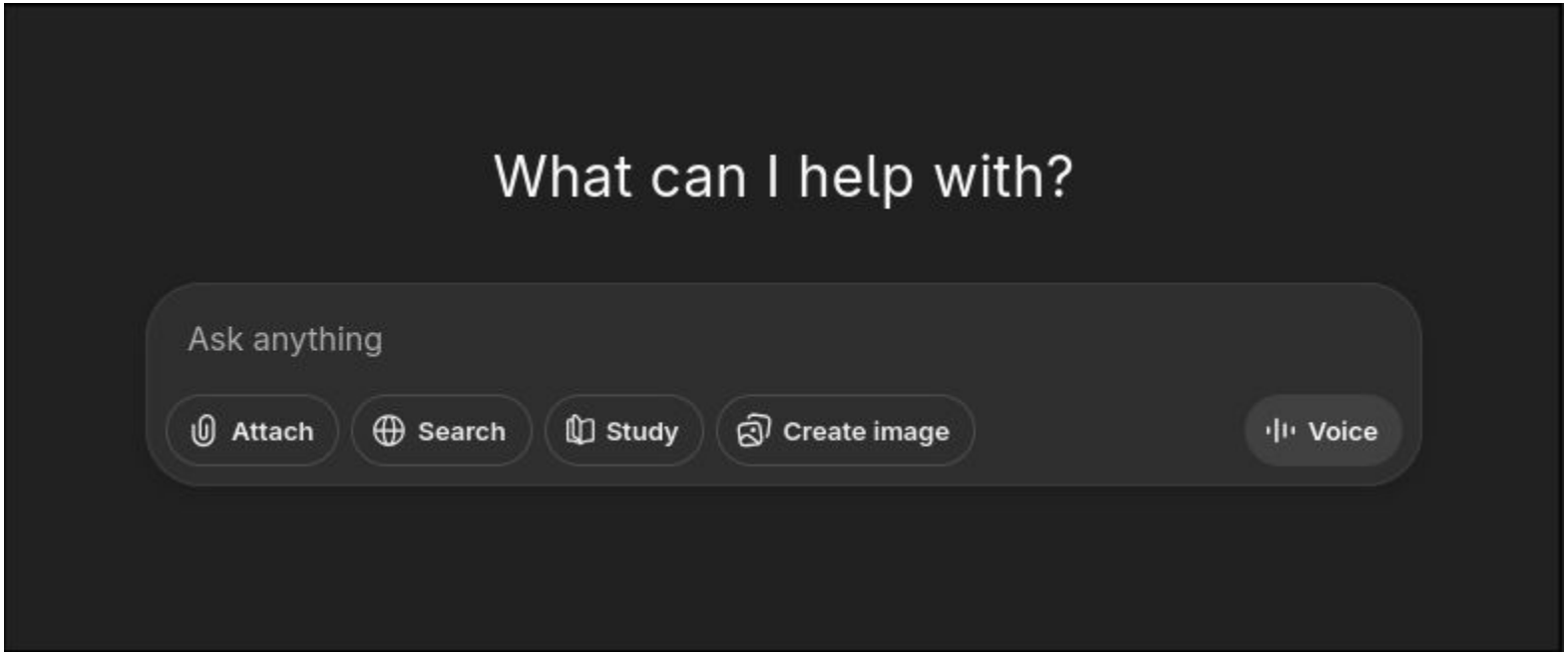
Local Inference

Jacob Matthews

HW00

- Posted on CMS
- Due 2/2 @ 11:59 PM
- It should not take you more than two hours at the absolute most
- Don’t use AI for this

What is your mental model of a chatbot?



What is your mental model of a chatbot/LLM?

- Where does the model “live”?
- How big is it?
- What’s happening on my device?

Frontier models are huge

Let’s say a contemporary frontier model is ~1 trillion (10^12) parameters in size*

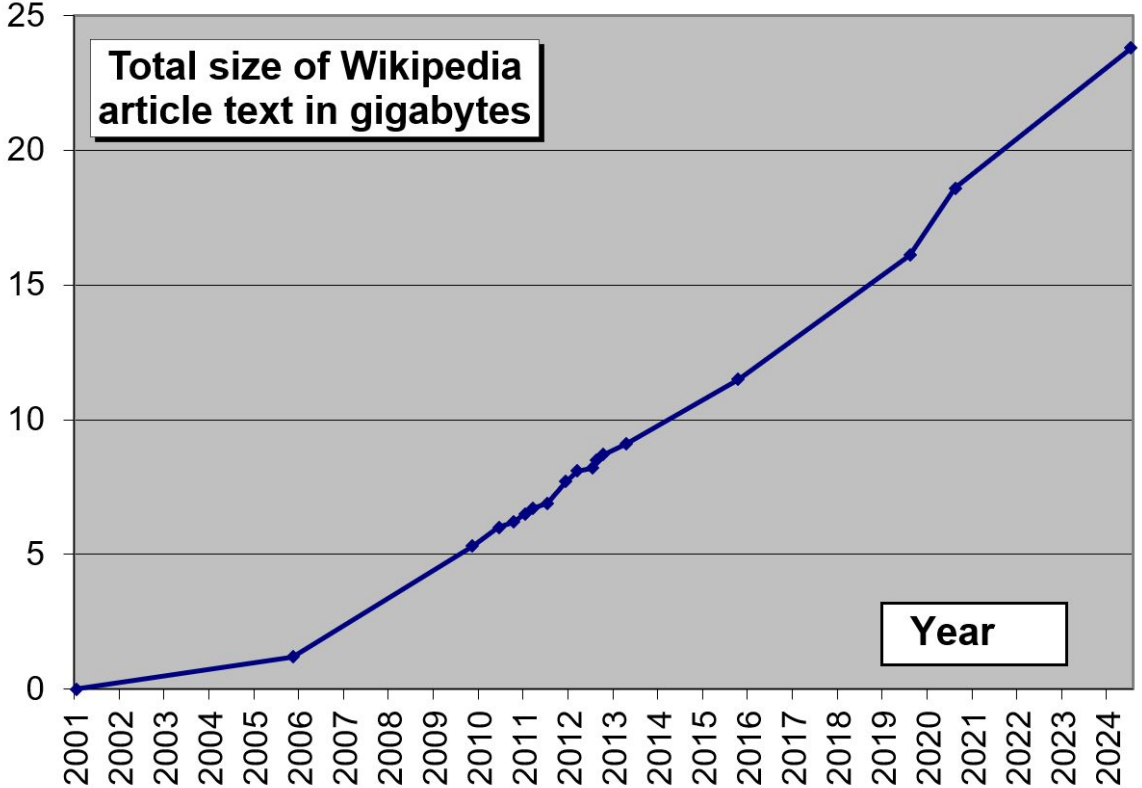
Say each parameter is 16 bits = 2 bytes

2 bytes * 10^12 = **2 TB**

Do you even have that much disk space available (let alone RAM)?

** we’ll talk more about what this means later*

For reference



1 Wikipedia ≅ 12.5B @ FP16

How LLMs “work”

HTTP request

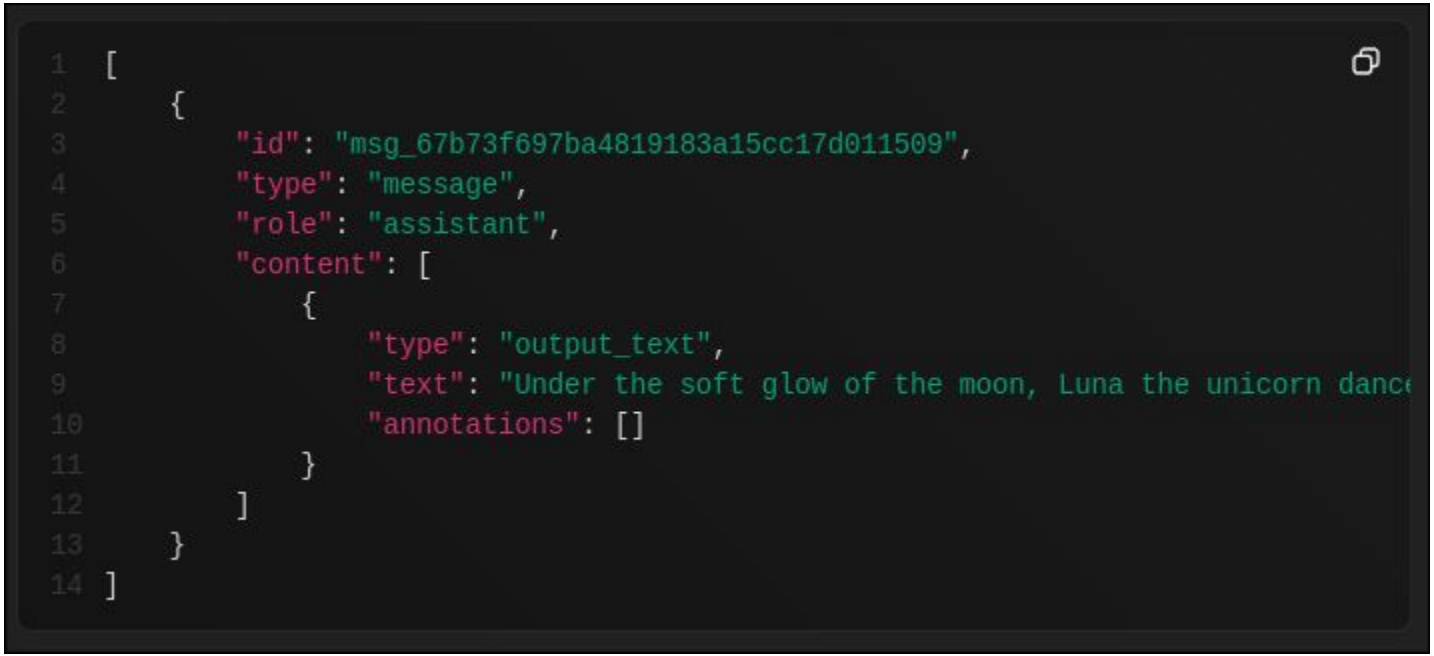
Generate text from a simple prompt

curl

```curl
curl "https://api.openai.com/v1/responses" \
    -H "Content-Type: application/json" \
    -H "Authorization: Bearer $OPENAI_API_KEY" \
    -d '{
        "model": "gpt-5-nano",
        "input": "Write a one-sentence bedtime story about a unicorn."
    }'
```

How LLMs “work”

HTTP response



```
[
    {
        "id": "msg_67b73f697ba4819183a15cc17d011509",
        "type": "message",
        "role": "assistant",
        "content": [
            {
                "type": "output_text",
                "text": "Under the soft glow of the moon, Luna the unicorn danc
                "annotations": []
            }
        ]
    }
]
```

What this means for us

- Large, SOTA models from big name providers (think OpenAI, Google, Anthropic) run entirely on servers (not your machine)
- You pay based on your usage
- You have no (or limited) access to:
  - The actual prompt the model receives
  - Model internals (hidden states, weights)
  - Logits
  - Hyperparameters

What this means for us

- These tradeoffs don’t really matter if
  - You already know how an LLM works
  - You just want to build something without much hassle/math/code
  - You have money for API fees
- In our case,
  - You don’t know how an LLM works (that’s why you’re here!)
  - We want to learn how LLMs work and are ok with some hassle
    - HTTP requests != LLM internals
  - I don’t want you to spend money

What we will be doing

- Running small LLMs locally or on Colab
  - Don’t pay for Colab unless you really want to
  - I am not designing this course to require it
  - If you have a relatively recent Macbook w/ Apple Silicon or similar, you should be fine running things locally

A quick tour of Hugging Face