

An Open AI Stack:

From scaling AI workloads to LLM evaluation

Ion Stoica,
July 31, 2025

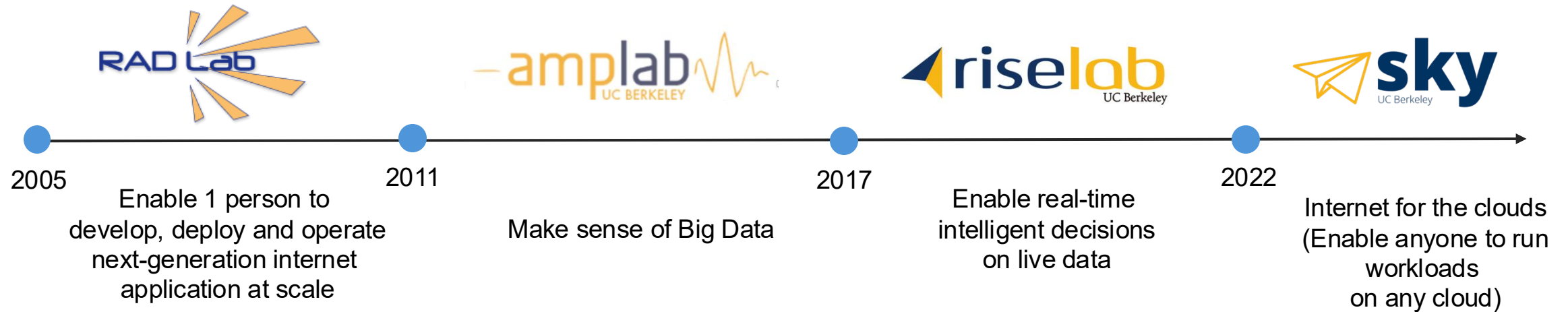


Berkeley five-year Labs



- Started in 1980s by Dave Patterson, Randy Katz, ...
- Created around a mission
- 6-10 faculty
- 30-60 PhD students & post-docs
- Cross-disciplinary, e.g., systems, ML/AI, databases, networking, etc
- Industry and governmental funding
- Building real software artifacts: all open-source
- Huge impact: RISC, RAID, NOW, ...

Berkeley Labs I've been part of

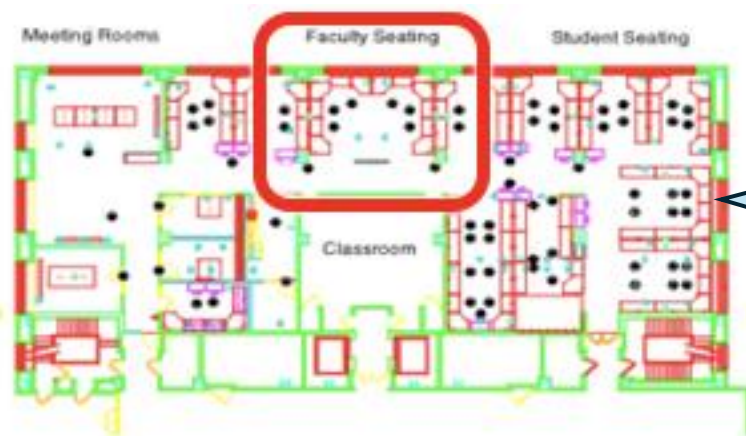




2011-2017

Mission: Make sense of Big Data

- 8 faculty
- ~50 students and postdocs
- Bridge systems, ML, databases
- Collaborative environment
- Successor of RAD Lab



ML, System, Database researchers,
working in the same open space

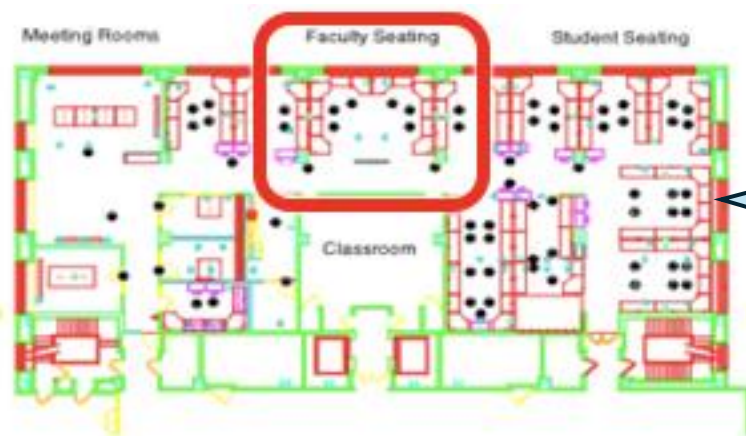




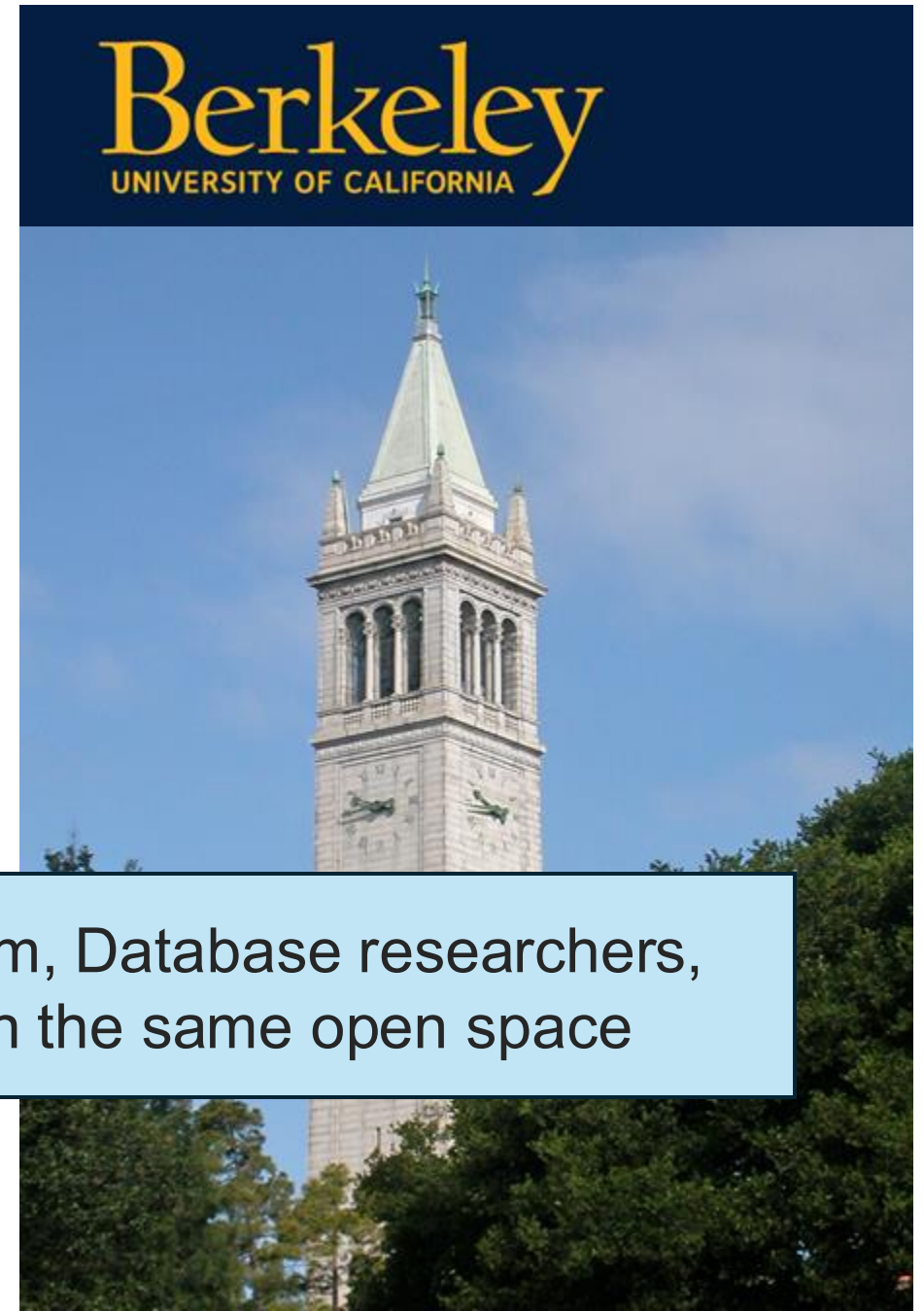
2022-2027

Mission: Create the internet for the clouds

- 8 faculty
- 50-60 students and postdocs
- Bridge AI, systems, databases
- Collaborative environment
- Successor of RISE Lab



AI, System, Database researchers,
working in the same open space



Some project I've been involved in



MemGPT



Chatbot Arena



Vicuna

Applications

Alpa

AlpaServ
e

LLM



Optimizers
(optimize ML workloads)



Distributed Frameworks
(dev & run distributed apps)



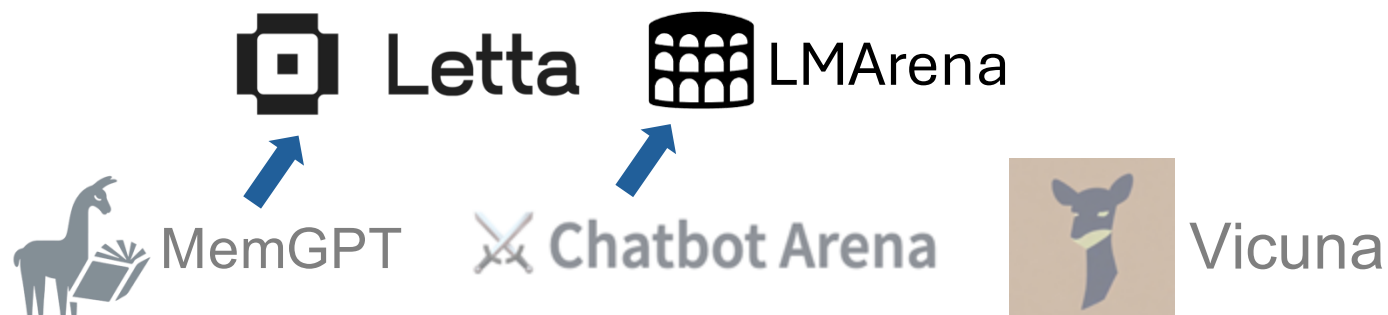
Apache
MESOS



Skyplane

Infra & Orchestration Layer
(allocate resources)

Many led to startups



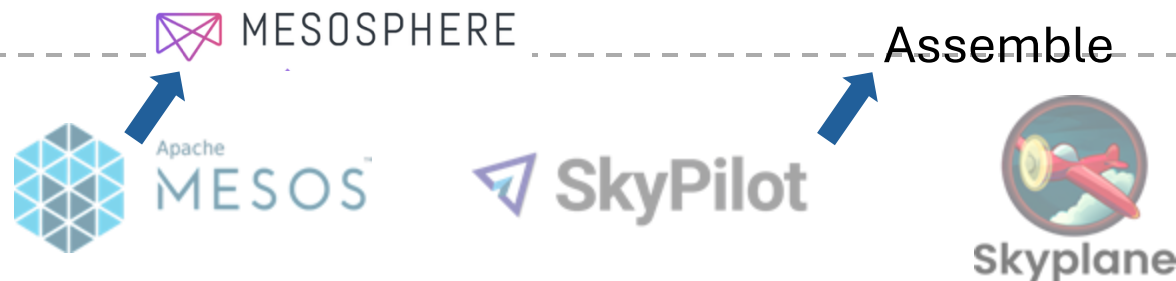
Applications



Optimizers
(optimize ML workloads)

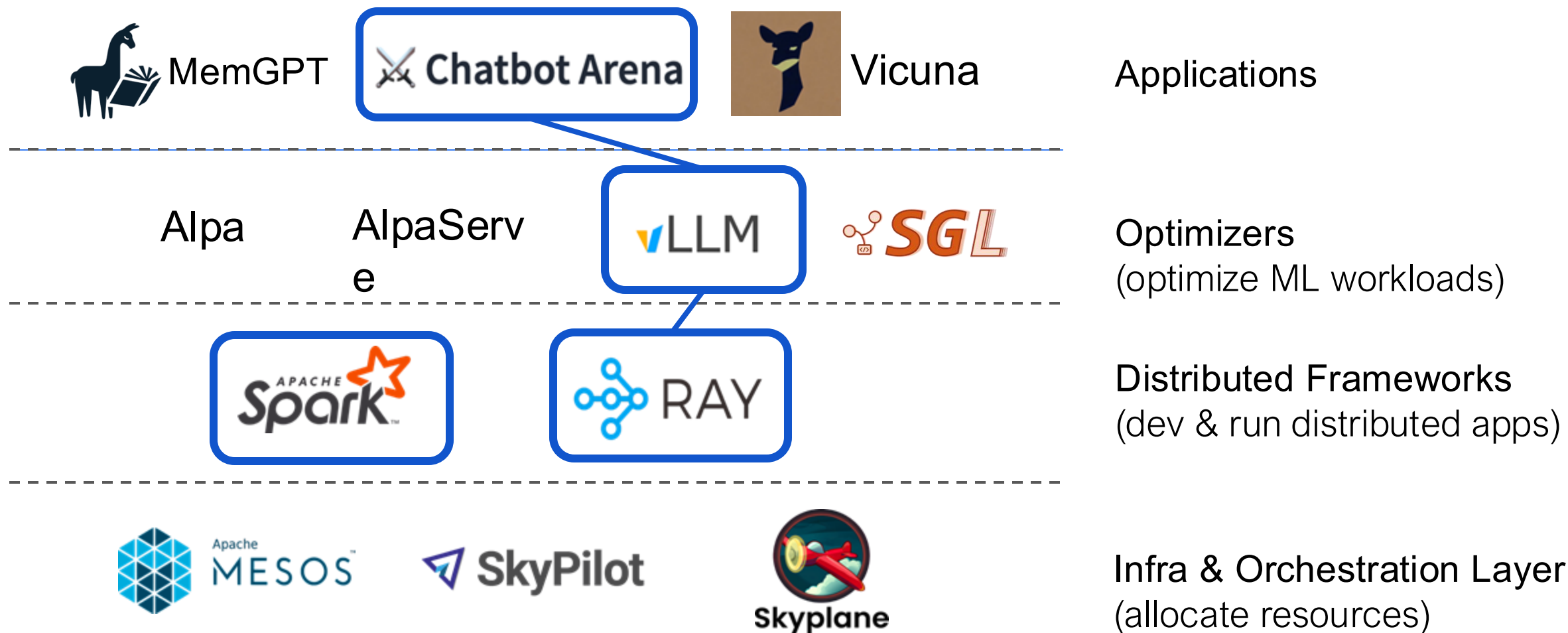


Distributed Frameworks
(dev & run distributed apps)



Infra & Orchestration Layer
(allocate resources)

Today's Talk



Template

Trends

Challenges

Key idea

Impact

Apache Spark (2009 -)

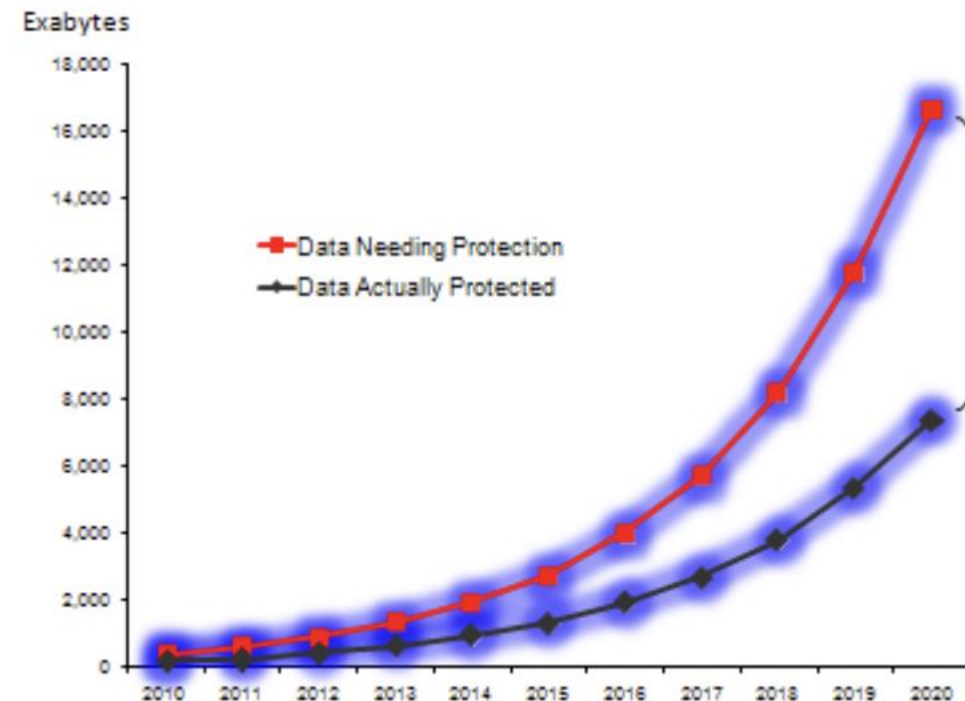
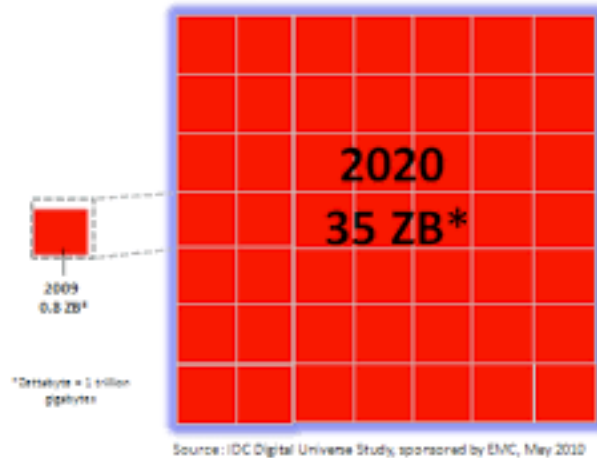


Trends (2009)

Data (e.g., machine and user logs) growing exponentially

Figure 1: The Digital Universe 2009 – 2020

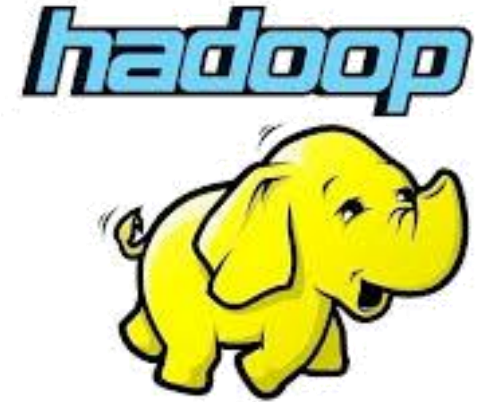
Growing by a Factor of 44



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

State-of-the-art in Big Data (2009)

- Apache Hadoop
 - Large scale, flexible data processing engine
 - Batch computation (e.g., 10s minutes to hours)
 - Open Source
- Getting rapid industry traction:
 - High profile users: Facebook, Twitter, Yahoo!, ...
 - Distributions: Cloudera, Hortonworks
 - Many companies still in austerity mode



Workload Trends

Interactive computations, e.g., ad-hoc analytics

- SQL engines like Hive and Pig drove this trend

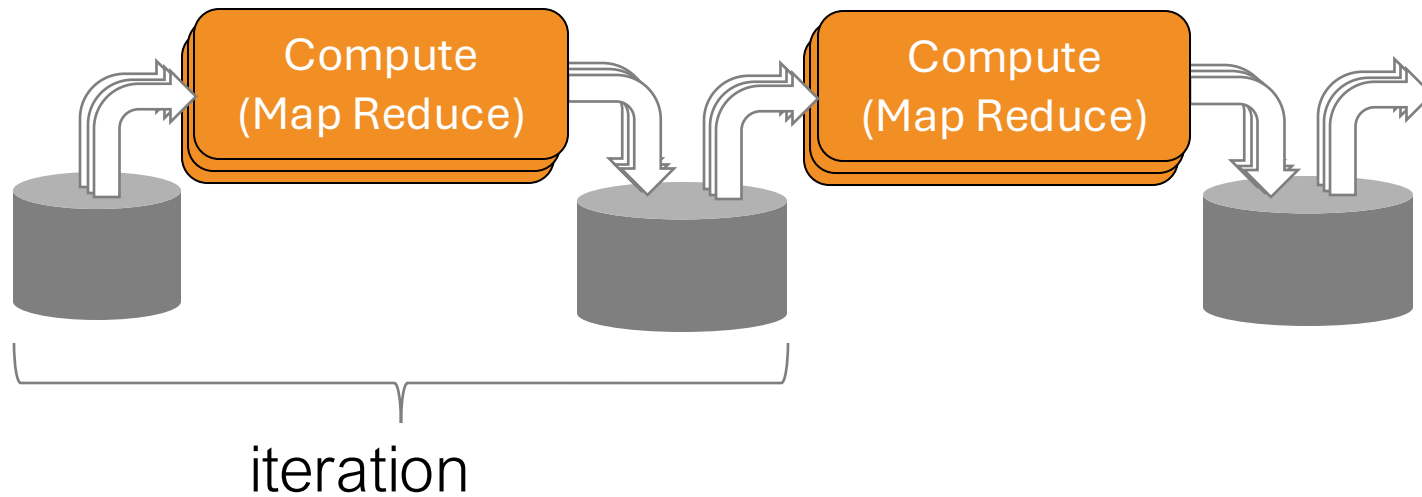
Iterative computations, e.g., Machine Learning

- Get deeper insights from data

Challenge: Hadoop too slow

Each iteration MapReduce jobs reads/writes from/to disk

- Scalable and fault-tolerant but slow



A story: Netflix challenge

ML students in RAD Lab wanted to participate



Lester
(ML postdoc)



Prize

anonymized movie rating dataset
best recommendation algorithm

\$1m

A story: Netflix challenge



Lester
(ML postdoc)

I need to scale
collaborative
filtering



Matei
(ML PhD)

Use Hadoop!

A story: Netflix challenge



It is too slow!

Lester
(ML postdoc)



Let me see
whether I can
help!

Matei
(ML PhD)

Key observation (2009)

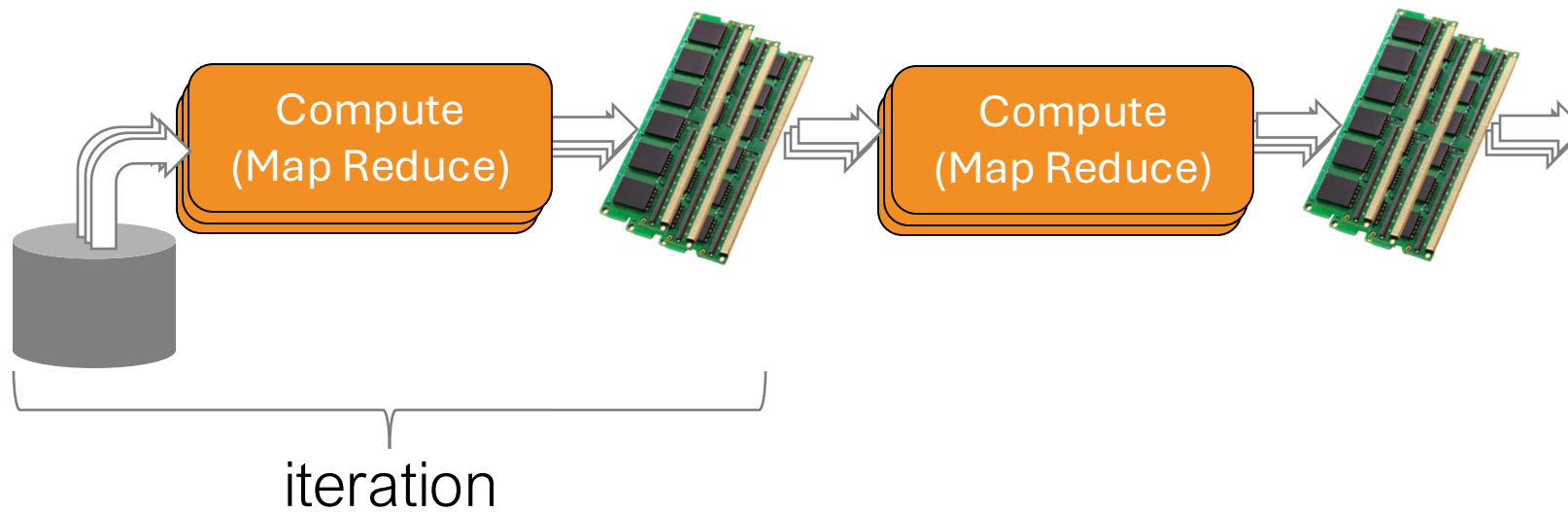
Despite huge amounts of data, many working sets in big data clusters **fit in memory**



Apache Spark: Key Idea

Store intermediate data in memory

- Great for interactive queries and machine learning

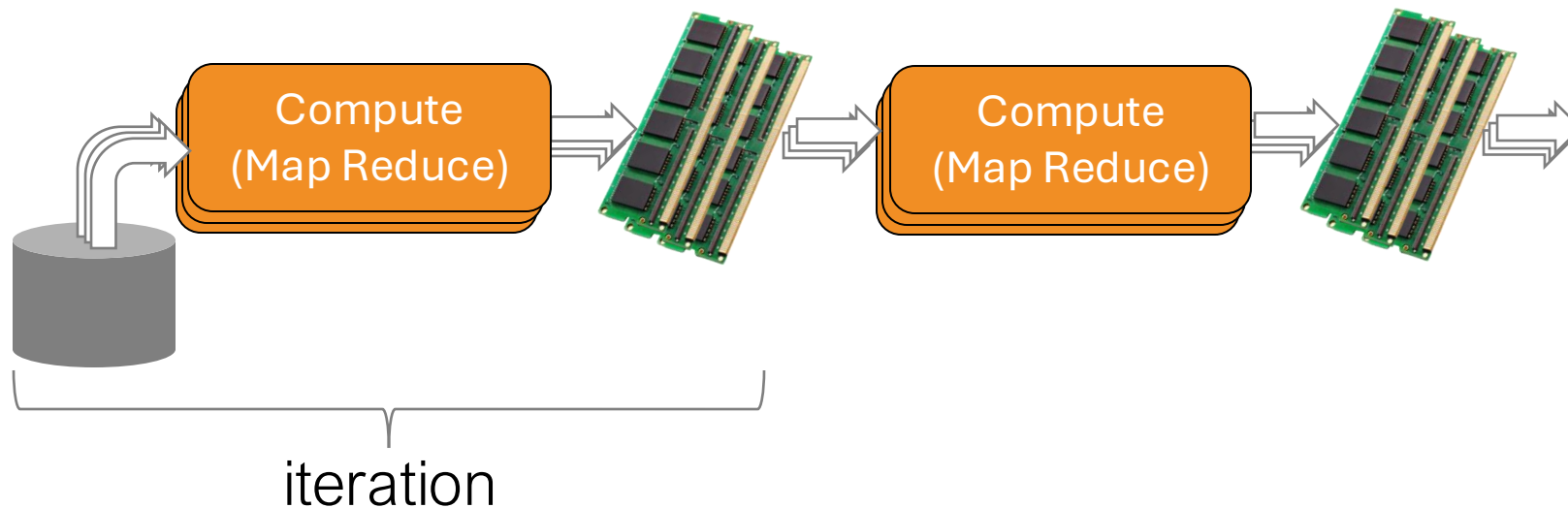




Apache Spark: Key Idea

Store intermediate data in memory

- Great for interactive queries and machine learning



Ensure reliability by executing the computation that produced the lost in the first place

Netflix Prize

COMPLETED

[Home](#)[Rules](#)[Leaderboard](#)[Update](#)[Download](#)

Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Display top leaders.

tied for
best score

20 mins late

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Impact

Apache Spark, de-facto standard for big data processing



Databricks, one of the most successful companies at the intersection between Data + AI



Ray (2016 -)

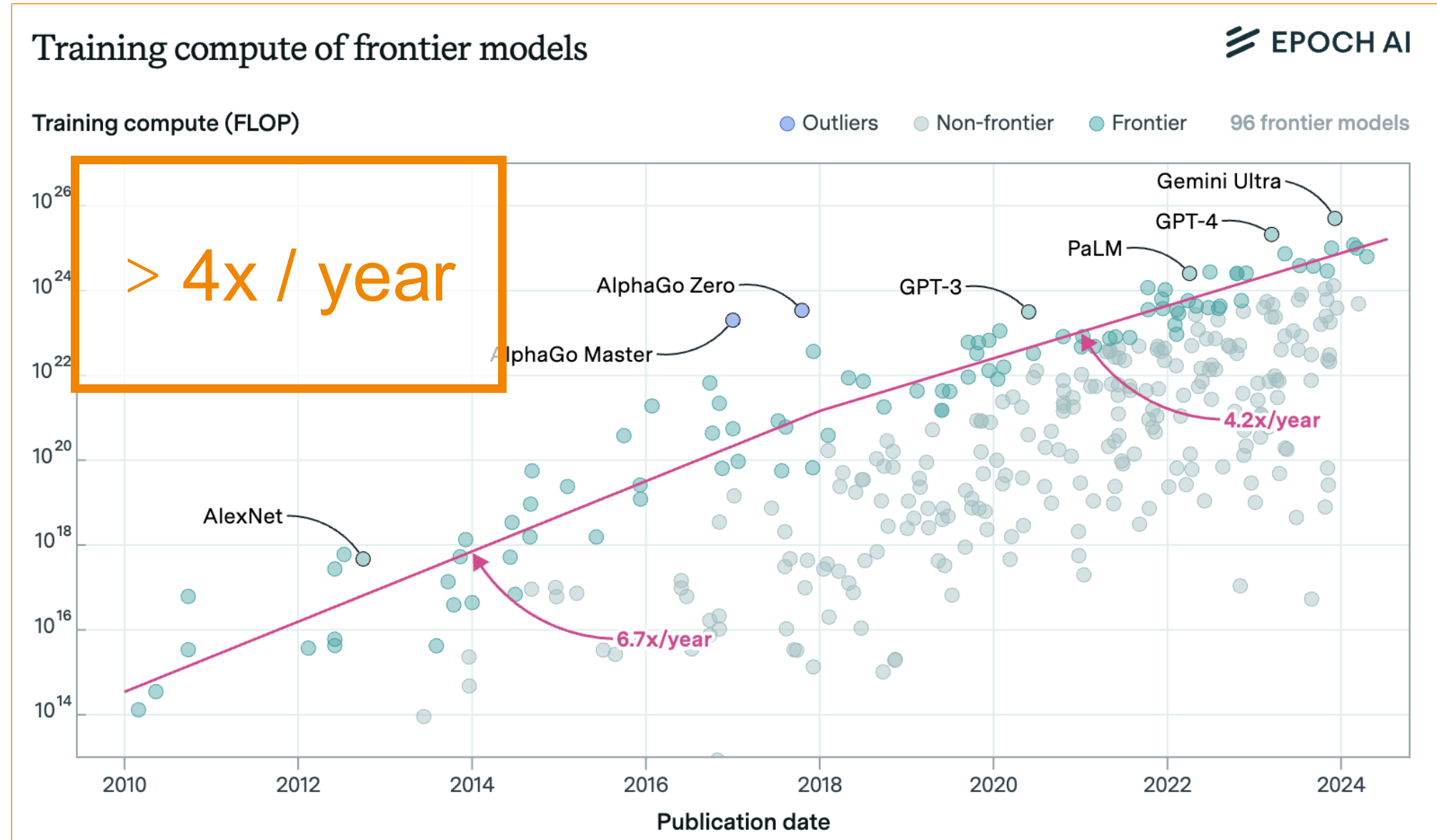


Trends

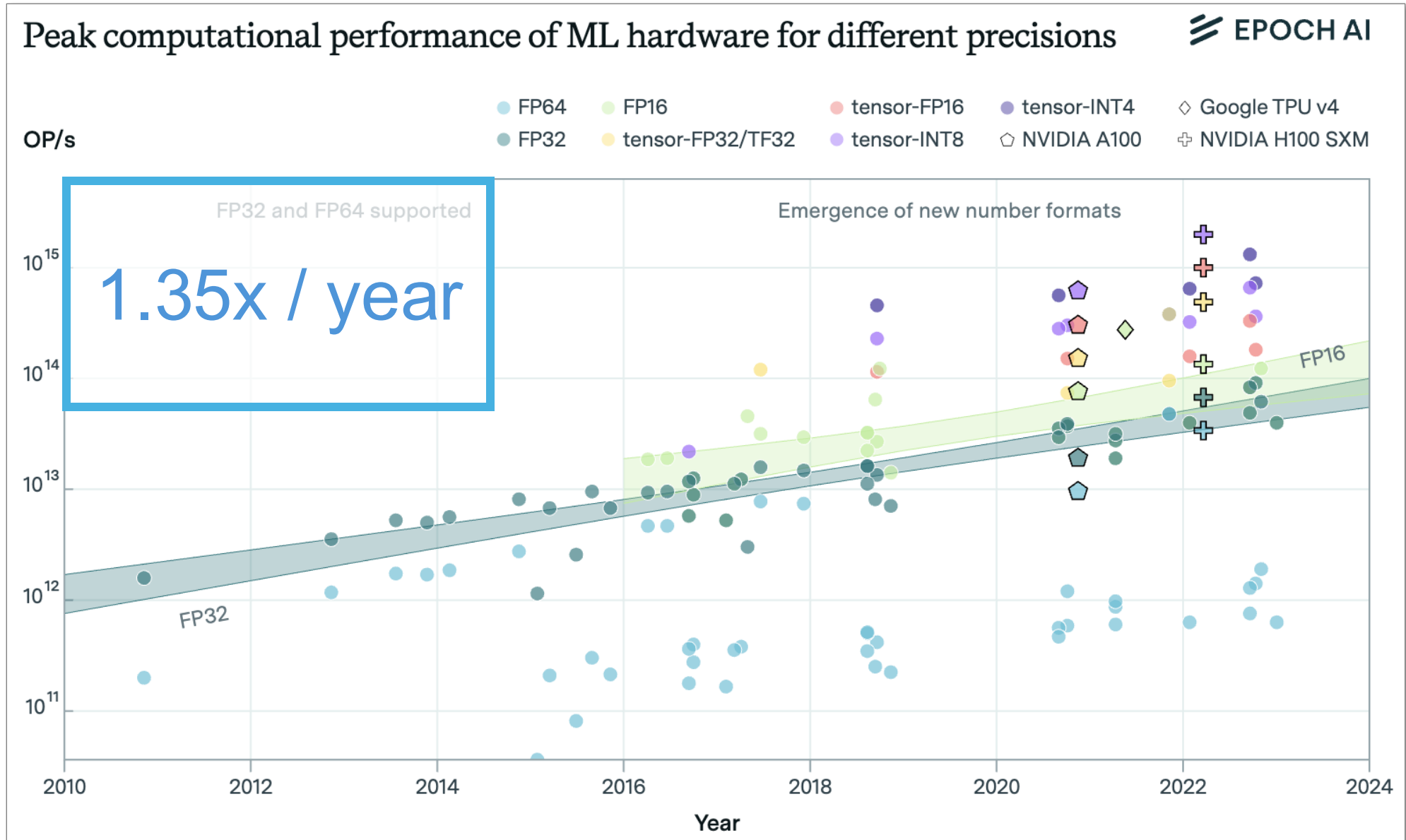
AI demands growing faster than single node capabilities

AI workloads becoming more complex

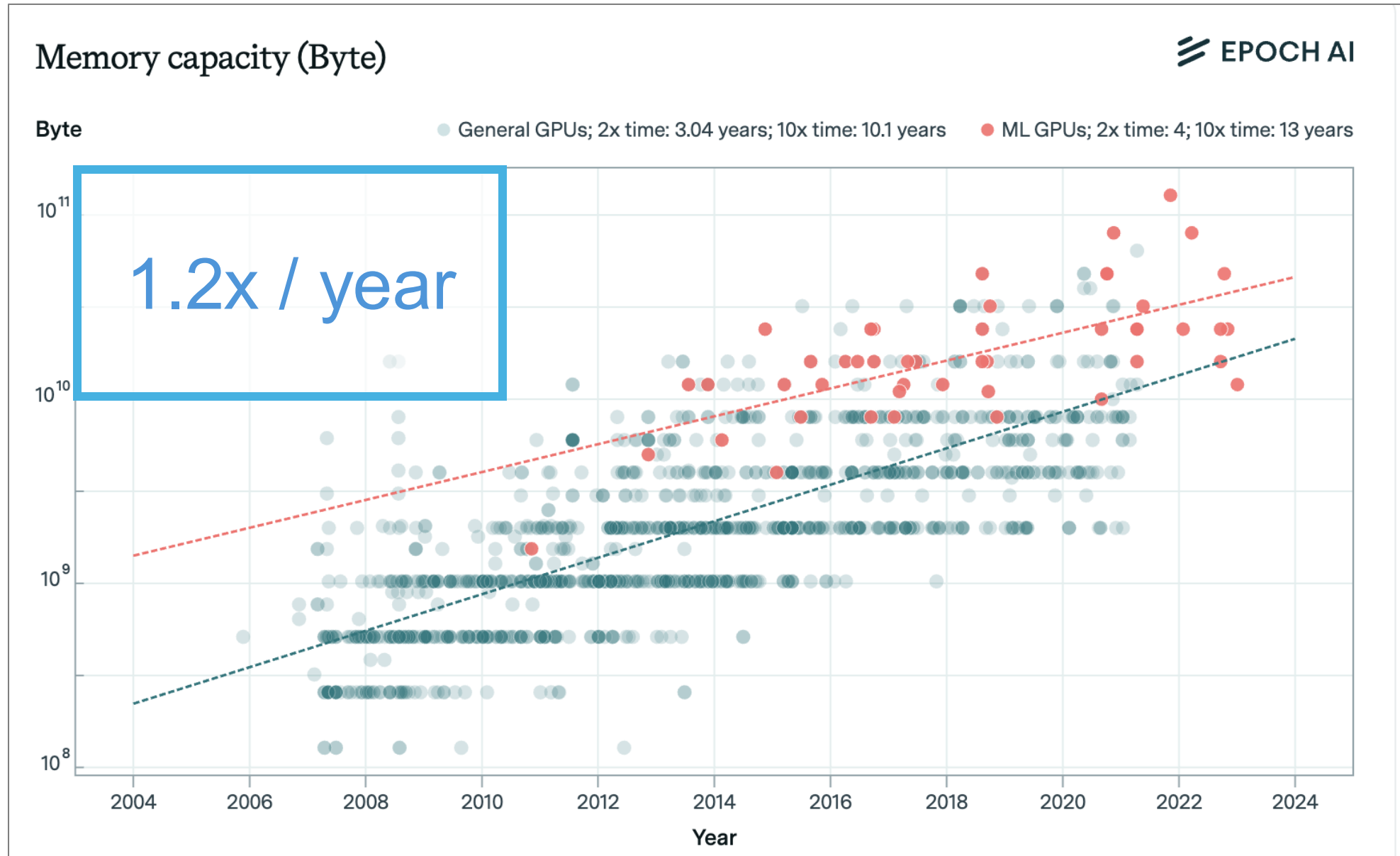
AI demands growing faster ...



AI demands growing faster than node's capabilities



AI demands growing faster than node's capabilities



AI demands growing faster than node's capabilities

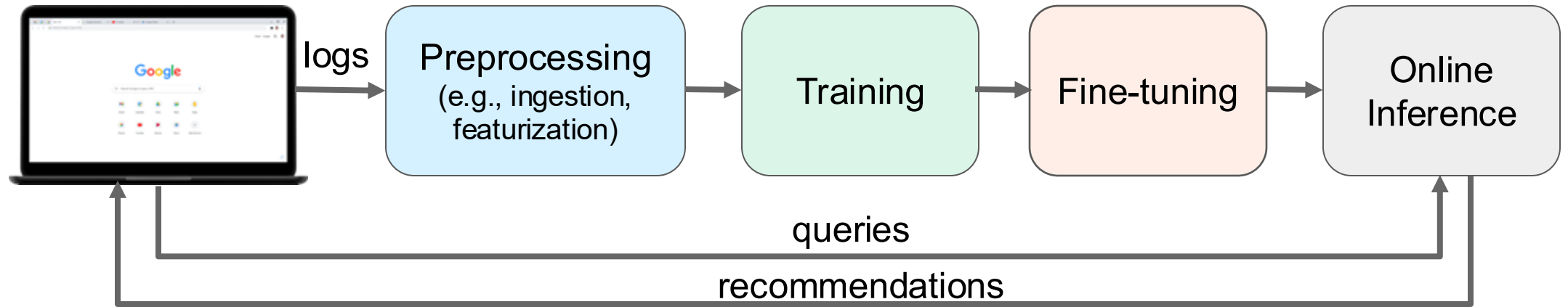
Compute demands are growing $\sim 3x / \text{year}$ faster than compute and memory capacity/bandwidth

Even if demand stops growing, it would take 10s of years to train a model on a single machine at current growth rates!

Heterogeneous distributed computing becoming the norm

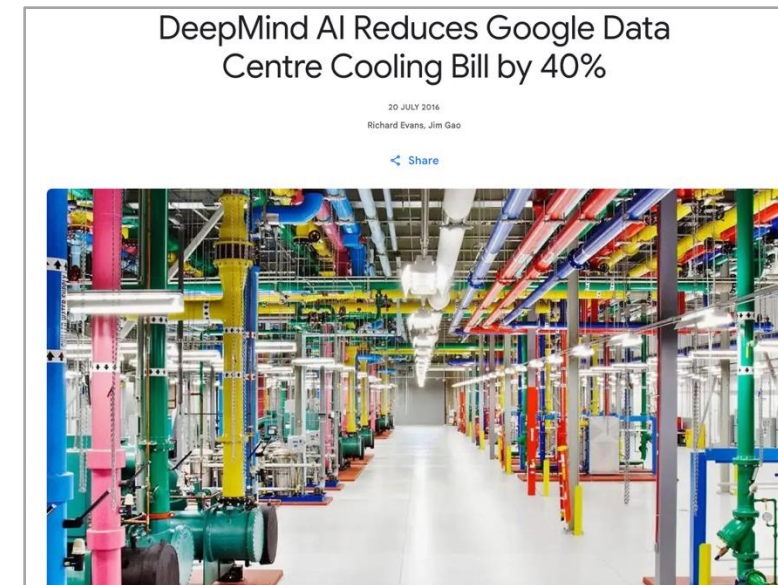
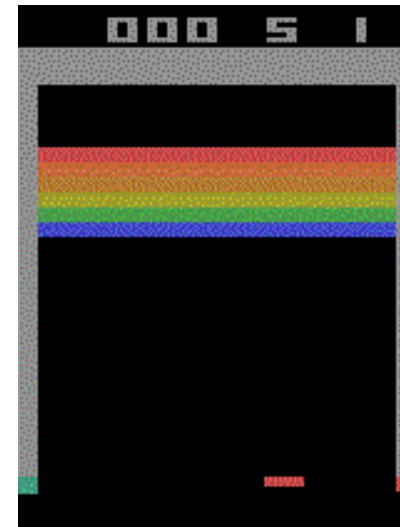
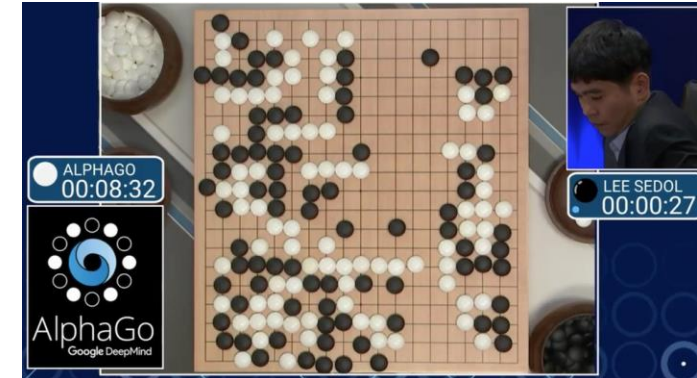
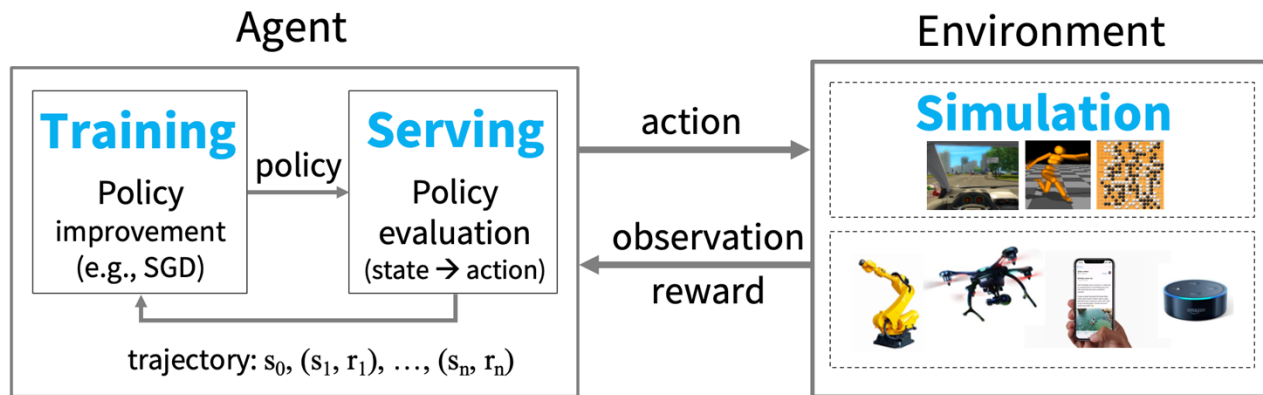
AI workloads becoming more complex

Recommendation system



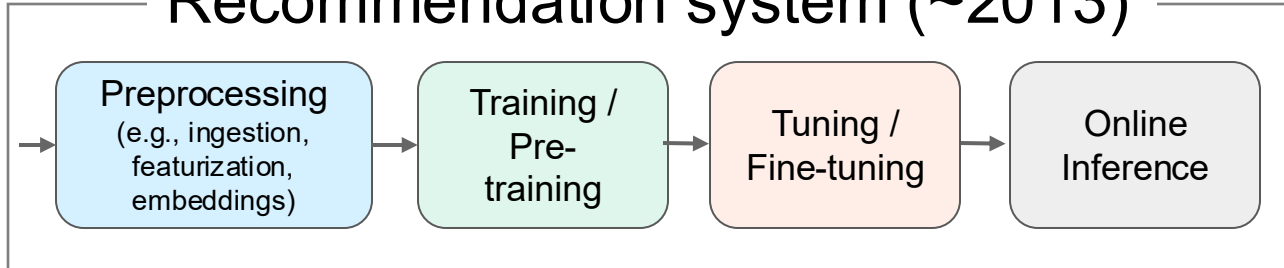
AI workloads becoming more complex

Reinforcement learning

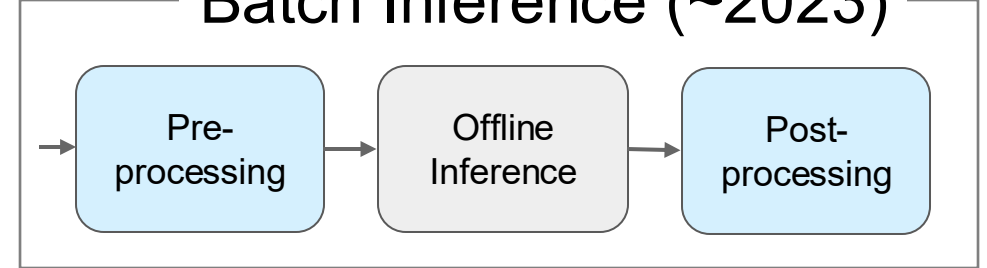


AI workloads becoming more complex

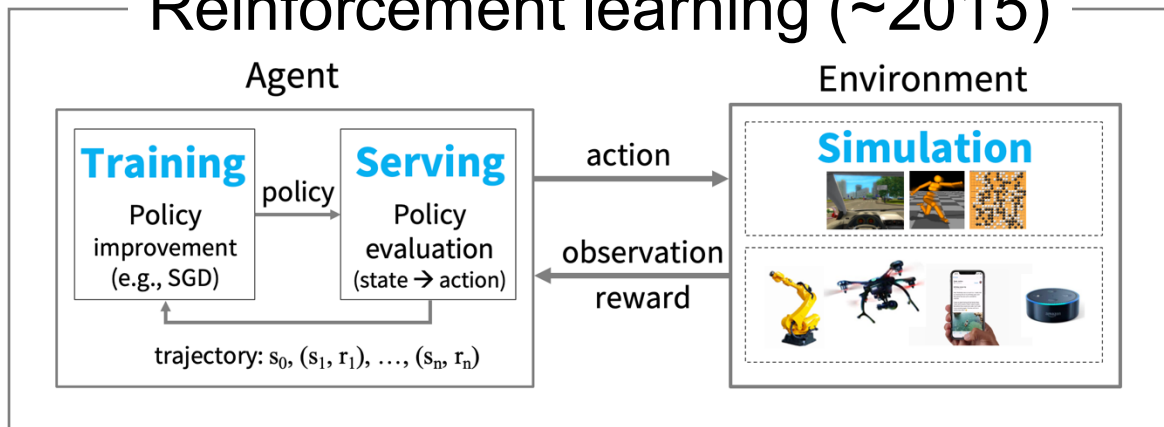
Recommendation system (~2013)



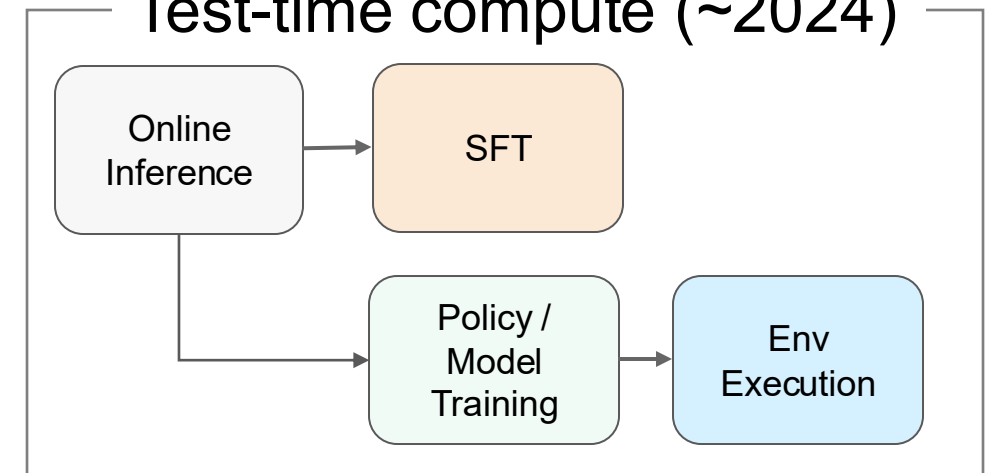
Batch Inference (~2023)



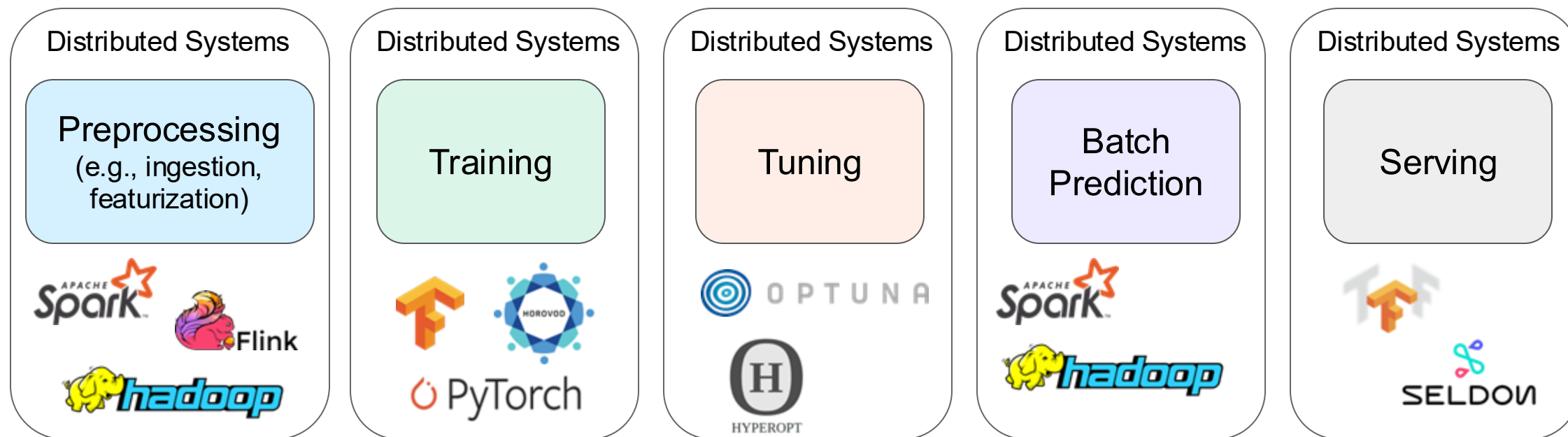
Reinforcement learning (~2015)



Test-time compute (~2024)



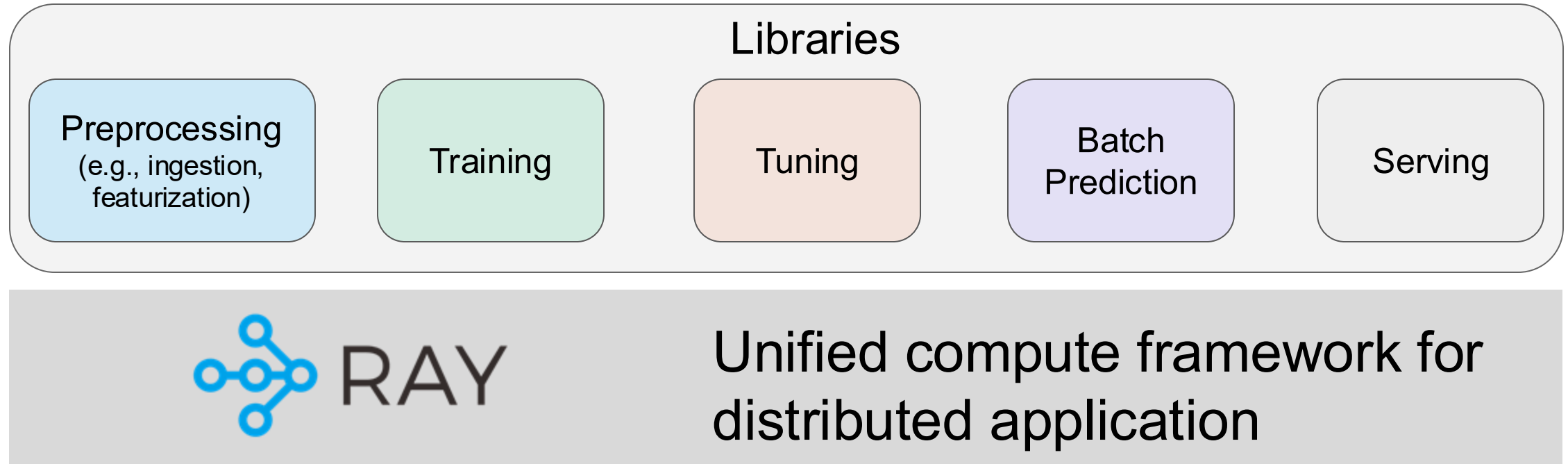
Challenge: need to scale every component!



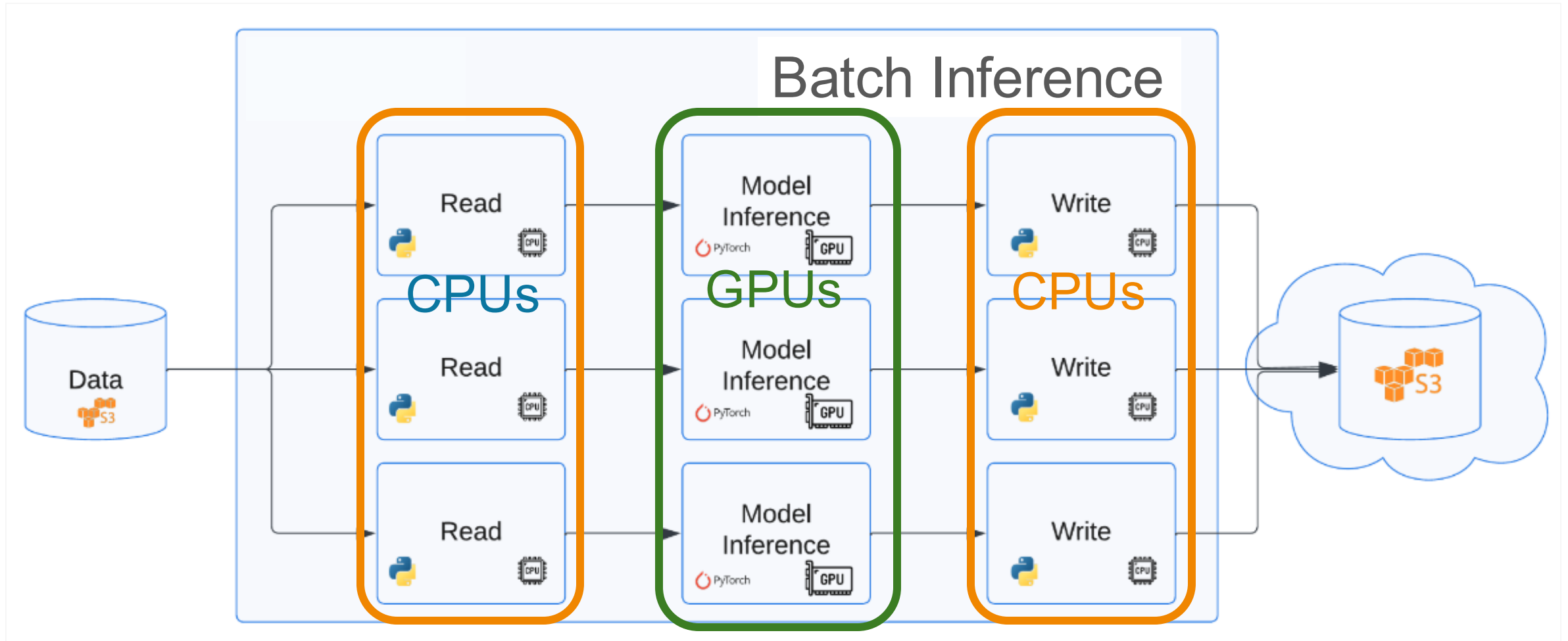
Need to stitch together a bunch of disparate systems

- Hard to develop
- Hard to deploy & manage
- Hard to use resources efficiently (interleave different stages)
- Slow

Ray: **One** system to support all these workloads

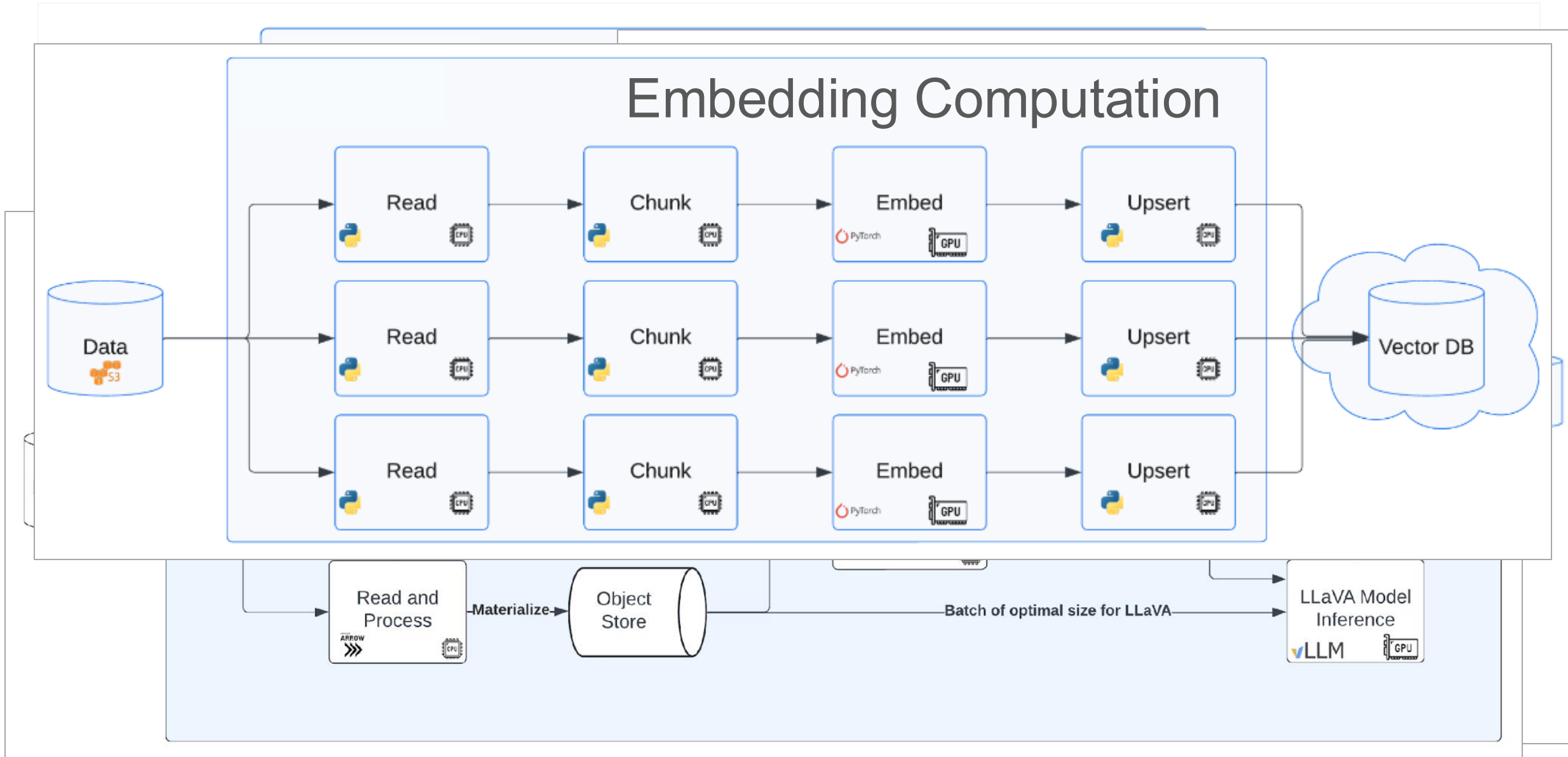


Workload examples



Exploiting CPU/GPU heterogeneity can reduce cost by 10x !

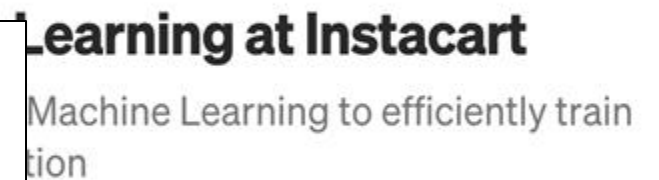
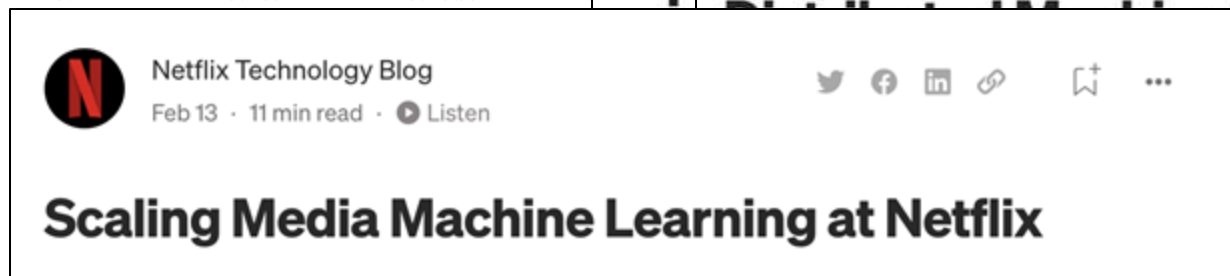
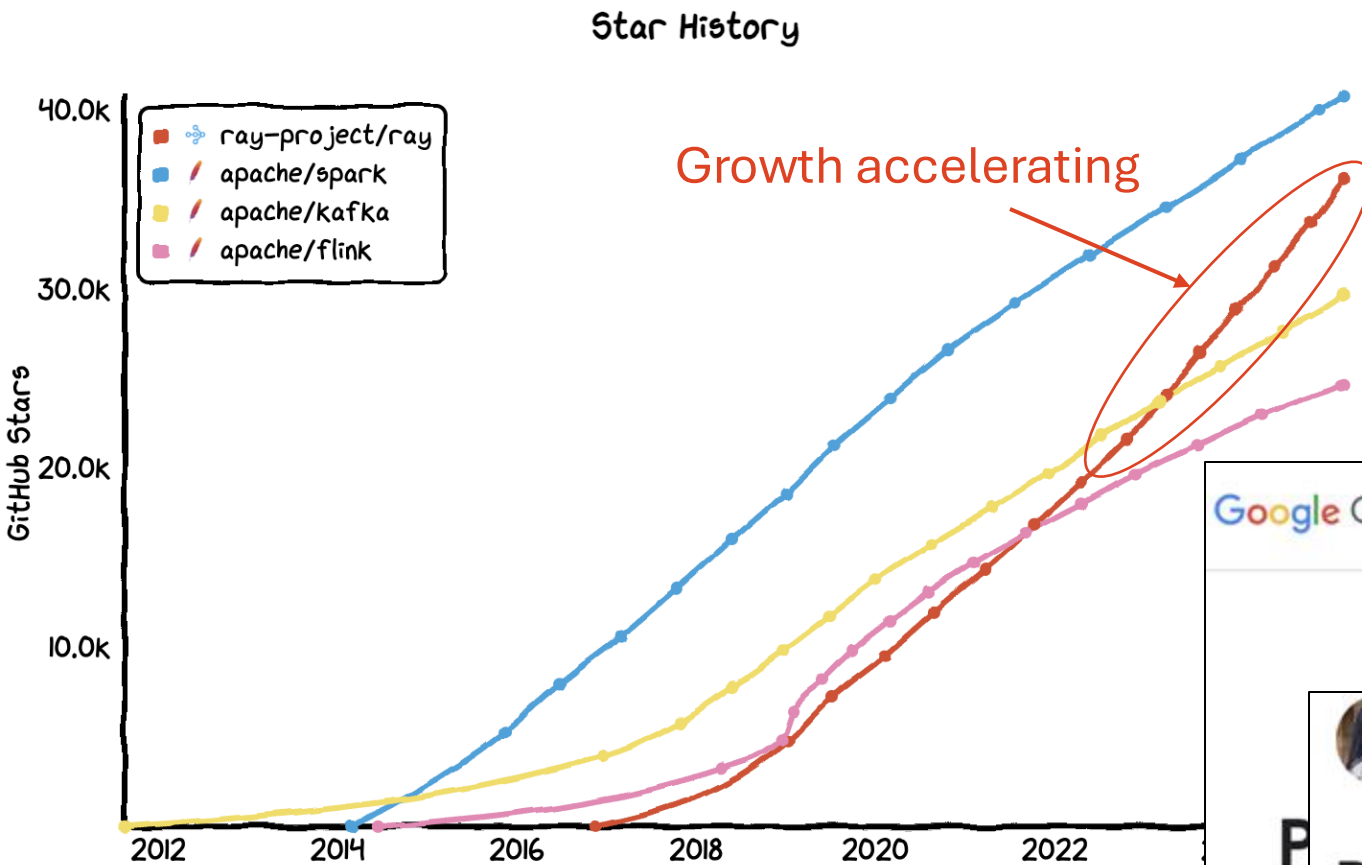
Workload examples



Ray dominates OSS post-training

Framework	Training backend / Algorithms	Serving engine	Orchestrator
TRL (Hugging Face)	Hugging Face Trainer / SFT, DPO, GRPO, PPO, others	Hugging Face, vLLM	—
Verl (ByteDance)	FSDP, DeepSpeed, Megatron / SFT, DPO, GRPO, PPO	vLLM, SGLan	Ray
OpenRLHF	DeepSpeed / SFT, DPO, GRPO, PPO, others	Hugging Face, vLLM	Ray
RAGEN	Verl backend / GRPO, PPO	Hugging Face, vLLM, SGLang	Ray
AReal (Ant Group)	DeepSpeed, Megatron / GRPO, PPO	vLLM, SGLang	Ray
Verifiers	Hugging Face Trainer / GRPO	vLLM, OpenAI	—
ROLL (Alibaba)	DeepSpeed, Megatron / GRPO, PPO, others	vLLM, SGLang	Ray
NeMo-RL (Nvidia)	FSDP, Megatron / SFT, DPO, GRPO	vLLM	Ray
SkyRL (UC Berkeley)	FSDP, DeepSpeed / GRPO, PPO	vLLM, SGLang, OpenAI	Ray

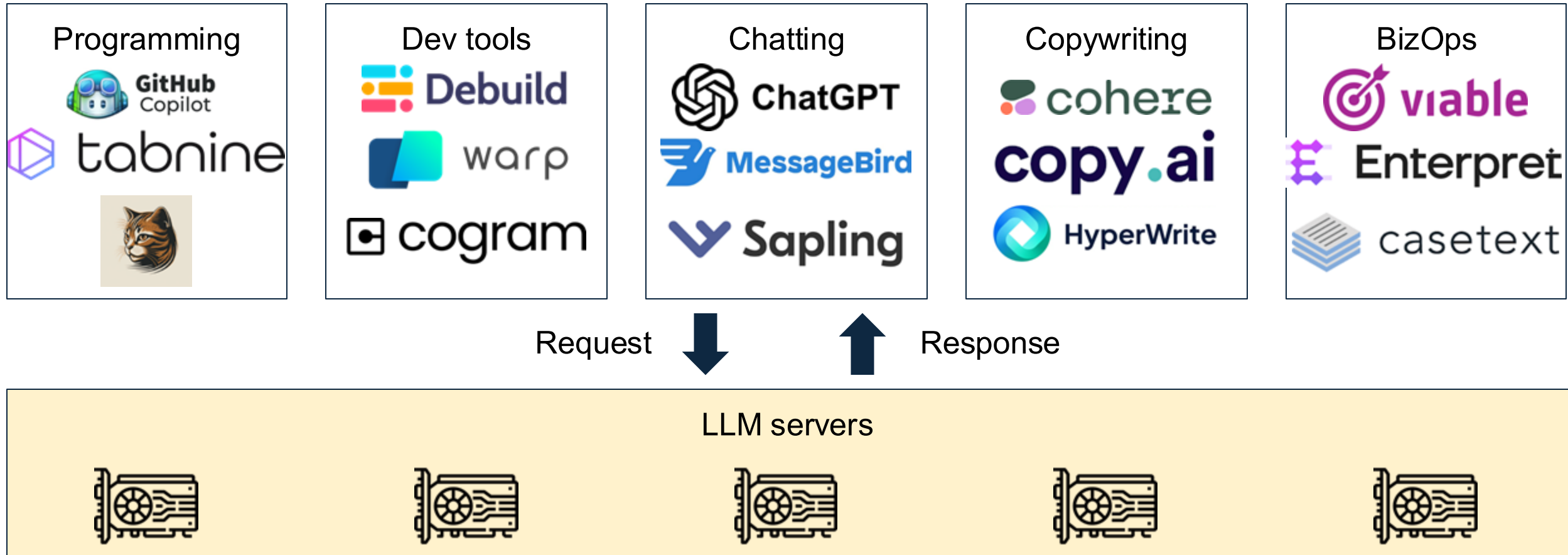
Impact: popular compute framework for scaling AI



vLLM (2023 -)



Trend: Growing number of LLM-powered services



Challenge: Serving LLMs very expensive (2023)

LLMs run on high-end GPUs such as NVIDIA A100 and H100

Each GPU can only serve a handful of requests/sec

- For LLaMA-13B, one A100 can process just a few requests/sec

Huge # of GPUs are required for production-scale LLM services



Inference on LLMs is slow (frustratingly high latency), expensive (multiple GPUs or TPUs), & engineering intensive (requires specialized skills to do it well)



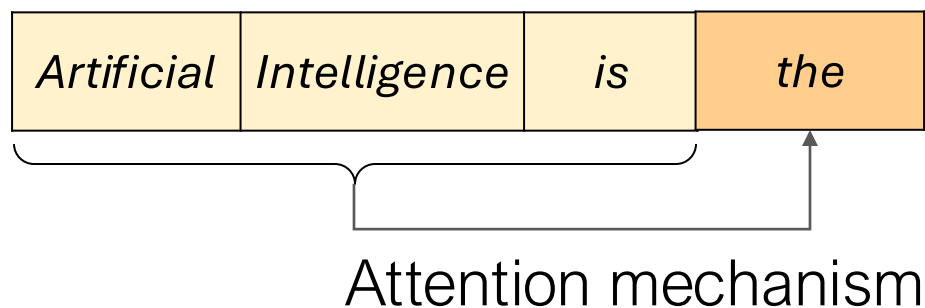
Join

Is local LLM cheaper than ChatGPT API?

ChatGPT api only costs 0.002 dollar for 1k token. I found that LLMs like llama output only 10-20 tokens per second, which is very slow. And such machines costs over 1 dollar per hour. It seems that using api is much cheaper. Based on these observations, it seems that utilizing the ChatGPT API might be a more affordable option.

Low throughput → high cost

Generating one token at a time doesn't use GPU parallelism



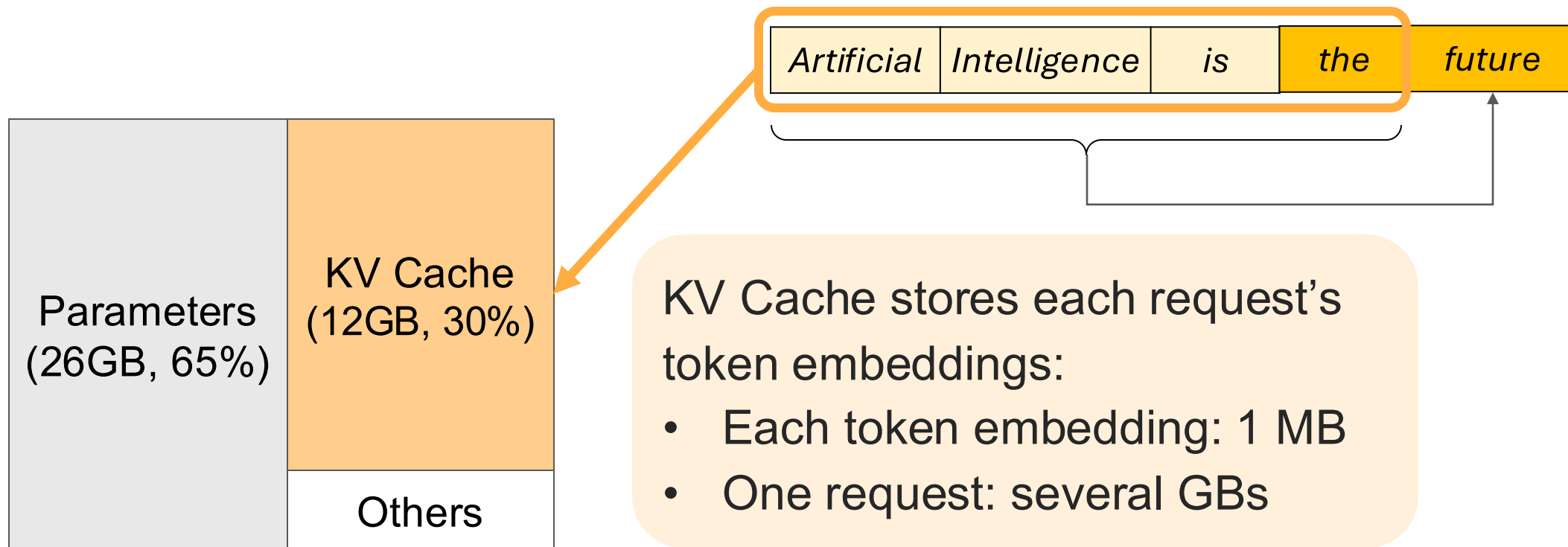
Low throughput → high cost

Generating one token at a time doesn't use GPU parallelism

Solution: process multiple requests at the same time (batch)

Challenge: run out of memory!

Where does the memory go?



13B LLM on A100-40GB

Memory management in previous systems

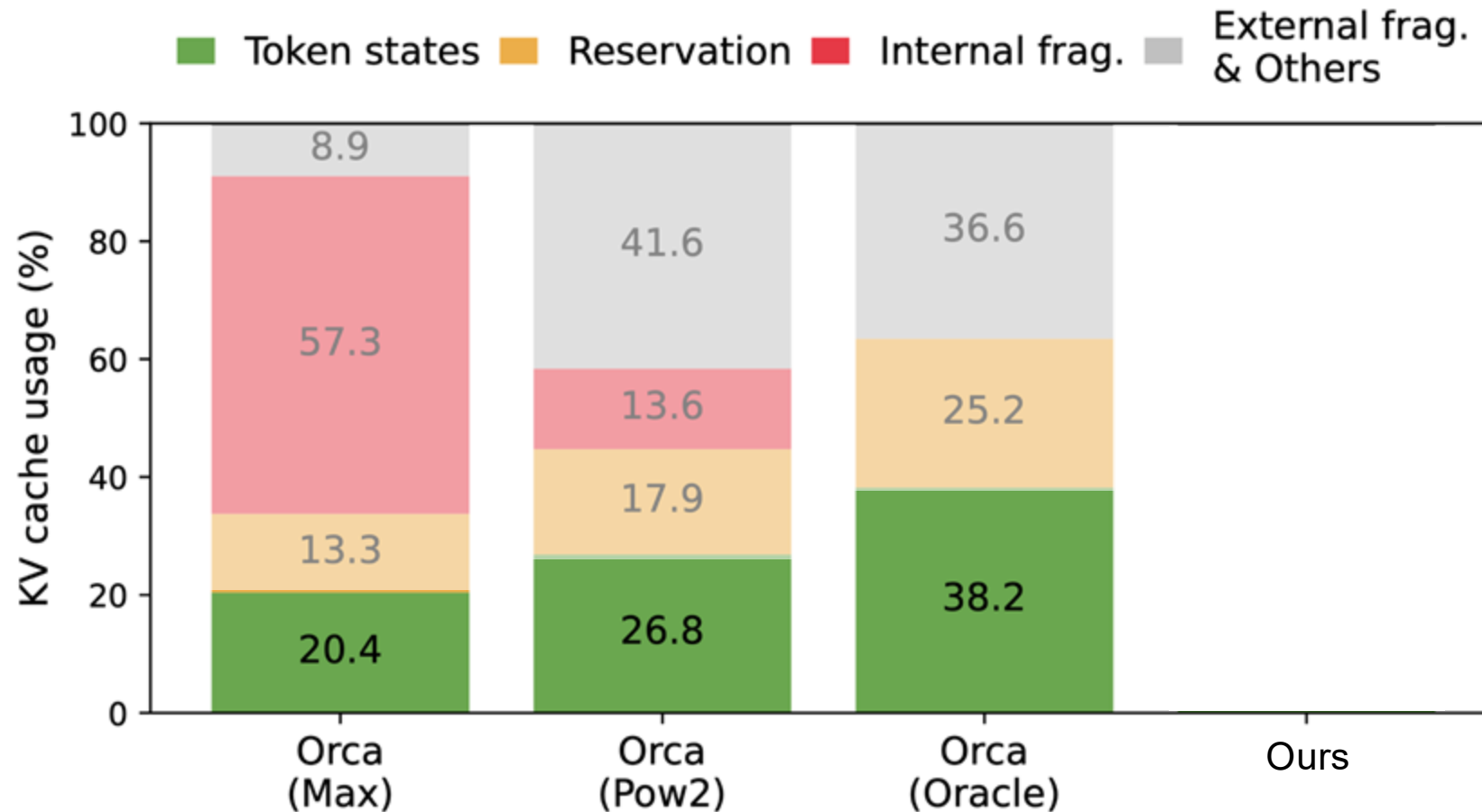
Store a sequence's token states in **contiguous** memory space

- Convention in traditional DNN workloads

Pre-allocate (reserve) the space for future token states, to handle unknown output lengths

Significant memory waste

Only 20–40% of KV cache is utilized to store token states

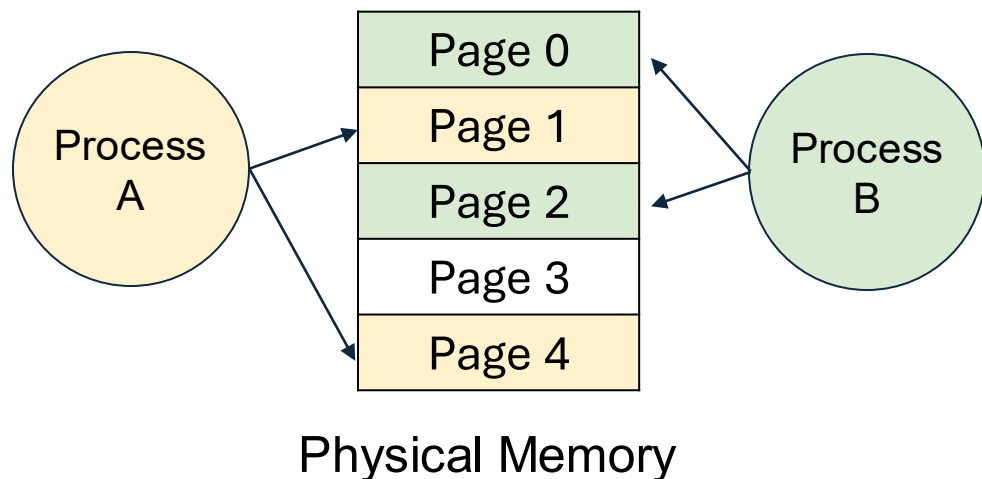


* Yu et al. "Orca: A Distributed Serving System for Transformer-Based Generative Models" (OSDI 22).

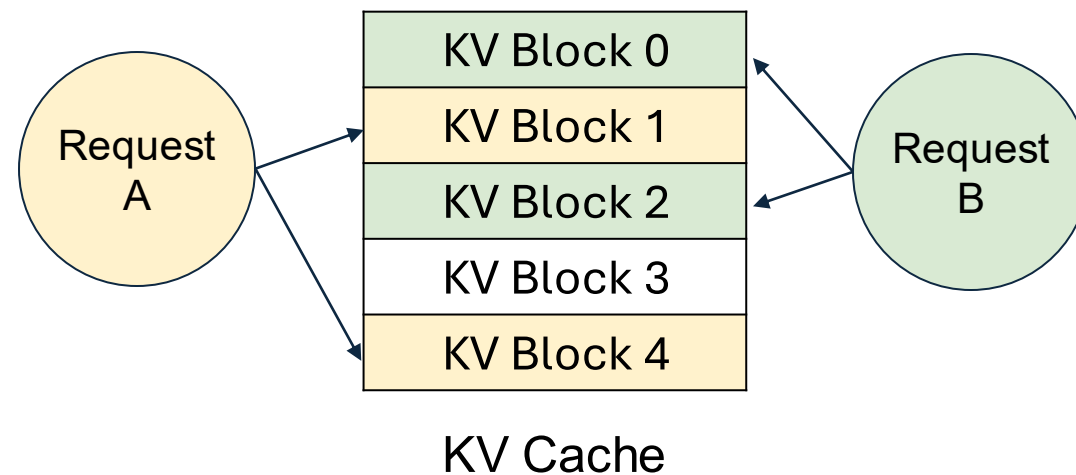
Key Idea: PagedAttention

Inspired by **virtual memory** and **paging**

Memory management in OS



Memory management in vLLM

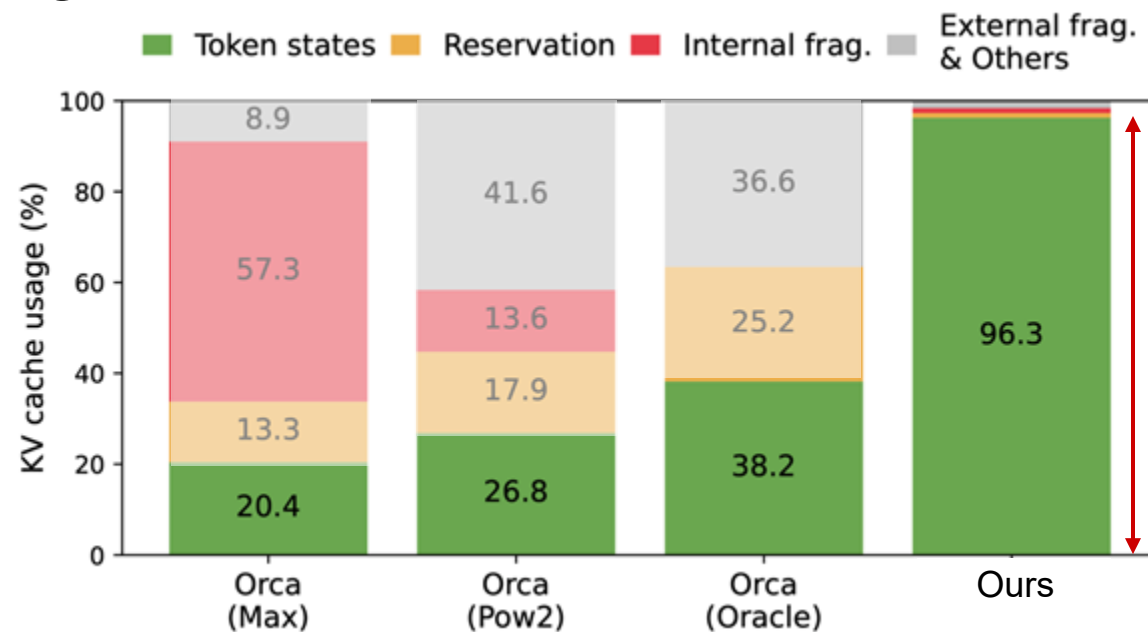


PagedAttention: Memory Efficiency

- Minimal internal fragmentation
 - Only happens at the last block of a sequence
 - $\# \text{ wasted tokens} / \text{seq} < \text{block size}$
 - Sequence: $O(100) - O(1000)$ tokens
 - Block size: 16 or 32 tokens
- No external fragmentation

Alan	Turing	is	a
computer	scientist	and	mathemati cian
renowned			

Internal fragmentation

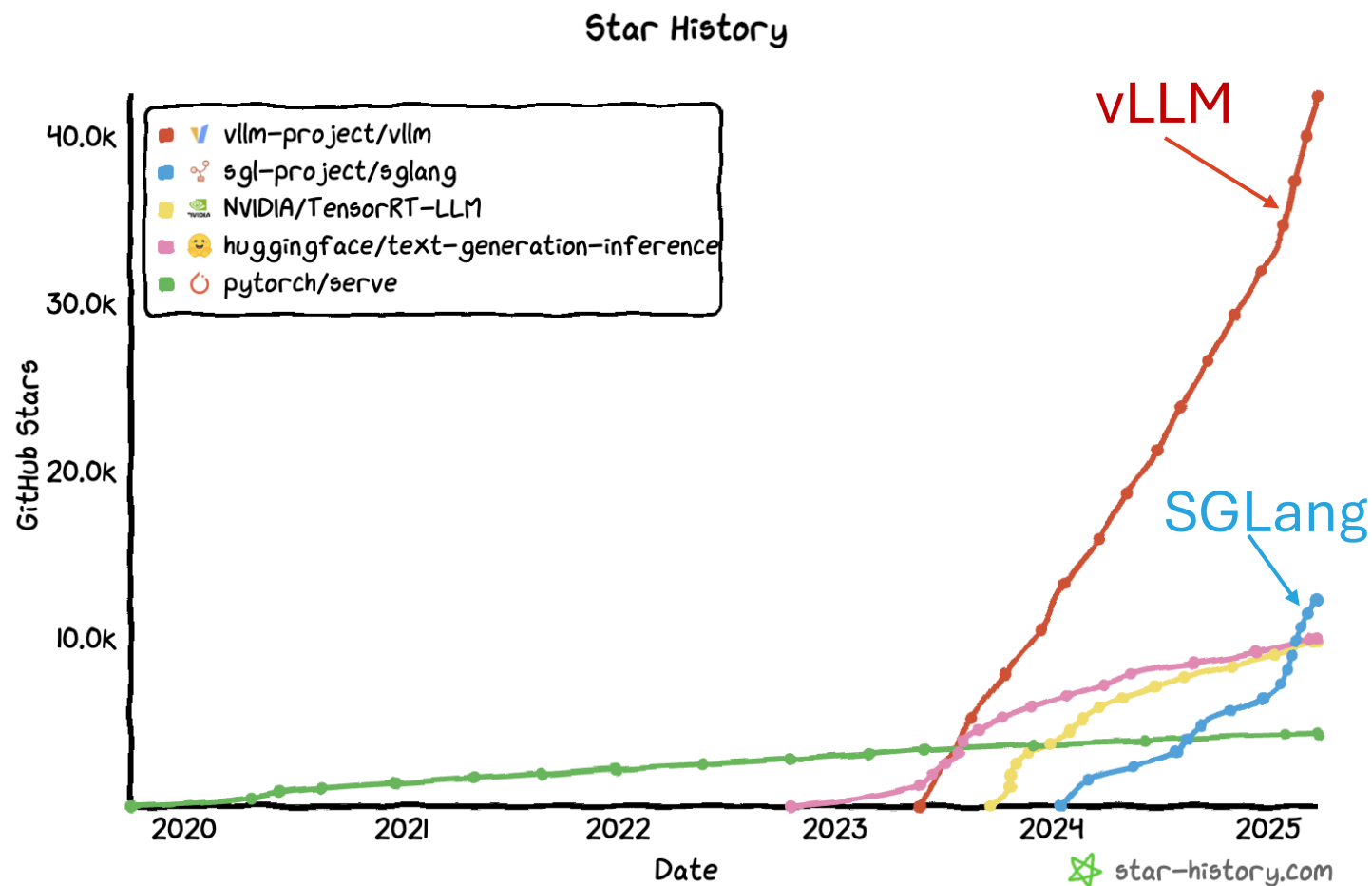


2.5-5x improvement

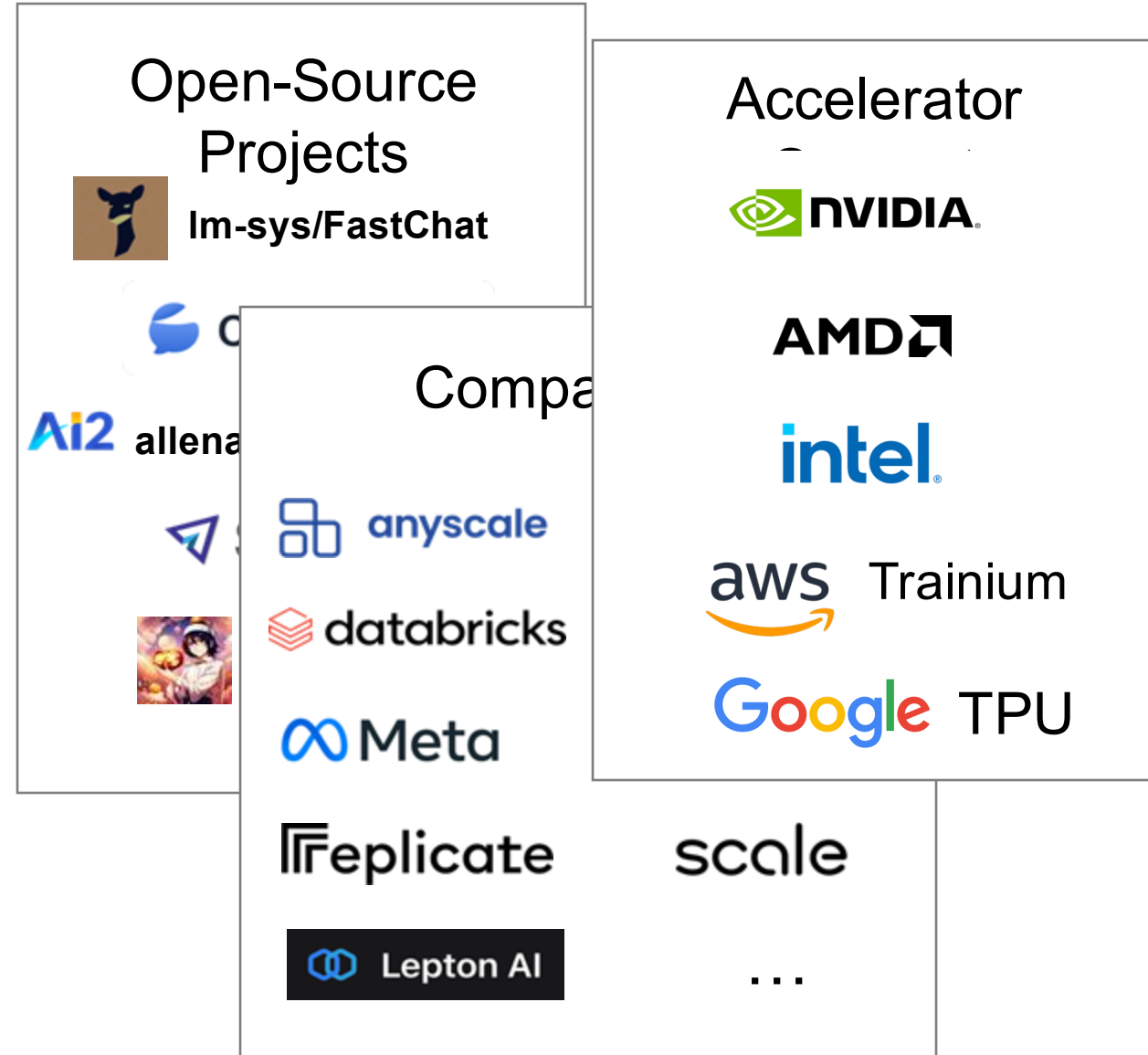
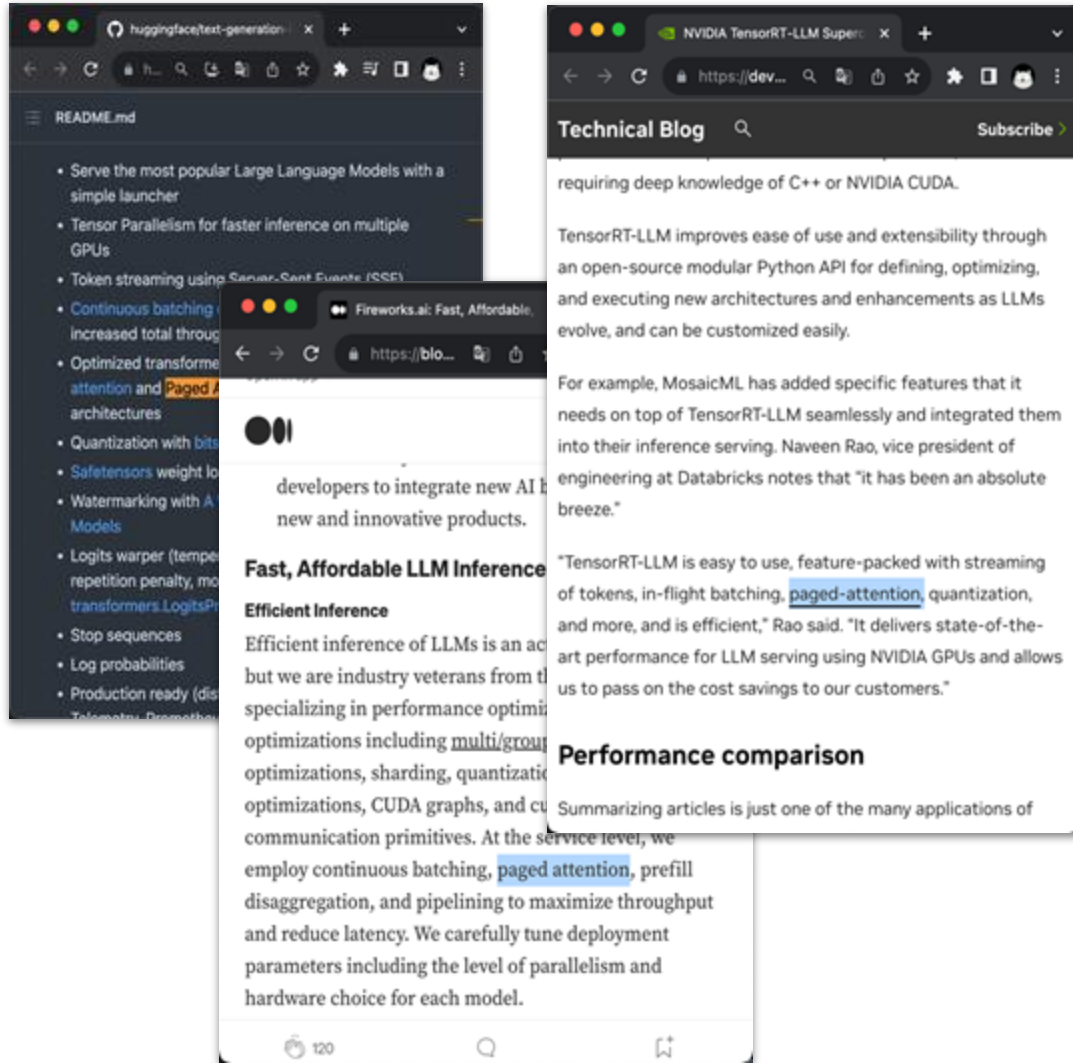
Impact: Most popular OSS LLM inference engines

PagedAttention → vLLM

RadixAttention → SGLang



Impact: widely used across industry



ChatBot Arena: Rethinking LLM Evaluation (2023-)



*“For better or worse,
benchmarks shape a field”*

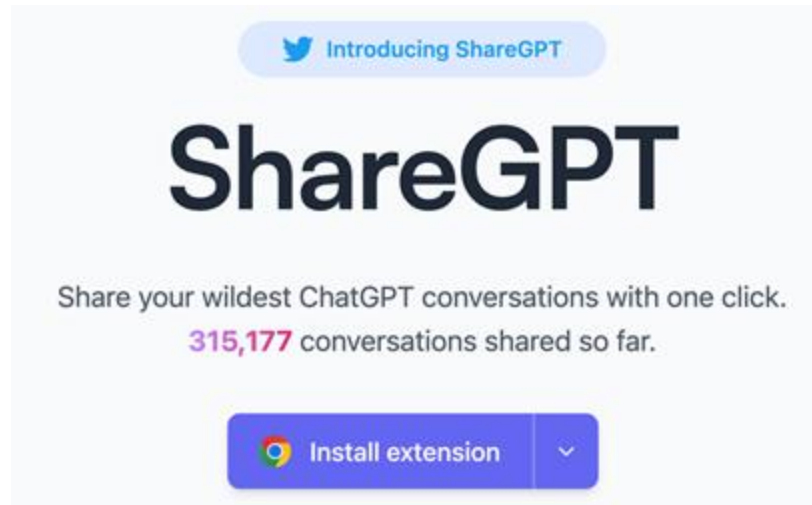
- David Patterson, Turing Award 2017



The beginning

One month after Meta released LLaMa in February 2023...

... we released Vicuna in March
fine tuned LLaMa using SharedGPT data
(70K conversations)



How to evaluate Vicuna?

Many benchmarks, e.g., MMLU, HELM, ...

But they

- are static → prone to contamination
- do not capture human preference

Contamination



Horace He
@cHHillee

...

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

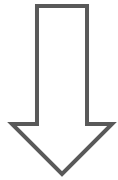
This strongly points to contamination.

1/4

g's Race	implementation, math	🚩	🌟		greedy, implementation	🚩	🌟
nd Chocolate	implementation, math	🚩	🌟	Cat?	implementation, strings	🚩	🌟
triangle!	brute force, geometry, math	🚩	🌟	Actions	data structures, greedy, implementation, math	🚩	🌟
	greedy, implementation, math	🚩	🌟	Interview Problem	brute force, implementation, strings	🚩	🌟
lumbers	brute force	🚩	🌟	yers	brute force, implementation, strings	🚩	🌟
line Line	implementation	🚩	🌟	nd Suffix Array	strings	🚩	🌟
r or Stairs?	implementation	🚩	🌟	ther Promotion	greedy, math	🚩	🌟
Loves 3 I	math	🚩	🌟	iForces	greedy, sortings	🚩	🌟
s	implementation, math	🚩	🌟	1 and Append	implementation, two pointers	🚩	🌟
	greedy, implementation, sortings	🚩	🌟	1g Directions	geometry, implementation	🚩	🌟

Need human evaluation

Get some students to evaluate it



Slow, does not scale!



User

Develop a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences

Which one is better?

User

Develop a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences.

Assistant A

Sure, here's a Python program that should accomplish what you're looking for:

```
import os
from collections import Counter

def get_top_5_words(directory):
    # Create a list of all text files in the directory
    file_list = [f for f in os.listdir(directory) if f.endswith('.txt')]

    # Initialize a dictionary to store word counts
    word_counts = {}

    # Loop through each text file and count the number of words
    for file in file_list:
        with open(os.path.join(directory, file), 'r') as f:
            file_contents = f.read()
            words = file_contents.split()
            for word in words:
                if word not in word_counts:
                    word_counts[word] = 1
                else:
                    word_counts[word] += 1

    # Sort the word counts in descending order
    sorted_word_counts = sorted(word_counts.items(), key=lambda x: x[1], reverse=True)

    # Return the top 5 words with the most number of occurrences
    return [word for word, count in sorted_word_counts[:5]]

# Call the function and print the top 5 words
directory = '/path/to/directory'
top_words = get_top_5_words(directory)
print(top_words)
```

Assistant B

Sure, here's a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences:

```
import os
import re
from collections import Counter

def get_top_words(directory):
    # Initialize a Counter object to keep track of word occurrences
    word_counter = Counter()

    # Loop through all files in the directory
    for filename in os.listdir(directory):
        # Check if the file is a text file
        if filename.endswith(".txt"):
            # Open the file and read its contents
            with open(os.path.join(directory, filename), "r") as file:
                contents = file.read()

            # Use regular expressions to split the contents into words
            words = re.findall(r'\b\w+\b', contents)

            # Update the word counter with the words from the file
            word_counter.update(words)

    # Return the top 5 words with the most occurrences
    return word_counter.most_common(5)

# Example usage
top_words = get_top_words("path/to/directory")
print(top_words)
```

This program uses the `os` module to loop through all files in the specified directory, and the `re` module to

User

Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs of each stage

Which one is better?

User

Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage?

Assistant A

Of course, I'd be happy to help! Photosynthesis is indeed a crucial process that supports life on Earth, and it occurs in specialized organelles called chloroplasts found in plant cells. The two main stages of photosynthesis are the light-dependent reactions and the light-independent reactions, also known as the Calvin cycle.

1. Light-dependent reactions: These reactions take place in the thylakoid membranes of the chloroplast and involve the conversion of light energy into chemical energy. The primary input for this stage is light, which is absorbed by pigments such as chlorophyll and other accessory pigments. The energy from light is used to generate ATP and NADPH, two energy-rich molecules that are essential for the light-independent reactions. The light-dependent reactions also produce oxygen as a byproduct.

2. Light-independent reactions (Calvin cycle): These reactions occur in the stroma of the chloroplast and are responsible for fixing CO₂ into organic molecules using the energy and reducing power generated during the light-dependent reactions. The primary input for this stage is CO₂, which enters the leaf through the stomata. The Calvin cycle produces glucose, a simple sugar that can be used by the plant for energy and growth. Oxygen is also produced as a byproduct of this stage.

Here's a summary of the inputs and outputs for each stage:

Light-dependent reactions:

- Inputs: Light, H₂O, CO₂
- Outputs: ATP, NADPH, O₂

Light-independent reactions (Calvin cycle):

- Inputs: CO₂, ATP, NADPH
- Outputs: Glucose, O₂

I hope this helps! Let me know if you have any further questions.

Assistant B

Sure, I'd be happy to help!

The two main stages of photosynthesis are the light-dependent reactions and the light-independent reactions (also known as the Calvin cycle).

The light-dependent reactions take place in the thylakoid membranes of the chloroplast. The primary inputs for this stage are light energy, water, and ADP (adenosine diphosphate) + Pi (inorganic phosphate). The primary outputs are oxygen, ATP (adenosine triphosphate), and NADPH (nicotinamide adenine dinucleotide phosphate).

During the light-dependent reactions, light energy is absorbed by pigments in the thylakoid membranes, which excites electrons and generates ATP and NADPH. Water is also split into oxygen and hydrogen ions, which contribute to the formation of ATP and NADPH.

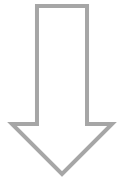
The light-independent reactions (Calvin cycle) take place in the stroma of the chloroplast. The primary inputs for this stage are carbon dioxide, ATP, and NADPH. The primary outputs are glucose (or other sugars) and ADP + Pi.

During the Calvin cycle, carbon dioxide is fixed into organic molecules using the energy from ATP and NADPH generated during the light-dependent reactions. This process ultimately results in the production of glucose (or other sugars) that can be used by the plant for energy or stored for later use.

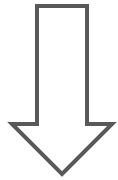
Challenge: how to scale human evaluation?

Key Idea 1: LLM-as-a-Judge

Get some students to evaluate it



Expensive, slow, does not scale!



Use GPT-4 (2 weeks after its release) → LLM-as-a-Judge

Which one is better?

User

Develop a Python program that reads all the

Assistant A

Sure, here's a Python program that should

```
import os
from collections import Counter

def get_top_5_words(directory):
    # Create a list of all text files in the directory
    file_list = [f for f in os.listdir(directory) if f.endswith('.txt')]

    # Initialize a dictionary to store word counts
    word_counts = {}

    # Loop through each text file
    for file in file_list:
        with open(os.path.join(directory, file), 'r') as file:
            file_contents = file.read()
            words = file_contents.split()

            for word in words:
                if word not in word_counts:
                    word_counts[word] = 1
                else:
                    word_counts[word] += 1

    # Sort the word counts in descending order
    sorted_word_counts = sorted(word_counts.items(), key=lambda x: x[1], reverse=True)

    # Return the top 5 words with their counts
    return [word for word, count in sorted_word_counts[:5]]

# Call the function and print the top 5 words
directory = '/path/to/directory'
top_words = get_top_5_words(directory)
print(top_words)
```

Game 1.

Judgment: The reference answer provides a comprehensive Python program that reads all text files in a directory, counts the occurrences of each word, and returns the top 5 words with the most occurrences. It uses the `os` module to list all files in the directory and open each text file, the `re` module to split the file contents into words, and the `Counter` class from the `collections` module to count the occurrences of each word and find the top 5 words.

Assistant A's answer is similar to the reference answer, but it does not handle case sensitivity or punctuation. It splits the file contents into words using the `split()` method, which does not remove punctuation from the words. This could lead to incorrect word counts if the text files contain punctuation. Also, it does not convert the words to lowercase before counting them, which could lead to separate counts for the same word in different cases.

Assistant B's answer is very similar to the reference answer. It uses the `os` module to list all files in the directory and open each text file, the `re` module to split the file contents into words, and the `Counter` class from the `collections` module to count the occurrences of each word and find the top 5 words. However, it does not convert the words to lowercase before counting them, which could lead to separate counts for the same word in different cases.

In conclusion, both Assistant A and Assistant B provided correct and helpful answers, but Assistant B's answer is more accurate because it handles punctuation correctly. However, both assistants failed to handle case sensitivity, which is a minor issue that could be easily fixed. Therefore, the final verdict is: `[[B]]`

This program uses the `os` module to loop through all files in the specified directory, and the `re` module to

Which one is better?

User

Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage?

Assistant A

Of course, I'd be happy to help! Photosynthesis occurs in specialized organelles called chloroplasts. The two main stages of photosynthesis are the light-dependent reactions and the Calvin cycle.

1. Light-dependent reactions: These reactions involve the conversion of light energy into chemical energy. Light is absorbed by pigments such as chlorophyll, which then generate ATP and NADPH, two energy carriers used in the Calvin cycle. The light-dependent reactions take place in the thylakoid membranes of the chloroplast.

2. Light-independent reactions (Calvin cycle): These reactions are responsible for fixing CO₂ into glucose. They take place in the stroma of the chloroplast. The Calvin cycle uses the ATP and NADPH produced in the light-dependent reactions. Stomata on the leaf surface allow CO₂ to enter and O₂ to exit. The Calvin cycle produces glucose and releases oxygen as a byproduct.

Here's a summary of the inputs and outputs for each stage:

Light-dependent reactions:

- Inputs: Light, H₂O, CO₂
- Outputs: ATP, NADPH, O₂

Light-independent reactions (Calvin cycle):

- Inputs: CO₂, ATP, NADPH
- Outputs: Glucose, O₂

I hope this helps! Let me know if you have any further questions.

Game 1.

Judgment: Both Assistant A and Assistant B provided accurate and detailed responses to the user's question about the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage. They both correctly identified the light-dependent reactions and the light-independent reactions (Calvin cycle) as the two main stages of photosynthesis, and correctly described where these stages take place within the chloroplast. They also correctly identified the primary inputs and outputs for each stage.

However, Assistant A made a mistake by stating that CO₂ is an input for the light-dependent reactions and that O₂ is an output of the Calvin cycle. In reality, CO₂ is an input for the Calvin cycle, and O₂ is an output of the light-dependent reactions. Assistant B correctly identified the inputs and outputs for each stage.

Therefore, Assistant B provided a more accurate response to the user's question.

Final verdict: [[B]]

the light-independent
chloroplast. The primary inputs
inorganic phosphate). The
dinamide adenine dinucleotide
the thylakoid membranes,
oxygen and hydrogen ions,
chloroplast. The primary
are glucose (or other sugars)
the energy from ATP and
ely results in the production of
r later use.

But how does GPT-4 compare with humans?

Still need a way to scale human evaluation!

Key Idea 2: ELO-like rating for Human Eval

PIF ATP RANKINGS

Singles Doubles Race To Turin Doubles Race Race to Jeddah

Live ☐ Top 100 All Countries 2025.0

Rank	Player	Age	Official Points
1	 Jannik Sinner	23	9,730
2	 Alexander Zverev	28	8,085
3	 Carlos Alcaraz	22	7,850
4	 Taylor Fritz	27	4,815
5	 Jack Draper	23	4,440
6	 Novak Djokovic	37	4,130

Last update: 12 May 2025,

Classical	Rapid	Blitz	Juniors						
#	▲	↑↓	◆	Name	◆	Rating	◆	+/-	◆
1				Carlsen, Magnus		2837.0		0.0	
2				Nakamura, Hikaru		2804.0		0.0	
3		↑1		Erigaisi Arjun		2782.0		0.0	
4		↑1		Caruana, Fabiano		2779.5		+3.5	
5		👑		Gukesh D		2778.2		-8.8	
6				Abdusattorov, Nodirbek		2764.4		-6.6	
7				Praggnanandhaa R		2762.2		+4.2	
8				Wei, Yi		2758.0		0.0	
9				Nepomniachtchi, Ian		2757.0		0.0	
10		↓1		Firouzja, Alireza		2756.3		-0.7	

ChatBot Arena: Benchmarking LLMs in the Wild

Benchmark platform for LLMs that features anonymous, randomized battles in a crowdsourced manner

Model A

Correct grammar: I are happy.

I'm happy to help you with your English grammar. Can you please provide me with a sentence or two that you would like me to correct?

Model B

Correct grammar: I are happy.

The correct grammar would be "I am happy."

👉 A is better

👉 B is better

👉 Tie

👉 Both are bad

<https://arena.lmsys.org>

Leaderboard

Use Bardley-Terry model

Category		Apply filter		Overall Questions			
Overall		☐ Style Control	☐ Show Deprecated	#models: 219 (100%) #votes: 2,811,226 (100%)			
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1443	+8/-13	2540	Google	Proprietary
2	4	Grok-3-Preview-02-24	1404	+5/-6	10398	xAI	Proprietary
2	2	GPT-4.5-Preview	1398	+7/-6	10615	OpenAI	Proprietary
4	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1381	+4/-3	22659	Google	Proprietary
4	4	Gemini-2.0-Pro-Exp-02-05	1380	+5/-5	20293	Google	Proprietary
4	3	ChatGPT-4o-latest (2025-01-29)	1374	+5/-4	22517	OpenAI	Proprietary
7	5	DeepSeek-R1	1360	+5/-5	12772	DeepSeek	MIT
7	12	Gemini-2.0-Flash-001	1355	+4/-4	18327	Google	Proprietary
7	4	o1-2024-12-17	1351	+4/-4	25044	OpenAI	Proprietary
10	12	Qwen2.5-Max	1340	+5/-3	17124	Alibaba	Proprietary
10	12	Gemma-3-27B-it	1340	+7/-6	6974	Google	Gemma
10	9	o1-preview	1335	+4/-4	33184	OpenAI	Proprietary
12	12	o3-mini-high	1326	+6/-5	14274	OpenAI	Proprietary
13	15	DeepSeek-V3	1318	+5/-4	22845	DeepSeek	DeepSeek
13	19	QwQ-32B	1315	+9/-8	4050	Alibaba	Apache 2.0
14	16	Command A (03-2025)	1311	+9/-8	3415	Cohere	CC-BY-NC-4.0
14	20	GLM-4-Plus-0111	1310	+6/-8	6034	Zhipu	Proprietary
14	18	Qwen-Plus-0125	1310	+7/-7	6059	Alibaba	Proprietary

<https://lmarena.ai/?leaderboard>

Leaderboard

Category		Apply filter		Overall Questions			
Overall		☐ Style Control ☐ Show Deprecated		#models: 219 (100%) #votes: 2,811,226 (100%)			
16	✓ Overall			Arena Score	95% CI	Votes	Organization License
	Math						
	Instruction Following						
	Multi-Turn						
	Creative Writing						
	Coding						
	Hard Prompts						
	Hard Prompts (English)						
	Longer Query						
	English						
	Chinese						
	French						
	German						
	Spanish						
	Russian						
	Japanese						
	Korean						
	Exclude Ties						
	Exclude Short Query (< 5 tokens)						
	Exclude Refusal						
14	16						
14	20						
14	18						

<https://lmarena.ai/?leaderboard>

Impact

2M+ monthly users

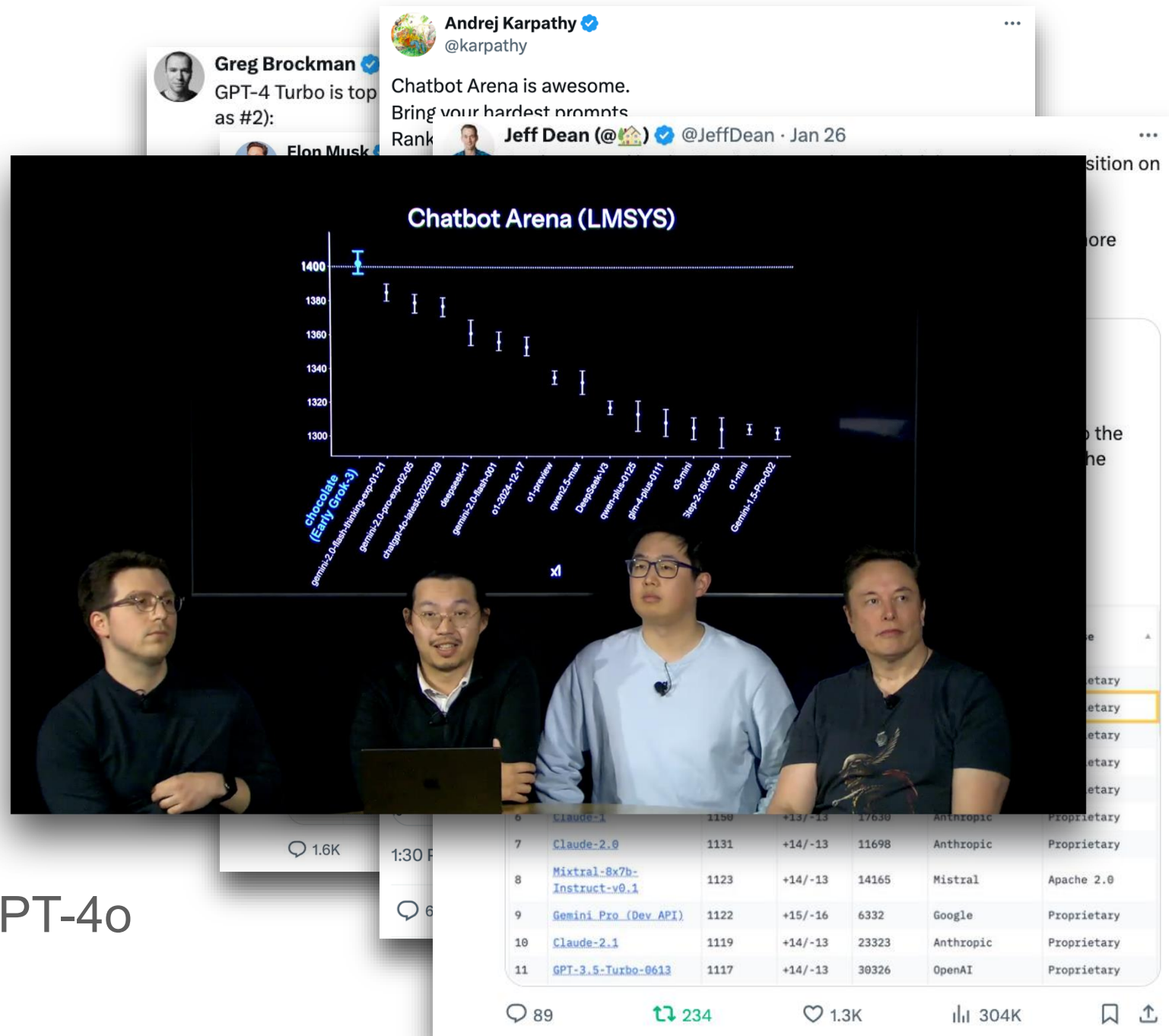
300M+ user queries

4M+ votes

600+ models

Private testing

Gemini, O1, O1-mini, GPT-4o



Summary

The open-source is the future of AI/LLM stack

Projects:

- Apache Spark: scale data processing including classic ML
- Ray: scale and productize AI workloads
- vLLM: high throughput LLM inference engine
- Chatbot Arena: how to evaluate LLMs?

Just scratching the surface: many more challenges at the intersection between systems & AI/LLMs

Some thoughts about open-source

Some say AI is going to be bigger than industrial revolution or the Internet

But this revolution is very different than the Internet

- Internet was the results of a three-way partnership: government, academia, and industry → enabled everyone to work together!
- AI happens in silos in the US: each company developing its own proprietary model (except Meta); academia and government are (have been) minor players

Some thoughts about open-source

If we want to speed up AI progress in a responsible way, we need to enable all AI researchers to collaborate

- But they cannot do that without shared artifacts, infrastructure

This means fully open-source and shared infrastructure

- Fully open-source: open data, open code, open weights, open evals
- Shared infrastructure: virtually everyone should access to it to develop fully open-source models

Thank you