GLM-4-Plus	1275	+3/-4	25601	Zhipu AI	Proprietary	Unknown
10	18	GPT-4o-mini-2024-07-18	1272	+3/-3	48407	OpenAI	Proprietary	2023/10
10	18	Gemini-1.5-Flash-002	1272	+4/-4	18112	Google	Proprietary	Unknown
10	26	Llama-3.1-Nemotron-70B-Instruct	1269	+6/-5	7263	Nvidia	Llama 3.1	2023/12
10	7	Meta-Llama-3.1-405B-Instruct-fp8	1267	+4/-3	48804	Meta	Llama 3.1 Community	2023/12

Source: <https://lmarena.ai/>, Snapshot 2024-11-18

Large Language Models

Practical Coding: First LLM Interaction

1.

API Key Setup

groq



<https://platform.openai.com/api-keys>
<https://console.groq.com/keys>

2.

Package Installation



LangChain

3.

LLM Use
Python Script



Large Language Models

Message Types

System Message

- defines how the model should react
- personality, behavior, and limitations throughout conversation
- works like role-play
- Example: „You are a helpful AI assistant designed to provide accurate, concise, and polite responses“
- not seen by user

User Message

- user input
- could be a request, inquiry, or command

AI Message

- corresponds to model response
- different properties,
- mainly „content“ relevant
- more information on input and output tokens available, ...

Large Language Models

Message Types: Example Customer Support

System Message

Example:

„You are a helpful customer support assistant for an online electronics store. Your role is to provide polite and clear responses, assist customers with product inquiries, shipping information, and troubleshooting. Never provide financial or legal advice. If you're unsure about something, kindly ask the customer to contact support for further assistance.“

User Message

- „Hi, I need help tracking my order. I ordered a laptop last week, and I haven't received a shipping confirmation yet.“

AI Message

Large Language Models

Message Types: Example Movie Critic

System Message

Example:

„You are a distinguished film critic with a passion for analyzing movies shown in cinemas. Your responses should be insightful, emphasizing cinematic techniques, character development, themes, and direction. Maintain a professional tone with a flair for the artistic. Avoid colloquial or overly casual language. “

User Message

- „Hey, I just saw *Oppenheimer* and, honestly, it felt kinda long. Why does everyone think it's so great? Can you break it down?”

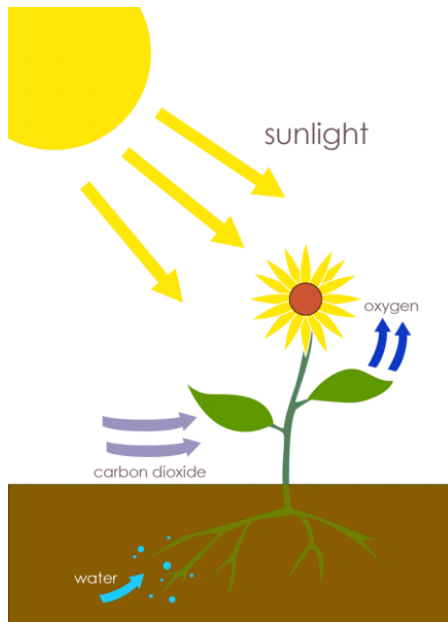
AI Message

Large Language Models

Exercise: Photosynthesis

Go to OpenAI playground

set up system,
and user message



Photosynthesis



Persona:
11 year old

Background:
school presentation

Large Language Models

LLM-Parameters

Temperature

- controls randomness in the process
- 0...model very focused, deterministic result (repeatedly same response)
- 1...increased randomness, broader distribution of tokens is selected; allows for more creative and unexpected outputs

Top p

- controls the probability to consider the next token
- E.g. top-p = 0.9: cumulative probability of tokens which add up to 90% and chooses smallest set of tokens

Max Tokens

- number of tokens to return
- limit due to cost reasons

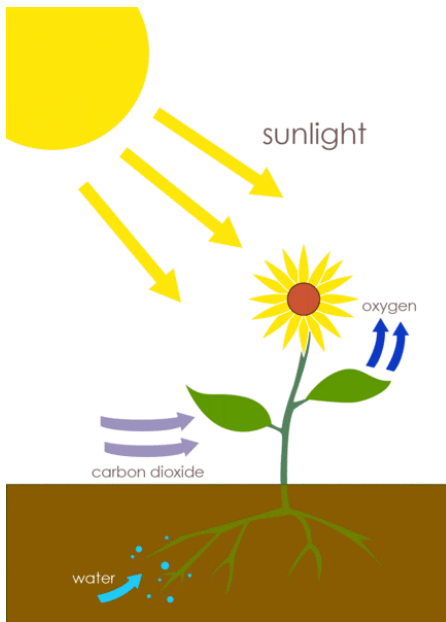
Large Language Models

Exercise: Photosynthesis

Go to OpenAI playground

set up system,
and user message

check impact of
temperature, top p, max
tokens



Photosynthesis

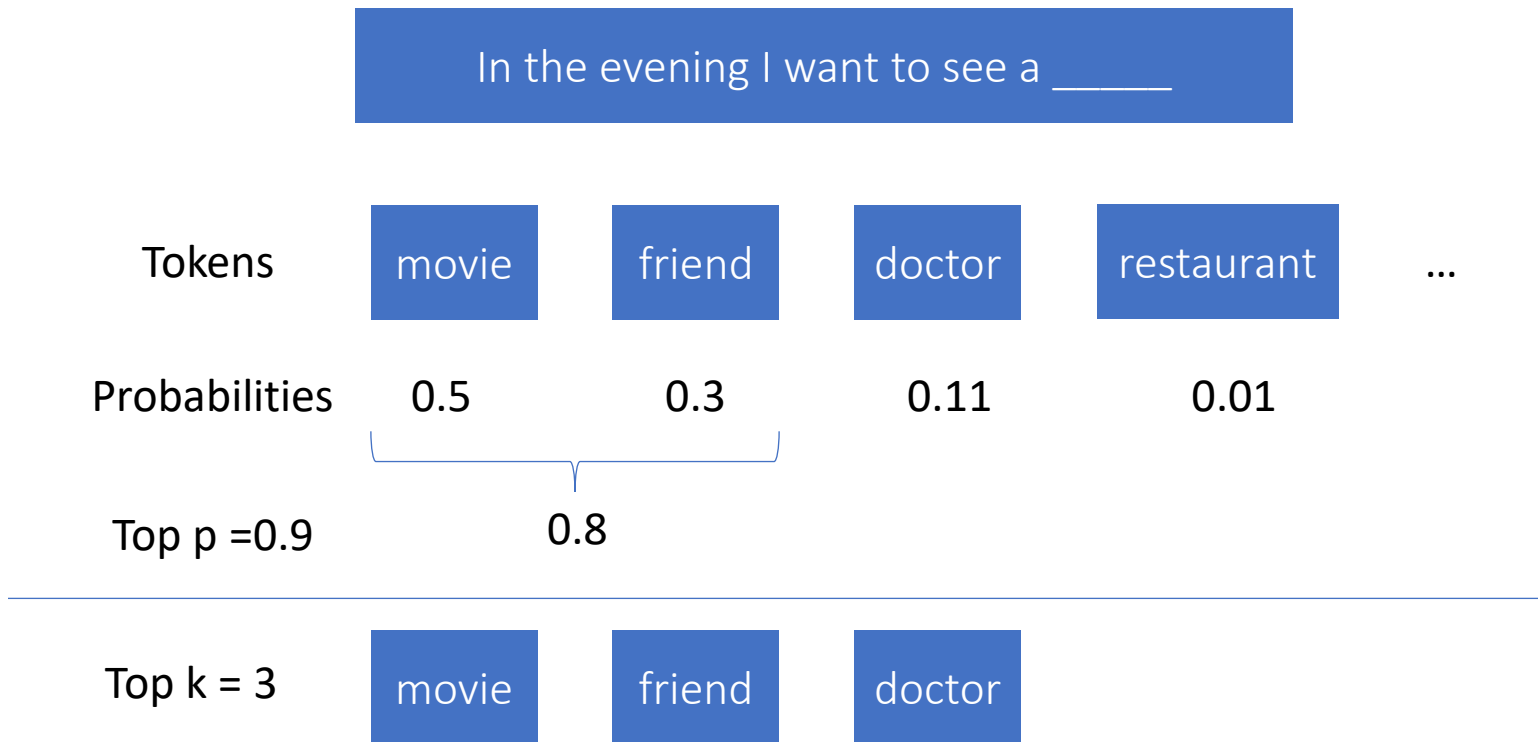


Persona:
11 year old

Background:
school presentation

Large Language Models

LLM-Parameters



Large Language Models

Model Selection



Price



On-Prem vs. Cloud



Performance



Closed Source vs.
Open Weight



Knowledge-Cutoff



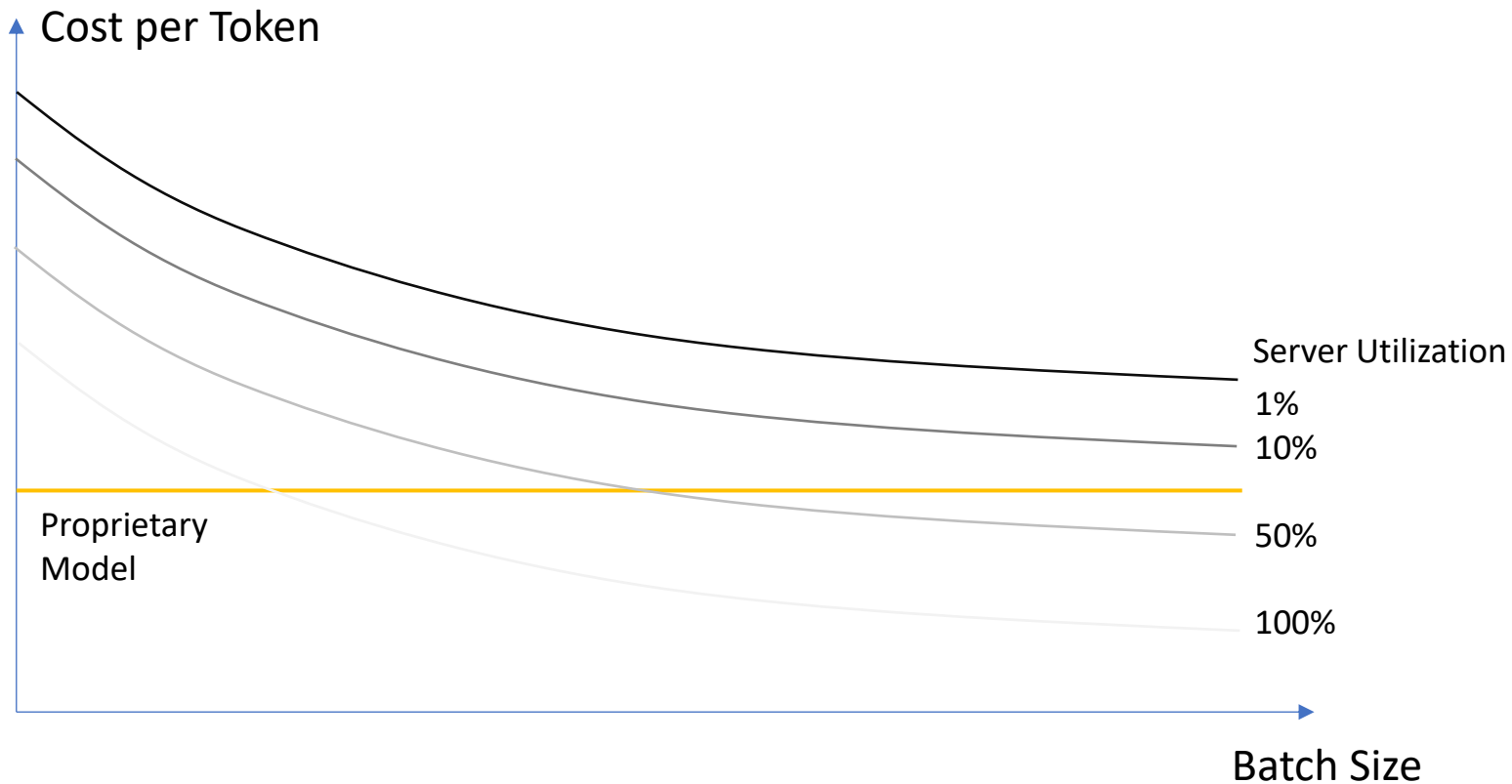
Context-Window



Latency

Large Language Models

Model Selection: Cost vs. Utilization



Large Language Models

Introduction

Artificial Narrow Intelligence (ANI)

- Designed for a specific task
- Limited to scope to well-defined task-specific applications

Artificial General Linguistic Intelligence (AGLI)

- Advanced general capabilities specifically in language understanding and generation
- Examples: GPT-4, Claude, Gemini, Llama, Mistral

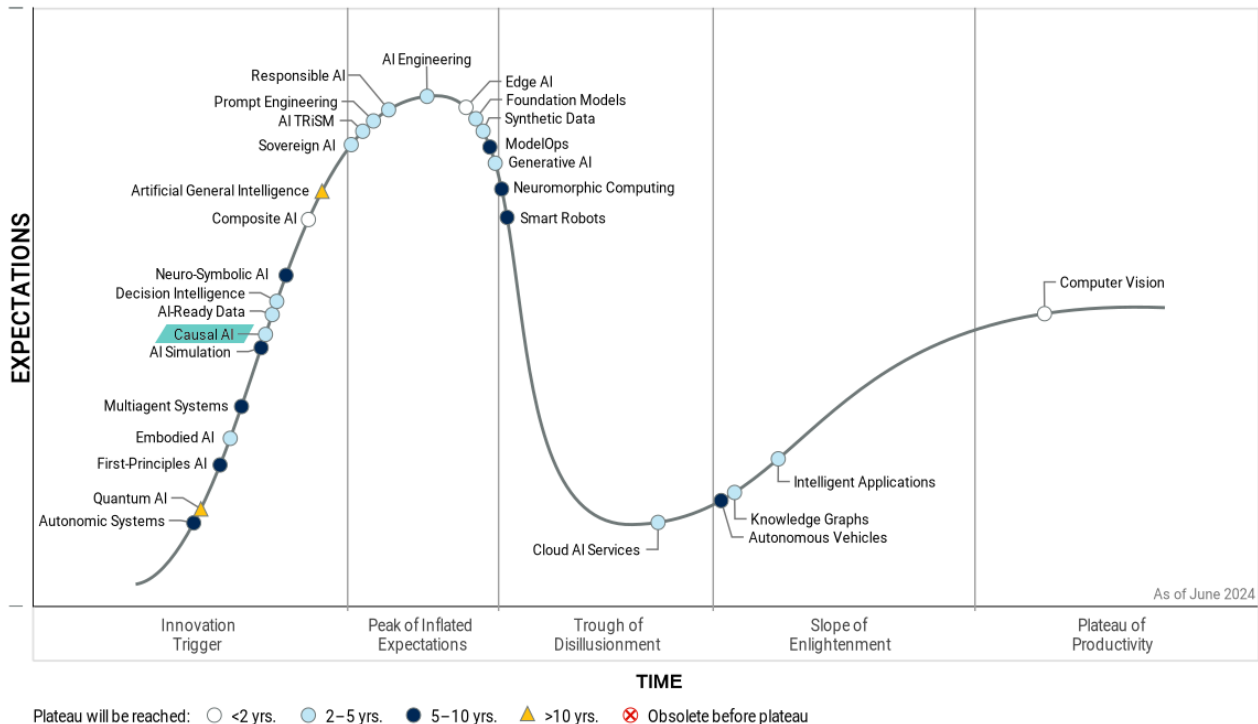
Artificial General Intelligence (AGI)

- AI systems with ability to understand, learn, and apply knowledge across broad range of tasks
- Targets all cognitive tasks, generalize knowledge

Large Language Models

AI Hype Cycle

Hype Cycle for Artificial Intelligence, 2024





Source: <https://xplain-data.de/gartner-ai-hype-cycle-2024/>


Gartner


Large Language Models

Using Local LLMs: OpenWebUI

 New Chat


 Workspace

 Search

OpenAI / GPT 4  +

Set as default

SUCCESS Open WebUI - On a mission to build the best open-source AI user interface.




OpenAI / GPT 4

How can I help you today?

Suggested


Help me study

vocabulary for a college entrance exam

Prompt 


Give me ideas

for what to do with my kids' art

Prompt 

Overcome procrastination



give me tips

Prompt 

Tell me a fun fact

about the Roman Empire

Prompt

+ Send a Message

Timothy J. Baek

LLMs can make mistakes. Verify important information.

