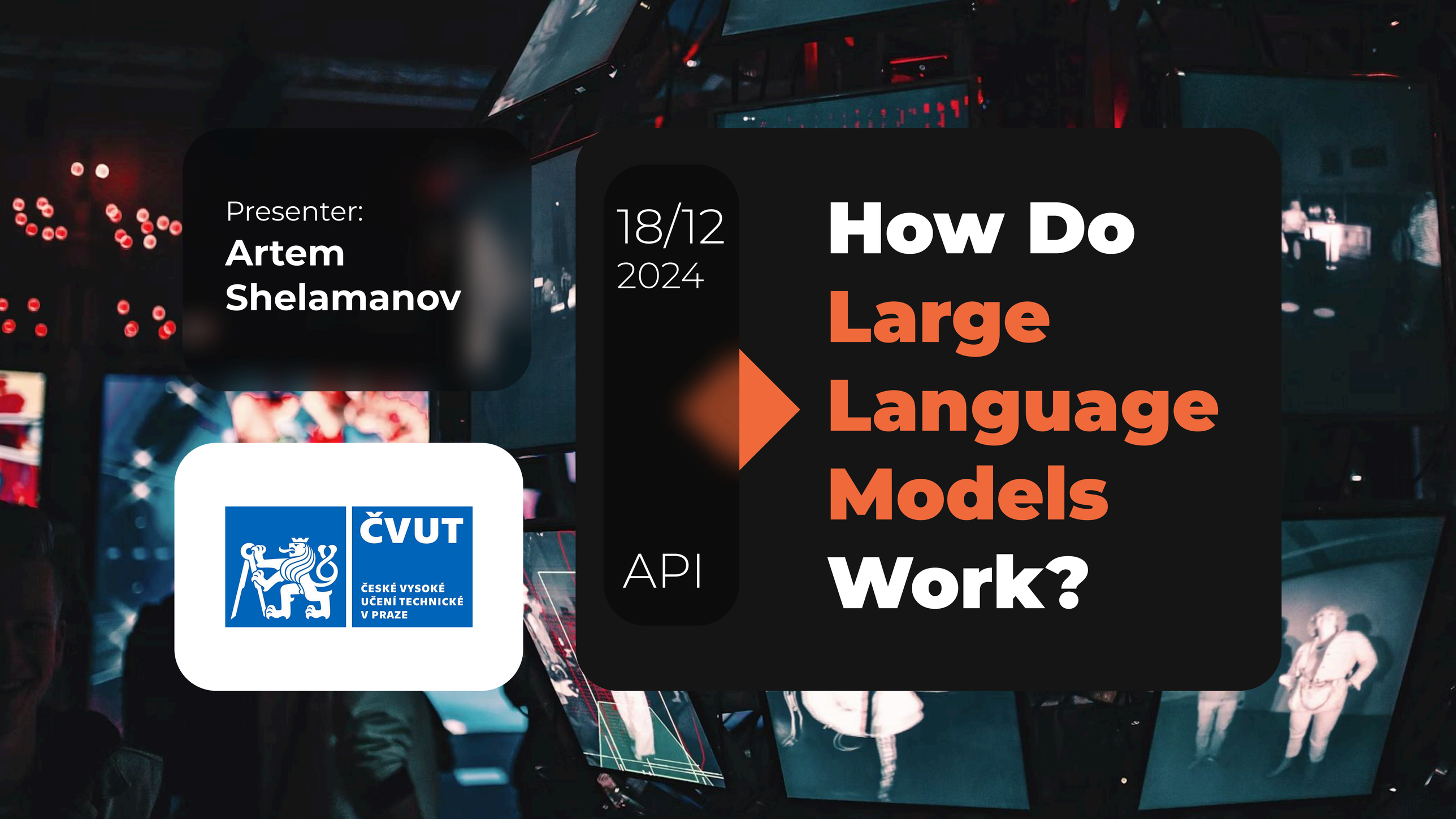


Presenter:
Artem Shelamanov

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

18/12
2024

API

How Do Large Language Models Work?

We will cover today:

- What are LLMs, transformers;
- Technical details about ChatGPT;
- Some practical advice;
- Interesting facts.

Have you ever used ChatGPT?

Raise your hand, if yes; no “AI-shaming” here.

Do you know how ChatGPT works?

What are the main ideas behind it?

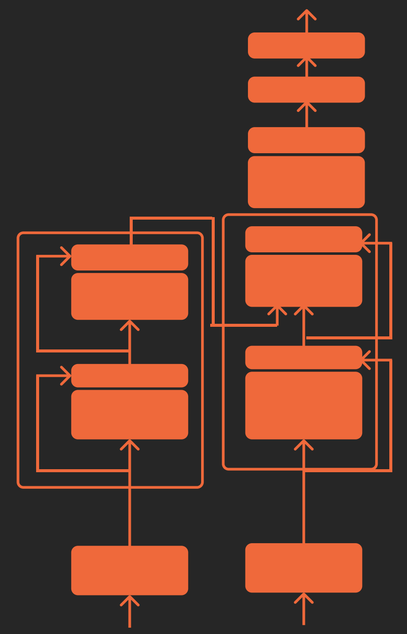
Let’s start from the basics...

ChatGPT
(app/website)

is
using

GPT3/GPT4
LLM

is
based
on



Transformer
Architecture
[1]

What are the inputs and outputs of the model?

You might think...

Input: text
Output: text

It’s not the case!

What are the inputs and outputs of the model?

You might think...

Input: text
Output: text

It’s not the case!

Actual inputs and outputs

Input: sequence of numbers
Output: sequence of numbers

What are the inputs of the model?

Input prompt

“What is AI?”

What are the inputs of the model?

Input prompt

“What is AI?”

-> tokens

[5735,
5,
1231,
752]

What model receives

What are the inputs of the model?

Input prompt

“What is AI?”

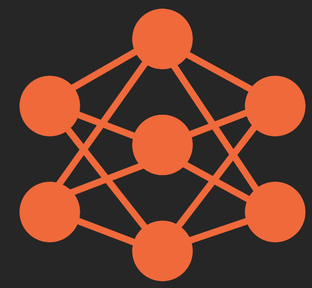
-> tokens

[5735,
5,
1231,
752]

What model receives

-> embeddings -> Net Layers

[[1.31, ... -0.2],
[0.53, ..., 3.1],
...
[54.1, ..., 12.5]]

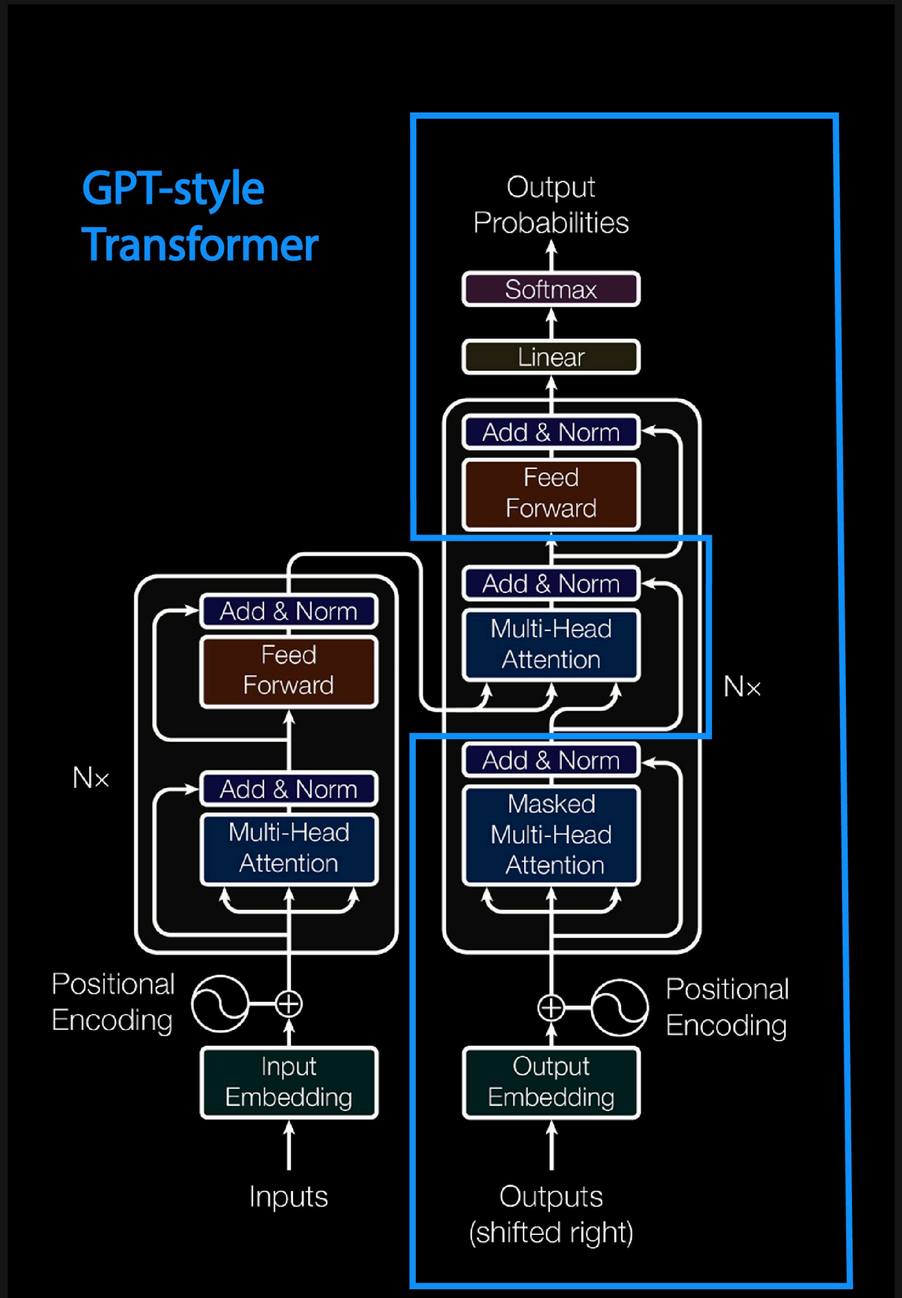


The Transformer

Transformer architecture

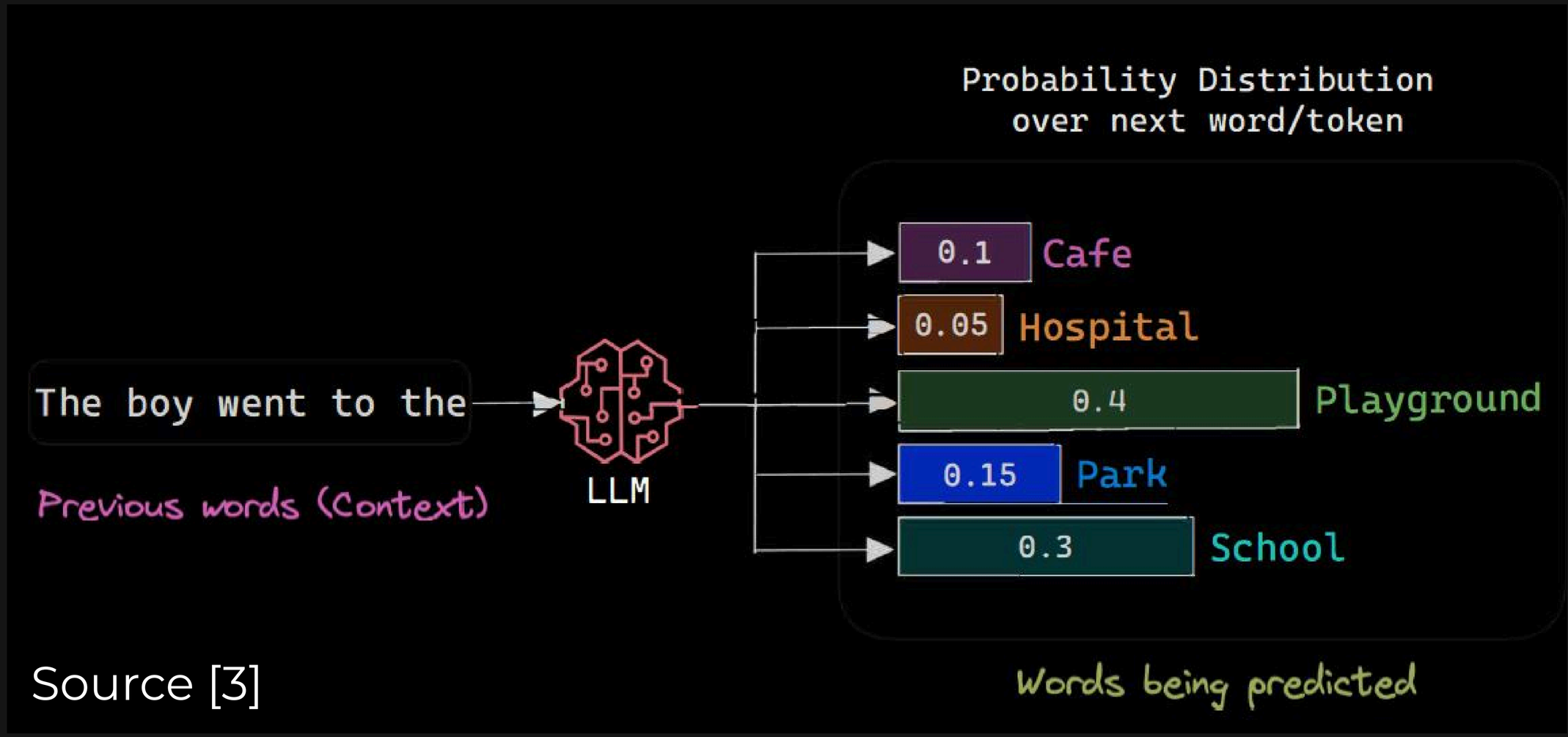
It consists of:

- Encoder or Decoder
- Transformer Blocks
- Attention Mechanism
- Feedforward Neural Networks
- Layer Normalization
- Output Layer



[1] Transformer Architecture

What are the outputs of the model?



Source [3]

Statistics

180 Million

Amount of ChatGPT users, November 2024 [4].

2 billion

visits per month - this was a peak recorded in April 2024 [4].

ChatGPT alternatives?

Commercial:

- Claude
- Gemini
- Grok

Open source:

- Llama
- Mistral
- Phi
- Gemma
- ...

Built on top:

- Perplexity
- CharacterAI
- and many more...

You can run an LLM on your laptop

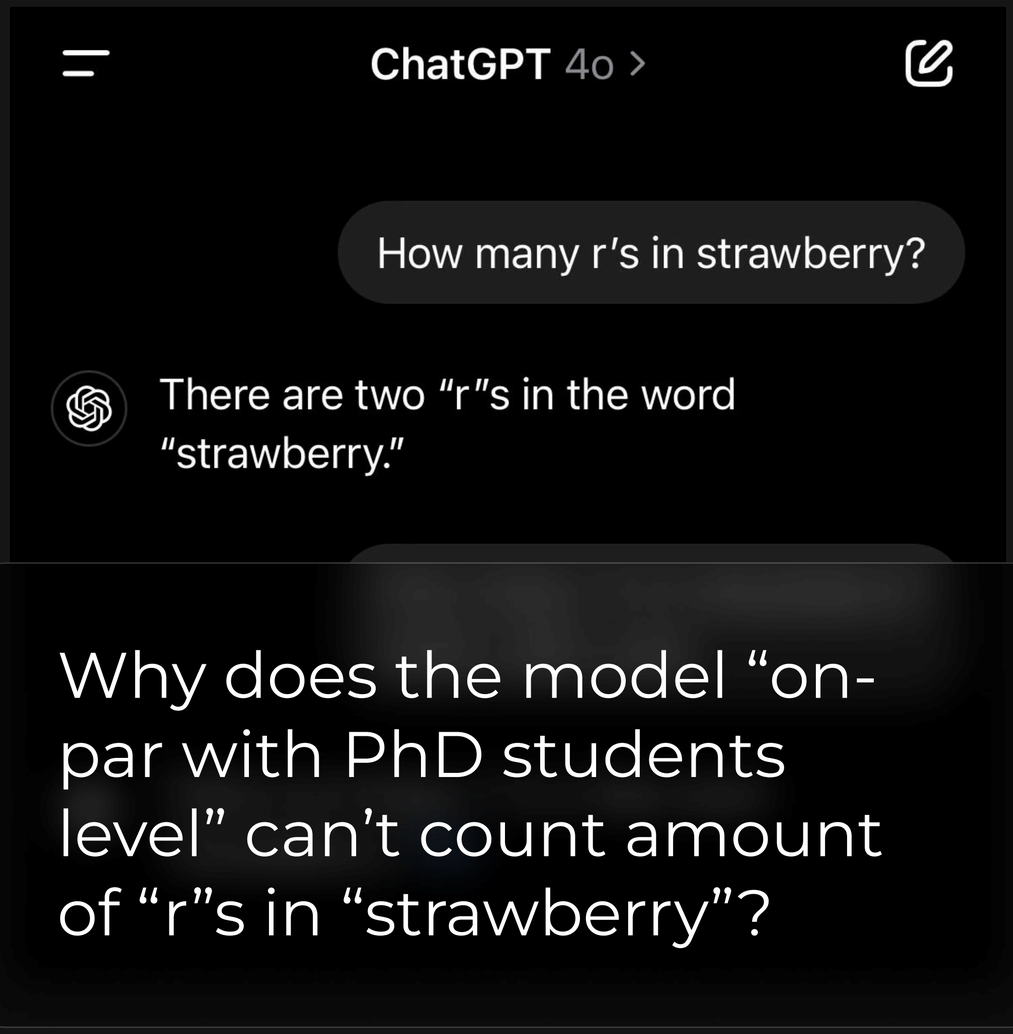
- Free
- Completely private
- Easy to setup
- ≈Can run on CPU

Ollama [2]

Get up and running with large language models.

Run Llama 3.2, Phi 3, Mistral, Gemma 2, and other models. Customize and create your own.

Why is ChatGPT bad at counting?

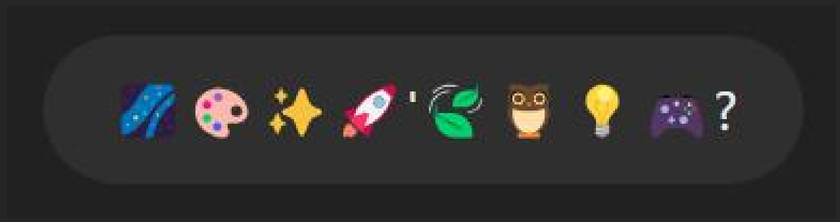


Why is ChatGPT bad at counting?

The reason: tokenization+vector embeddings.

How many r's in strawberry?

What we see



What ChatGPT sees

Do you have any questions?

Thank You!

**“AI will not replace you.
People who use AI will.”**

Handout:
presentation.artem.yazero.io

References

Pictures are from Unsplash under free use license - no attribution is required.

1. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need" Advances in Neural Information Processing Systems. 30. Curran Associates, Inc.;
2. https://ollama.com - availability checked at 11.13.2024;
3. https://www.linkedin.com/pulse/how-do-language-modelsllm-work-we-call-chatgpt-mishra-fdqsc/ - availability checked at 11.14.2024;
4. https://explodingtopics.com/blog/chatgpt-users%C2%A0