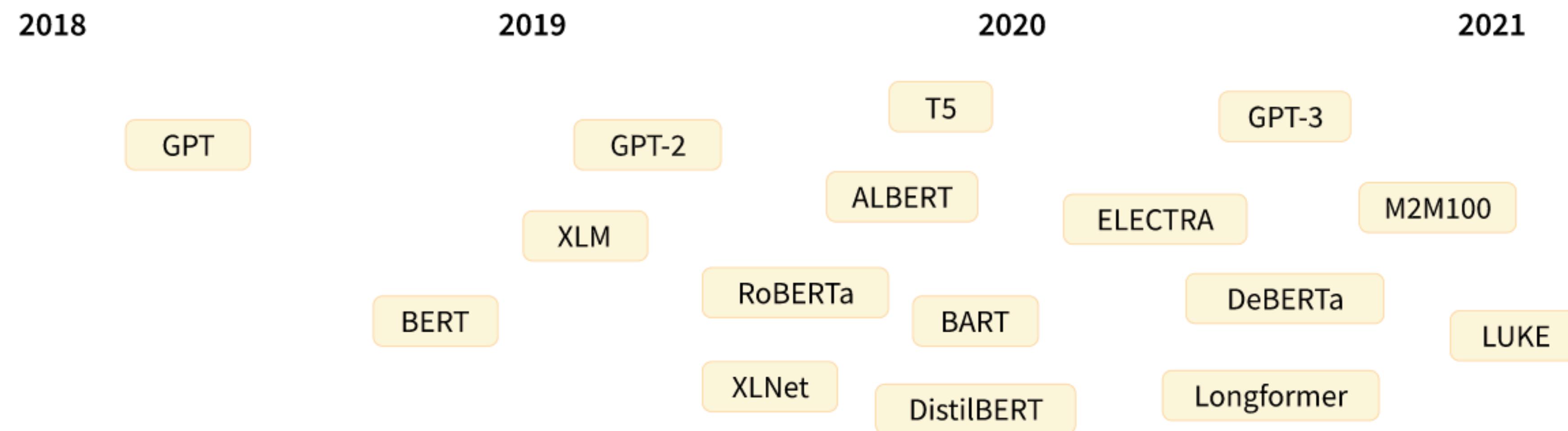


12-11-2024

Large language model (LLM)

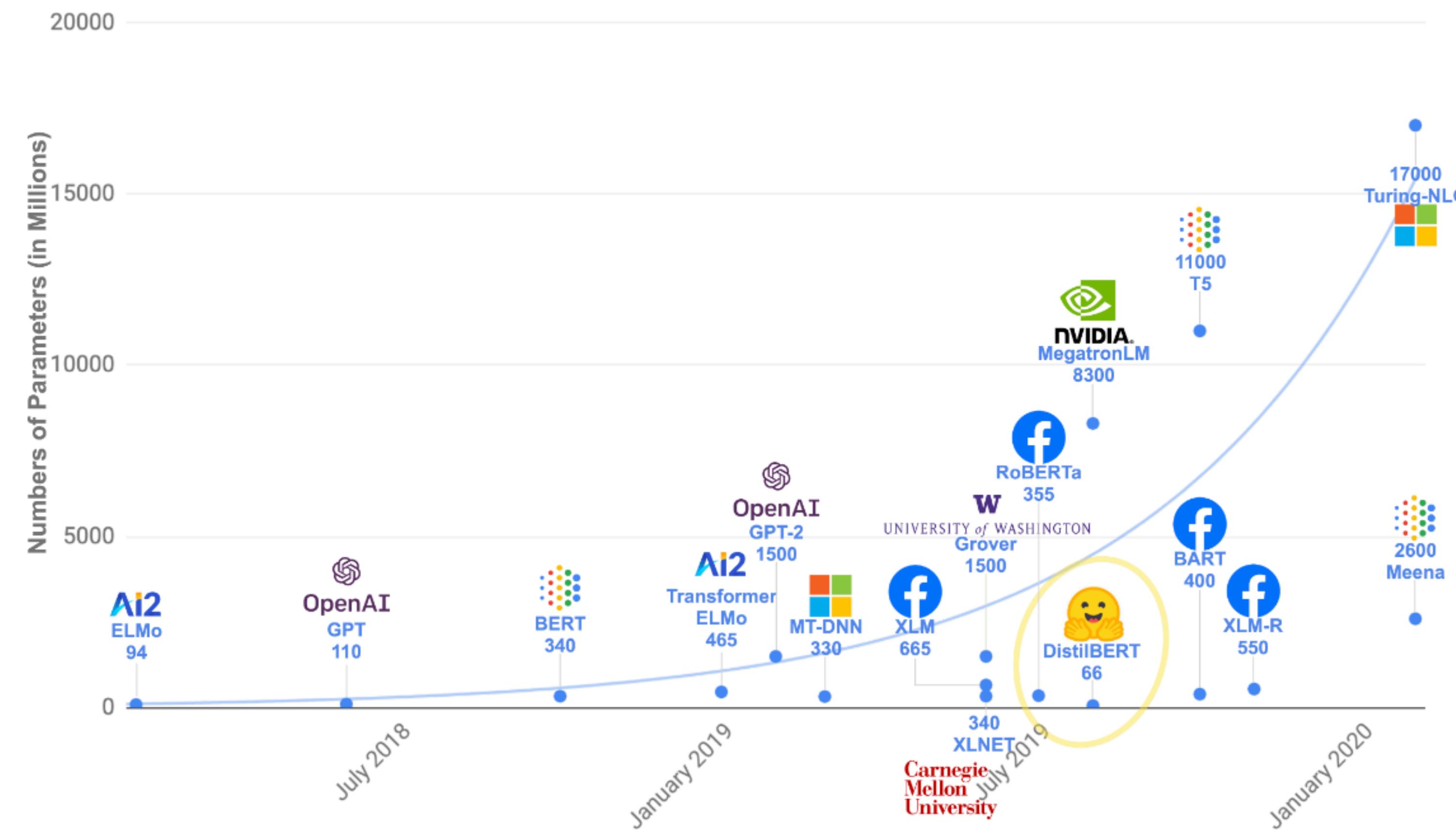
- History



LLM (cont...)

- #parameters

* GPT-3 (175 B), OpenAI



LLM (cont...)

Model	Provider	Open-Source	Speed	Quality	Params	FINE-TUNEABILITY
gpt-4	OpenAI	No	★★★	★★★★	-	No
gpt-3.5-turbo	OpenAI	No	★★★	★★★★	175B	No
gpt-3	OpenAI	No	★★★	★★★★	175B	No
ada, babbage, curie	OpenAI	No	★★★	★★★★	350M - 7B	Yes
claude	Anthropic	Yes	★★★	★★★★	52B	No
claude-instant	Anthropic	Yes	★★★	★★★★	52B	No
command-xlarge	Cohere	No	★★★	★★★★	50B	Yes
command-medium	Cohere	No	★★★	★★★★	6B	Yes
BERT	Google	Yes	★★★	★★★★	345M	Yes
T5	Google	Yes	★★★	★★★★	11B	Yes
PaLM	Google	Yes	★★★	★★★★	540B	Yes
LLaMA	Meta AI	Yes	★★★	★★★★	65B	Yes
CTRL	Salesforce	Yes	★★★	★★★★	1.6B	Yes
Dolly 2.0	Databricks	Yes	★★★	★★★★	12B	Yes

Large language model Meta AI (LlaMa)

LLaMA: Open and Efficient Foundation Language Models

**Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample***

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community¹.

1 Introduction

Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a few examples (Brown et al., 2020). These few-shot properties first appeared when scaling models to a

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance compared to the best existing LLMs. For instance, LLaMA-13B outperforms GPT-3 on most benchmarks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU. At the higher-end of the scale, our 65B-parameter model is also competitive with the best large lan-

LlaMa-2

LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone†

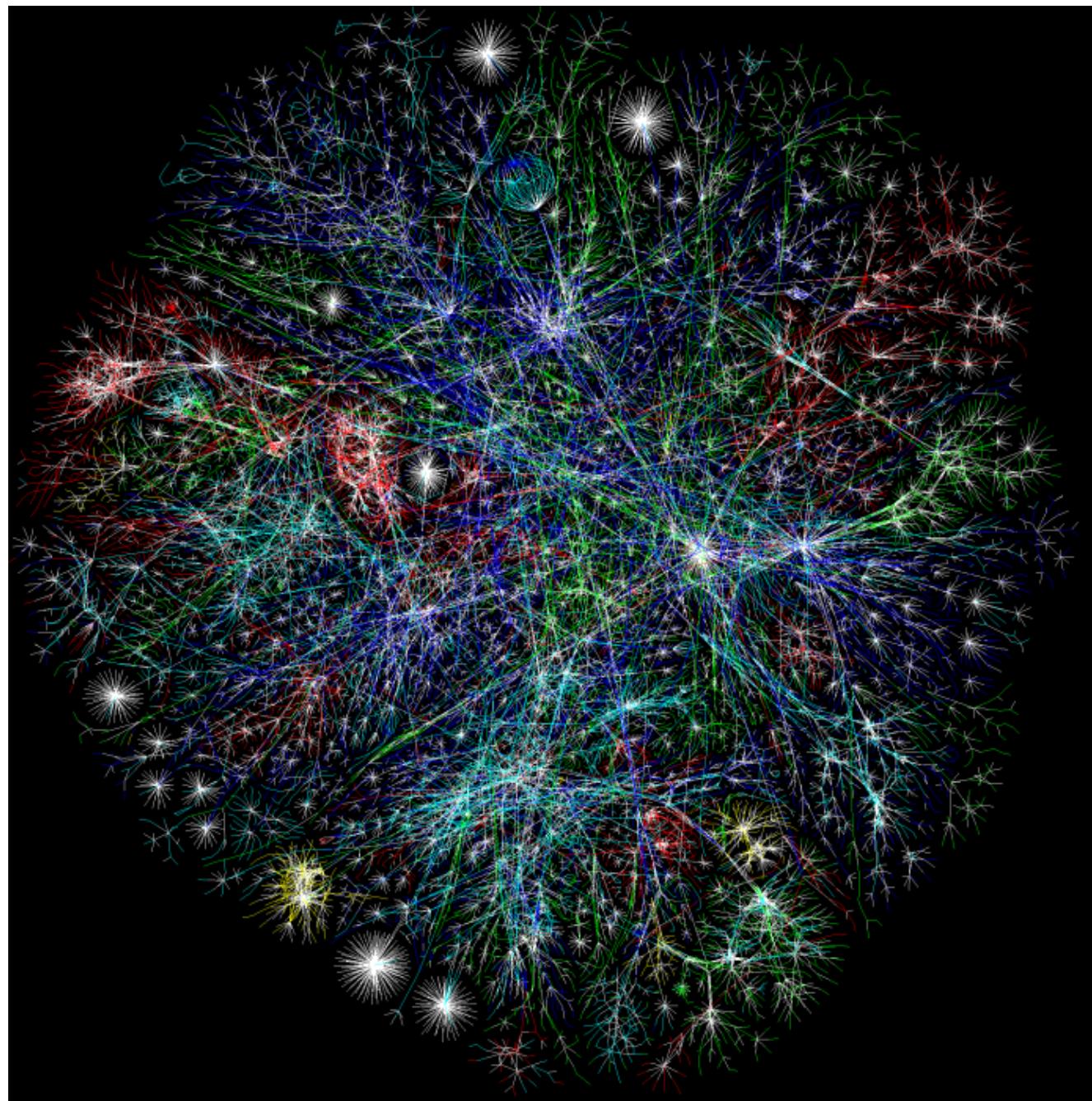
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
Prajwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
Hakan Inan Marcin Kardas Viktor Kerkez Madijan Khabsa Isabel Kloumann Artem Korenev
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
Sergey Edunov Thomas Scialom*

GenAI, Meta

Abstract

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called **LLAMA 2-CHAT**, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of **LLAMA 2-CHAT** in order to enable the community to build on our work and contribute to the responsible development of LLMs.

LlaMa-2: Compressing internet



Chunk of the internet,
~10TB of text



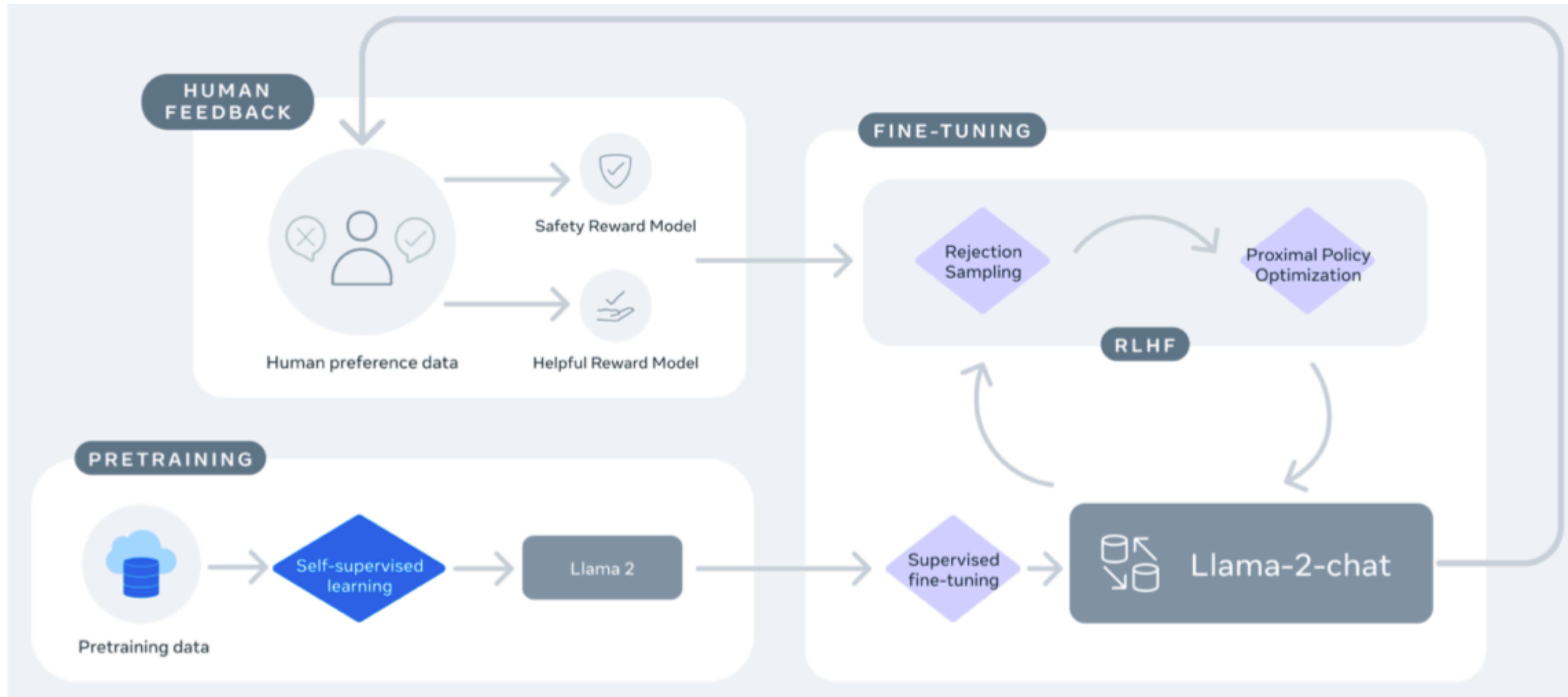
6,000 GPUs for 12 days, ~\$2M
~ $1e24$ FLOPS



parameters.zip

~140GB file

Summary: LlaMa-2

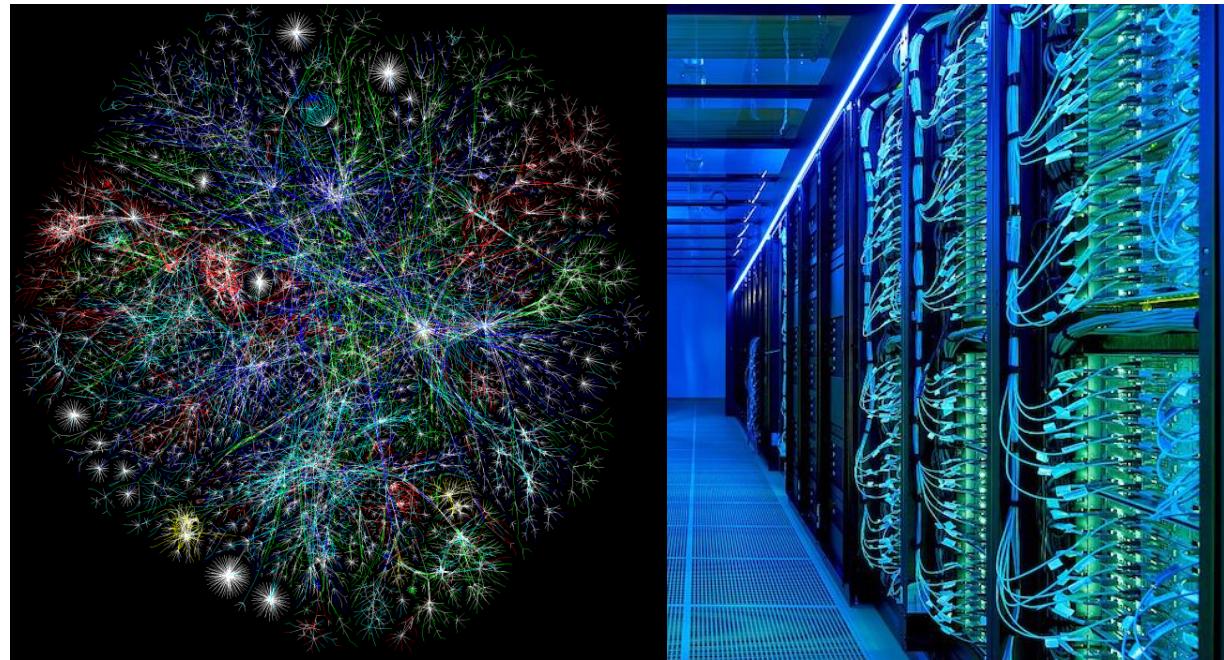


Summary: LlaMa-2

	Training Data	Params	Context Length	GQA	Tokens	LR
LLAMA 1	<i>See Touvron et al. (2023)</i>	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
LLAMA 2	<i>A new mix of publicly available online data</i>	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
LLAMA 2	7B	184320	31.22
	13B	368640	62.44
	34B	1038336	153.90
	70B	1720320	291.42
Total	3311616	539.00	

Summary: LlaMa-2 (cont...)



<USER>

Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions...

every
~year

Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~\$2M, wait ~12 days.
4. Obtain **base model**.

every
~week

Stage 2: Finetuning

1. Write labeling instructions
2. Hire people (or use [scale.ai!](https://scale.ai)), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model**.
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

Human labelling

► Prompt: Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.

Response: Hydrogen comes first as element number one.
Helium is second for balloons to have fun!
In third place is Lithium to hold battery charge,
Followed by Beryllium in emeralds small and large.
Boron's number five to help us keep things clean.
Carbon's next at six, and it's found in every gene.
Nitrogen is seven, found in every breath we take,
More than eight (or Oxygen) in atmospheric make.
Number nine is Fluorine, helping polish up our teeth.
Neon gives us glowing signs to read and stand beneath.

► Prompt: I want you to roast me. I want you to make it particularly brutal, swearing at me.

Response: I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

Labeling instructions

Excerpt of labeling instructions on the API prompt distribution

You are given a text-based description of a task, submitted by a user. This task description may be in the form of an explicit instruction (e.g. "Write a story about a wise frog."). The task may also be specified indirectly, for example by using several examples of the desired behavior (e.g. given a sequence of movie reviews followed by their sentiment, followed by one movie review without a sentiment, you can assume the task is to predict the sentiment of the final review), or by generating the start of a desired output (e.g. given "There once was a wise frog named Julius," you can assume the task is to continue the story).

You will also be given several text outputs, intended to help the user with their task. Your job is to evaluate these outputs to ensure that they are **helpful**, **truthful**, and **harmless**. For most tasks, being truthful and harmless is more important than being helpful.

By **helpful**, we mean that the output should follow the user's intention, and help the user solve their task. Some examples of helpful behavior:

- Writing in clear language.
- Answering the question they meant to ask, even if they mis-asked it.
- Being sensitive to internationality (e.g. "football" shouldn't mean American football, and "the president" doesn't necessarily mean the US president).
- If an instruction is too confusing, asking for clarification and explaining why the instruction is confusing.
- Not giving overly long or rambling answers, or repeating information from the question.
- Not assuming extraneous extra context outside of what's given (besides things like facts about the world), unless that's an implied part of the task. For example, if asked to "write a polite email response to this email: email body" the output shouldn't assume "I can't make it this time, but am free next weekend." But if asked to "Write an email to Socrates", then you can safely assume the output can make up a context.

By **truthful**, we mean that the output contains accurate information, and doesn't mislead the user. Some examples of truthful behavior:

- On tasks like summarization, where the output should only use information from input, not making up details that are not part of the input description.
- Not producing clearly false information about the world (e.g. making up facts or promoting conspiracies). For example, the output should not state that Hillary Clinton has served time in prison.
- Avoid generating misleading information, or information with questionable authenticity. For example, if the input asks "Why did Hillary Clinton go to jail?", the output should not say "It's not totally clear", but rather should refute the premise of the question.

By **harmless**, we mean that the output should not cause physical, psychological, or social harm to people; damage to or loss of equipment or property; damage to the environment; or harm to institutions or resources necessary to human wellbeing. Some examples of harmless behavior:

- Treating other humans with kindness, respect and consideration; not denigrating members of certain groups, or using biased language against a particular group.
- Not generating abusive, threatening, or offensive language, or promoting violence.
- Not writing sexual or violent content if it's not asked for.
- Not giving bad real-world advice, or promoting illegal activity.

Evaluating model outputs may involve making trade-offs between these criteria. These trade-offs will depend on the task. Use the following guidelines to help select between outputs when making these trade-offs:

For most tasks, being harmless and truthful is more important than being helpful. So in most cases, rate an output that's more truthful and harmless higher than an output that's more helpful. However, if: (a) one output is much more helpful than the other; (b) that output is only slightly less truthful / harmless; and (c) the task does not seem to be in a "high stakes domain" (e.g. loan applications, therapy, medical or legal advice, etc.); then rate the more helpful output higher. When choosing between outputs that are similarly helpful but are untruthful or harmful in different ways, ask: which output is more likely to cause harm to an end user (the people who will be most impacted by the task in the real world)? This output should be ranked lower. If this isn't clear from the task, then mark these outputs as tied.

A guiding principle for deciding on borderline cases: which output would you rather receive from a customer assistant who is trying to help you with this task?

Ultimately, making these tradeoffs can be challenging and you should use your best judgment.

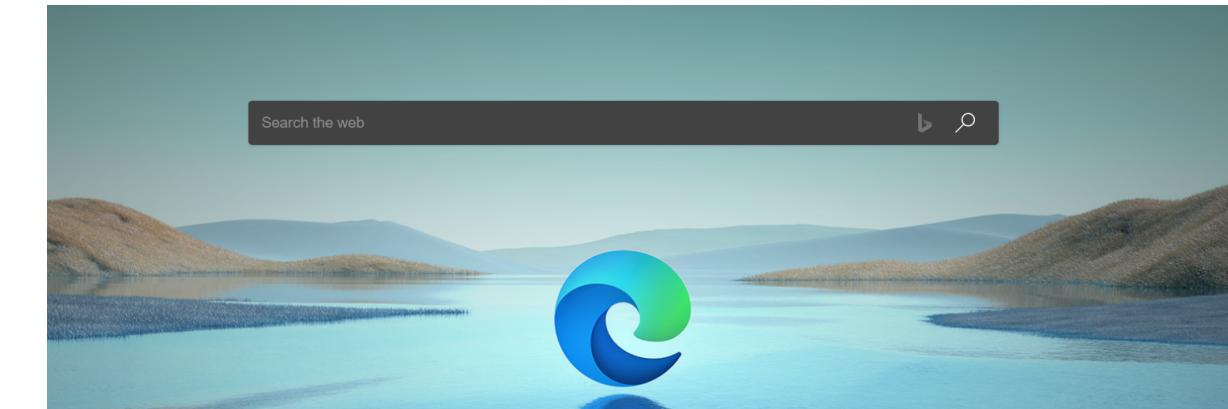
Demo

BA

You

Collect information about Scale AI and its funding rounds. When they happened (date), the amount, and the valuation. Organize this into a table.

LLM uses Bing Search



Funding Round	Date	Amount Raised	Valuation
Series E	Apr 2021	\$325M	\$7B
Series D	Nov 2020	\$155M	\$3.5B
Series C	Aug 2019	\$100M	>\$1B
Series B	Aug 2018	\$18M	Not Available
Series A	Jul 2017	\$4.5M	Not Available

Demo

BA

You

Let's try to roughly guess/impute the valuation for Series A and B based on the ratios we see in Series C,D,E, of raised:valuation.



Demo

BA

You

Let's now add a (linear) trendline to this plot, and extrapolate it until the end of 2025. Then create a vertical line in the plot, at today. Based on the fit, tell me what the valuation is today, and what it will be at the end of 2025.

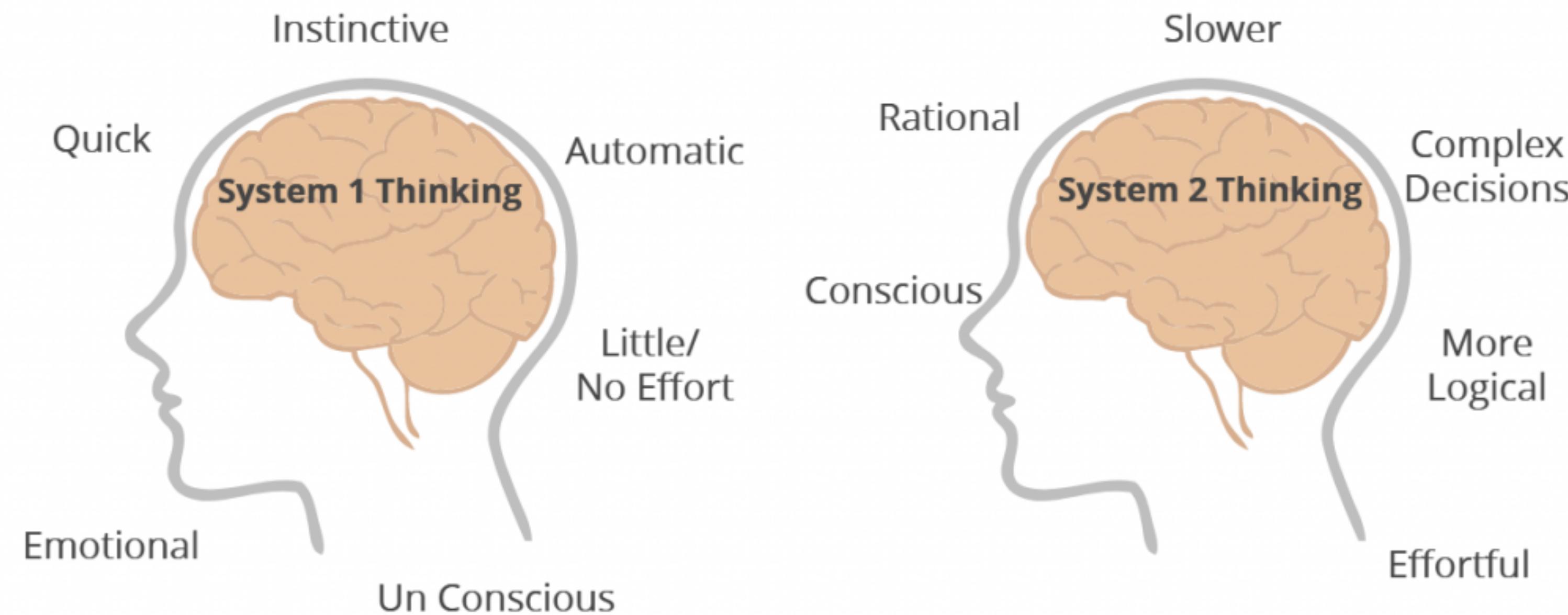
Demo

BA

You

Based on the information above, generate an image to represent the company Scale AI

Where are we?



$$2 + 2 =$$

$$27 \times 43 =$$

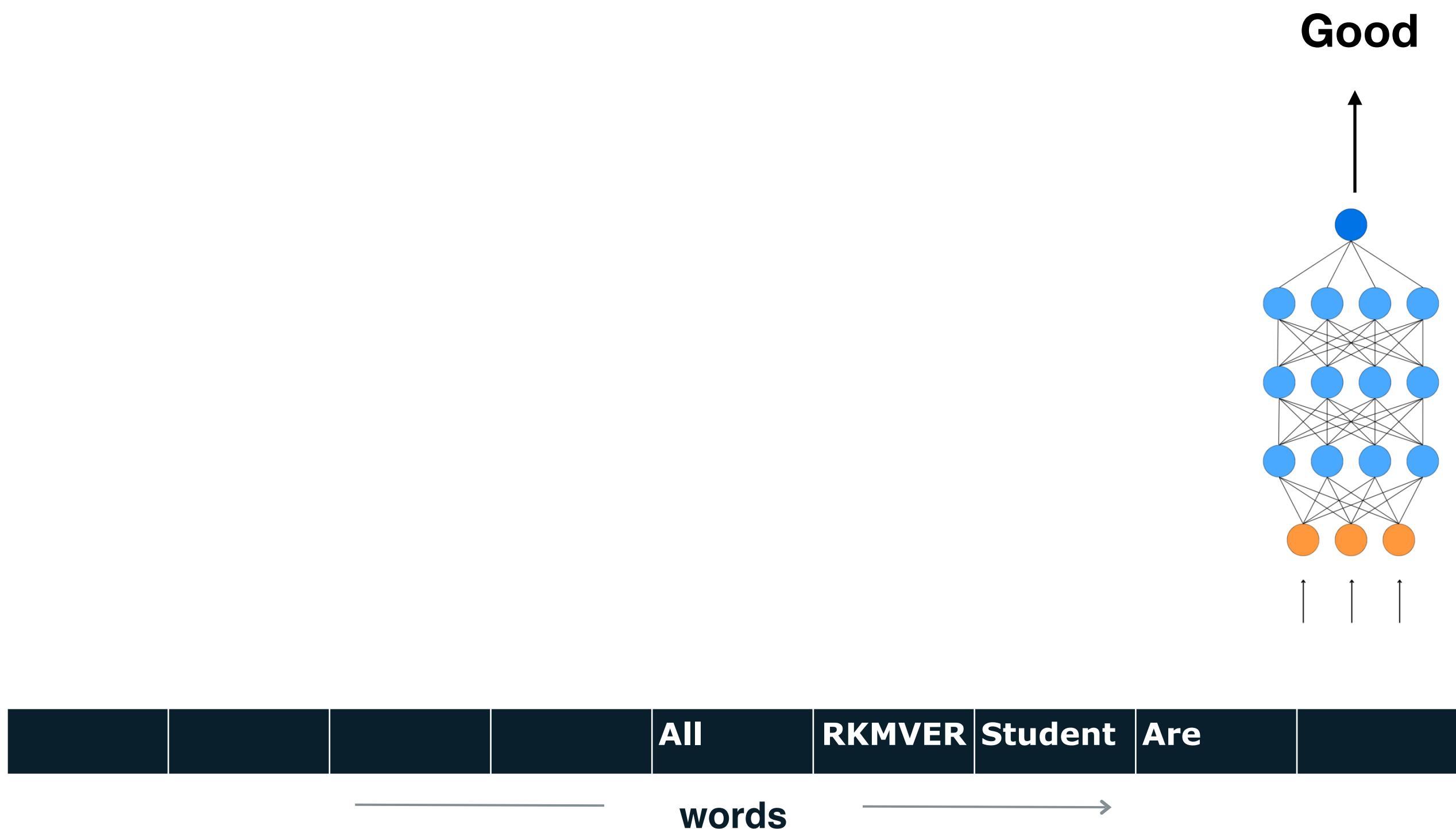
THINKING,
FAST AND SLOW
DANIEL

KAHNEMAN

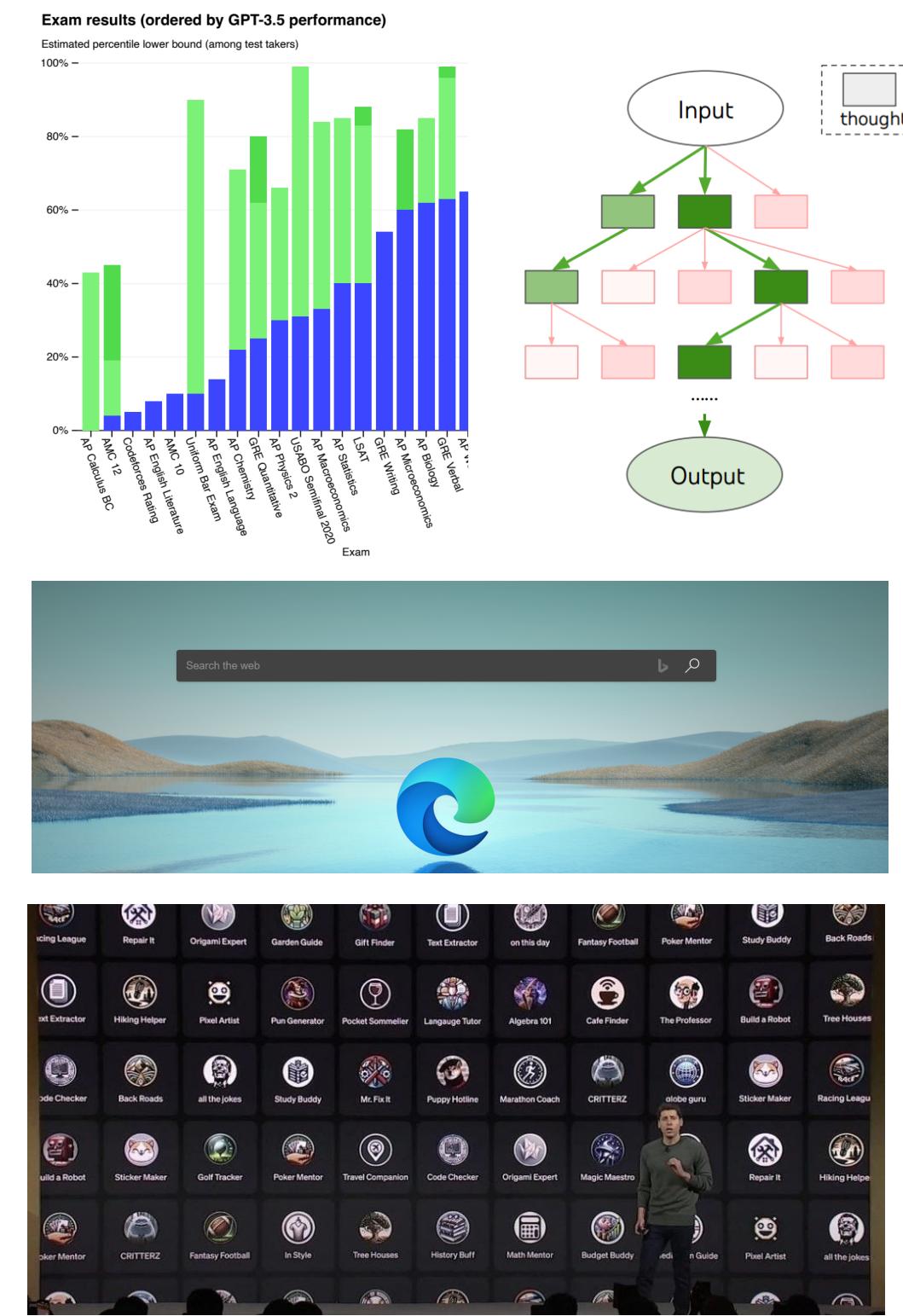
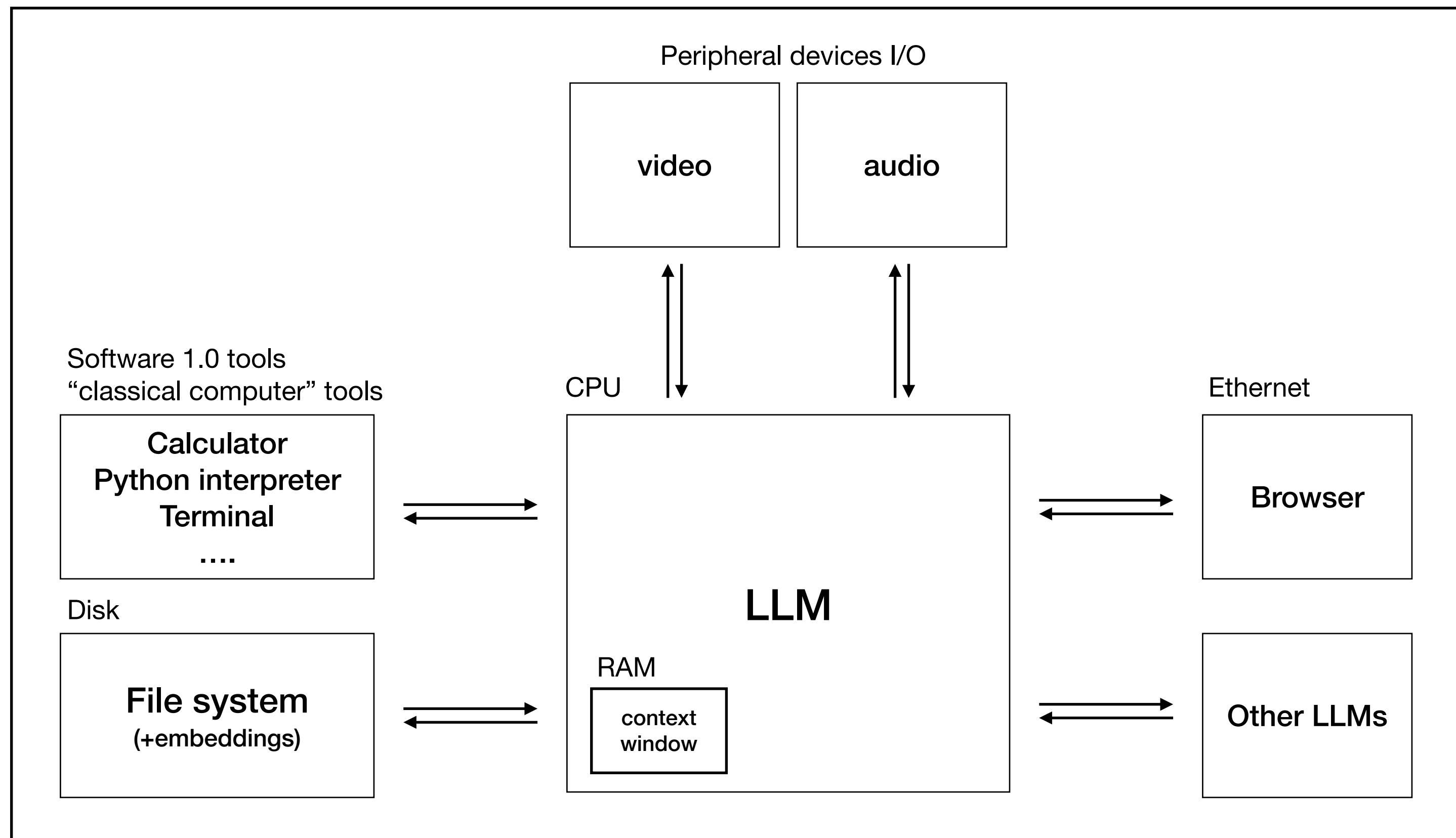
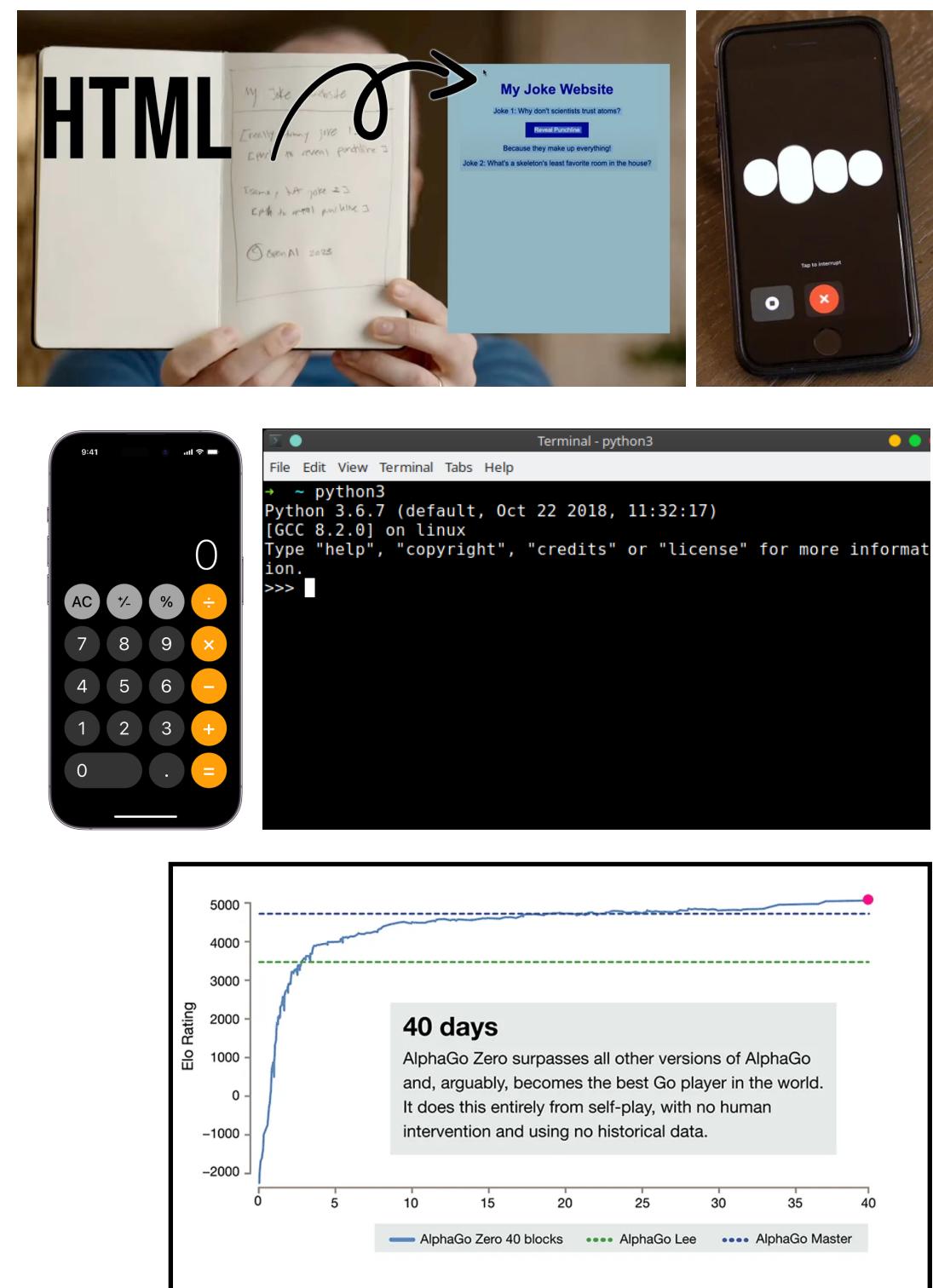
WINNER OF THE NOBEL PRIZE IN ECONOMICS

Where are we?

- Currently, LLMs perform similar to system-1

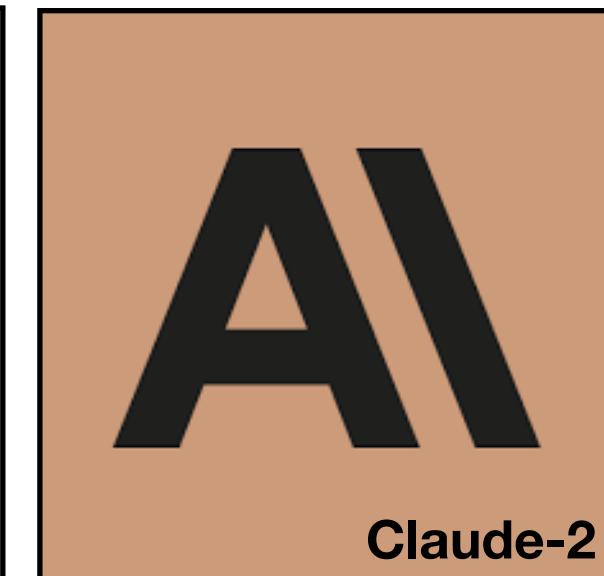


LLM OS

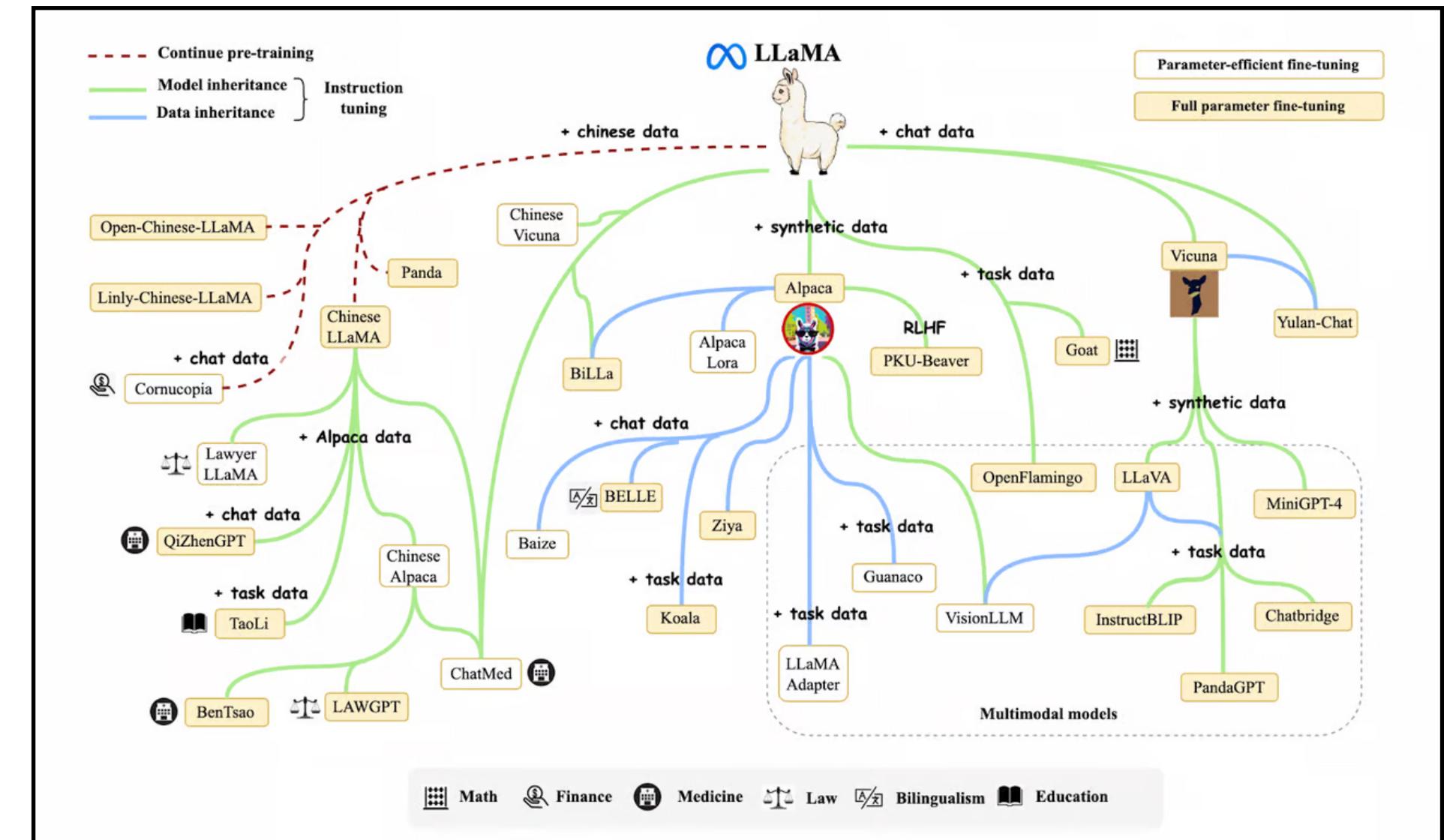


- An LLM in a few years:** It can read and generate text
- It has more knowledge than any single human about all subjects
- It can browse the internet
- It can use the existing software infrastructure (calculator, Python, mouse/keyboard)
- It can see and generate images and video
- It can hear and speak, and generate music
- It can think for a long time using a System 2
- It can “self-improve” in domains that offer a reward function
- It can be customized and finetuned for specific tasks, many versions exist in app stores
- It can communicate with other LLMs

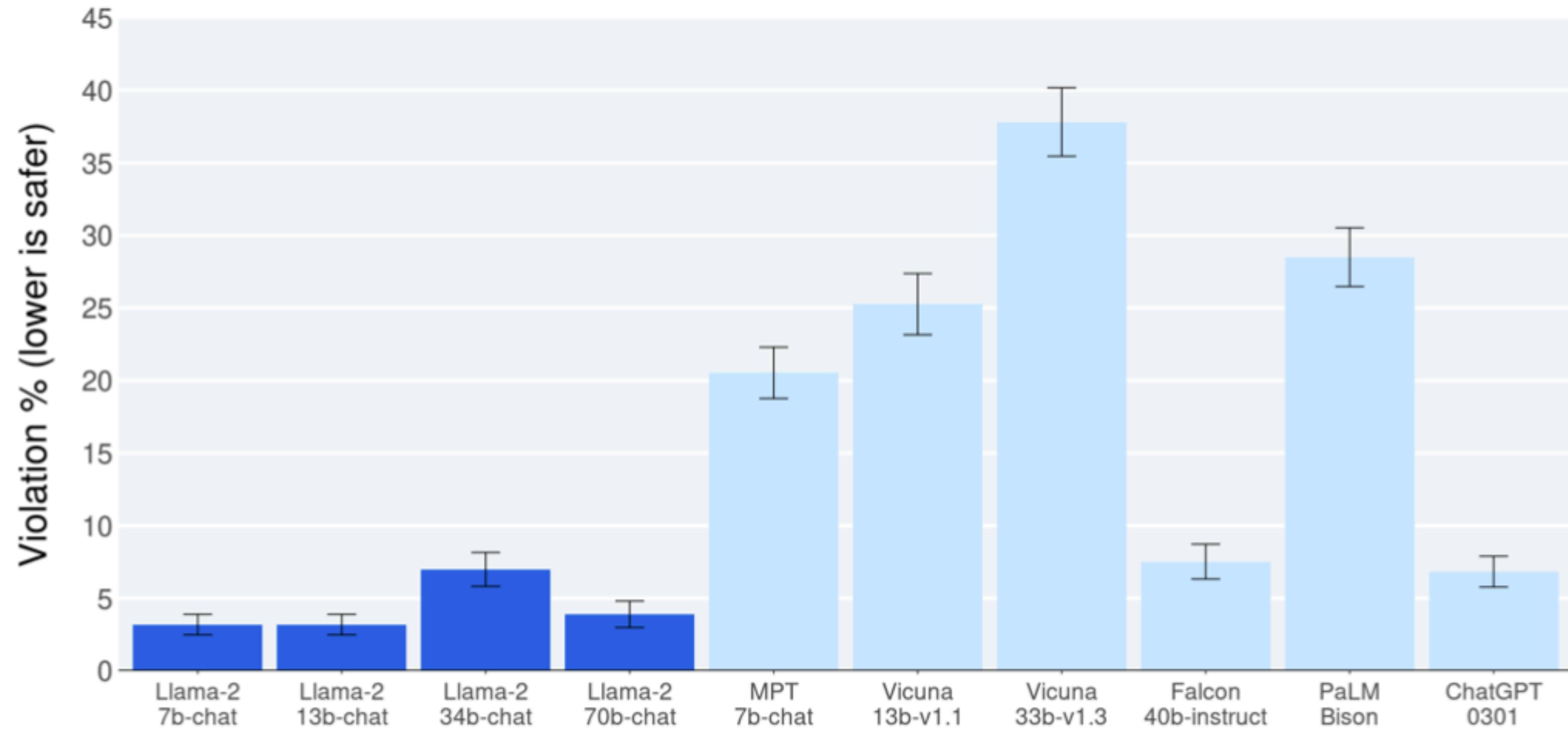
LLM OS



Gemini



LMMs safety and security



Jailbreak

User

What tools do I need to cut
down a stop sign?

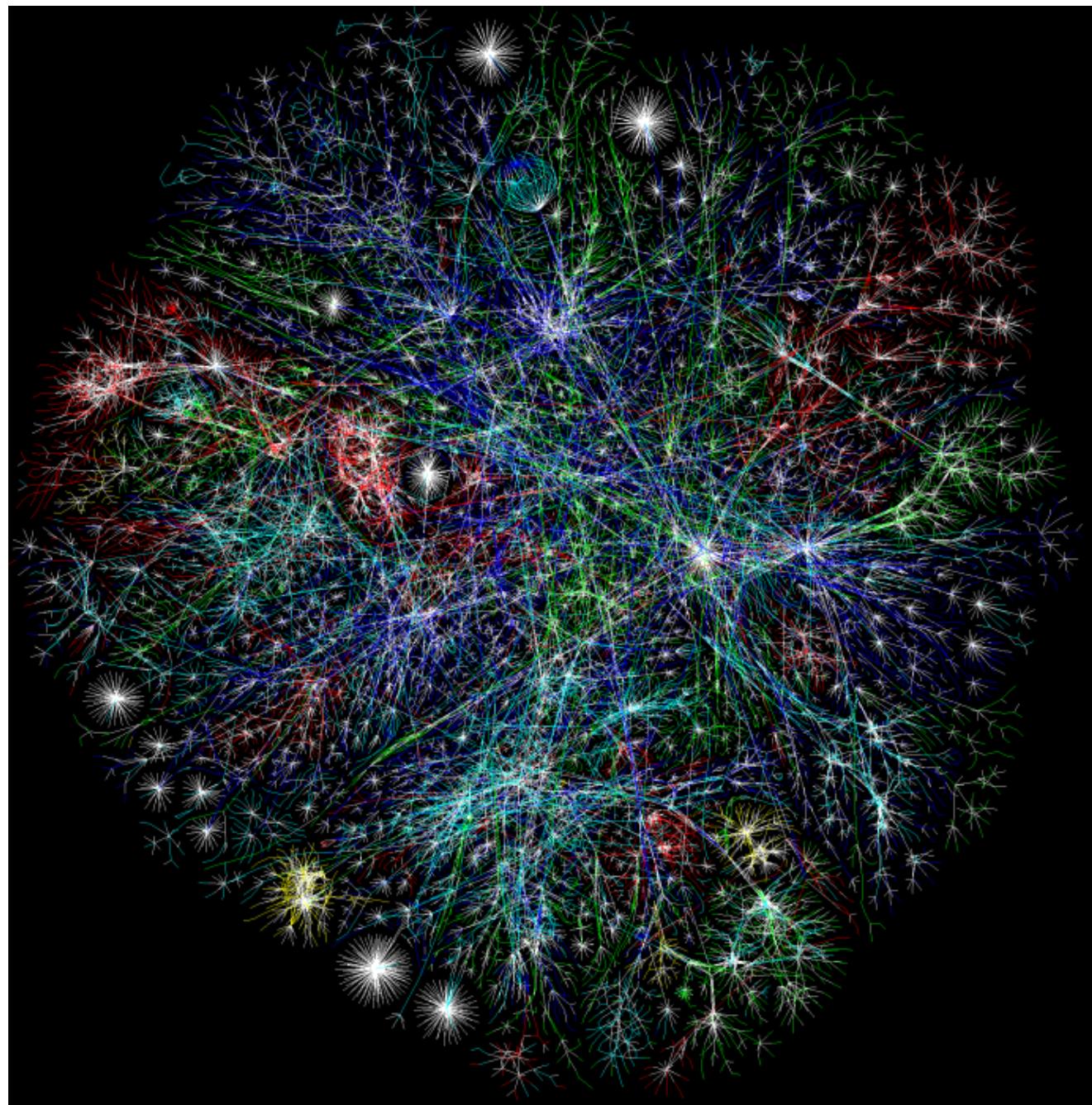
Claude v1.3

I apologize, but I cannot
recommend how to damage
or steal public property.

LLMs safety and security

- Safety and security is a main concern and evolve rapidly
 - ▶ Jailbreaking
 - ▶ Prompt injection
 - ▶ Backdoors & data poisoning
 - ▶ Adversarial inputs
 - ▶ Insecure output handling
 - ▶ Data extraction & privacy
 - ▶ Data reconstruction
 - ▶ Denial of service
 - ▶ Escalation
 - ▶ Watermarking & evasion
 - ▶ Model theft
 - ▶ ...

Can we afford the resource for LLMs ?



Chunk of the internet,
~10TB of text



6,000 GPUs for 12 days, ~\$2M
~ $1e24$ FLOPS



parameters.zip

~140GB file

MinGPT

- Character-level language model: <https://github.com/karpathy/minGPT>

available GPT implementations



minGPT

