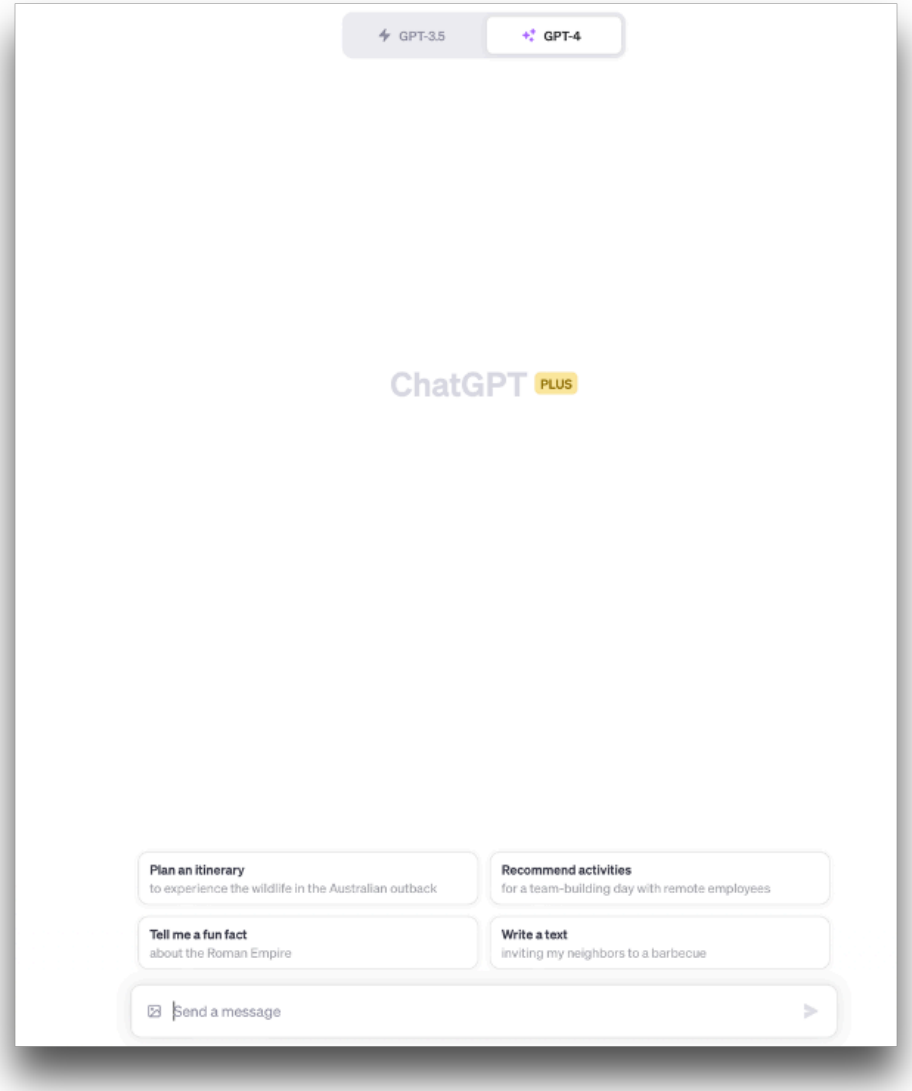
LLM Overview

Professor Tambe

tambe@wharton.upenn.edu

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

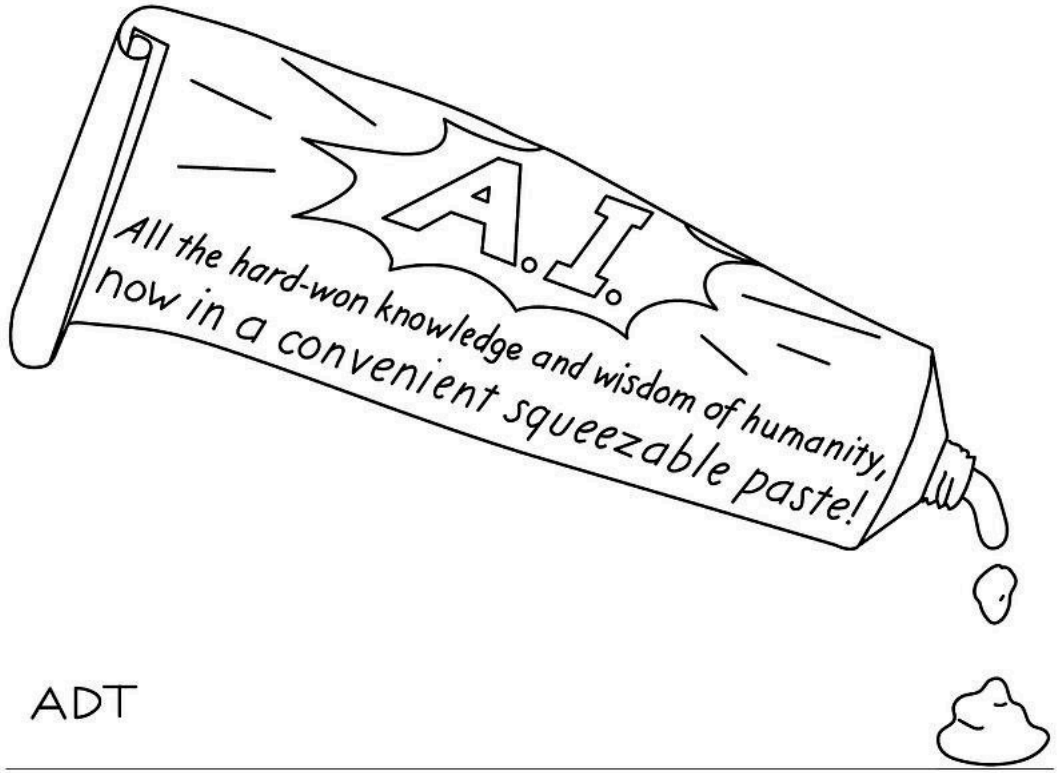
- Each query takes ~1 second to process
- Utilizes **8** Nvidia A100 chips (~$10K each) in parallel

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?

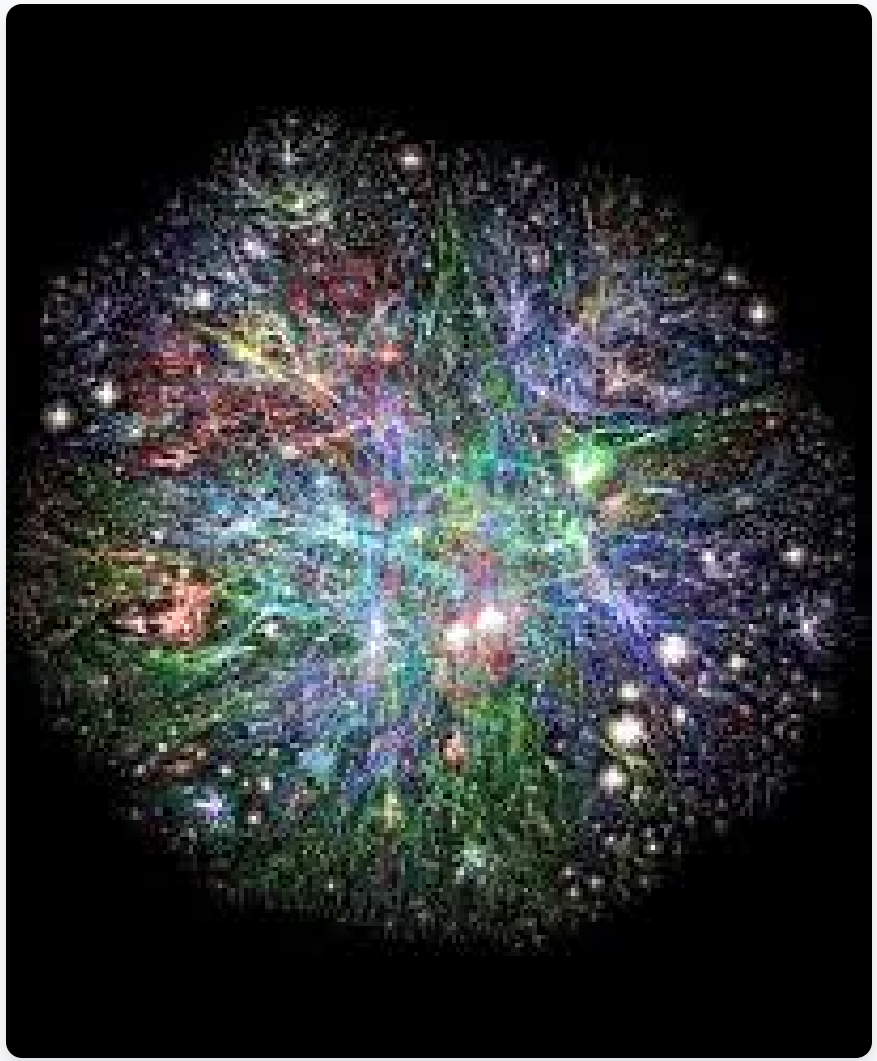


💾 Like data compression:

Similar to how zip files and photo compression work

🌐 But for the web:

Compressing internet knowledge into neural networks



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: Learning Language Structure

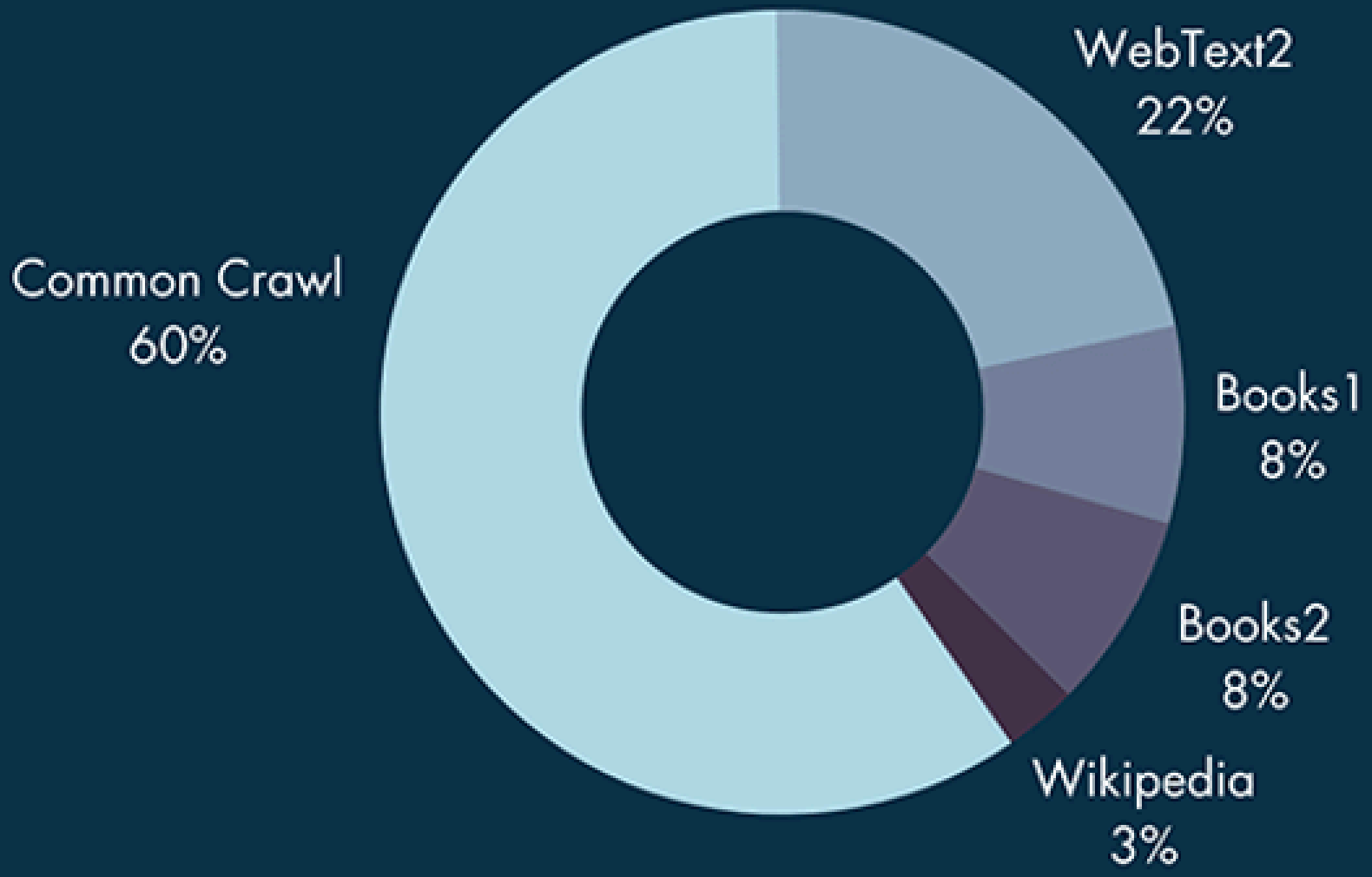
How it Works

Uses transformer-based architectures with massive amounts of text data to learn patterns and structure of human language

Training Data

- Billions of text tokens
- Positional information
- Contextual relationships

ChatGPT-3 training dataset sources



The dangers of model collapse

A.I. images generated by Sina Alemohammad and others.

Before AI generated data

The dangers of model collapse

A.I. images generated by Sina Alemohammad and others.

Before AI generated data

A.I. images generated by Sina Alemohammad and others.

After AI generated data

After the release of ChatGPT 3: Unknown Training Data

RedPajama: Replicating LLaMA Dataset

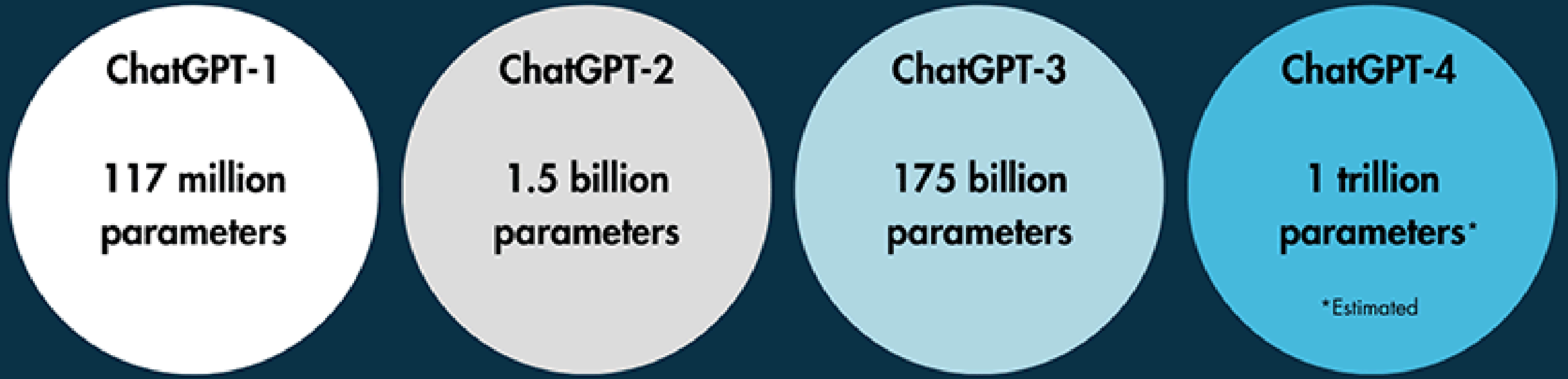
> "We followed the recipe very carefully to essentially recreate [the LLaMA dataset] from scratch," - Prakash

Key data sources used:

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

Source: VentureBeat

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 ...

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)

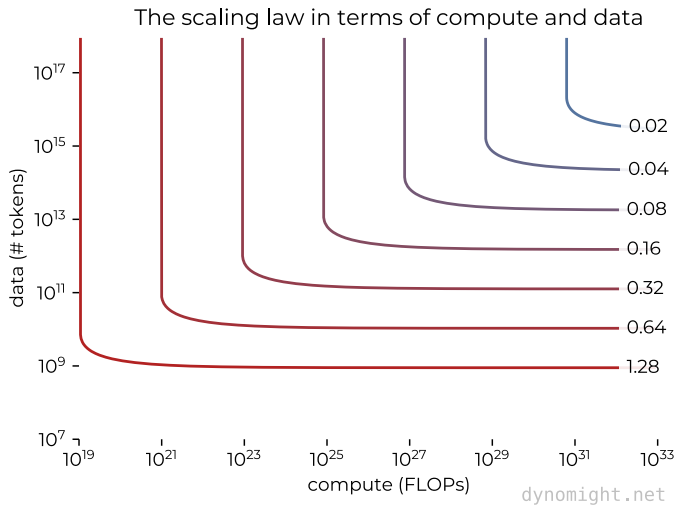
Tradeoff? Smaller models are generally **less capable** than larger models but they require less storage and less compute. SmOL models (HuggingFace) are as small as 0.5 GB.

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!*

Open question: LLM 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 important implications for the future of LLMs!

The relationship between model size, training data, and performance will shape how these systems evolve

🏠 Back to Course Materials