

Machine Learning

for CSC196

Chris J. Maddison, October 2025

About Me

- UofT Bachelor's and Master's in this department
- Began my research career in Geoff Hinton's lab in 2011
- Co-founded the AlphaGo project that was the first computer agent to defeat a world master in the game of Go
- Joined UofT faculty in 2020

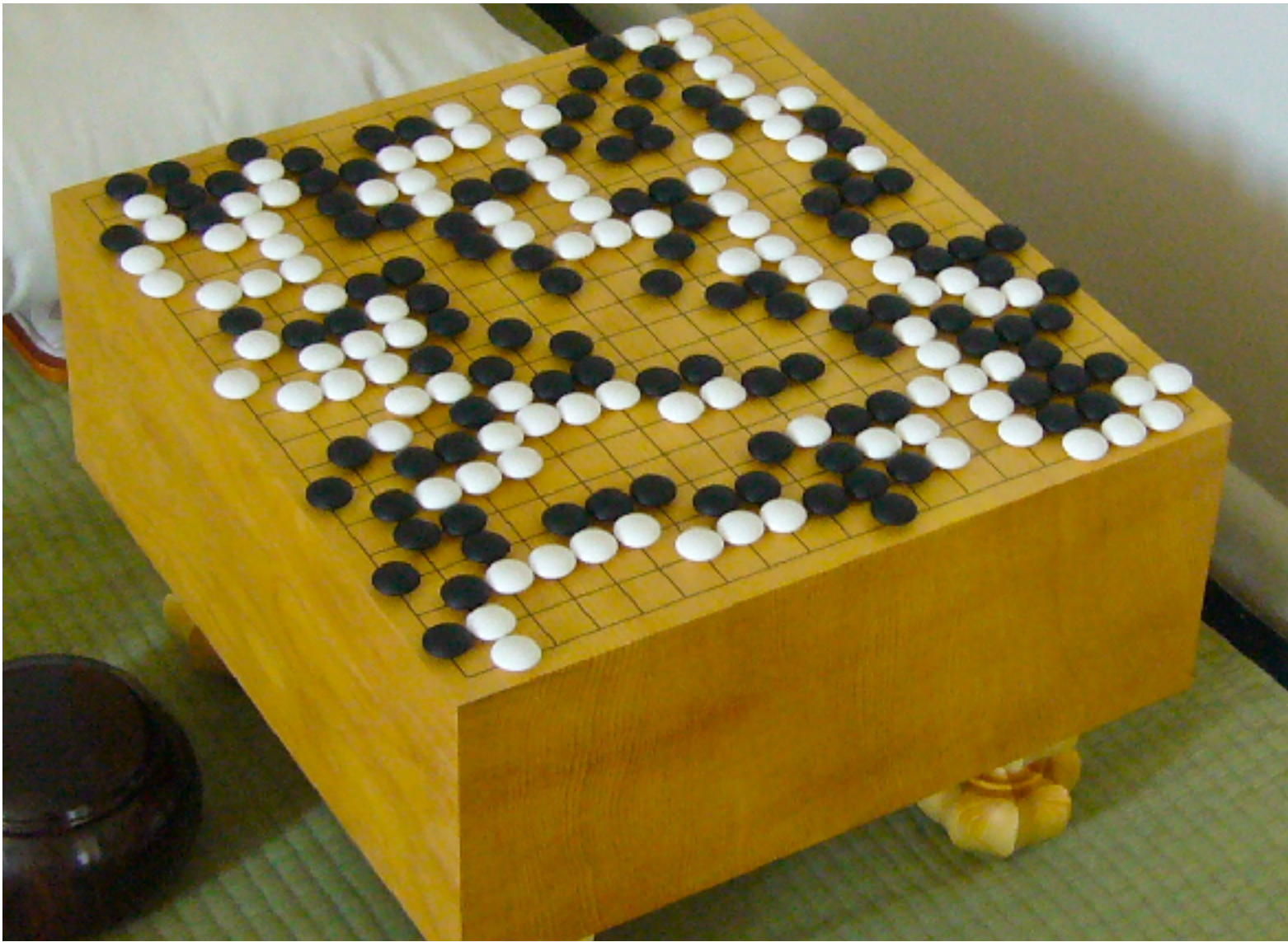
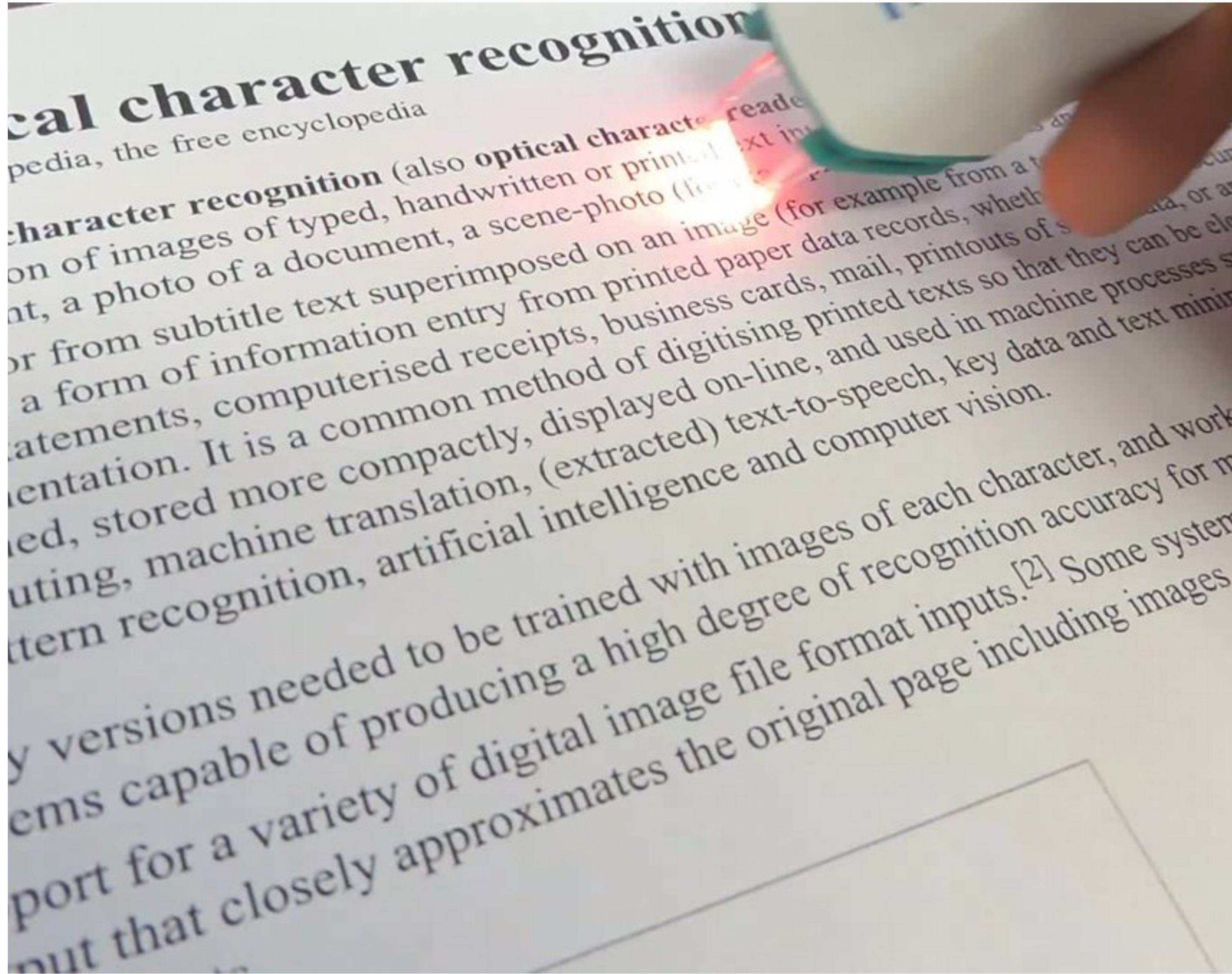
Agenda

- **What is machine learning?**
- What are large language models?
- What is driving the AI boom?

Computer Science and machine learning

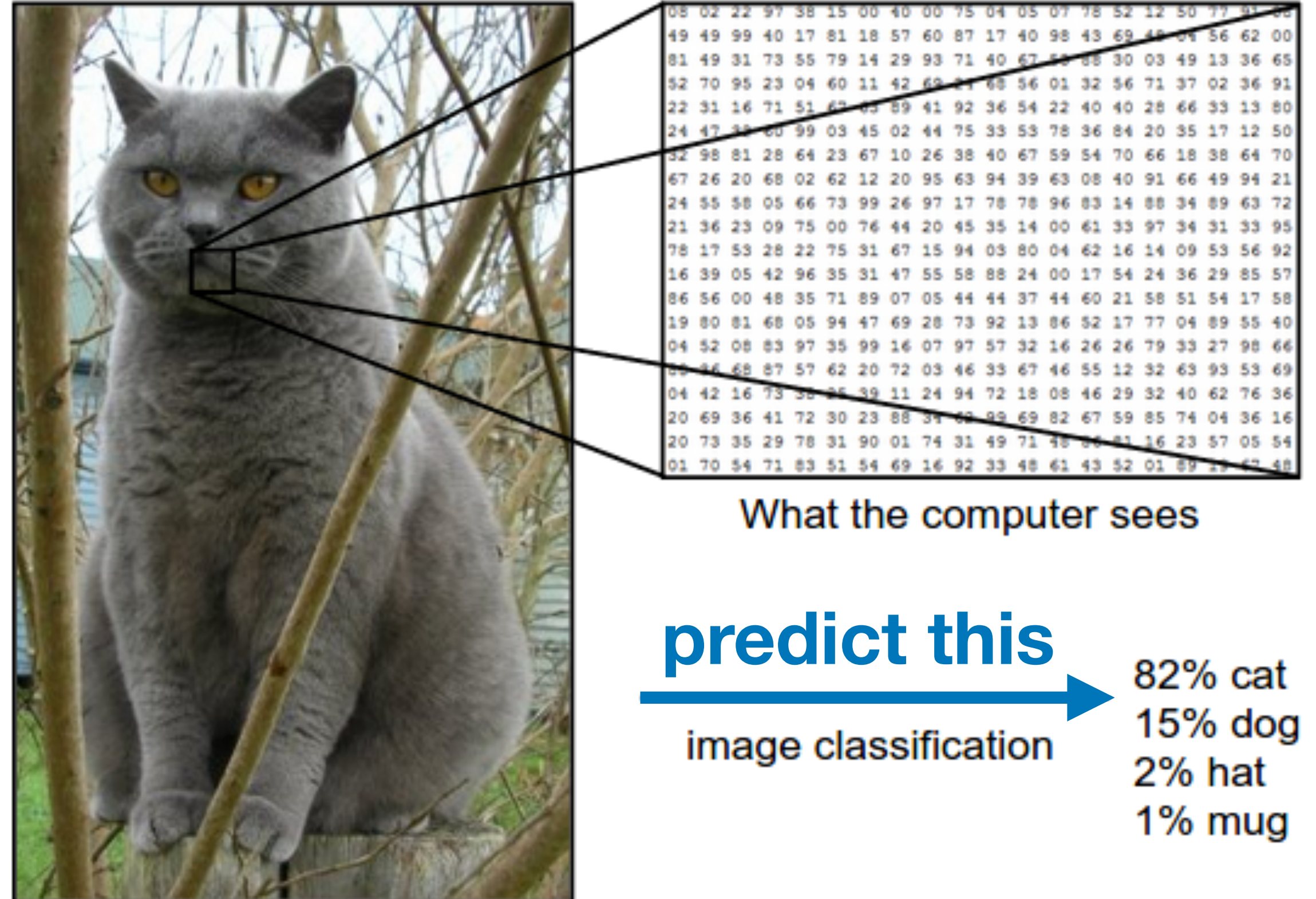
- CS: study problems whose solutions we can automate through systematic, mechanical means.
- What do we do if we don't know how to design the solution?
- ML: study ways to **learn solutions from examples or from experience.**





Learning to predict from examples starting from measurements

- Examples or experience are captured by measurement.
 - *E.g.*, silver crystals in film determining light intensity and your friend classifying an image.
- **Measurements are stored as data.**
- **Learn from the data to predict the solution.**
 - *E.g.*, image classification.

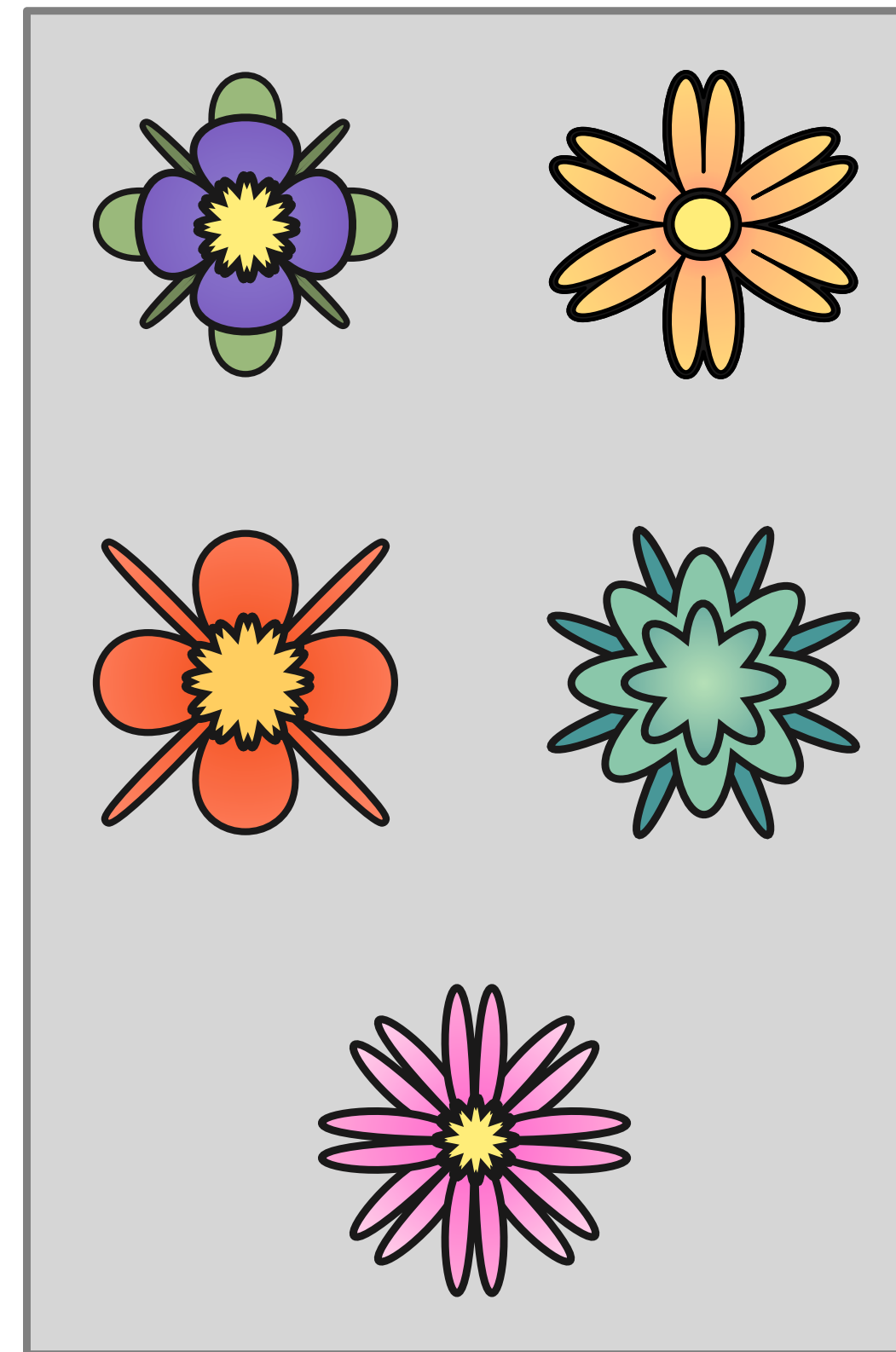


A Silly Example

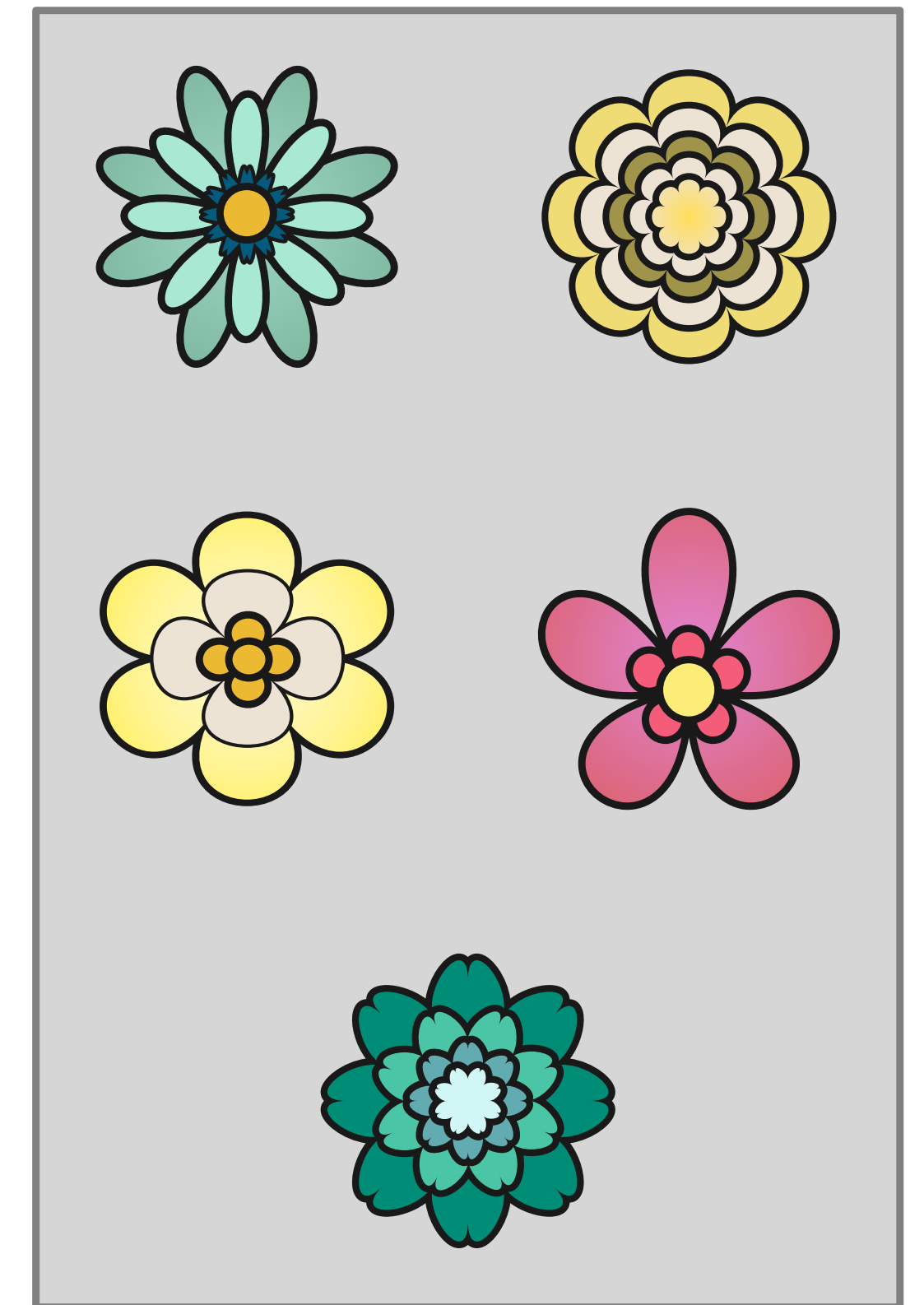
Setup

- Want to detect whether a flower is **pointy** or **round**.
- Given 10 example flowers, labelled by your botanist friend.
- To **classify a new flower**, we **want an automated procedure** that doesn't rely on our friend to label it.

pointy-petalled



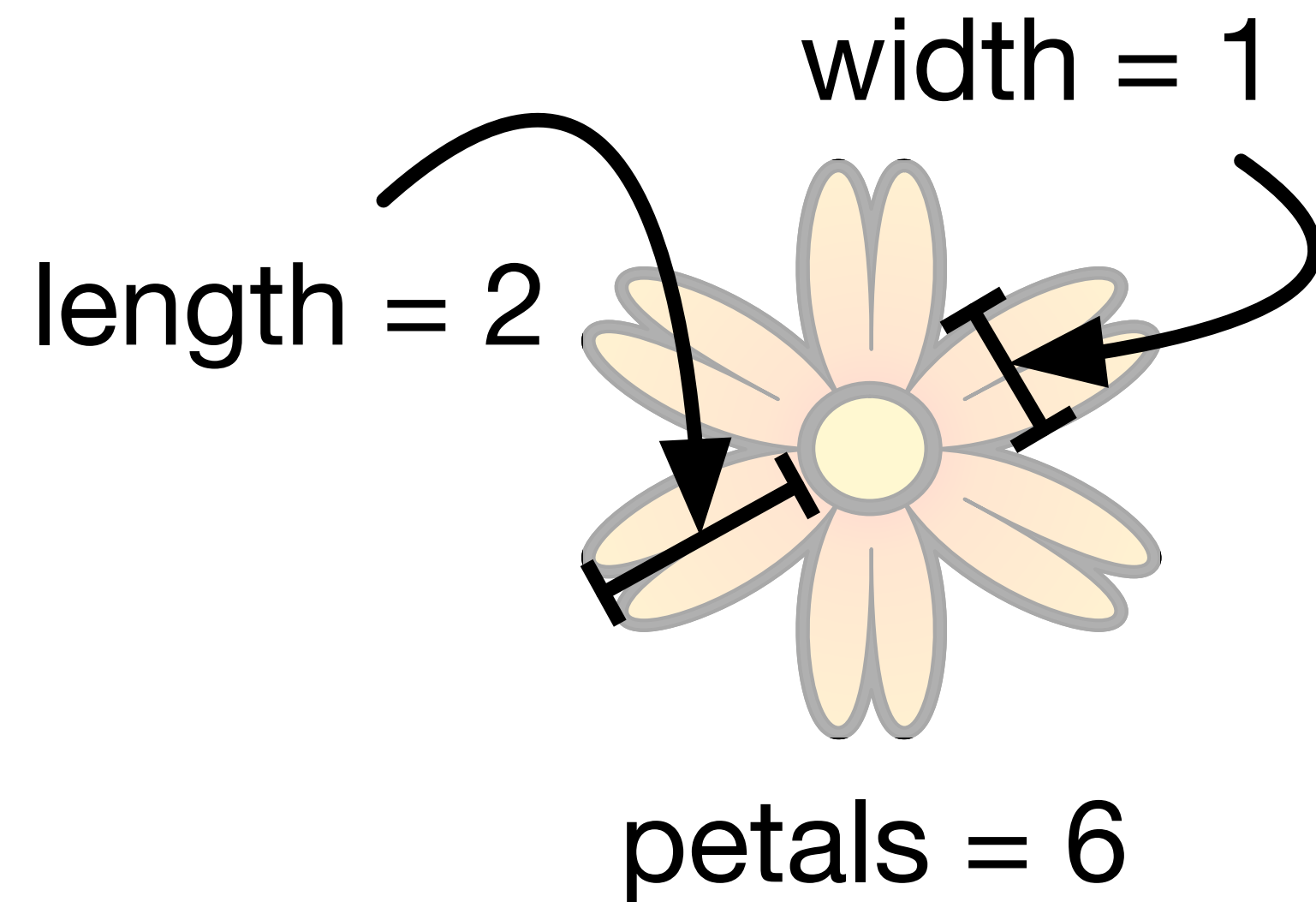
round-petalled



A Toy Example

Measurement

- First, we will **store each flower as a list of numbers.**
- E.g., for flowers:
 - the number of petals
 - the length of longest petal
 - the width of narrowest petal

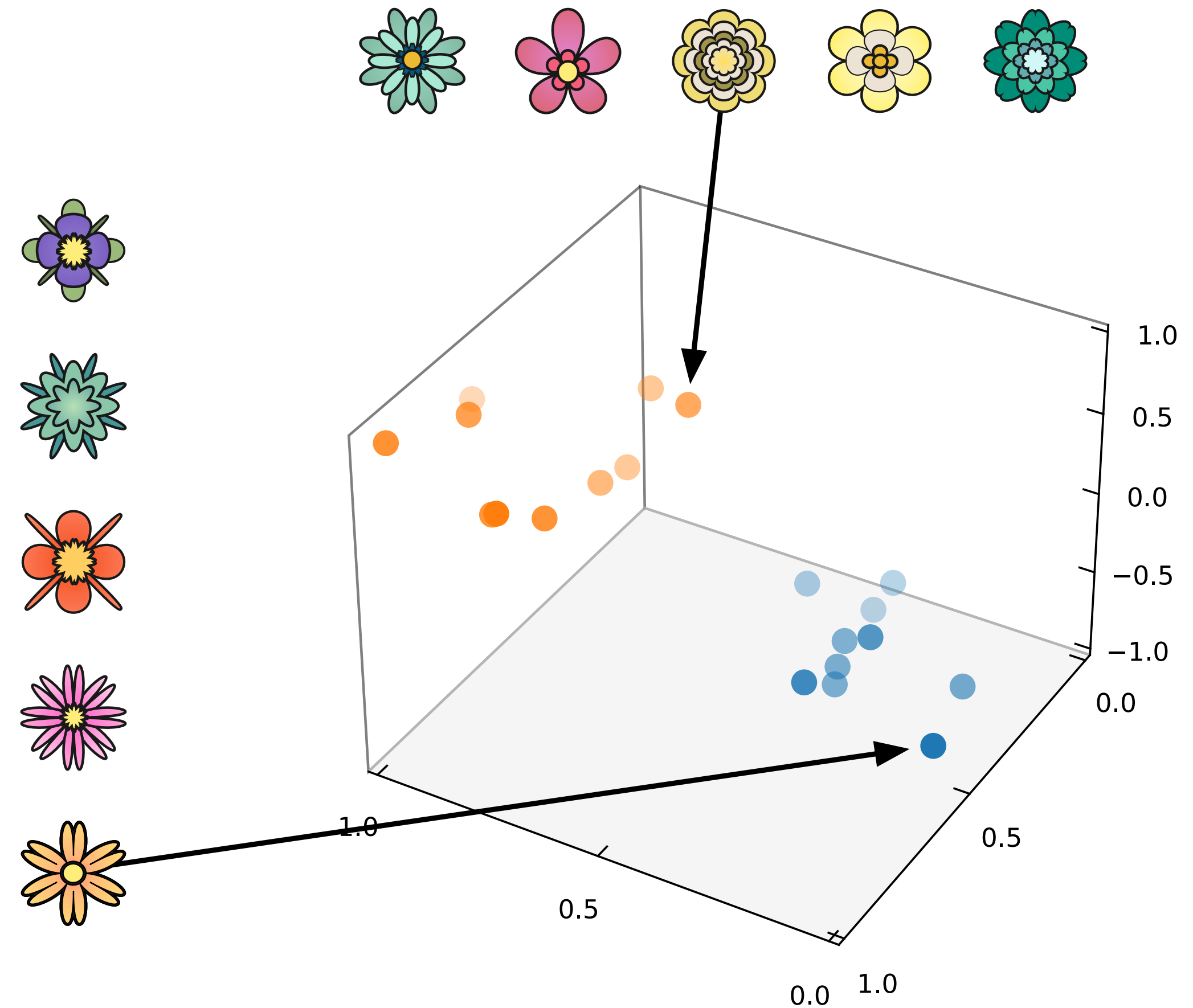


representation = (6, 2, 1)

A Toy Example

Inputs and labels

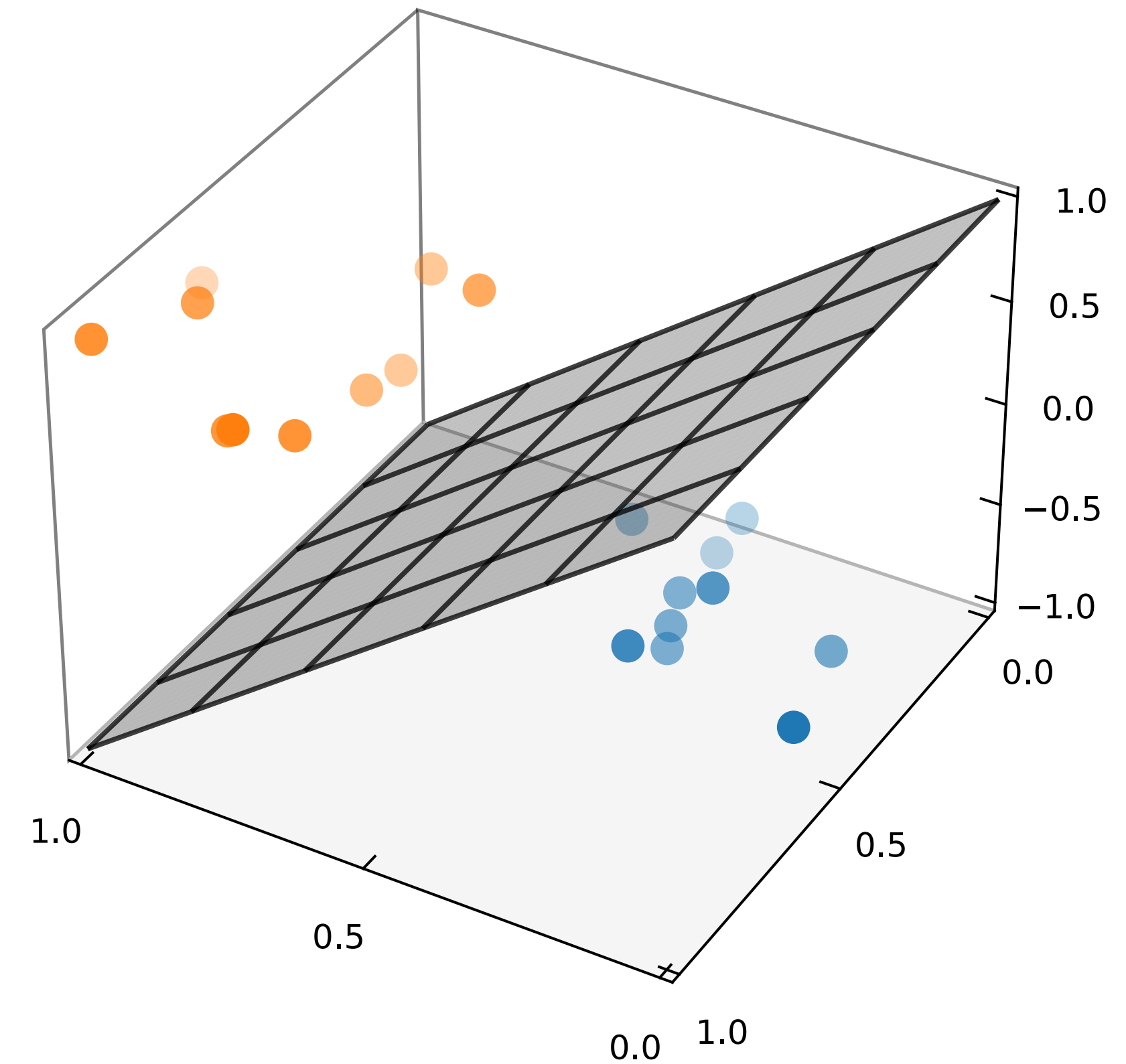
- A data point has a coordinate that represents it.
 - We call this the **input X**.
- Each data point we have is either a pointy or round flower, as labelled by our friend.
 - We call this the **label Y**.



A Toy Example

Logistic regression

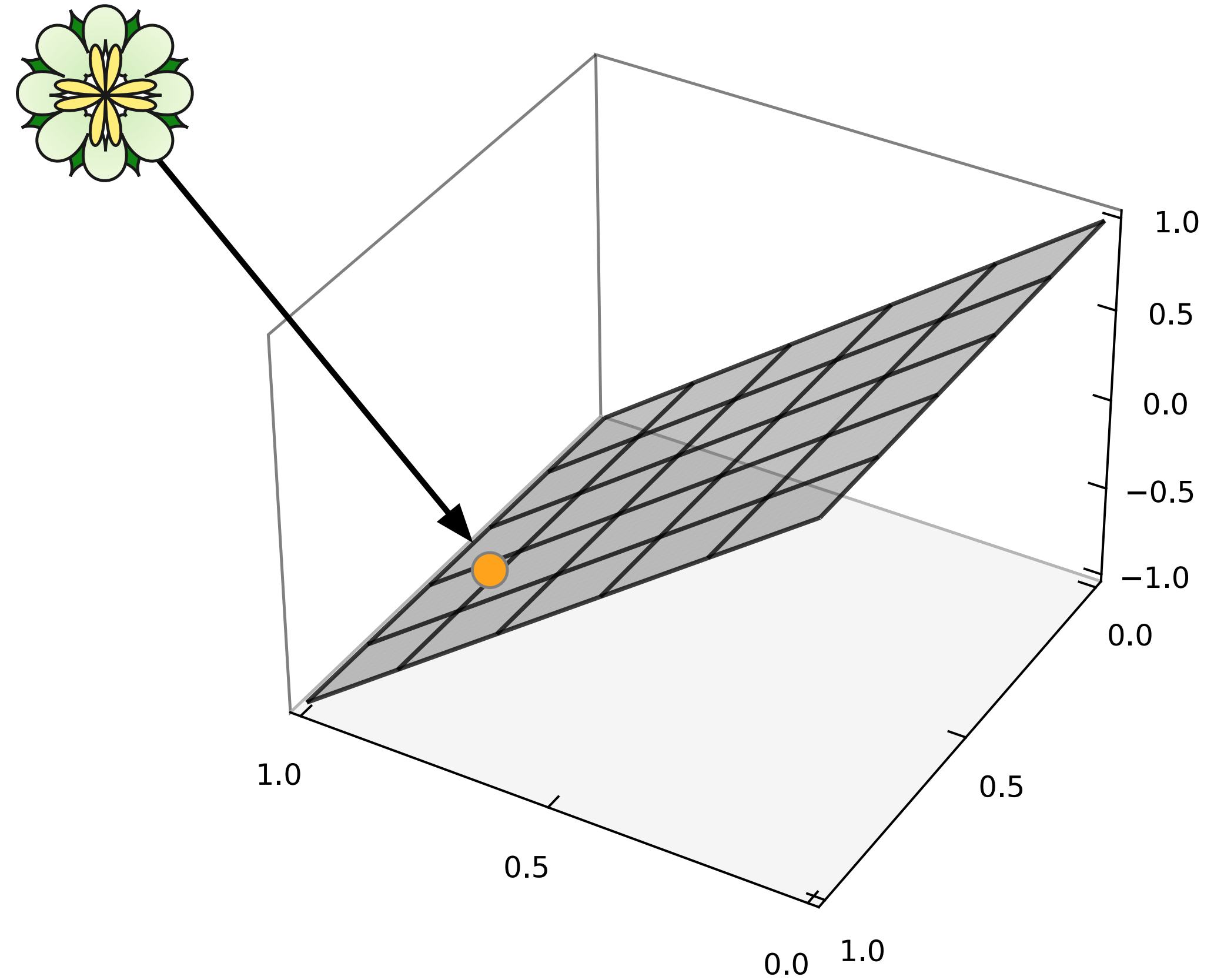
- What label do we predict for a **new unseen point**?
- We can use mathematics to find a **plane that separates our data**.
- The surface that separates the orange dots from the blue dots.



A Toy Example

Logistic regression

- Plane is our automated solution.
- For a new flower:
 1. measure it to get an input
 2. check if above or below plane
- We predict round flower!
 - That's logistic regression.



Recap

- ML is the study of algorithms that learn from examples or experience.
- A lot of machine learning boils down to predicting future measurements.
- When should you not use ML?
 - If you can afford to take the measurement.
 - If you can already make nearly perfect predictions, as in physics.

Agenda

- What is machine learning?
- **What are large language models?**
- What is driving the AI boom?

Next Token Prediction

- Early efforts at OpenAI were looking at models that were trained to predict text.
- Predict a span of text (“tokens”) given the preceding span of text.
- Trained on a corpus of books.
 - Remember:

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

The quick brown **predict this** → fox
input context **next token label**

Large Language Models

are next-token predictors

- We call these next-token predictors, **Large Language Models (LLMs)**.
 - First OpenAI next-token predictors were called GPTs.
- To test LLMs on a task, **you prompt them with text that encodes the task.**

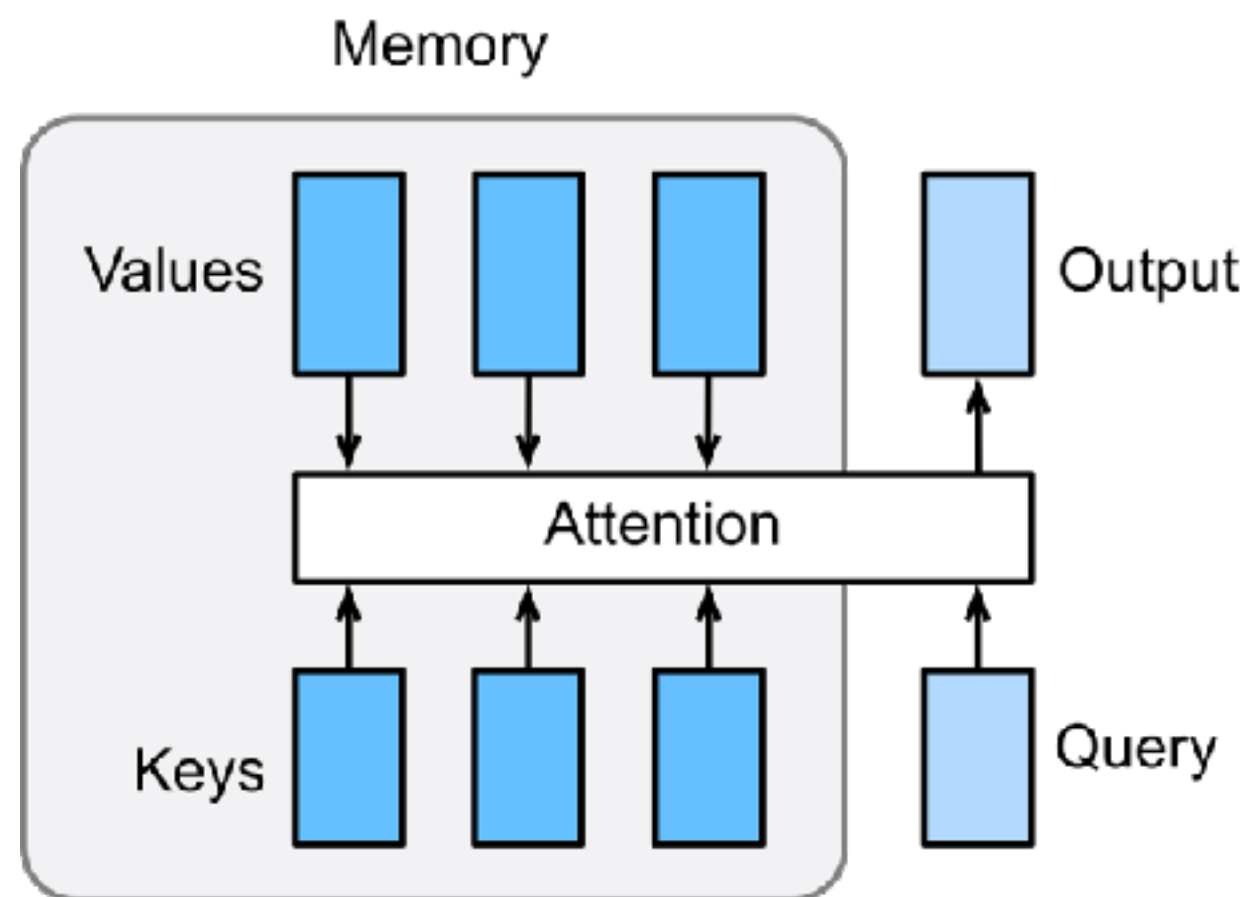
prompt		completion
Who wrote the book <i>The Origin of Species</i> ?	predict this 	Charles Darwin

- Fitting LLMs is very similar to logistic regression!

Logistic regression to ChatGPT

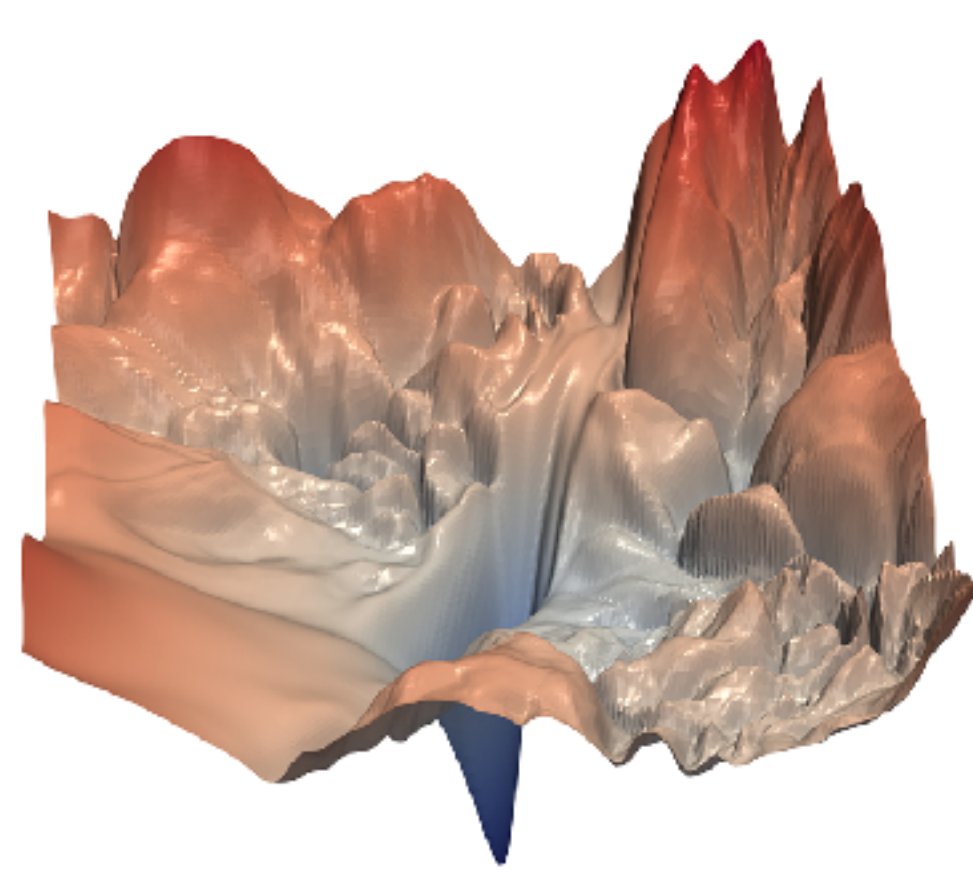
Decades of progress

More **complex predictors**



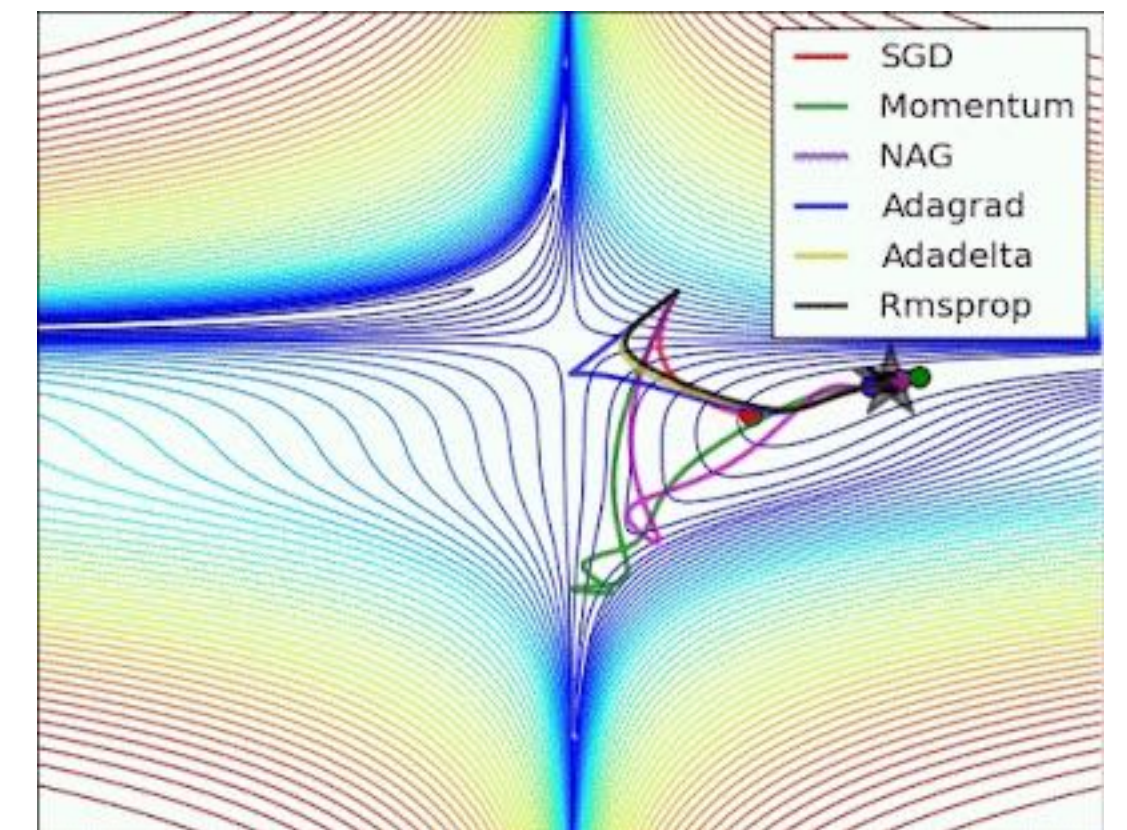
credit: Wikipedia

More challenging **fitting problems**



Li et al. 2018. Visualizing the Loss Landscape of Neural Nets.

Slightly **better algorithms**



credit: Deniz Yuret

But largely the same principles!

GPT-2

Training on internet text (WebText) results in multi-task LLMs

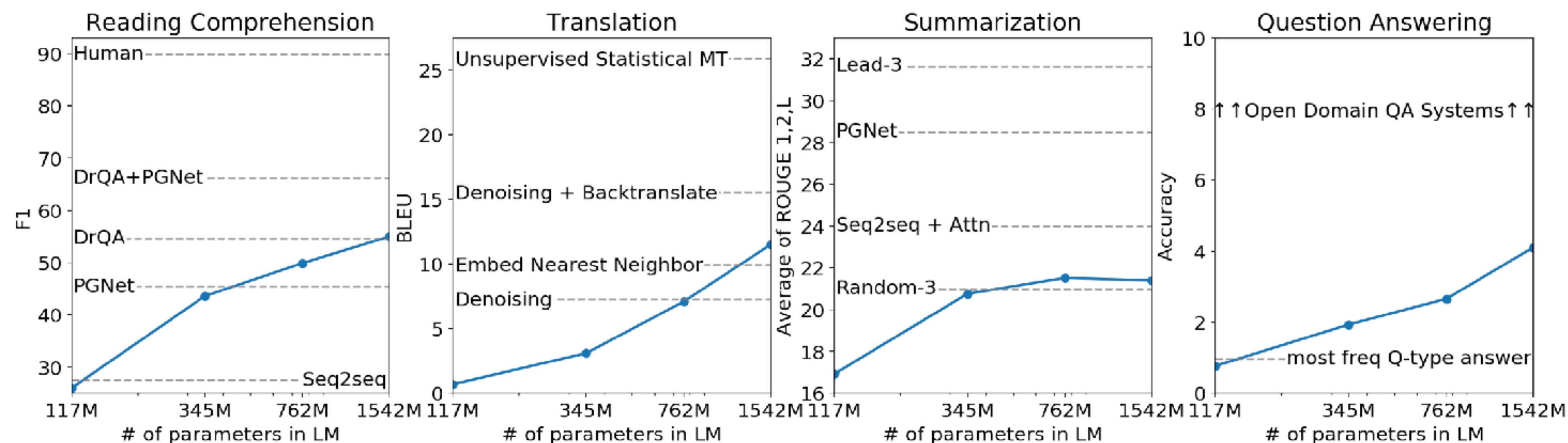


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

GPT-2

Is it just memorizing data?

- A data contamination study suggested that **data overlap between WebText and evaluation datasets provided a *small* but consistent benefit.**
- Comparable test vs. train overlap in common datasets as test vs. WebText train overlap.
- Therefore, memorization was not the dominant effect.

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	2.67%	0.66%	7.50%	2.34%	9.09%	13.19%
WebText train	0.88%	1.63%	6.31%	3.94%	2.42%	3.75%

Table 6. Percentage of test set 8 grams overlapping with training sets.

Radford et al. 2019. Language Models are Unsupervised Multitask Learners.

Why do LLMs do well on many tasks?

Maybe the natural structure of text gives LLMs task information

task

context

next token

My favourite movies: Alien, Star Wars,

Lion King

Movies from the 70s: Alien, Star Wars,

Taxi Driver

Sci. fiction movies: Alien, Star Wars,

Arrival

Why do LLMs do well on many tasks?

Maybe capabilities are shared between tasks

- Blakeney et al (2024) trained with and without mathematically enriched text.
- Controlling for the amount of data, found that **training on math-enriched data improved basic reading comprehension** in their experiments.

Benchmark	No DU	10% DU	
		With Math	Sans Math
MMLU (5-shot)	35.69	43.19	29.71
GSM8K (8-shot)	14.71	20.47	11.37
HumanEval (pass@1)	17.23	20.39	21.15
<i>Gauntlet v0.3</i>			
Core Average	35.37	38.46	32.54
World Knowledge	41.77	44.72	39.08
Commonsense Reasoning	38.38	42.33	31.76
Language Understanding	61.52	60.41	59.97
Symbolic Problem Solving	16.28	19.55	16.80
Reading Comprehension	37.02	43.35	26.48
Programming	17.23	20.39	21.15

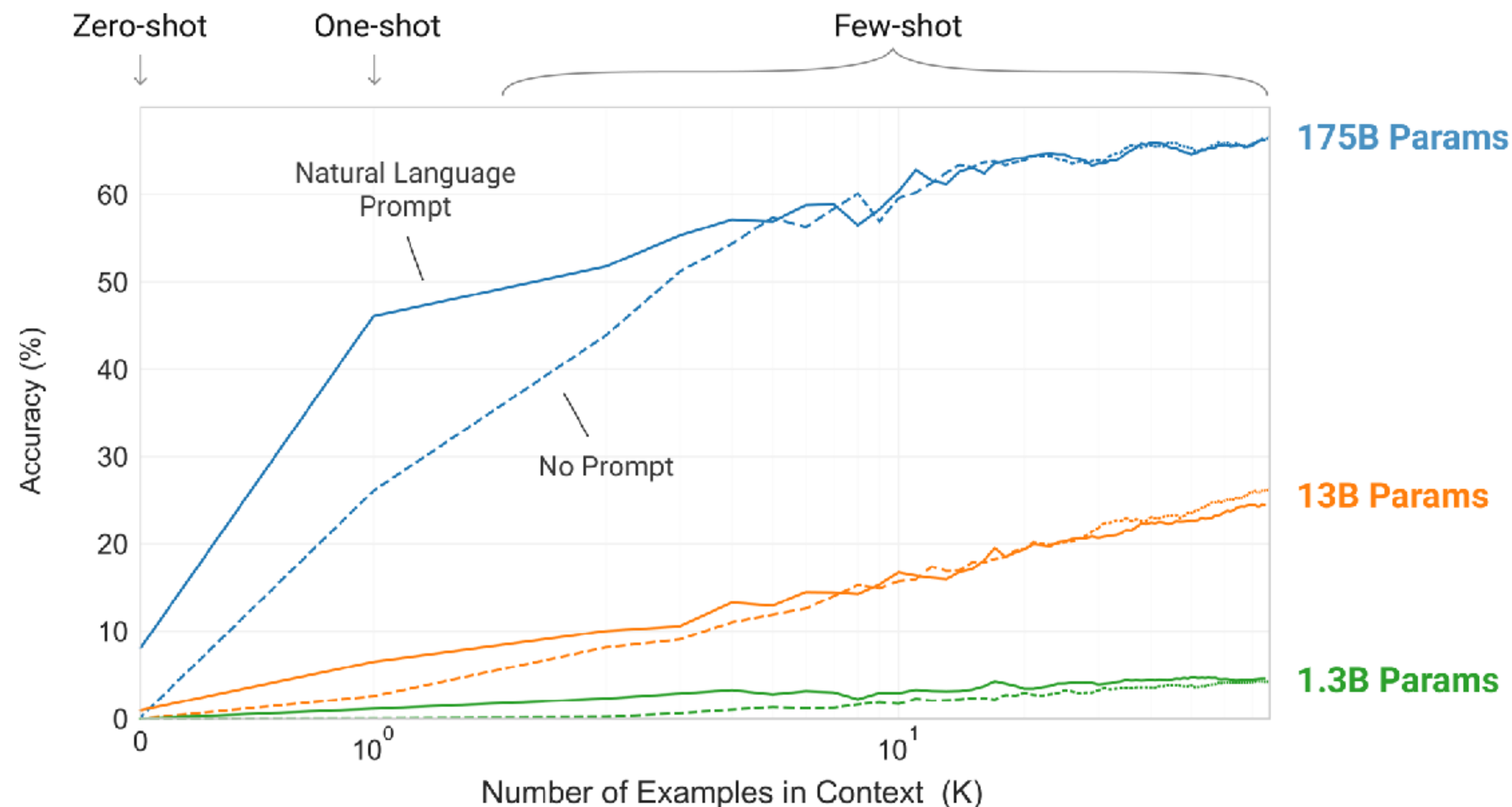
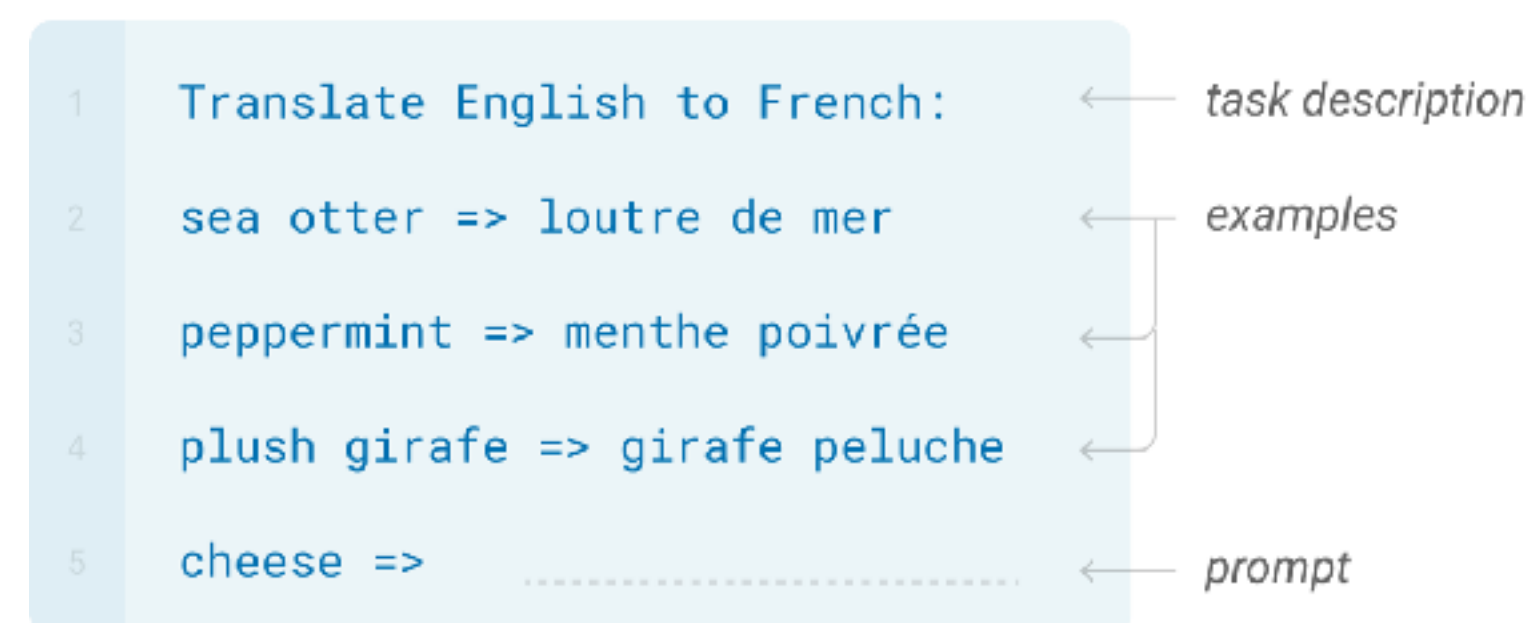
Blakeney et al. 2024. Does your data spark joy?
Performance gains from domain upsampling at the
end of training.

Result: LLMs learn to learn

Trained LLMs take advantage of examples given in-context

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Brown et al. 2020. Language Models are Few-Shot Learners.

Result: intuitive prompts improve performance

To improve LLM performance, you can literally talk to them

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

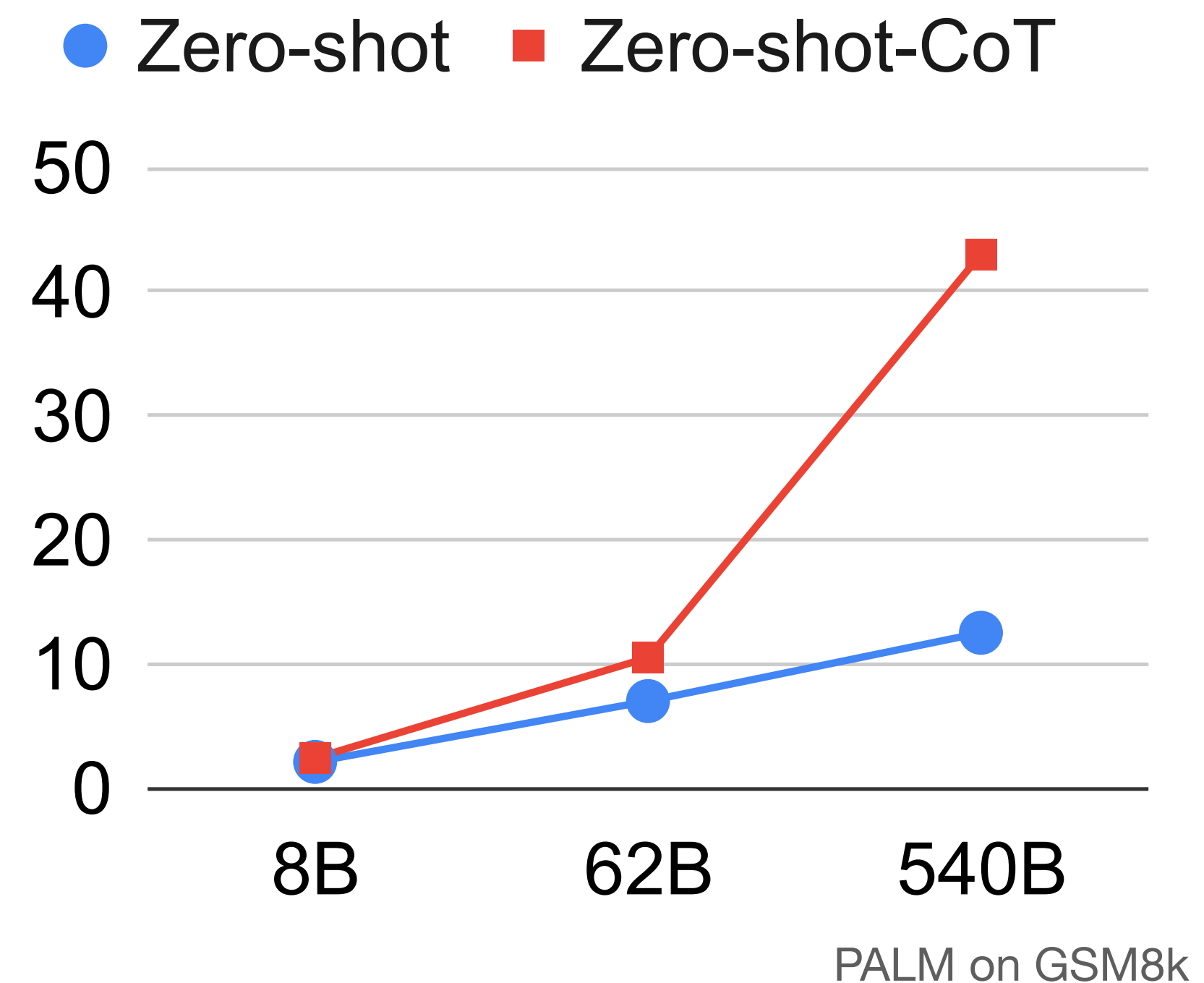
(Output) 8 **X**

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

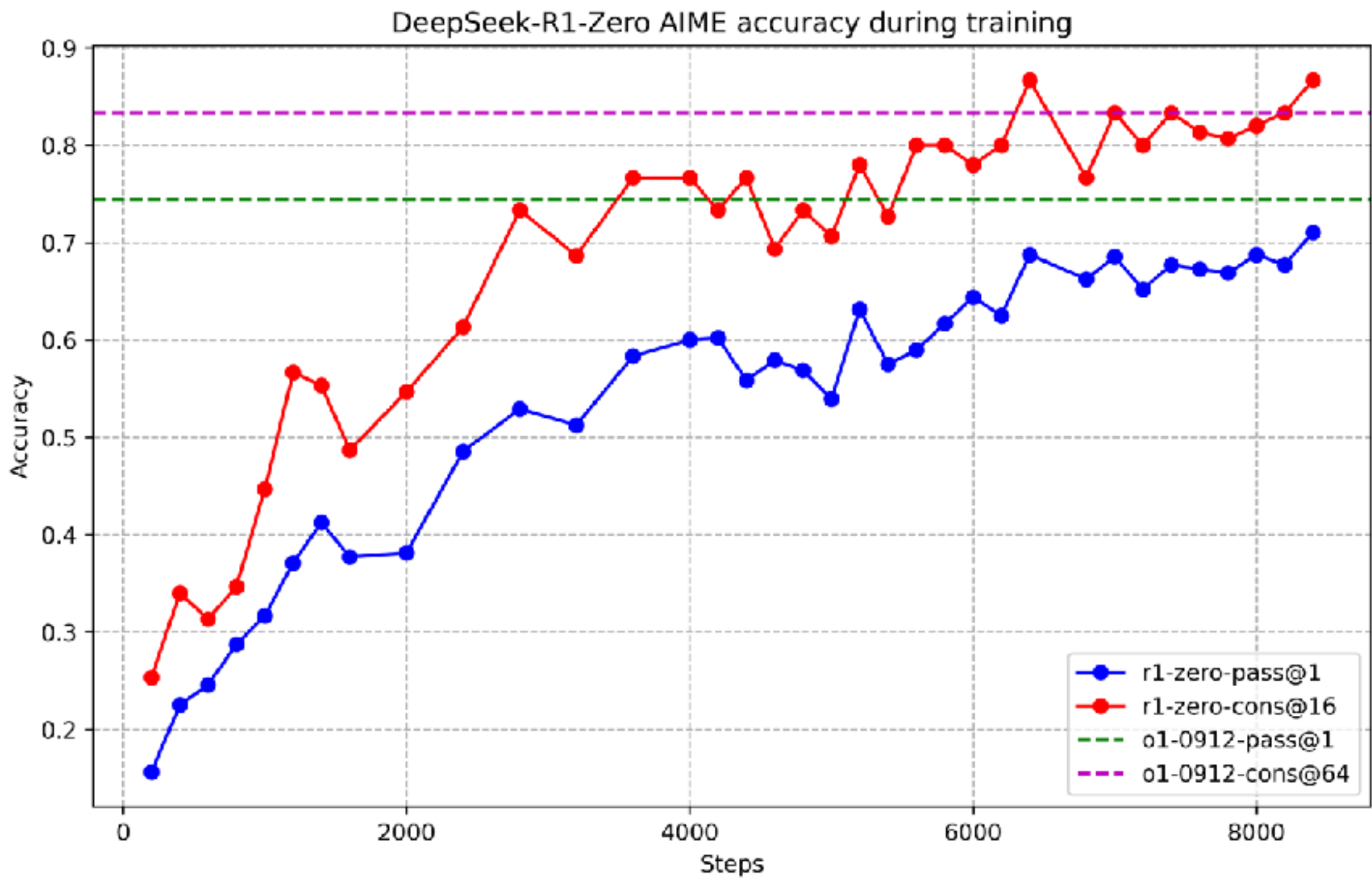
A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*



Result: trained LLMs can learn to reason

LLMs fine-tuned to maximize accuracy learn to reason



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$.
Rearrange to isolate the inner square root term:
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
...
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be ...
We started with the equation:
 $\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
 $a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: ...
...

Table 3 | An interesting “aha moment” of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

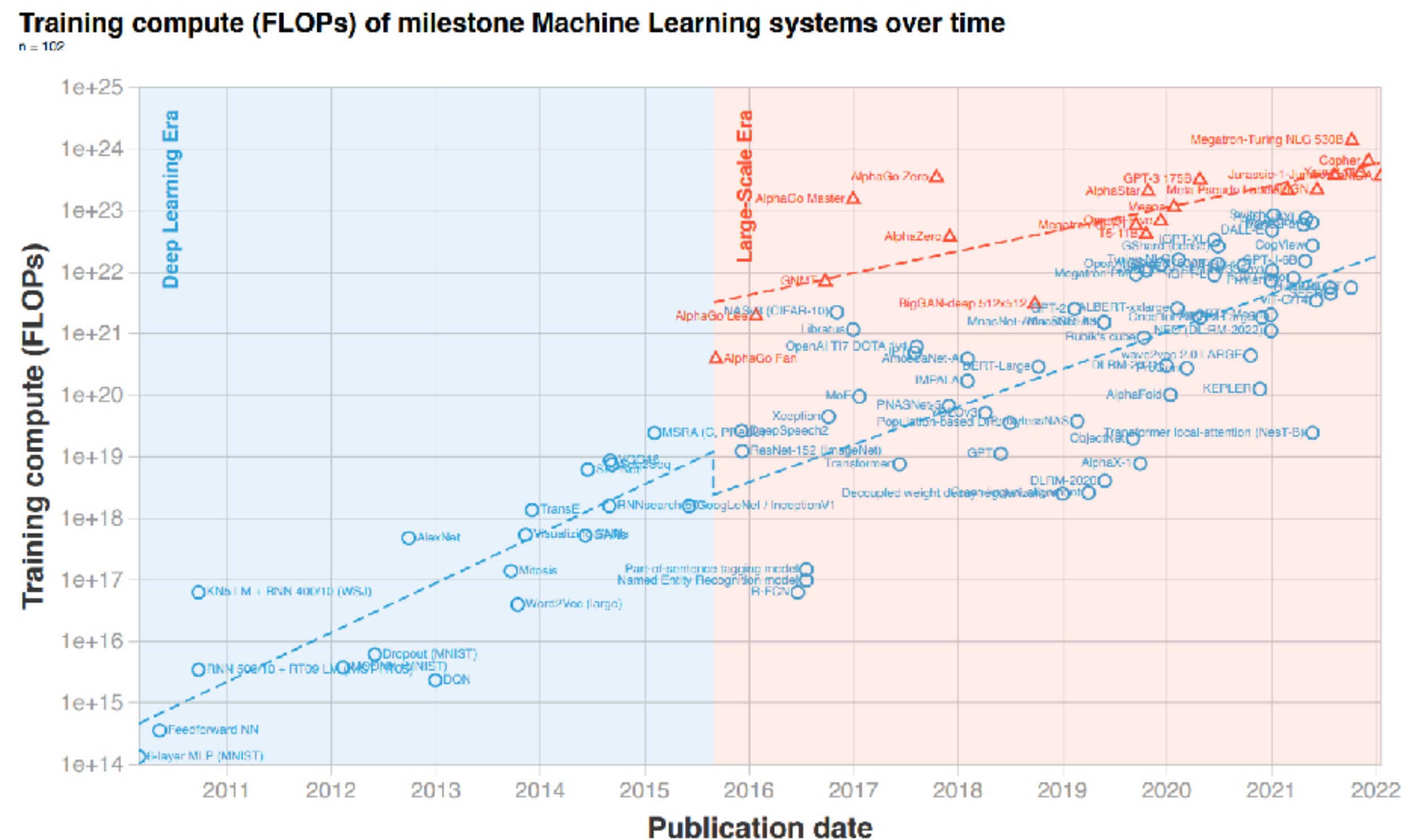
Recap

- LLMs train to predict the next token on web-scale text data
- Training to predict internet text results in massively multi-task predictors

Agenda

- What is machine learning?
- What are large language models?
- **What is driving the AI boom?**

On the back of LLMs, ML industrialized rapidly
We are consuming increasing amounts of compute



Sevilla et al., 2022. “Compute trends across three eras of machine learning”

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.

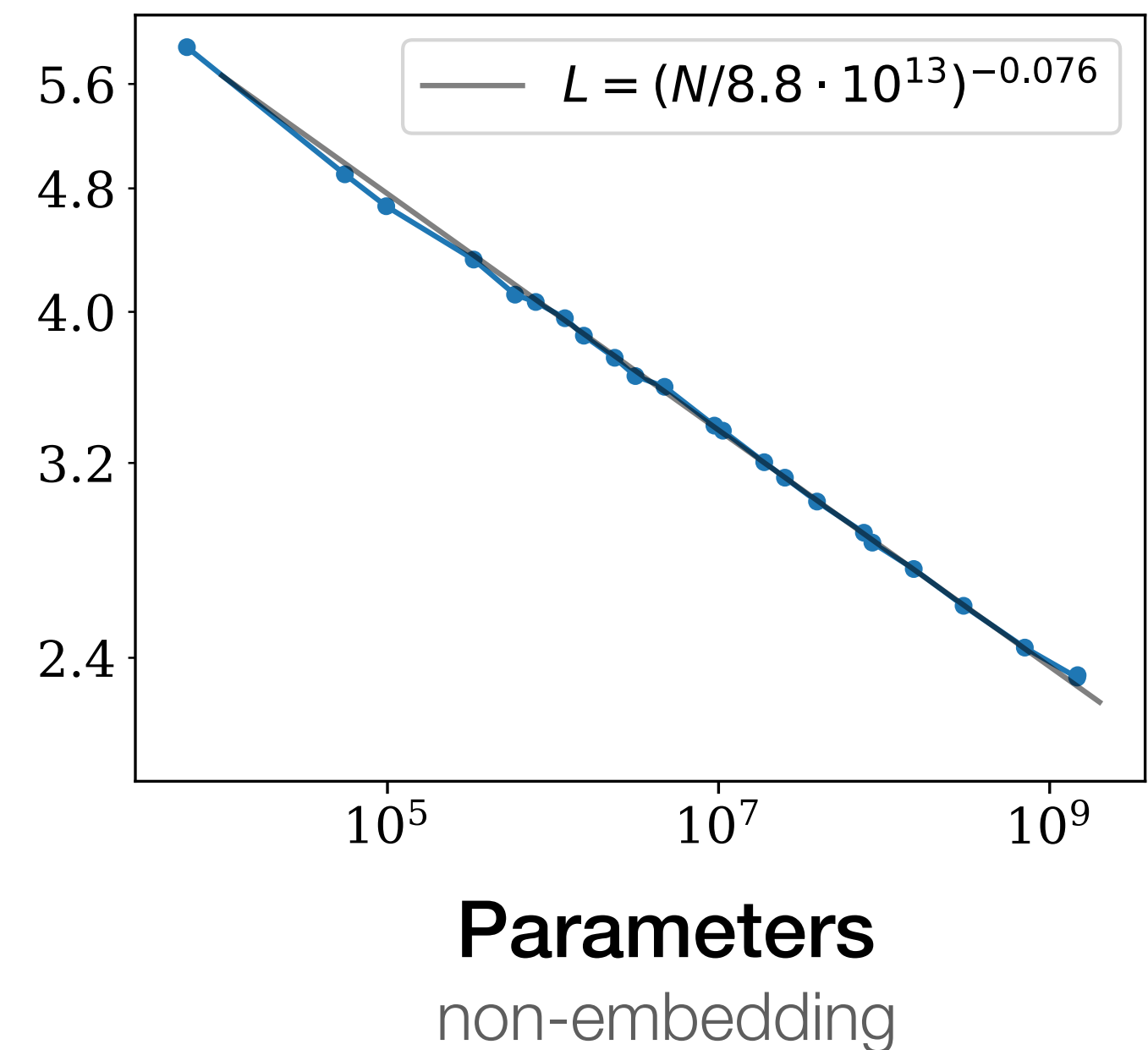
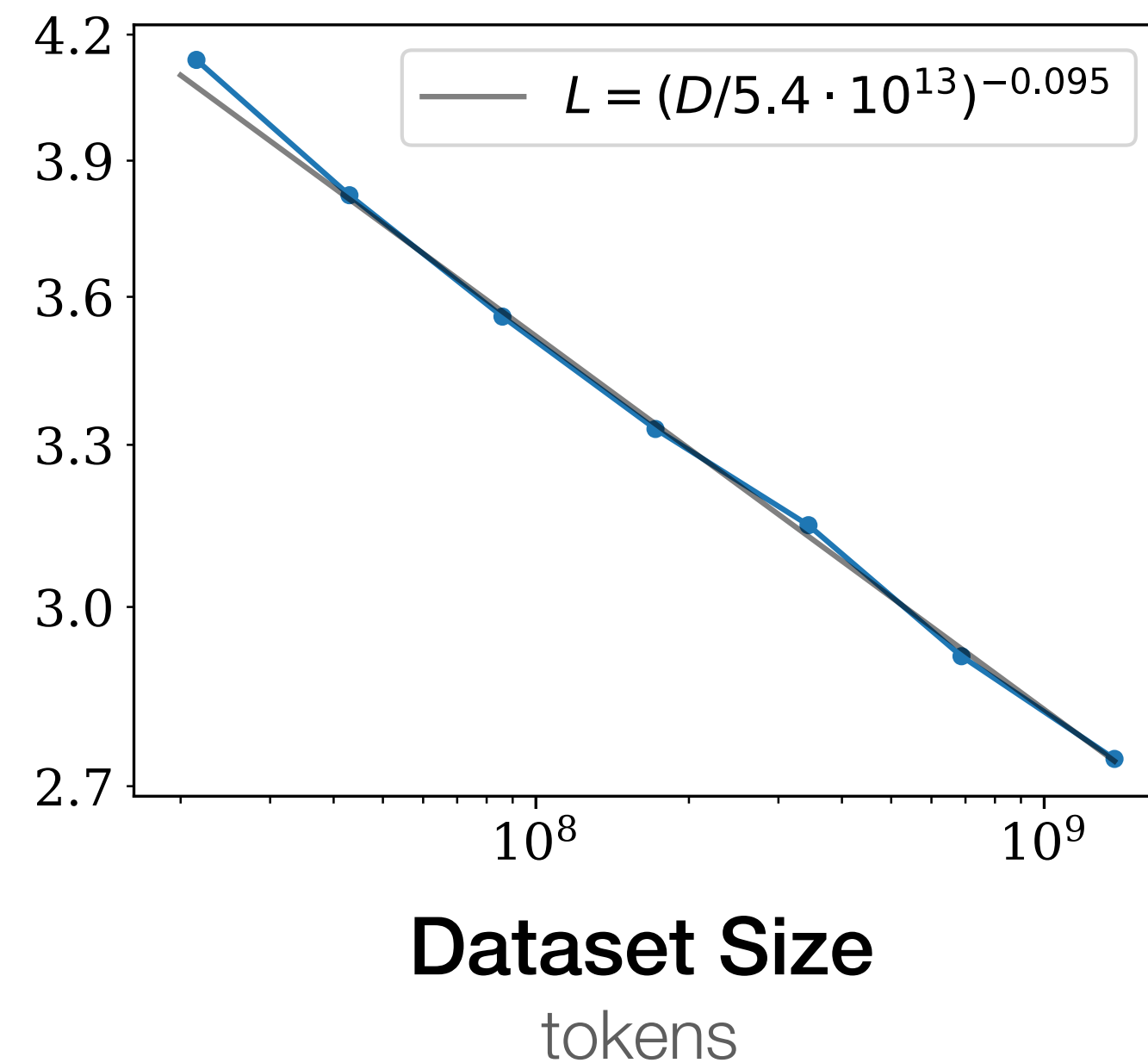
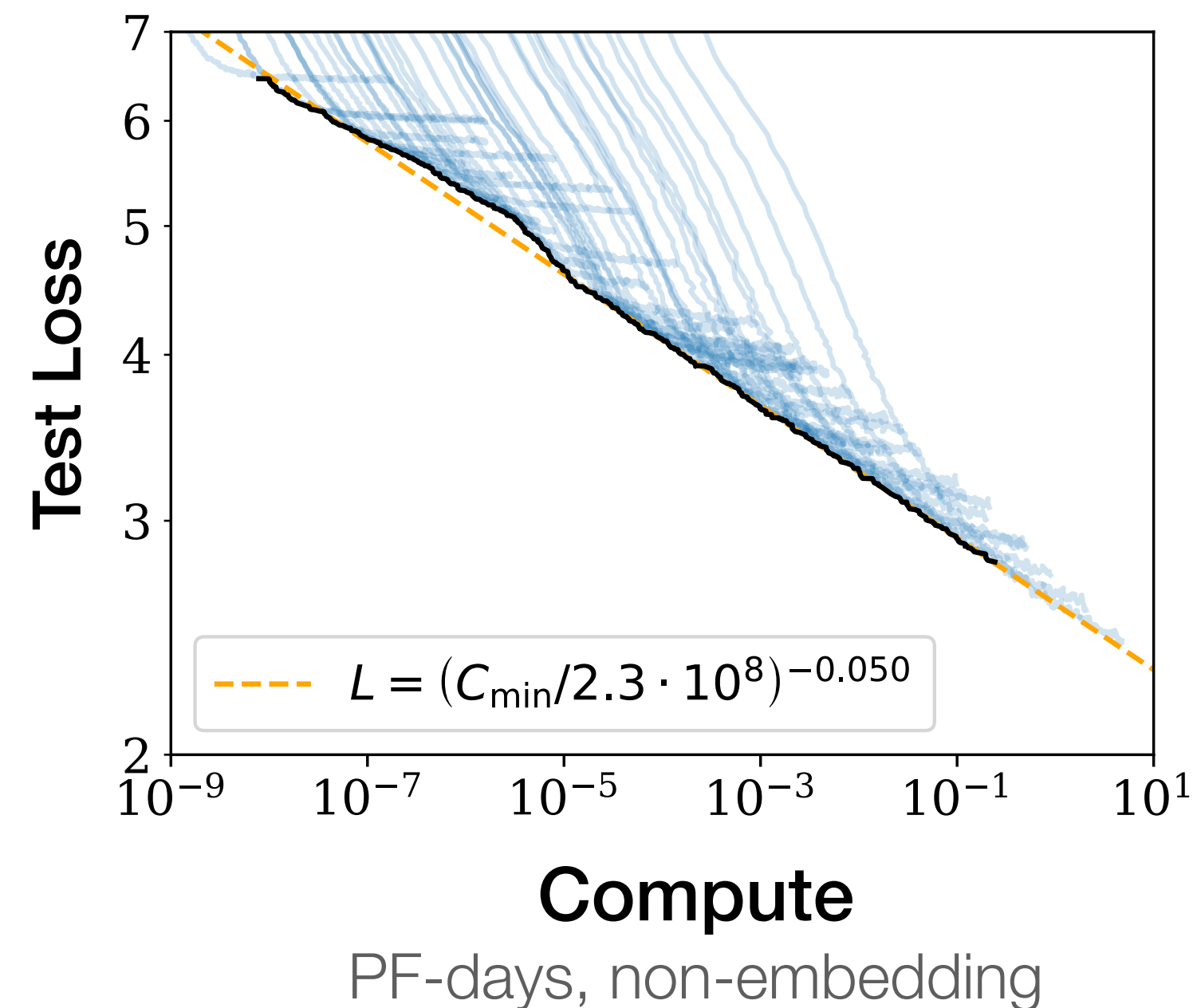
Elon Musk turns on xAI's new AI supercomputer: 100K liquid-cooled NVIDIA H100 AI GPUs at 4:20am

Elon Musk posts on X saying 'nice work by xAI and X team, NVIDIA and supporting companies getting Memphis Supercluster training started at 4:20am.'

A number of key results are driving this trend:

Why? loss improves *predictably* with scale

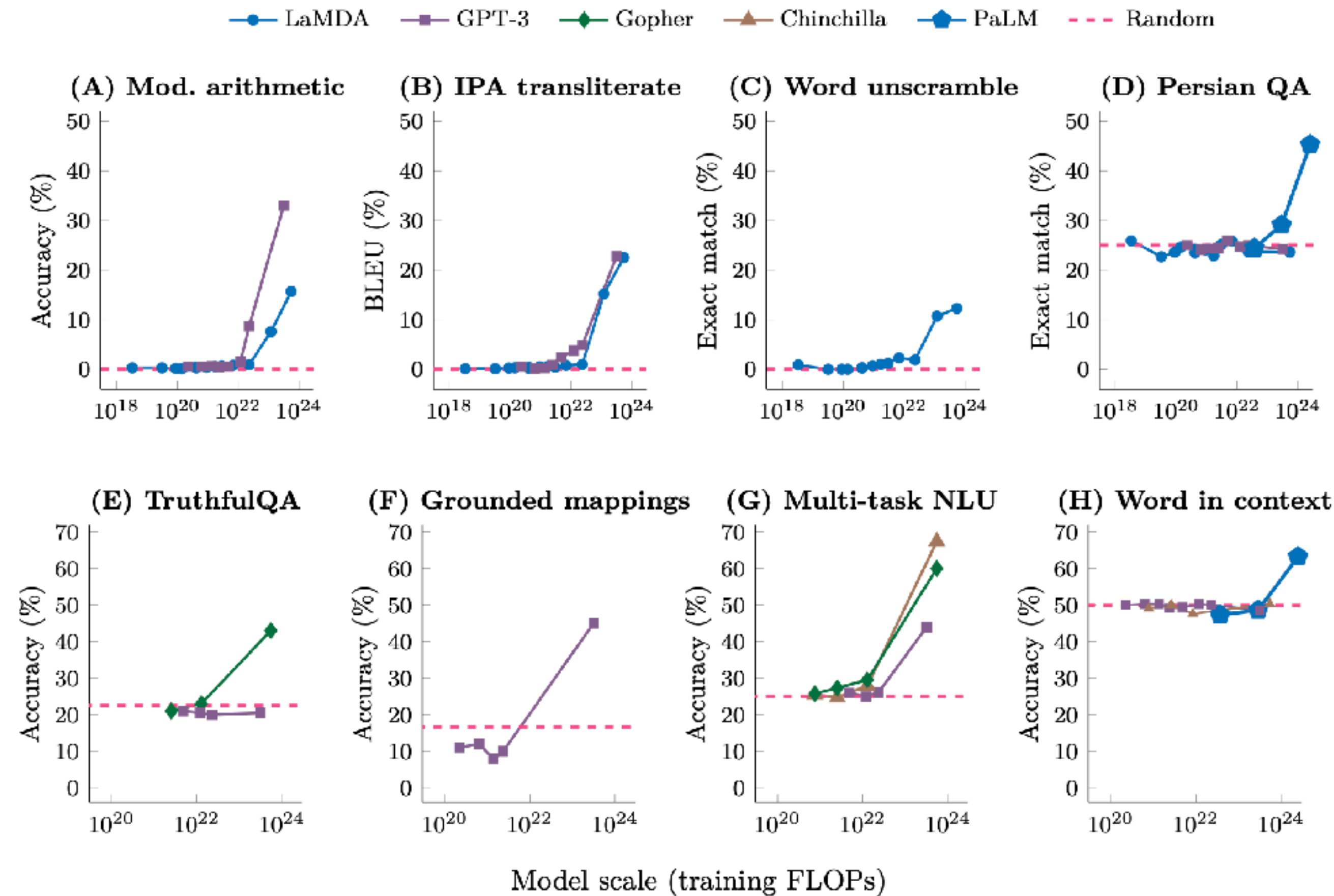
More data or parameters improves performance *in a predictable way*



Kaplan et al. 2020. Scaling Laws for Neural Language Models.

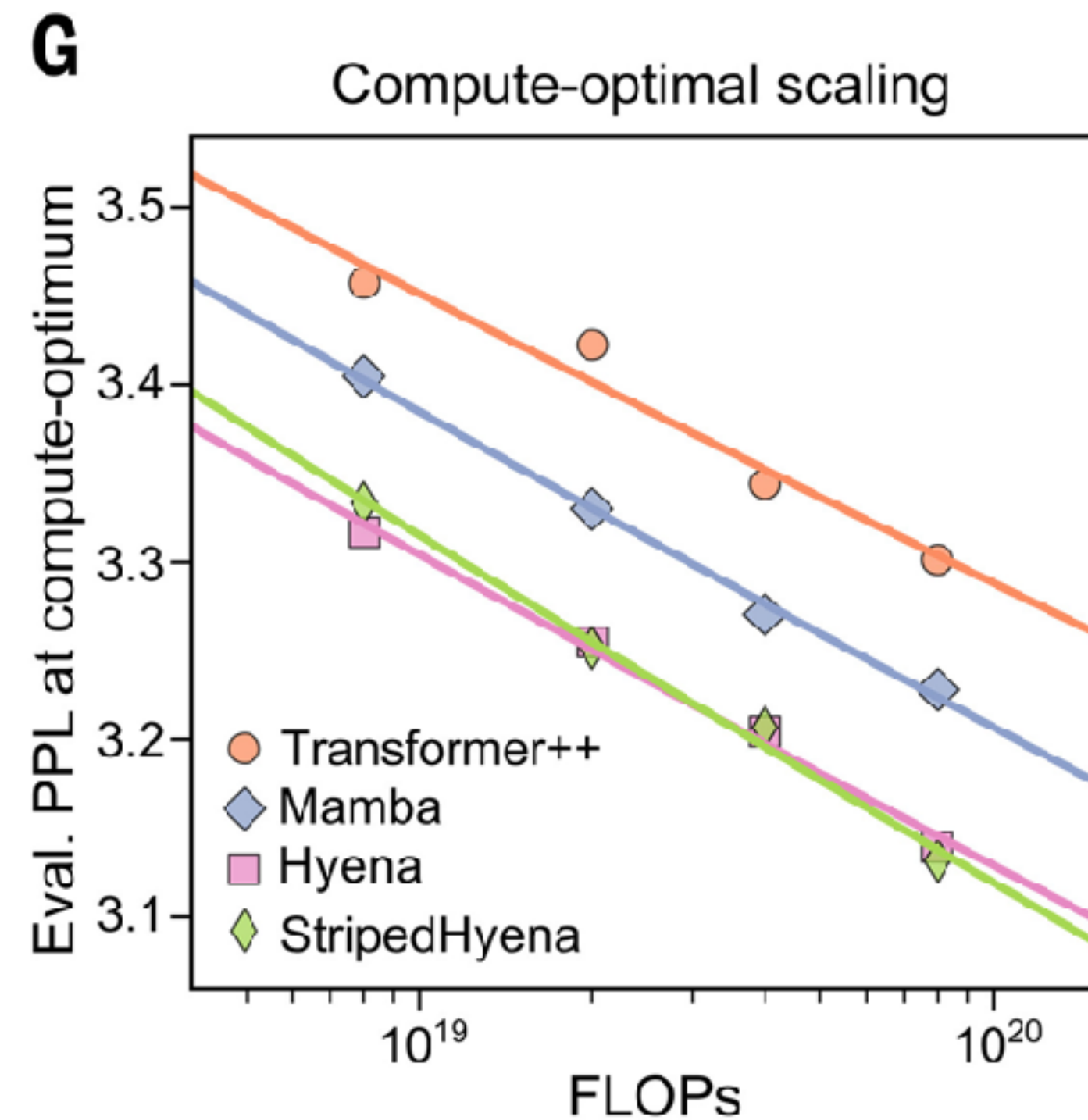
Why? capabilities emerge with scale

As models scale on internet data, they improve on very diverse set of capabilities



Why? improvement rates consistent across algorithms

Data and compute is like oil



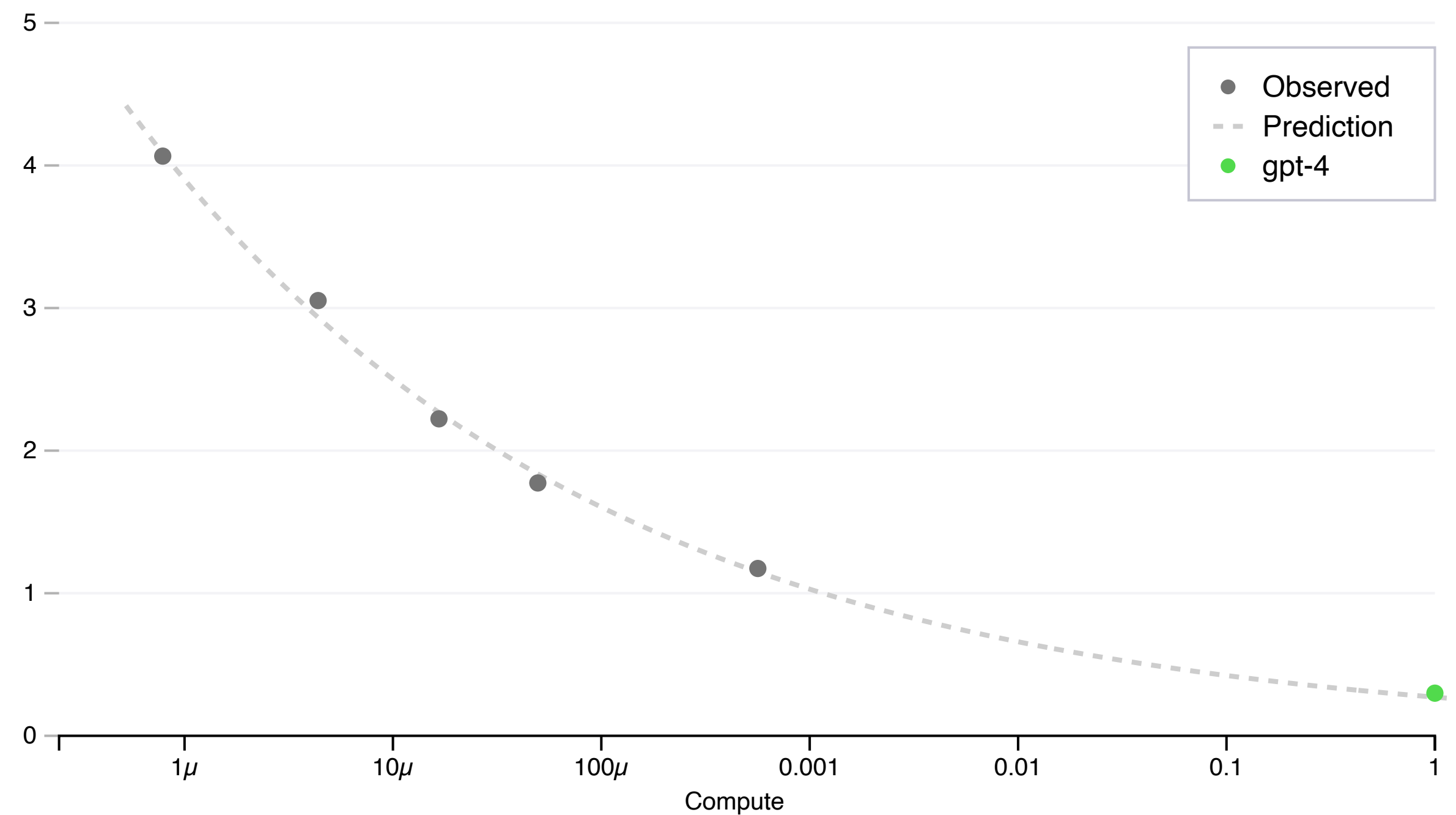
The results are predictable, the recipe simple

So, we industrialized

- The cost of fitting can be measured in compute (FLOPs).
- The performance of an LLM can be forecasted as a function of compute used to fit.
- **So far, we've been able to predict the performance return on compute investment.**

Capability prediction on 23 coding problems

– Mean Log Pass Rate



The Bitter Lesson

Could compute be the key driver?

- Rich Sutton wrote about this in a 2019 essay titled “The Bitter Lesson”. He was comparing **two approaches to progress**:
 - researchers designing clever methods that capture knowledge of the data
- vs.
- compute invested into general-purpose algorithms
- **The “bitter lesson”, he argues, is that compute-driven approaches are winning over longer time scales.**

The Bitter Lesson

Could compute be the key driver?

- Results are fairly consistent across LLMs and training algorithms.
- **Suggests our successes are determined by natural properties of human text.**
 - Text is where we store reasoning, knowledge, etc.
 - It is a very rich interface and text-predictors inherit that richness.

“We believe there are three key levers in the development of high-quality foundation models: data, scale, and managing complexity.”

Llama 3 Tech Report

Recap

- LLMs improve as you scale the compute used to train them
- Incredible capabilities and massive multi-task abilities emerge as you scale
- The specific training algorithms seemingly have less impact

Parting thoughts

- Machine learning is starting to look more like biology: everything is about rates.
 - Rate at which you can collect examples
 - Rate at which you can convert examples into intelligence
- Traditional computer science still has a role to play in improving rates, but there are a bunch of interesting empirical questions that look more like biology.
- Algorithms that learn from examples and experience inherit the richness of our world.

Thanks!