

# CS 6501 Natural Language Processing

## LLM Overview

Yangfeng Ji

Information and Language Processing Lab

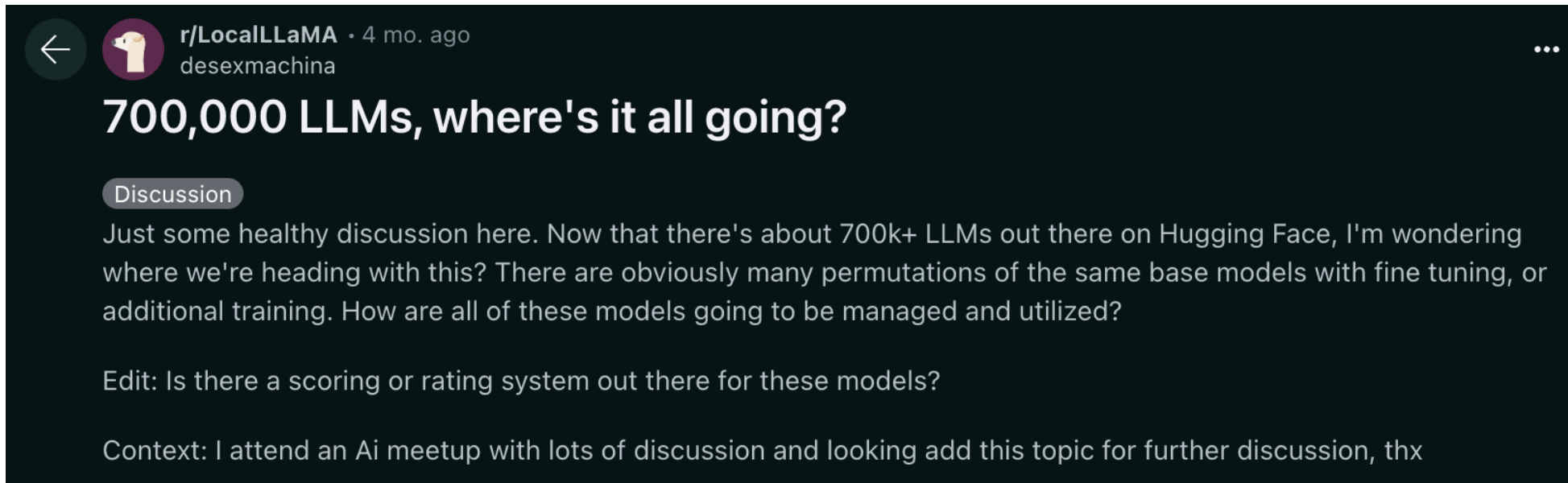
Department of Computer Science


University of Virginia

<https://uvanlp.org/>

# How Many LLMs?

- More than 1M models
  - Based on the Hugging Face website
  - Most of the models are fine-tuned with existing LLMs
- A few months ago



←  r/LocalLLaMA · 4 mo. ago  
desexmachina

## 700,000 LLMs, where's it all going?

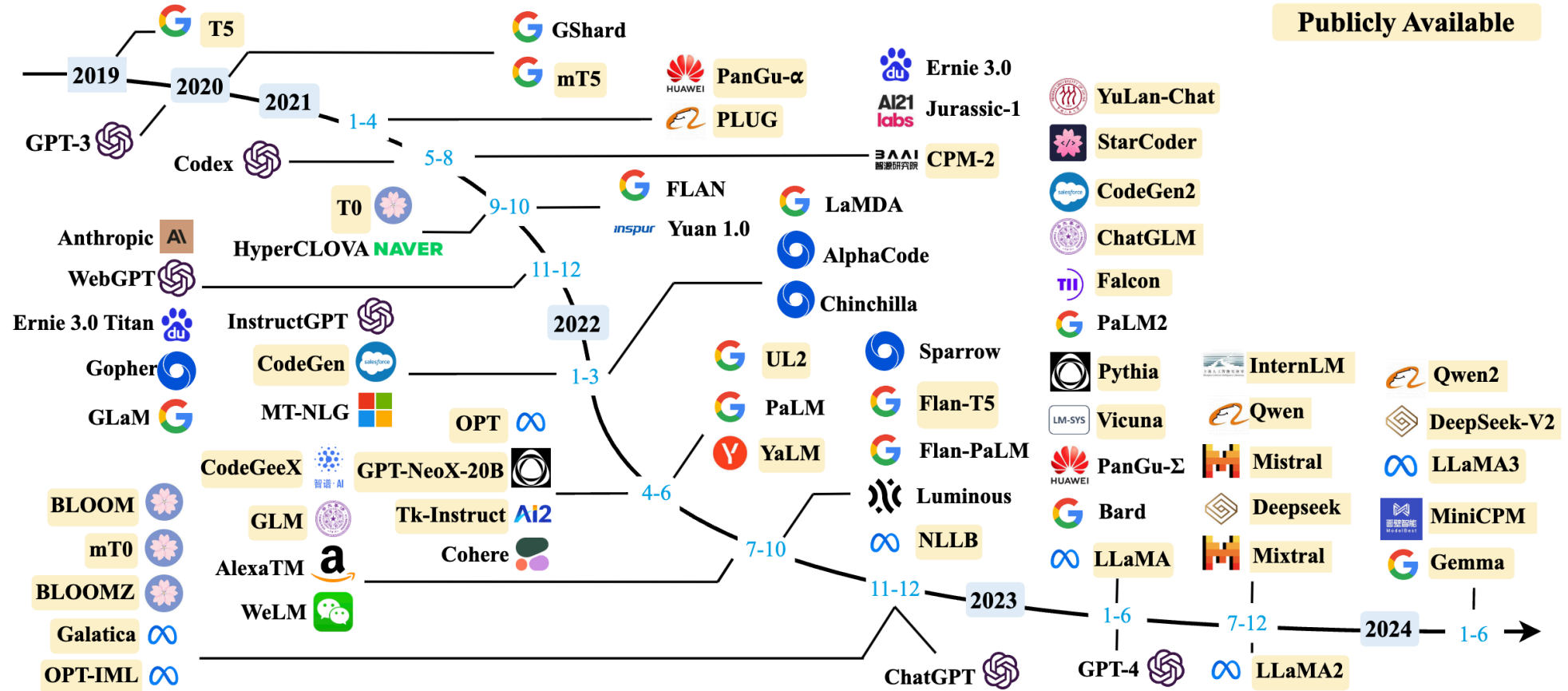
Discussion

Just some healthy discussion here. Now that there's about 700k+ LLMs out there on Hugging Face, I'm wondering where we're heading with this? There are obviously many permutations of the same base models with fine tuning, or additional training. How are all of these models going to be managed and utilized?

Edit: Is there a scoring or rating system out there for these models?

Context: I attend an Ai meetup with lots of discussion and looking add this topic for further discussion, thx

# A Timeline







Zhao et al., 2024

# Categorizing LLMs (I): Model Family

There are different dimensions of categorizing LLMs.

For example, by model family,  
**GPT family** (from OpenAI)

Kalyan 2023

 GPT-3	ada <sup>Raw</sup> text-ada <sup>SFT</sup>	babbage <sup>Raw</sup> text-babbage <sup>SFT</sup>	curie <sup>Raw</sup> text-curie <sup>SFT</sup>	davinci <sup>Raw</sup> text-davinci <sup>SFT</sup>
 GPT-3.5	code-davinci-002 <sup>Code</sup>	text-davinci-002 <sup>SFT</sup>	text-davinci-003 <sup>RLHF</sup>	
 ChatGPT	gpt-3.5-turbo <sup>RLHF-Chat</sup>	gpt-3.5-turbo-16k <sup>RLHF-Chat</sup>		
 GPT-4	gpt-4 <sup>RLHF-Chat</sup>	gpt-4-32k <sup>RLHF-Chat</sup>		

# Model Family: Claude

Claude family (from Anthropic)

## Claude 3.5 Sonnet

Enhanced reasoning, state-of-the-art coding skills, computer use, and 200K context window

Try Claude

Get API access

### Announcements

**NEW** Claude 3.5 Haiku and a new Claude 3.5 Sonnet Oct 22, 2024

Our updated version of Claude 3.5 Sonnet is now state-of-the-art for real-world software engineering tasks, agentic capabilities, and computer use in public beta.

[Read more](#)

**NEW** Developing a computer use model Oct 22, 2024

A discussion of how our researchers developed Claude's new computer use skills, as well as some of the relevant safety considerations.

✦ GPT o1 and Claude 3.5 Sonnet are both large language models, but they have some key differences:

**GPT o1:**

- Developed by OpenAI
- Known for its ability to generate code and perform complex reasoning tasks
- Can be more verbose and less concise in its responses
- May require more specific prompts to get the desired output
- More expensive to use

**Claude 3.5 Sonnet:**

- Developed by Anthropic
- Strong in coding tasks and reasoning abilities
- More cost-effective (4x cheaper than GPT o1)
- Can be more thoughtful and faster than GPT o1
- May be more stubborn and require adjustments to the system prompt or more concrete examples to change its output

# Model Family: Gemini

Belong to the same model family

- Bard
- PaLM
- LaMDA

# Open-source Model Family

- [Llama](#) (from Meta AI)
- [Pythia](#) (from Eleuther AI)
- [Falcon](#) (from Technology Innovation Institute, Abu Dhabi)
- [Mistral](#) (from Mistral AI)
- [OLMo](#) (from AI2)

# Categorizing LLMs (II): Model Size

By model size

- $< 1B$  parameters (e.g., OPT-350M, Pythia-160M)
  - Research toys, domain-specific, embedding
- Between  $1B$  and  $10B$  parameters (e.g., Llama3 8B)
  - Storytelling, writing, code generation, speech
- Between  $10B$  and  $100B$  parameters (e.g., Llama3 80B)
  - Reasoning, planning, world knowledge
- More than  $100B$  parameters (e.g., Llama-3.1 405B)
  - Comparable performance to GPT 4 (on some benchmarks)



# Categorizing LLMs (III): Accessibility

How can we access an LLM model?

- Closed-source models (via APIs), e.g.,
  - GPT-3, 4 (OpenAI)
  - Claude 3.5 (Anthropic)
- Open-source models, e.g.,
  - the Llama family
  - GPT-1, GPT-2
  - the OLMo family
  - the Mistral family

# Categorizing LLMs (IV): Data Types

What types of data they can process?

- Text only
  - Most of the open-source LLMs
- Code
  - Code Llama
- Multi-modal
  - LLama 3.2 11B and 90B (Open-sourced)
  - GPT-4 (Closed-source)

# How to Choose a Model?

- Research vs. Applications
- Performance vs. Efficiency
- Cost vs. Infrastructure
- Privacy (and many other concerns)

# Research vs. Applications

- Research: for research, highly recommend open-source models
- Application: closed-source models are easier to use

# Performance vs. Efficiency

- Performance
  - Open-source models are getting better and better
- It's better to find a trade-off between performance and efficiency
  - For example, working on embedded systems

# Cost vs. Infrastructure

- Hosting a GPU server requires non-trivial cost
- In many cases, money spent on closed-source model APIs is also considerably large

# Privacy and Other Concerns

- Not be able to share data to the public domain
- Identifying model biases is more like observational study, but mitigating model bias requires deeper understanding of model behaviors

**Thank You!**