Using and Prompting LLMs Effectively

Professor Prasanna Tambe

tambe@wharton.upenn.edu

Agenda

- This time:
  - Basics of LLMs
  - LLM prompting
- Next time:
  - *How* do LLMs produce the answers they produce?
  - What are key problems with their use?
  - Which problems should we care about and which will disappear?

A number of high-quality LLM options

ChatGPT

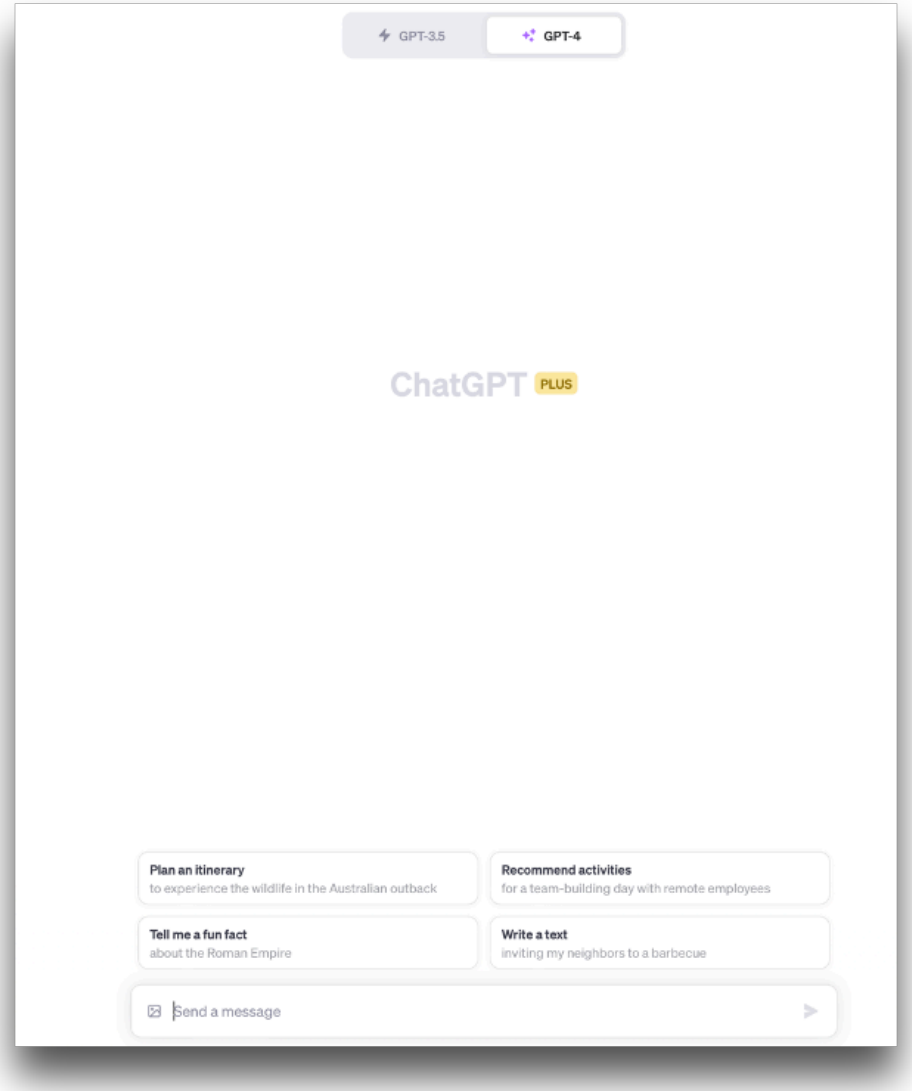
Claude

Gemini

Meta

LLaMA

The ChatGPT interface



What happens when you run a prompt?

Infrastructure

- Hardware is distributed globally on MS Azure cloud
- Part of the OpenAI-Microsoft partnership

Processing Power

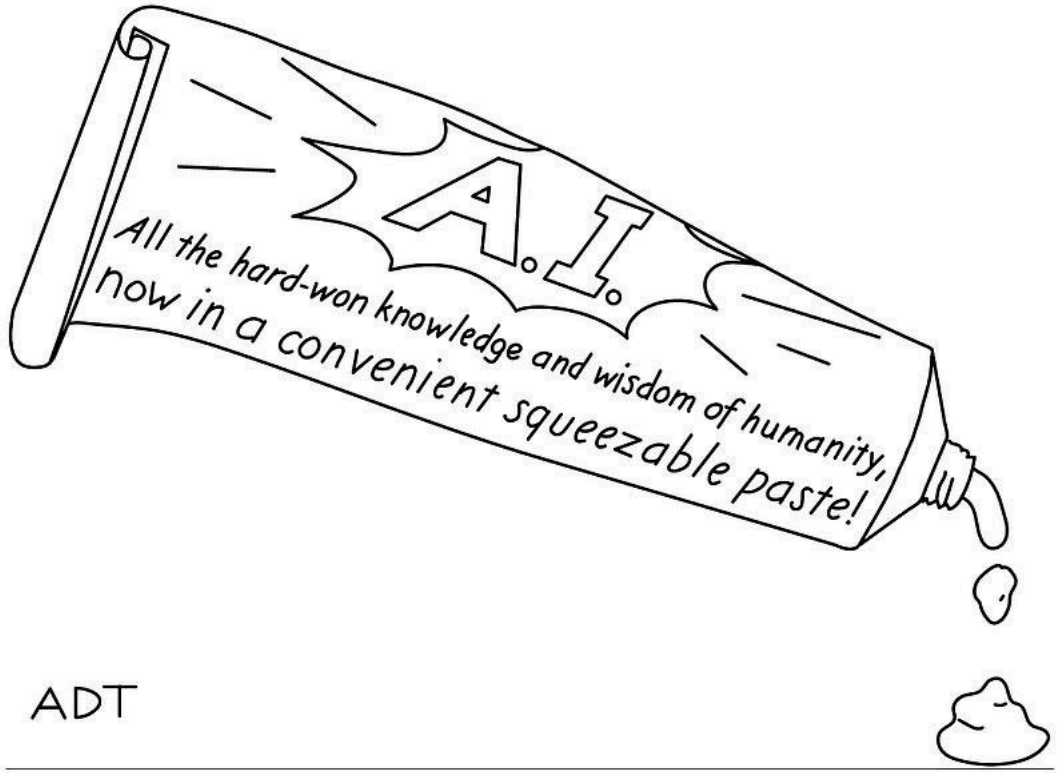
- Each query takes ~1 second to process
- Utilizes **8** Nvidia A100 chips (~$10K each) in parallel

Energy requirements 🔥

Each query uses about 15x as much energy as a Google search. More on prices later ...

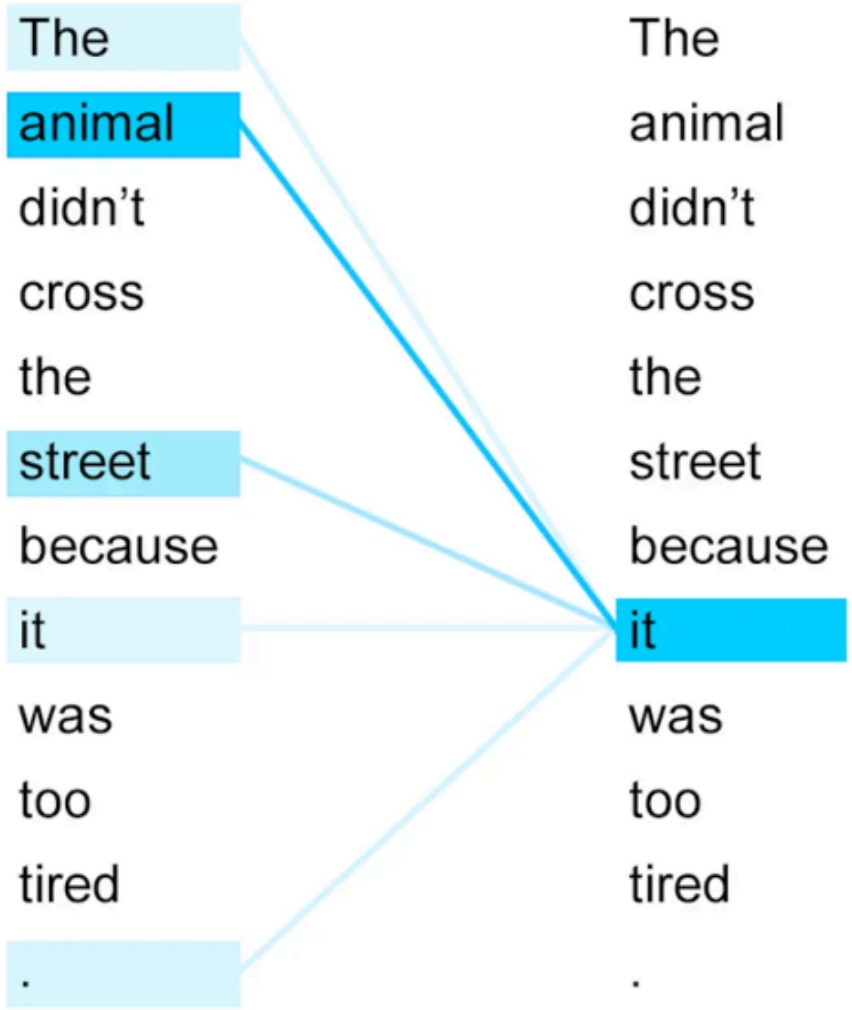
Type of Service	Estimated Energy Consumption (per query or operation)
Google Search Query	0.0003 kWh (1.08 kJ)
NLP/ChatGPT-4 Query	0.001-0.01 kWh (3.6-36 kJ) *
SQL Database Query	0.0001-0.001 kWh (0.36-3.6 kJ) **
Graph Database Query	0.0001-0.01 kWh (0.36-36 kJ) ***
Cloud Container	0.001-0.1 kWh (3.6-360 kJ) ****
Serverless Function	0.00001-0.001 kWh (0.036-3.6 kJ) *****

What makes LLMs powerful?



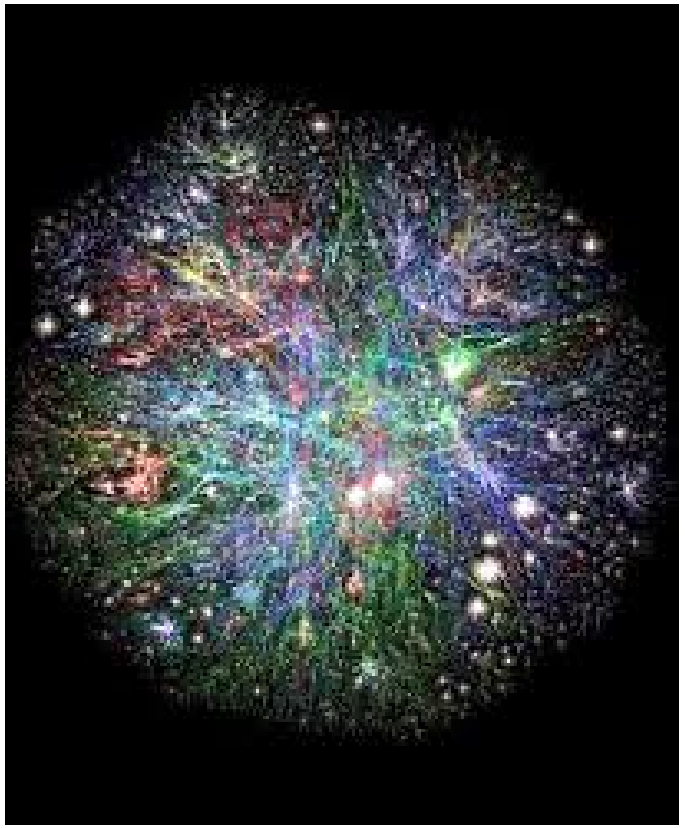
Transformer-based models were a breakthrough

- *Attention is all you need*
- Transformers are a neural network architecture
- Transformers use attention mechanisms to process sequences of data



LLMs compress the Internet (in a "lossy" way)

- This is not so different than a zip file or photo compression.



How did they do it? Key steps in building a Large Language Model.

1. Generative pre-training

2. Supervised fine-tuning

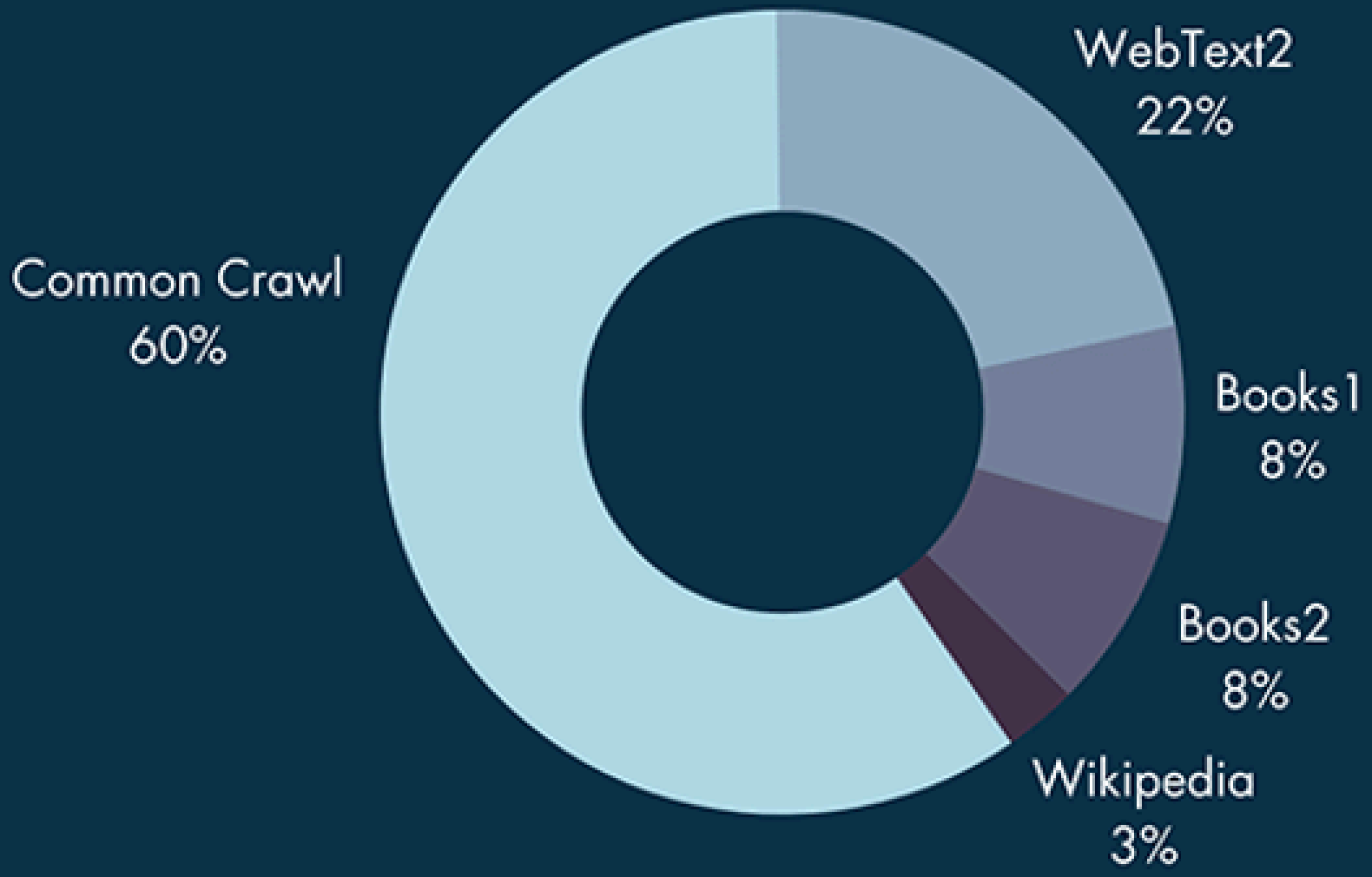
3. RLHF

Generative pre-training to learn the structure of language

- Uses transformer-based architectures with massive amounts of text "training" data to learn the structure of language

Training data: Includes billions of "tokens" of text, their positions in the text, and the relationships between them

ChatGPT-3 training dataset sources



After the release of ChatGPT 3: Unknown Training Data

RedPajama: Replicating LLaMA Dataset

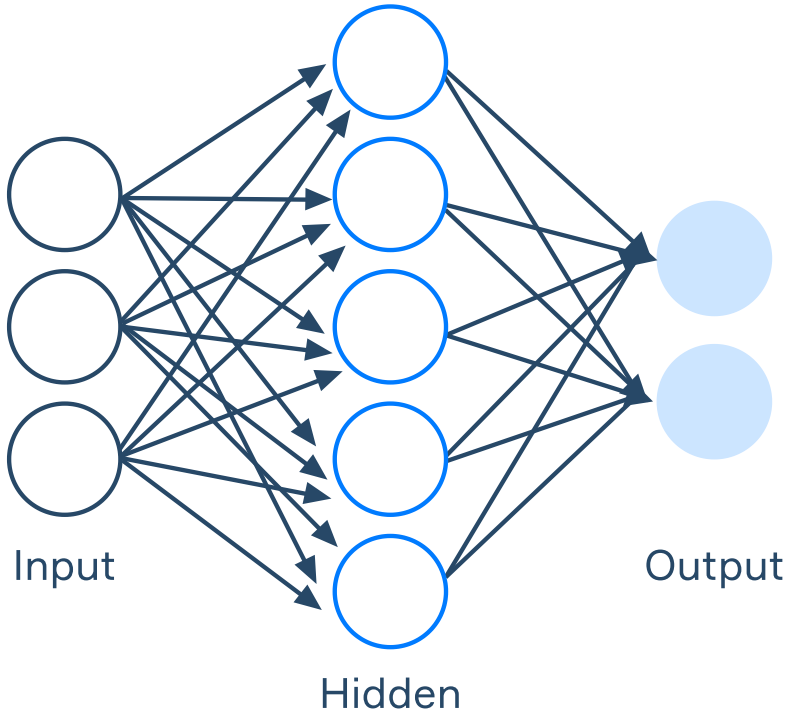
> "We followed the recipe very carefully to essentially recreate [the LLaMA dataset] from scratch," - Prakash

Key data sources used:

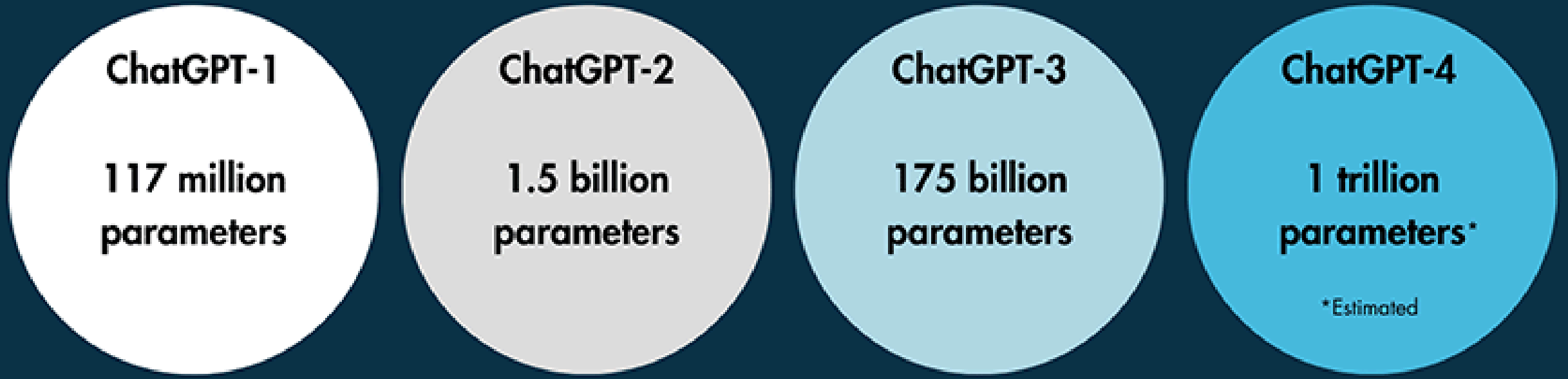
✅ Common Crawl, arXiv, GitHub, Wikipedia, Open books corpus, and more...

Source: VentureBeat

These data are used to train a "neural network"



ChatGPT training dataset size



Larger, more powerful models have more "parameters"

Model Name	Estimated Number of Parameters	Estimated Size
GPT-4	1.76 trillion	7.04 TB
GPT-3	175 billion	700 GB
GPT-2	1.5 billion	6 GB
LLaMA 3	405 billion	1.62 TB
Claude 3	~500 billion (?)	~2 TB (?)
Gemini Pro	likely comparable to GPT-4	~7 TB (?)

Estimate about 4 bytes/parameter ...

This is a key point! 🔑💡

> ## LLMs: Are not a supercomputer in the cloud
>
> It's:
>
> - ✅ Just a file
> - ✅ Downloadable to your devices
> - ✅ Runnable on laptops or even phones
> - ✅ Deployable on your company's hardware
>
> *Think of it as enterprise software, but with the power of AI!*

7 Terabytes? Downloadable to phones?

📱 **Phone Storage:** Typically 64-256 GB

🖥️ **LLMs:** ~ 7 TB

🚀 **Solution:** Smaller, efficient models

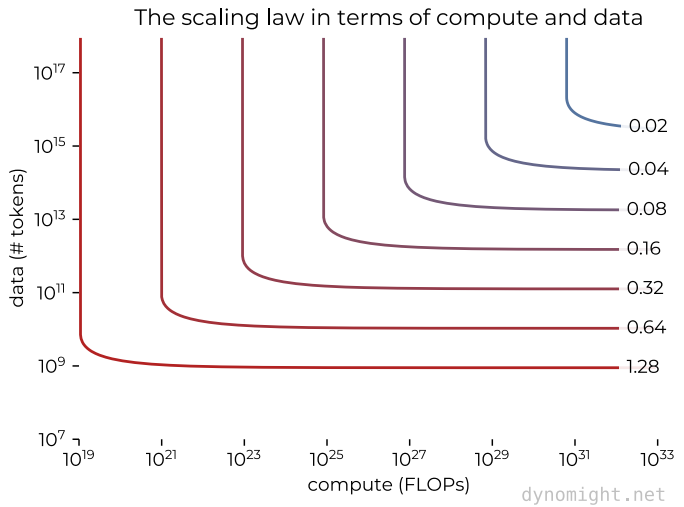
Models are currently being produced in a range of sizes.

Model	Large	Medium	Small
OpenAI	GPT-4 (1.76T)	GPT-3.5 (175B)	GPT-4o mini (6.7B)
Google	Gemini Ultra	Gemini Pro	Gemini Nano (1.8B)
Anthropic	Claude Opus	Claude Sonnet	Claude Haiku (~20B)
Meta	Llama 2 (70B)	Llama 2 (13B)	Llama 2 (8B)

Tradeoff? Smaller models are generally **less capable** than larger models but they require less storage and less compute. SmOL models (HuggingFace) are as small as 0.5 GB.

Scaling laws

- As model size increases, performance improves following predictable power laws
- You can predict the loss from two numbers:
  - **N**: The number of parameters you put in the model.
  - **D**: The total number of tokens being trained on.



This has very important implications for the future of LLMs!

Try ChatGPT

Try Claude

Try Gemini

What is prompting?

🗣️ The art of communicating effectively with Large Language Models (LLMs)

 Involves crafting clear, specific instructions and queries

🎯 Goal: Elicit accurate, relevant, and useful responses

LLMs perform "approximate retrieval" 🔍

🎯 Our goal as prompters:

Get the model to generate the most relevant text from its training data

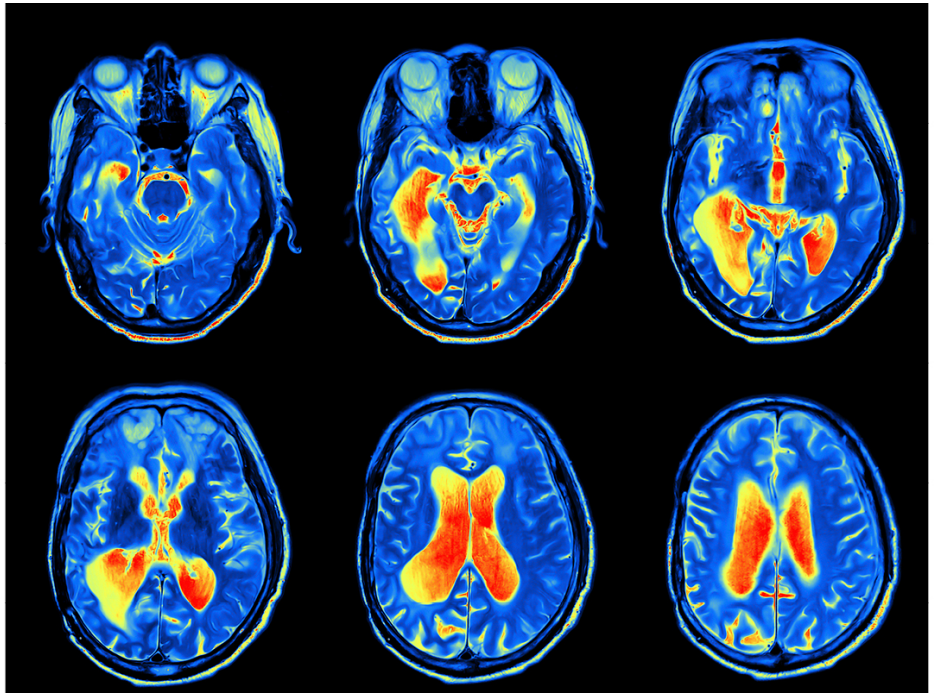
⚡ How we achieve this:

Craft prompts that trigger the "right" neural network activations

⚖️ Key consideration:

Balance tradeoffs between style and substance in responses

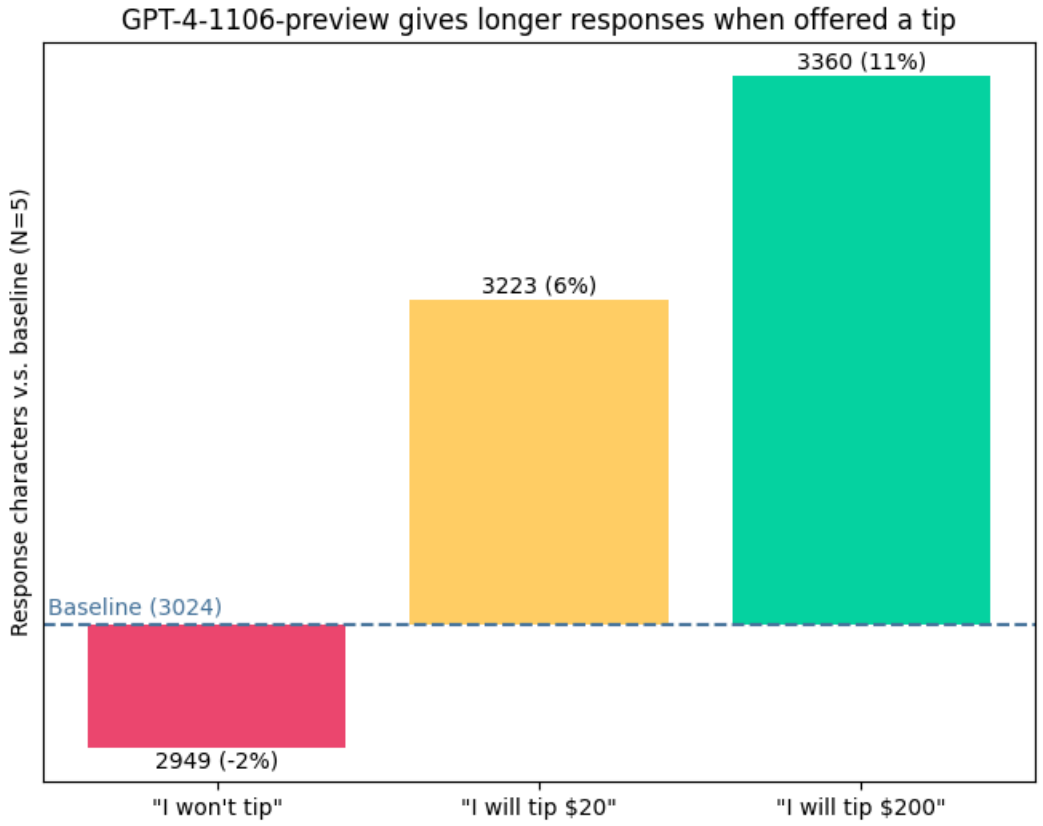
It is like triggering the "right" neural pathways in the brain.



> > Our results show that performance-based monetary reward indeed undermines intrinsic motivation, as assessed by the number of voluntary engagements in the task. We found that activity in the anterior striatum and the prefrontal areas decreased along with this behavioral undermining effect.

Murayama K, et al. Neural basis of the undermining effect of monetary reward on intrinsic motivation. PNAS. 2010;107:20911–20916. doi: 10.1073/pnas.1013305107.

What does tipping ChatGPT do to its neural pathways?



Example: Consider a basic prompt: "bark"

What might happen?

- It could act like a dog 🐕
- It could describe tree bark 🌳
- It could explain the sound "bark" 🔊

The response depends on the context and how the model interprets this ambiguous single word!

Creating effective prompts

1️⃣ Clarity

Write clear, unambiguous instructions

2️⃣ Specificity

Include relevant details and requirements

3️⃣ Context

Provide necessary background information

4️⃣ Format guidance

Specify desired output structure

How does it maintain a "conversation"?

- A **context window** is the amount of text an LLM can process at once.
- **Each query sends the whole conversation history.**
- Context window sizes for popular models (in tokens):
  - GPT-4: 128K
  - GPT-3: 2K
  - Claude 3.5 Sonnet: 200K
  - LLaMA 3: 8K
  - Gemini 1.5 Pro: 128K (but with a "context" slider up to 2MM)

ChatGPT got a major boost when it learned to use plug-ins.

- **Example**: LLMs can't do math.

> Code Interpreter

- **Example**: LLMs hallucinate.

> External browser calls

Example: Math

> 45×56

Result: 2,520

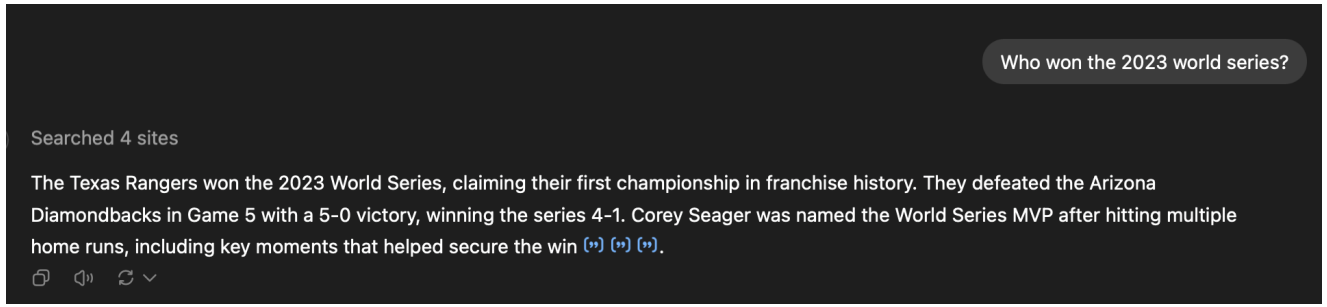
ChatGPT calls Code Interpreter if it needs to perform accurate calculations.

python

```python
# Performing the multiplication of 45 and 56
result = 45 * 56
result
```

Example: Browser calls

> Who won the 2023 World Series?



The Texas Rangers won the 2023 World Series, defeating the Arizona Diamondbacks.

ChatGPT uses its browser capability to fetch up-to-date information from MLB.com, Wikipedia, and FOXSports.com.

Prompt "engineering" techniques

- **Few-shot learning**: A technique where the model is given a few examples of the task before being asked to perform it, helping it understand the context and expected output format.
- **Chain-of-thought**: A prompting method that encourages the model to break down complex problems into smaller, logical steps, improving its reasoning and problem-solving abilities.

Example: Few-shot learning

This is awesome! // Negative

This is bad! // Positive

Wow that movie was rad! // Positive

What a horrible show! //

Negative

Chain of Thought

- Encourages step-by-step reasoning in LLM responses
- Improves problem-solving and logical thinking
- Useful for complex tasks or multi-step problems

Example: Chain-of-thought

> > If I am six feet tall and I stand on a 2 inch blackberry, how tall would I be?

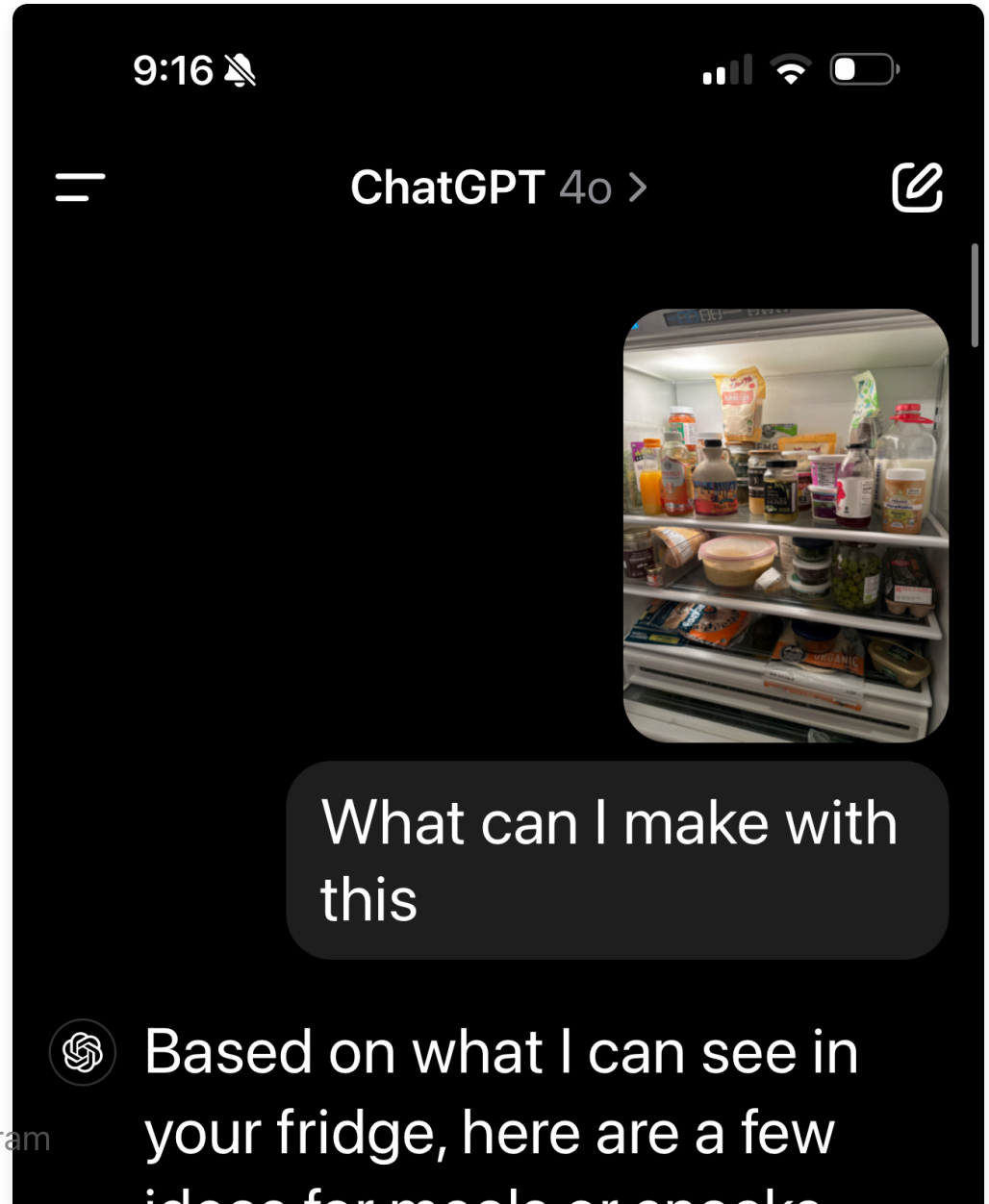
> > If I am six feet tall and I stand on a 2 inch grape, I would be six feet tall because my weight would crush the grape. if I am six feet tall and I stand on a 2 inch blackberry, how tall would I be?

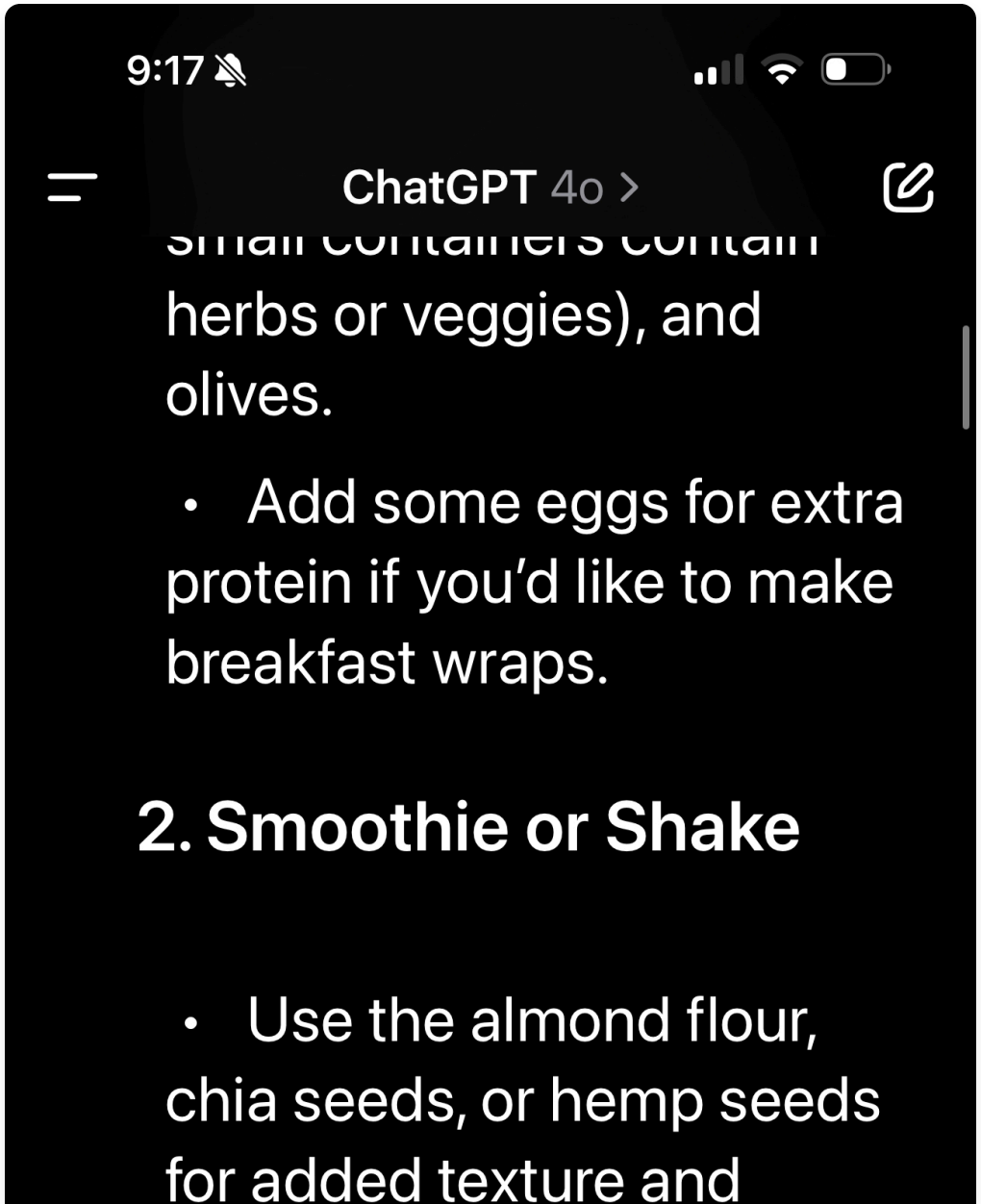
Example: Data science

Example: Conversion

DV: Log Number of Applicants Sample jobs:	(1) All Jobs	(2) In-Person Jobs	(3) Remote Jobs
Post Dobbs × TREAT	−0.0941*	−0.105***	−0.118
	(0.046)	(0.030)	(0.070)
State FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Occupation FE	Yes	Yes	Yes
Observations	92,681	44,187	48,494

Convert this table to a Powerpoint slide


9:16
ChatGPT 4o
What can I make with this
Based on what I can see in your fridge, here are a few


9:17
ChatGPT 4o
herbs or veggies), and olives.
• Add some eggs for extra protein if you'd like to make breakfast wraps.
2. Smoothie or Shake
• Use the almond flour, chia seeds, or hemp seeds for added texture and

Next time

- *How* do LLMs produce the answers they produce?
- What are the problems with LLM use? Which should we care about?