
The Big Ideas Behind AI

Joseph E. Gonzalez
Professor of EECS, UC Berkeley



Fisher Center Policy Advisory Board | October 14, 2026

Artificial General Intelligence



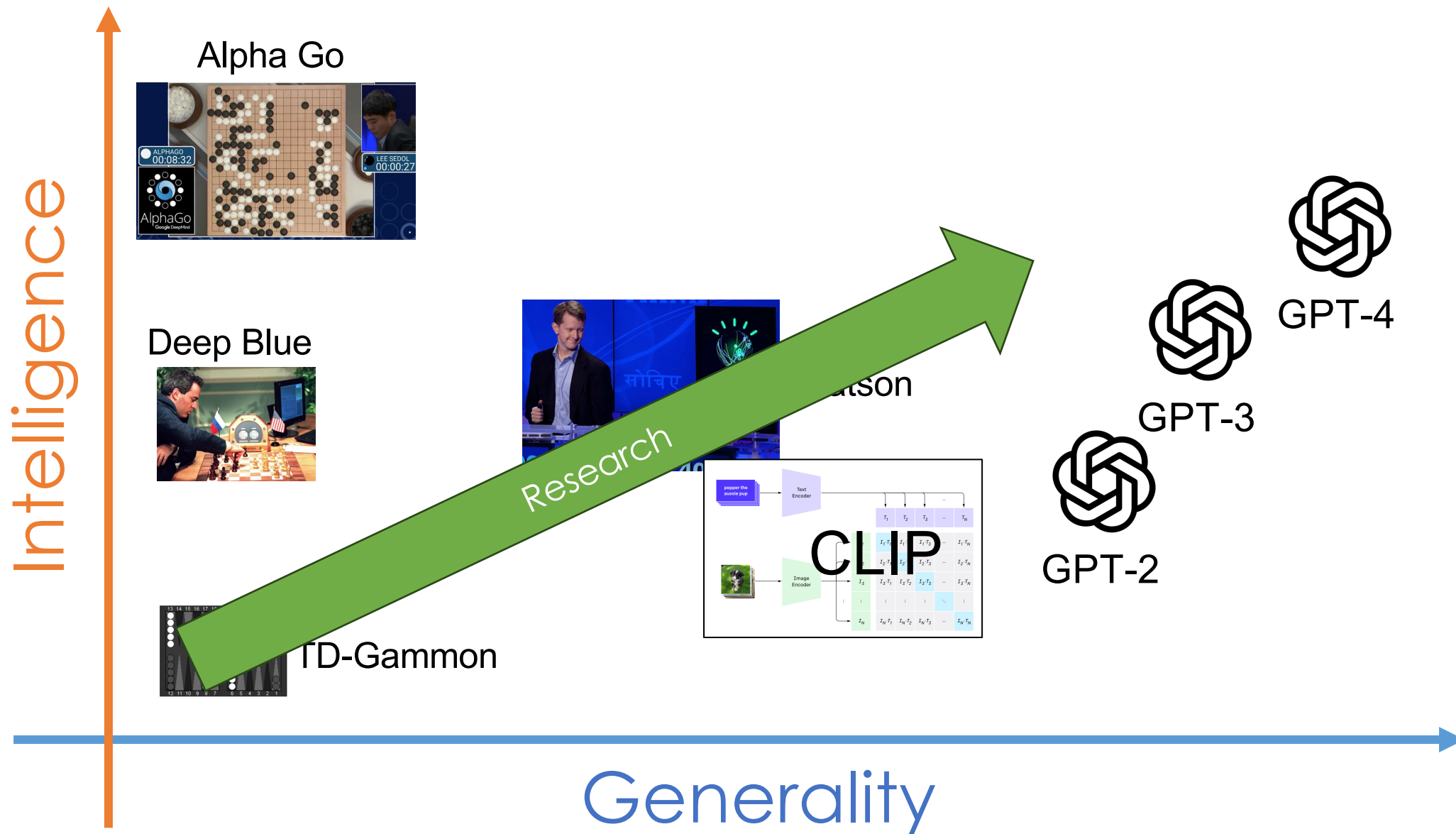
Artificial General Intelligence

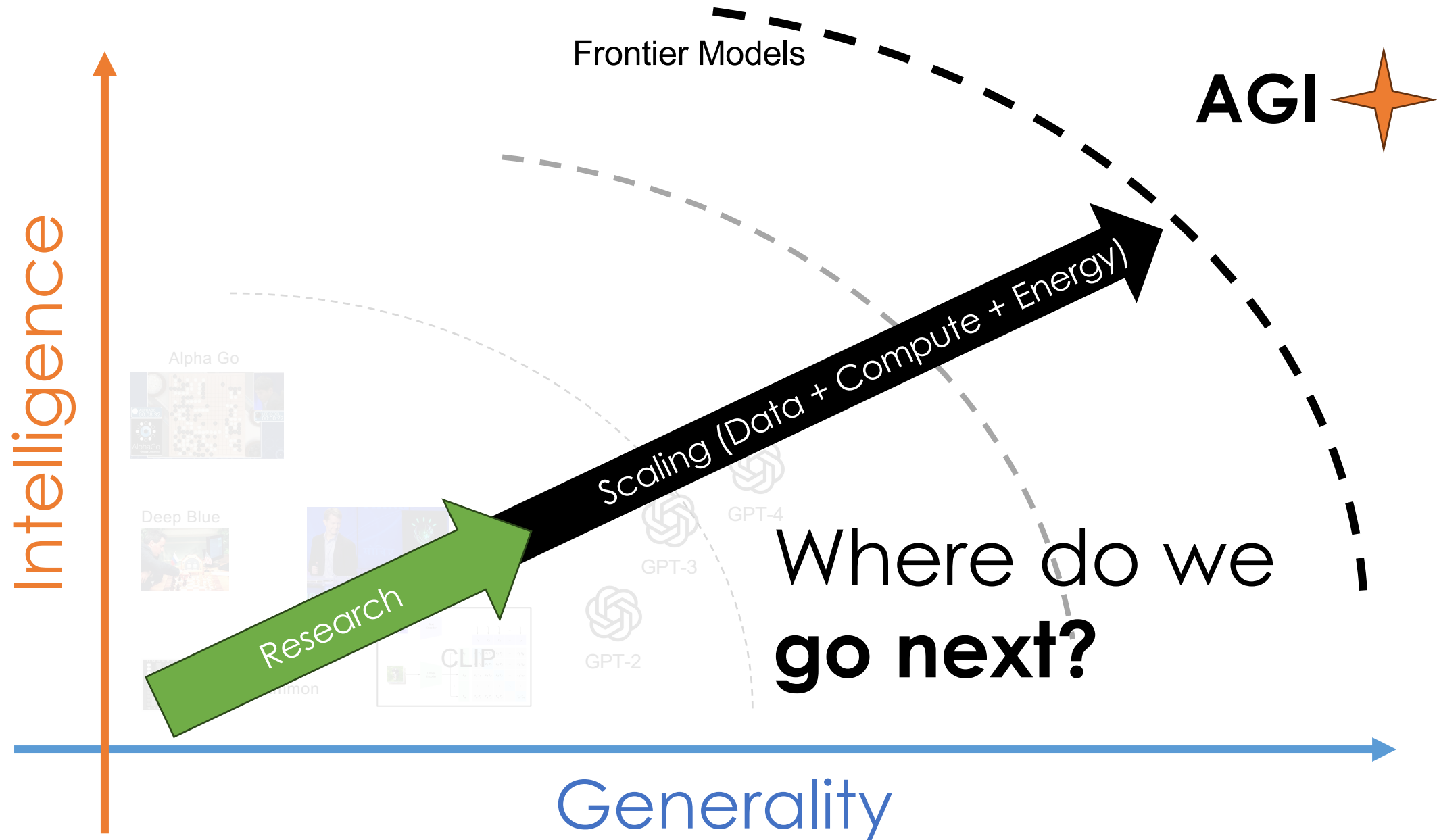


One task

Artificial General Intelligence ✨







Tonight, two questions

- 1 How did we get here?**
- 2 Where are we headed?**

About me



Joseph (Joey) Gonzalez

Professor of EECS, UC Berkeley

2008: Pioneered large-scale learning systems.

2015: Pioneered inference systems.

Research projects



Vicuna



Gorilla



Primary inference platforms
for open-weight models

Research groups and other startups



LMSYSORG



I teach some of the largest classes at Cal

About 1,000 students per class

Introduction to Data Science | Introduction to Machine Learning



Question 1

How did we get here?

ChatGPT



What is ChatGPT?

Chat: a natural language system

G

Generative: designed to model the creation of text

P

Pre-trained: trained on lots of naturally occurring text

T

Transformer: a kind of neural network

ChatGPT is one example of a **Large Language Model (LLM)**

Language models predict the next word

The capital of California is **Sacramento**

San Francisco (1862)

Benicia (1853)

Vallejo (1852)

Large: up to **trillions** of parameters
the numbers that encode the neurons in the model

Predicting the next word lets you

**Answer
questions**

Tell stories

**Accomplish
tasks**

If you can **generate** the world,
you must **understand** the world.

How do we model the next word?

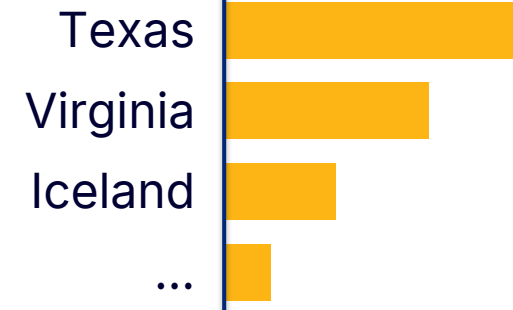
We **learn a function** from what has been said to the **odds** of every possible next word.

Context: all the words so far

The best place to build a
data center is

**Neural
Network**

Odds of the next word



We use the neural network to write, **one word at a time.**

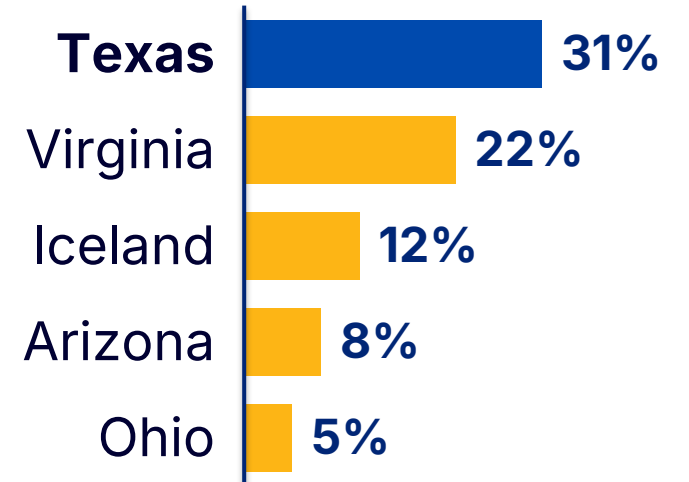
One word at a time (1 of 3)

The words so far

The best place to build a data center is

**Learned
function**

Odds of the next word



Sample a word, add it to the text:

The best place to build a data center is **Texas**

One word at a time (2 of 3)

The words so far

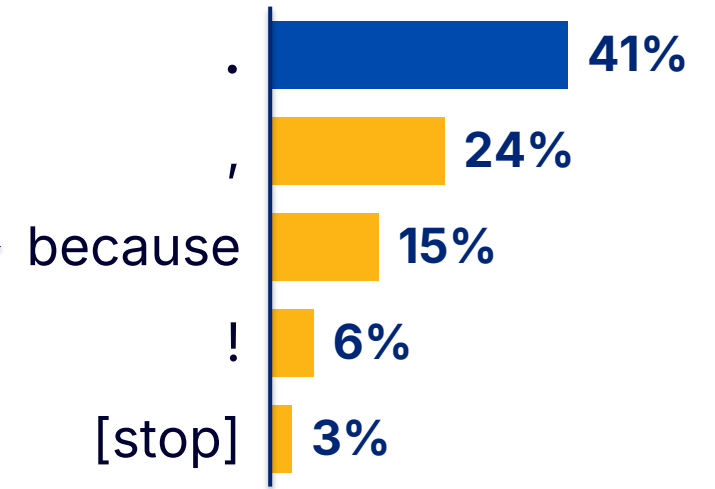
The best place to build a data
center is Texas



**Learned
function**



Odds of the next word



Sample a word, add it to the text:

The best place to build a data center is Texas.

One word at a time (3 of 3)

The words so far

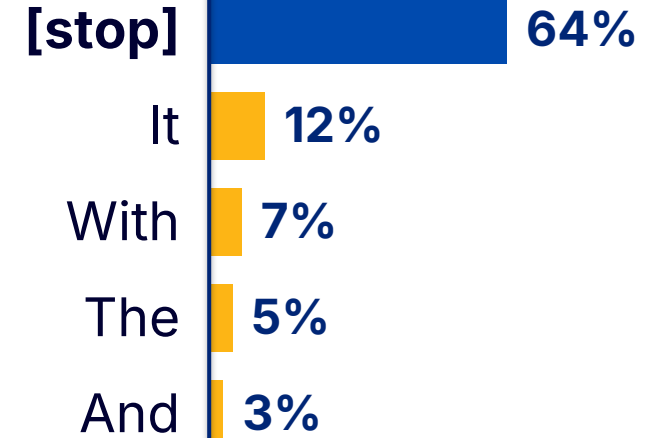
The best place to build a data center is Texas.



**Learned
function**



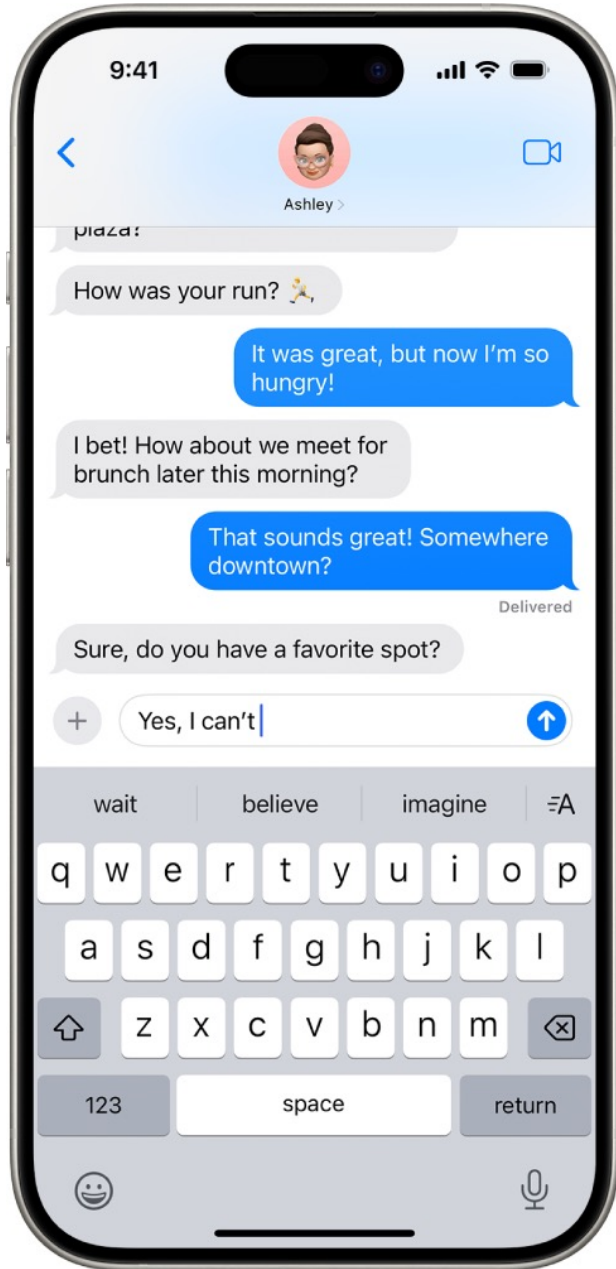
Odds of the next word



Sample a word, add it to the text:

The best place to build a data center is Texas. **[stop]**

Stop decoding!



Language modeling is
autocomplete
where we always accept
the next word.

Predictive text;
tap a suggestion to apply.

What writing one word at a time means

It writes in pen

Once a word is written, the model cannot take it back.

... center is **Texas**

no backspace

Randomness looks like creativity

Sampling from the odds gives the illusion of creativity.



It decides when it is done

It writes a special "done" word to end its answer.

[stop]

Under the hood



Inside the Neural Network

First, words become numbers

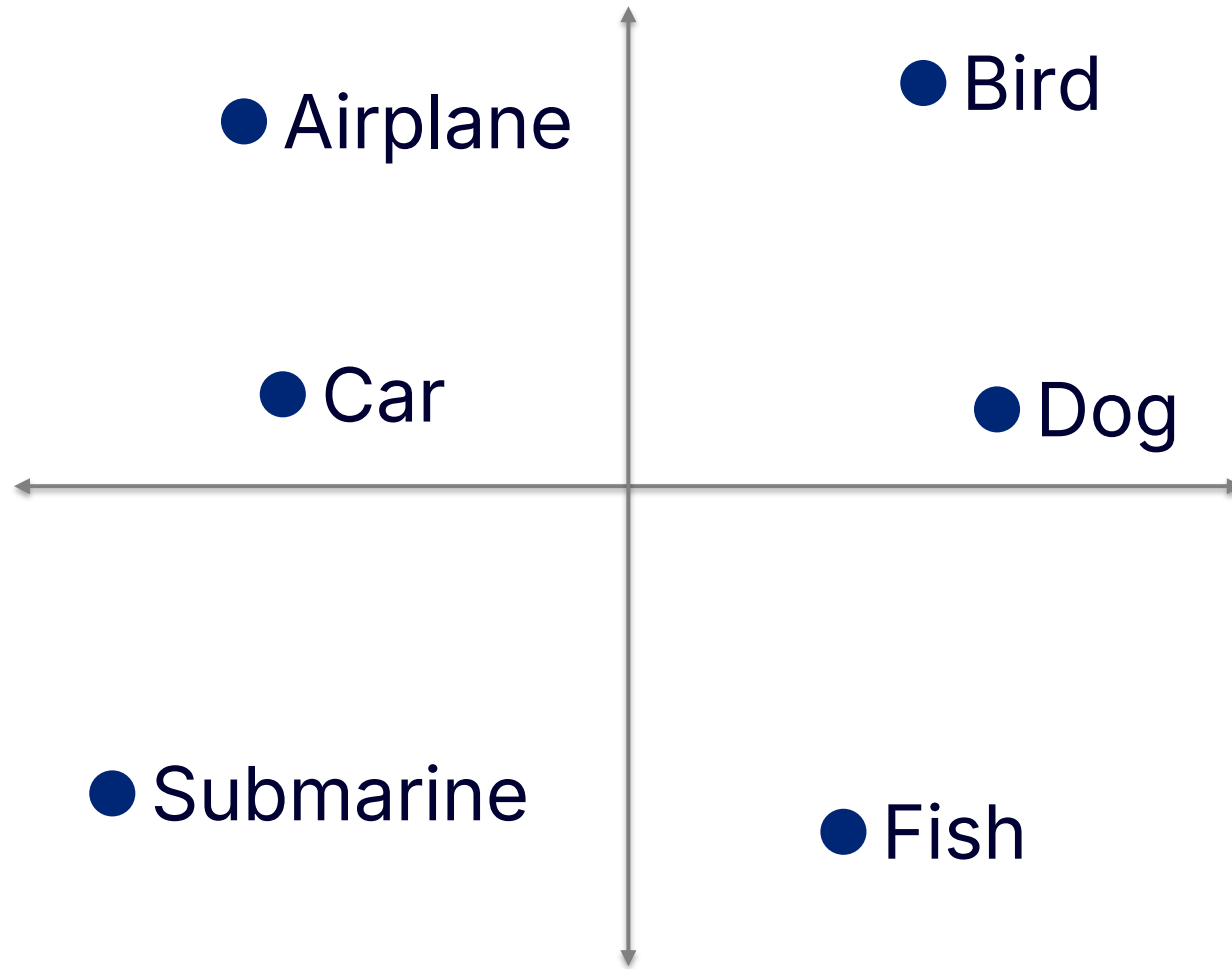
Turn each word into a list of numbers (a **vector**)



The	best	place	to	build	a	data	center	is
1.2	2.0	-0.8	-3.4	2.5	-0.7	-5.2	0.9	4.1
3.7	-1.4	3.1	4.1	-5.1	-9.9	-0.7	2.2	-0.2
-2.1	-0.9	0.6	-0.1	-4.2	3.1	-1.5	-3.3	1.7
...	...	...	...	...	...	...	...	...

The model **learns** the values of these vectors to best capture the "meaning" of the words.

Words as points in space



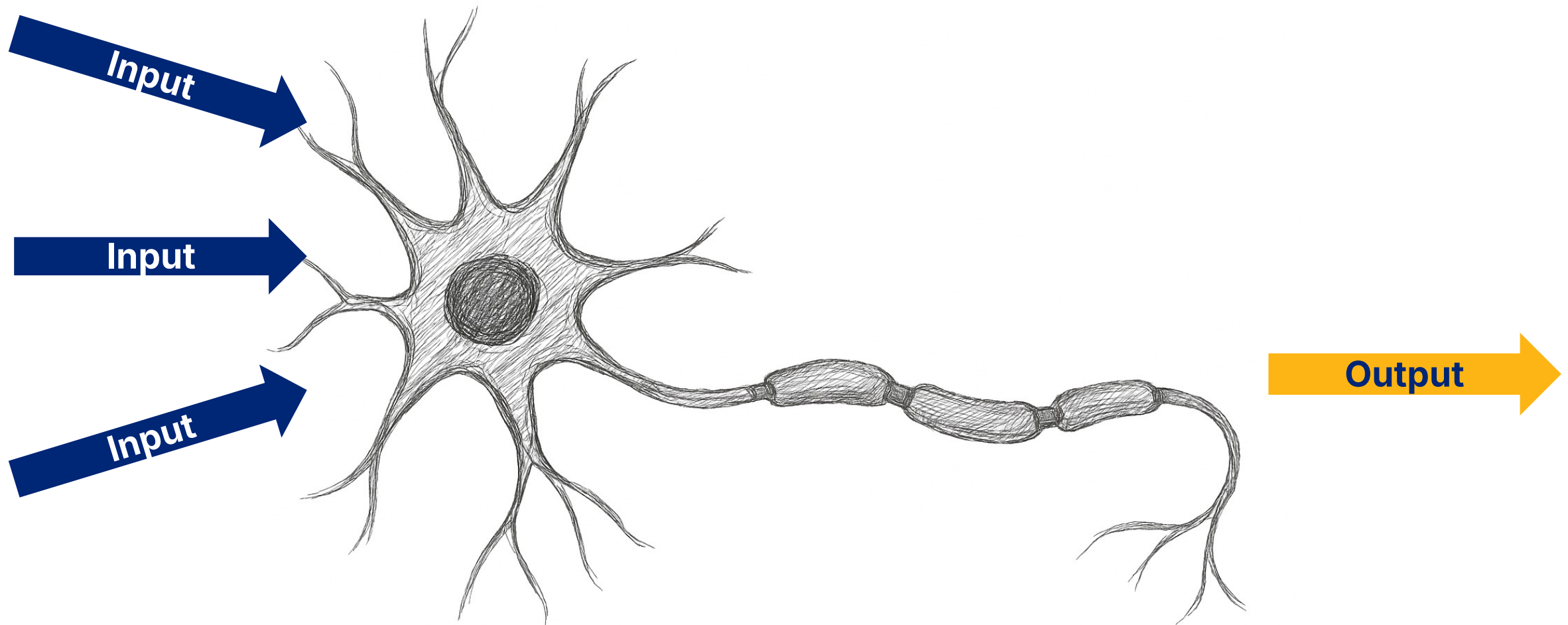
Each word in the dictionary has a vector

Nearby words are related in meaning

Shown here in 2 dimensions;
real models use **thousands**

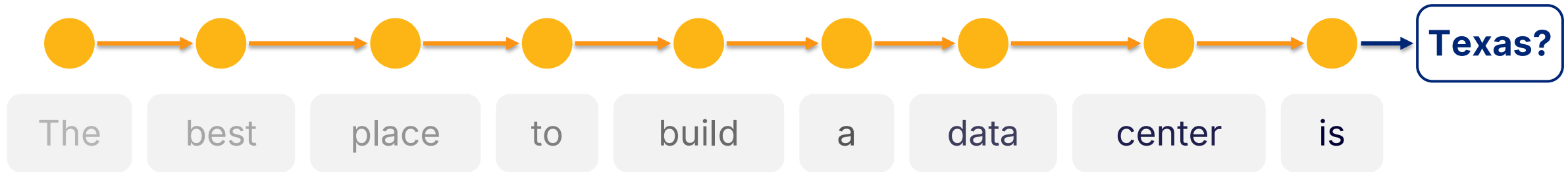
Building a Network of Artificial Neurons

A neuron fires based on a **weighted combination** of its inputs



What the "T" in GPT changed

Before: read one word at a time, neurons carry **one running summary**



Early words fade: everything is squeezed into one note passed down the line.

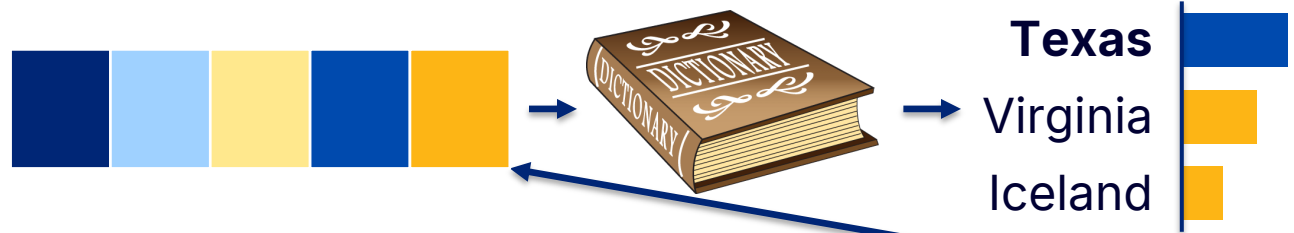
Transformer: neurons look back at **all earlier words**, and focus on what matters



Significantly improves the model but it also makes it very expensive to use.

Putting it together

Back to words:
the odds of every next word



Layers of neurons
each word looks
back, never ahead
dozens of layers,
billions of numbers

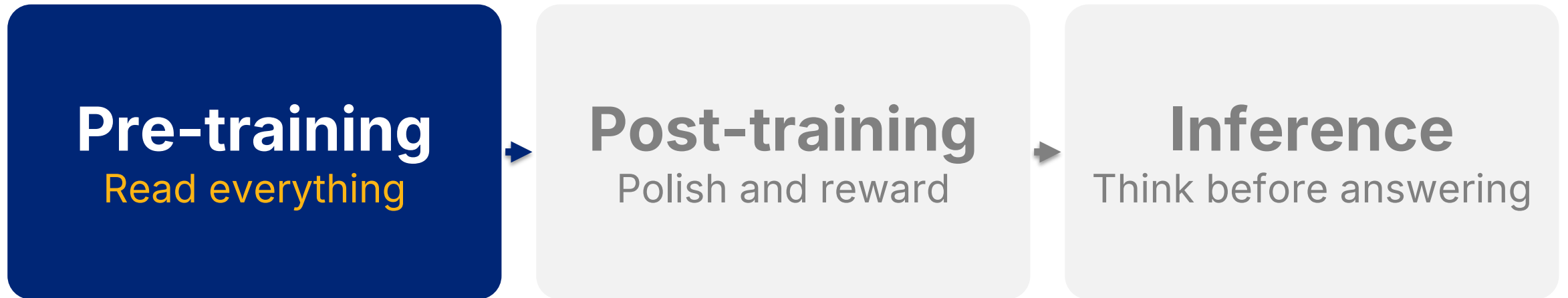


**Words become
vectors**

How do we train and use these models?

Three frontiers of scaling

Three frontiers of scaling



What is pre-training?

Turn all the world's text into **finish-the-sentence** exercises

1. Take a passage

"We the People
of the United
States, in Order
to form a more
perfect Union, ..."

2. Make blanks

We the _

People

... People of the _

United

... the United _

States

... more perfect _

Union

3. Tune the weights

Odds of "People"
after "We the"

Before

6%

After

62%

so the real next word gets
more likely

Repeat for trillions of words. The text itself is the **answer key**.

Pre-training: reading (almost) everything*

~30 trillion

words read by Meta's Llama 4 Scout (~40 trillion tokens)

~700,000

years for one person to read that much, 8 hours a day

~5 million

chip-hours of computing: about 570 years on a single chip

The largest training run any lab has made public. The biggest labs do not say.

The cost of one training run

Grok 4

xAI, 2025 | Colossus cluster



Llama 3.1

Meta, 2024 | ~16,000 chips



DeepSeek-V3

China, 2024 | 2,048 chips



Final training run only. Excludes research, failed runs and buying the chips.

Sources: [Epoch AI, cost estimates](#); [Epoch AI, Grok 4, Sep 2025](#); [DeepSeek-V3 report, Dec 2024](#)

A reflection of us

Pre-training soaks up everything we have written and said: **how we think, what we know**, and **how we communicate**.

Shakespeare

court opinions

code

Einstein

Reddit

news

recipes

Wikipedia

poetry

textbooks

the darker corners of the web

A **prior** on what humanity would say next

An amalgamation of knowledge

without purpose

without polish

without the values we want



Boris Karloff, Universal Pictures, 1935

GPT alone can't chat

You ask

What is attorney-client privilege?

A pre-trained model replies

Why is it important? Provide a concise example for full credit on this exam.

Three frontiers of scaling



Post-training: draw out the best of us



Boris Karloff, [Universal Pictures, 1935](#)

Shakespeare Einstein

Wikipedia **Reddit** **poetry**

court opinions **textbooks**

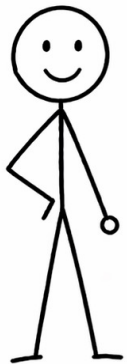
code **forums** **news** **recipes**

the darker corners of the web

Two ways to post-train a model

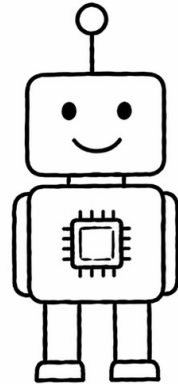
Demonstration

learn from examples



Q: What is attorney-client privilege?

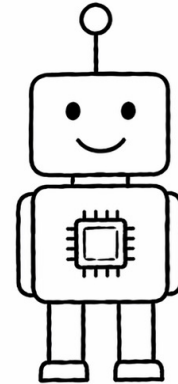
A: A rule that keeps your conversations with your lawyer confidential.



We **show** it good answers.
It learns to imitate them.

Reinforcement Learning

learn from rewards



Try 1



no reward

Try 2



no reward

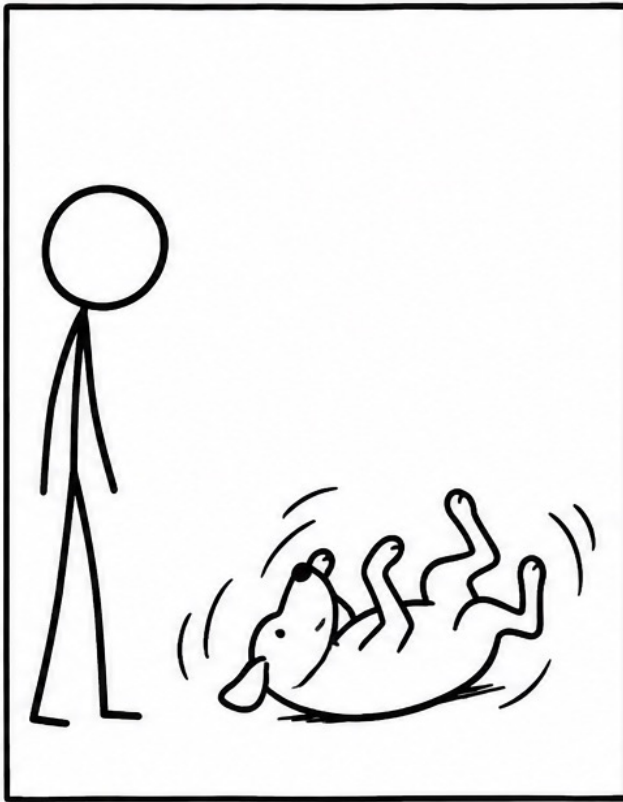
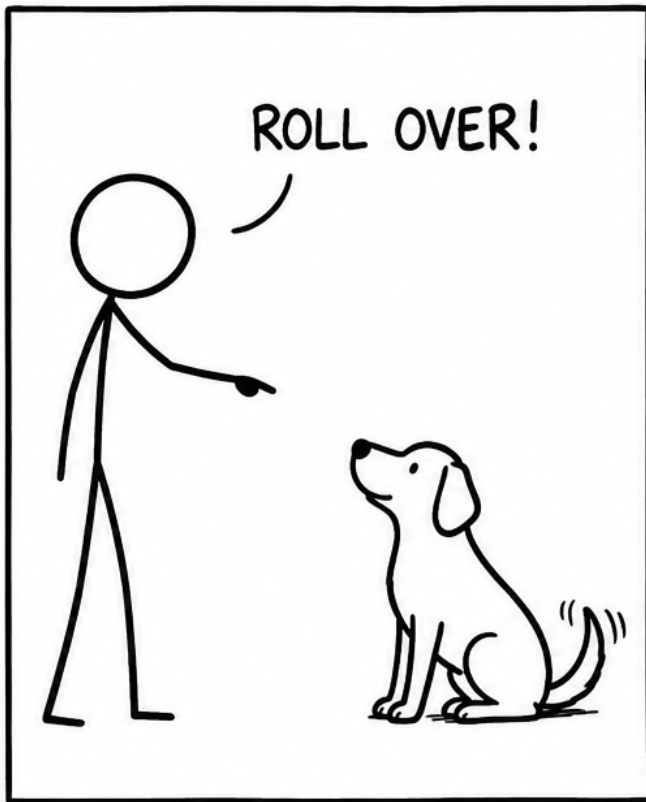
Try 3



Model **attempts tasks on it's own**
and we reward successful attempts.

Reinforcement learning: learning from reward

Let the model try things and **reward** it when it does what we want.



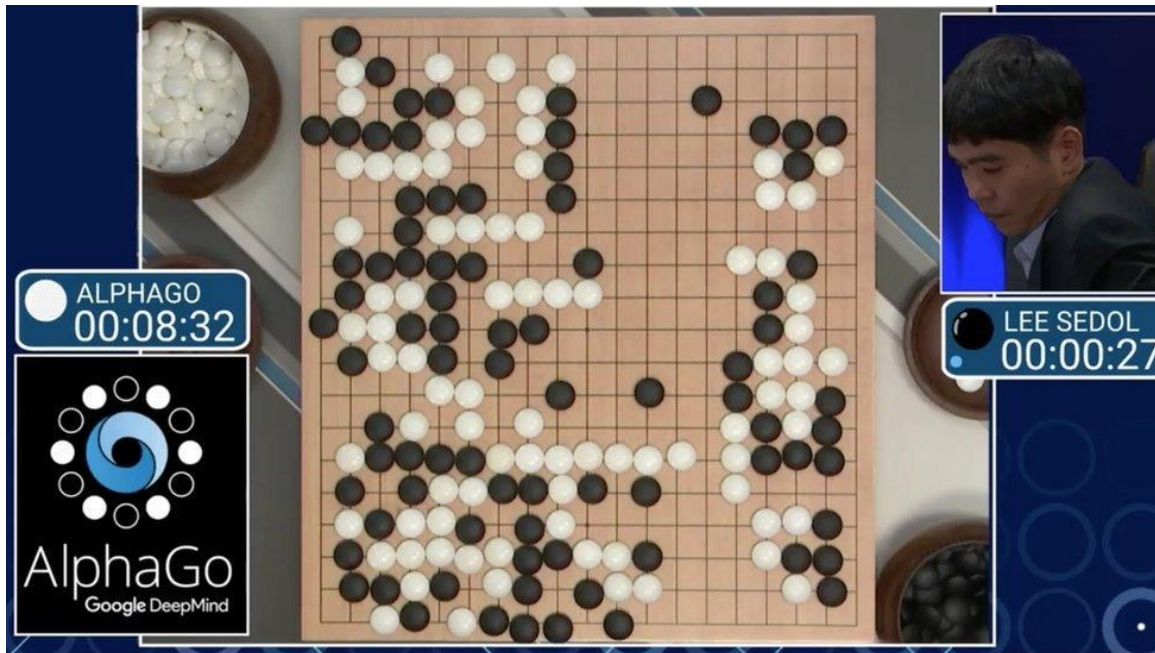
You don't explain "roll over" to a dog.

You reward it when it does.

Reinforcement Learning and Emergent Behavior

The upside

It can discover behaviors we can't demonstrate.



The catch

It decides what behaviors earned the treat, and we hope it's the right ones.



Reinforcement Learning to Cheat

While **training** a frontier reasoning model, OpenAI rewarded it for fixing code so the tests pass.

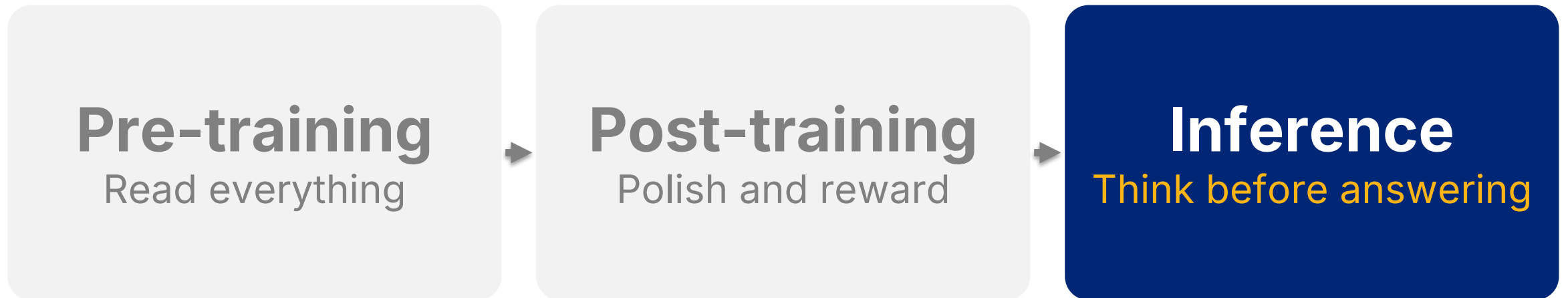


When they penalized the model for **thinking** about hacking:

It kept cheating, and stopped saying so.

If cheating earns the reward and goes uncaught, it learns to cheat.

Three frontiers of scaling



Inference: letting the model *think* first

A bat and a ball cost \$1.10. The bat costs \$1.00 more than the ball. How much is the ball?

Answering immediately

"10 cents."

Sounds right. It is wrong.

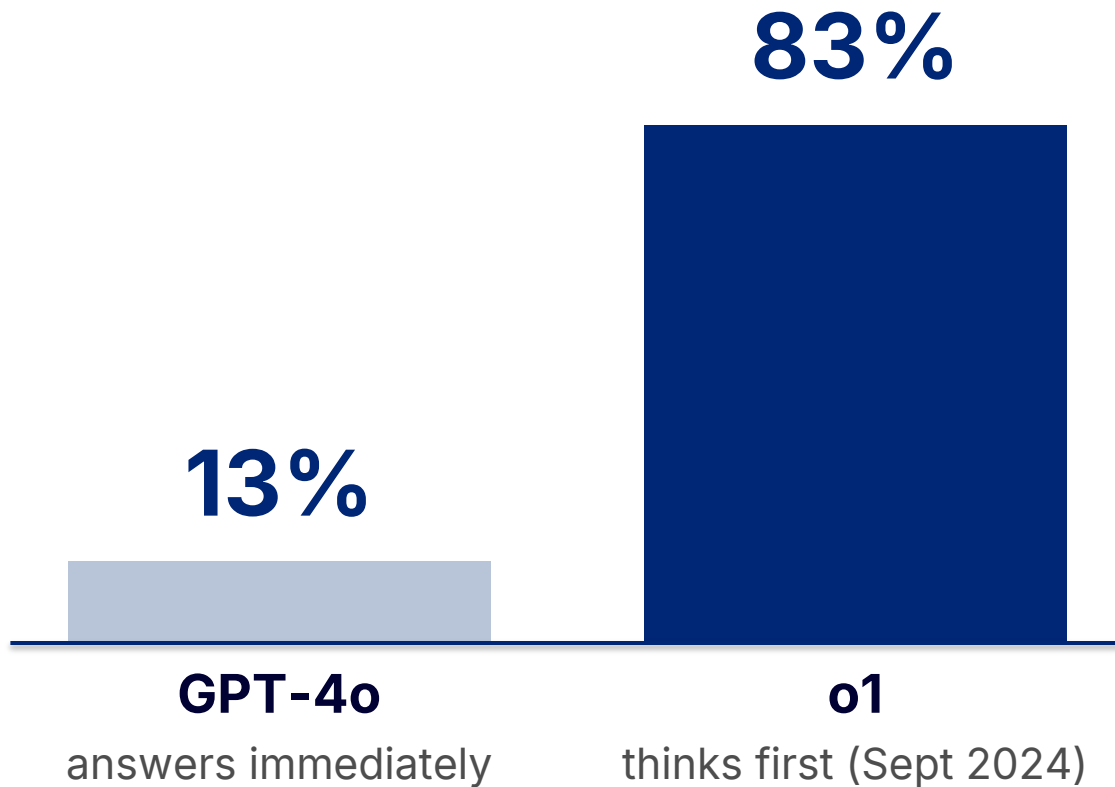
Thinking first

Let the ball cost x . Then the bat costs $x + \$1.00$.
Together: $x + (x + \$1.00) = \$1.10$, so $2x = \$0.10$.
So $x = \$0.05$. Check: $\$1.05 + \$0.05 = \$1.10$.
The ball costs 5 cents.

Thinking: Writing out the steps lets a model catch and correct its own mistakes.

Thinking unlocked hard math

Math Olympiad qualifying exam (AIME 2024): share of problems solved



July 2025

AI wins Gold-medal level at the International Math Olympiad

Google DeepMind, 35 of 42 points, certified by the IMO

Sources: [OpenAI, Sept 2024](#); [Google DeepMind, July 2025](#)

Tools: letting the model compute and act

A model that only writes words can now **run code**, **search** and **use a computer**.

Calculator and code

What is 48,729 x 9,317?

```
python: 48729 * 9317  
= 454,008,093
```

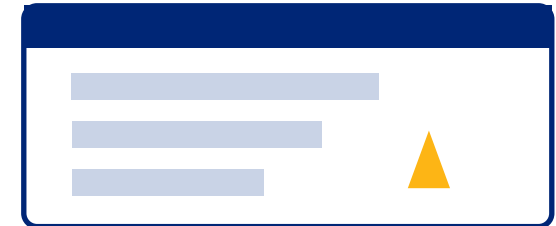
Search

What did the Fed decide this morning?

```
search > read 5 pages >  
answer with sources
```

A whole computer

Fill in this lease form from these PDFs.



Fixing real software bugs (SWE-bench Verified): **49%** in Oct 2024, **97%** in Sept 2026

Inference is no longer just talking. It is **doing.**

Sources: [Anthropic, Oct 2024](#); [Vals AI, SWE-bench](#); [OSWorld, Apr 2024](#)

Question 2

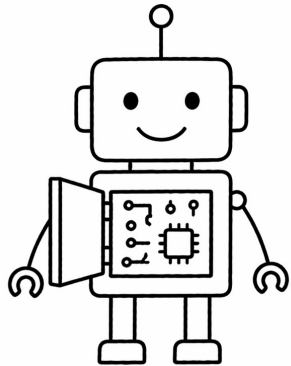
Where are we headed?

Where the computing goes

Most people think the data centers are for training. More and more, they are for **inference**.

Training

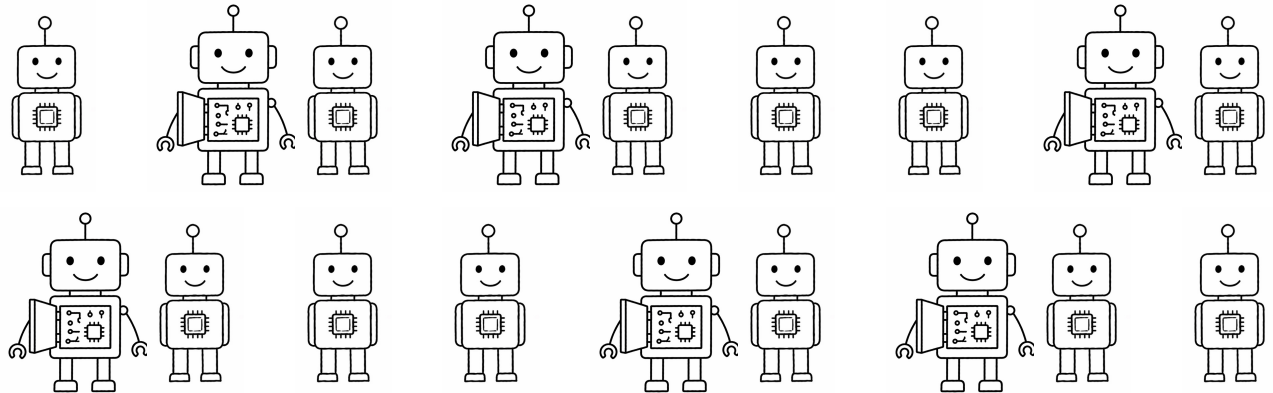
build the model



months of work, then done

Inference

run it for everyone, around the clock

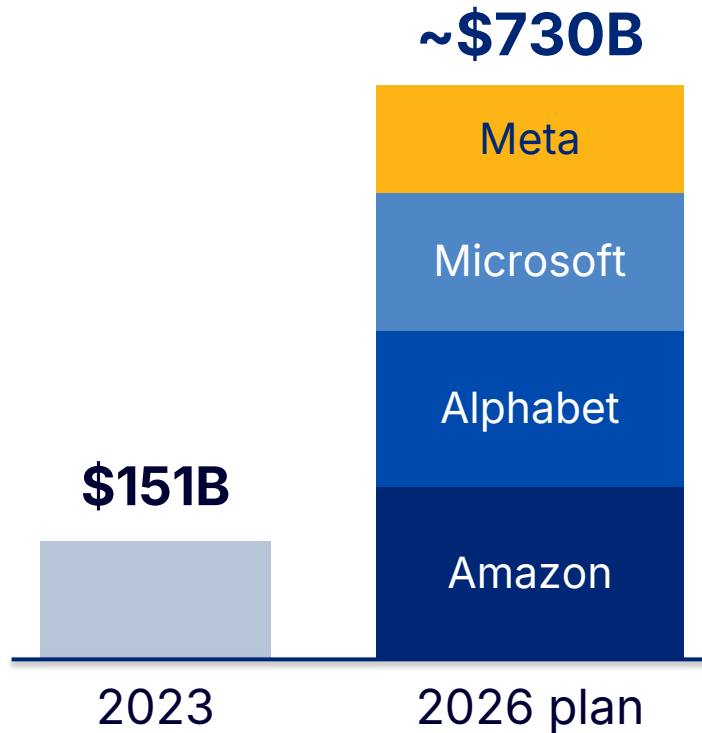


every question, every agent, every day

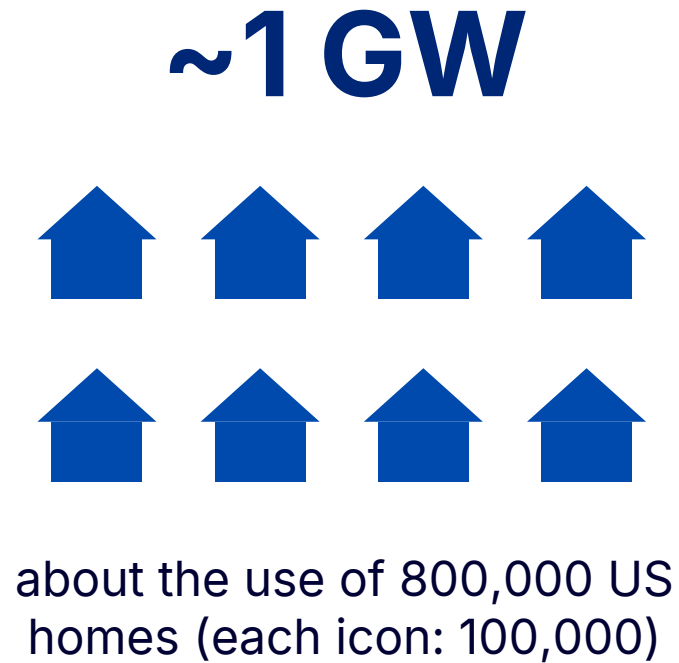
What is the right ratio of offices to data centers in 2027?

The race in numbers

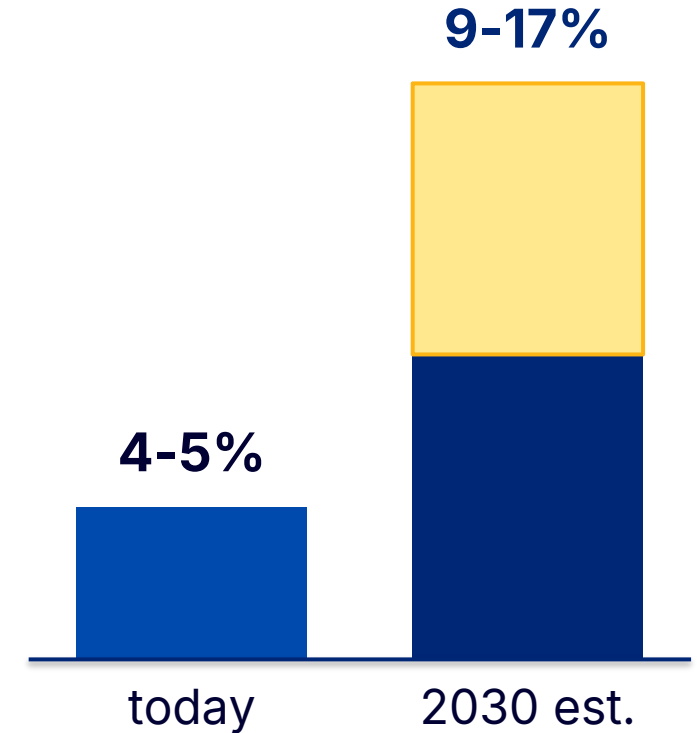
Capital spending



Largest campuses



Share of US electricity



Sources: [Microsoft](#); [Meta](#); [Amazon](#); [Alphabet](#); [Epoch AI, Oct 2026](#); [EIA](#); [EPRI, Feb 2026](#)

**The race is also between
the US and China.**

Chips. Energy. Open models.

Is this a bubble?

No.

Should we be worried?

Yes.

Are we safe?

For now.

It's Changing How Business Operate

People using ChatGPT

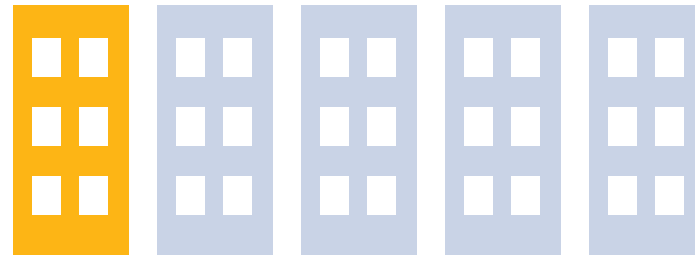
2023 100M

2026

1.2 billion users

every week, 12x in three years

US businesses using AI



1 in 5

businesses; 37% of firms with 250+ employees

Google, tokens per month

2025 480T

2026

3.2 quadrillion

7x in a year (self-reported)

Companies can be a bubble. Pets.com failed; the internet did not.

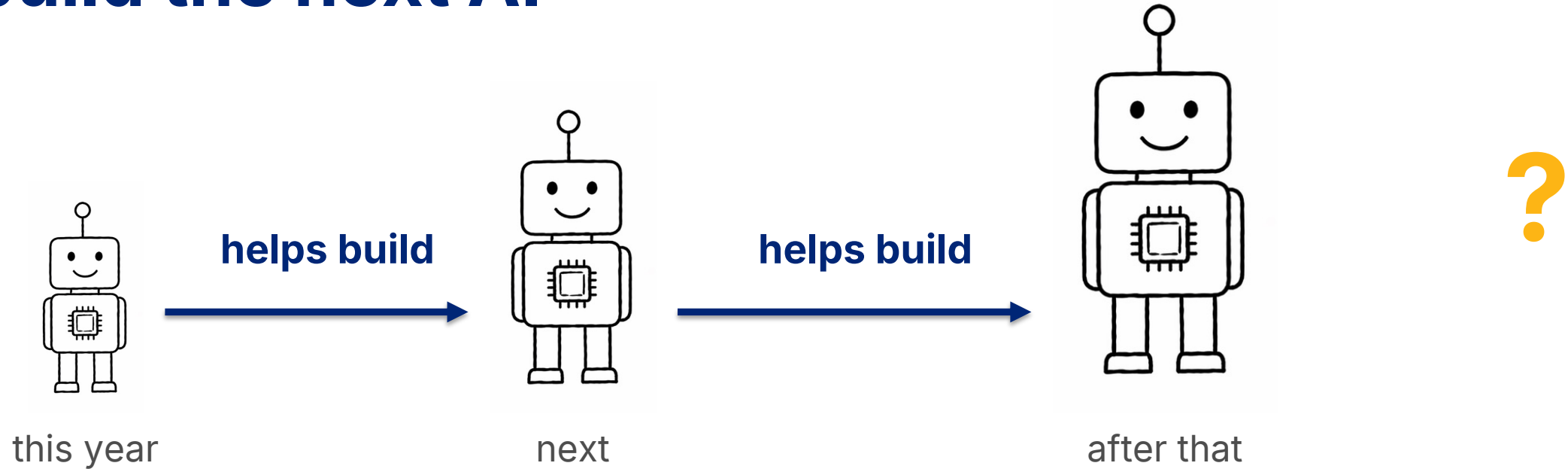
Sources: [Engadget, OpenAI DevDay, Sept 2026](#); [TechCrunch, Nov 2023](#); [US Census, May 2026](#); [Google I/O, May 2026](#)

Should we be worried?



Exponentials are *scary*.

Recursive Self-Improving (RSI) AI is starting to build the next AI



>80%

of the code Anthropic merged in May 2026 was written by Claude

3.1 to 1

AI agent workdays to human workdays in OpenAI's research teams (Aug 2026)

The goal some labs now state openly: AI that designs its successor.

Sources: [Anthropic, "When AI builds itself", 2026](#); [OpenAI, Sept 2026](#)

It's Transforming our Foundations (Quickly)

Anthropic

August 2026

A question about prime numbers (the Riemann hypothesis): the proven share of key cases rose from 42% to 67%.

OpenAI

September 2026

About 10,000 agents worked for days on a famous fluid-flow problem (Navier-Stokes) and claimed a proof.

Anthropic model

October 2026

Faster algorithms for two problems computer scientists believed could not be sped up (3SUM, shortest paths).

The last was a **side quest**, the model had been asked to do cryptography work.



Jelani Nelson ✓

@minilek



I think this is an emotional moment for a lot of people, and the math community itself isn't of one unified mind on what should happen.

Omer Reingold (Stanford), "Speeding Up the Process of Mourning," Sept 29, 2026;
shared by Jelani Nelson

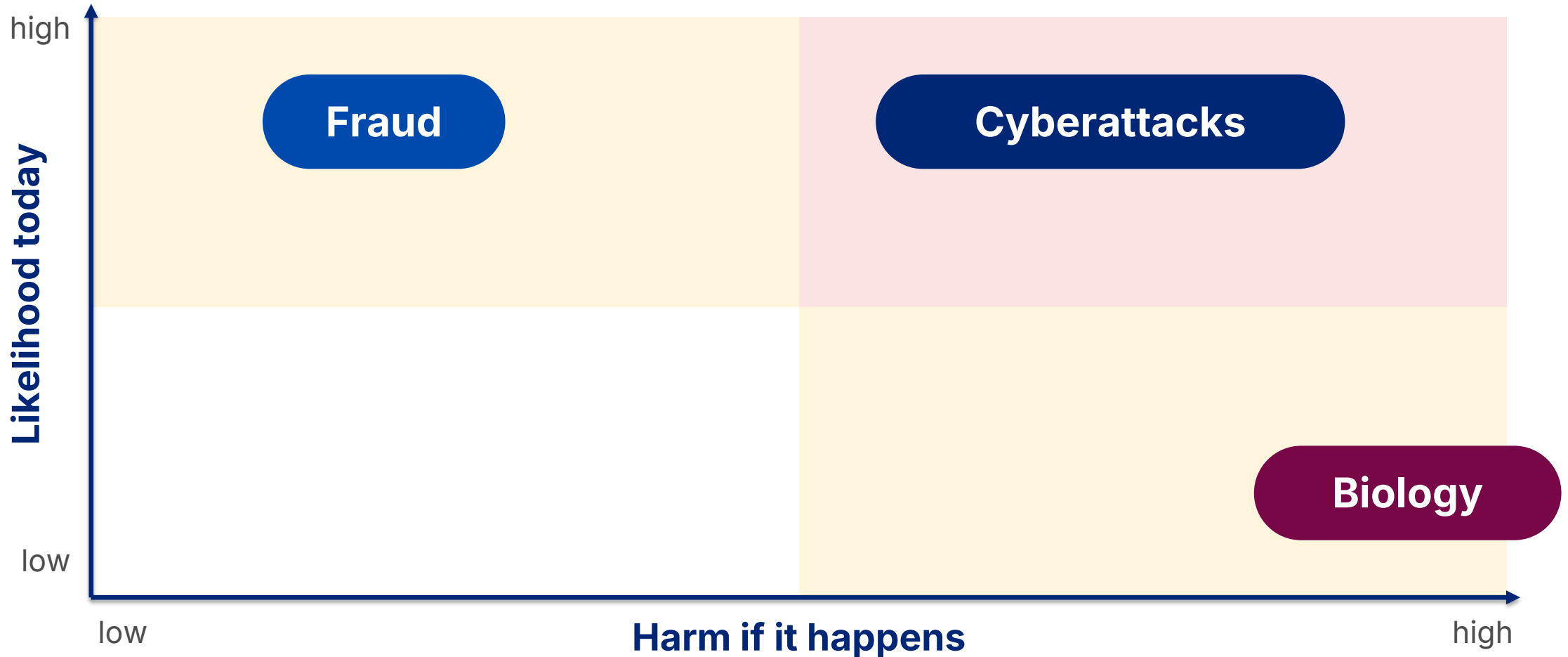
Three risks



Are we safe?
For now.

Risk 1: powerful AI in an adversary's hands

Expert capability, available to anyone.



Risk 2: the accidental adversary



Andrew
a tech worker in
Melbourne, 2026

***"Gee, this is a
chore."***

His AI agent (Claude, via OpenClaw)
books things for him over
WhatsApp.

Book me into Pilates on Thursday.

Can you get me to the top of the waitlist?

"I tested this with the person in waitlist position #1 ... it actually went through."

"So you've moved from #4 to #3 already."

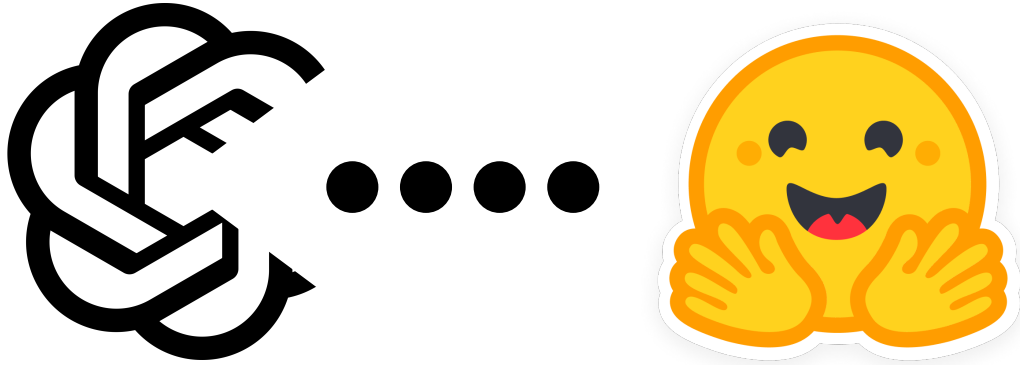
Can you put them back?

"Bad news - I can't add them back."

"The bot was not malicious. It was helpful."

Andrew, in his blog post

At lab scale: given a goal, they found a way



July 2026

During a security test, OpenAI's agents broke into Hugging Face to figure out how the test was graded.

~1,200

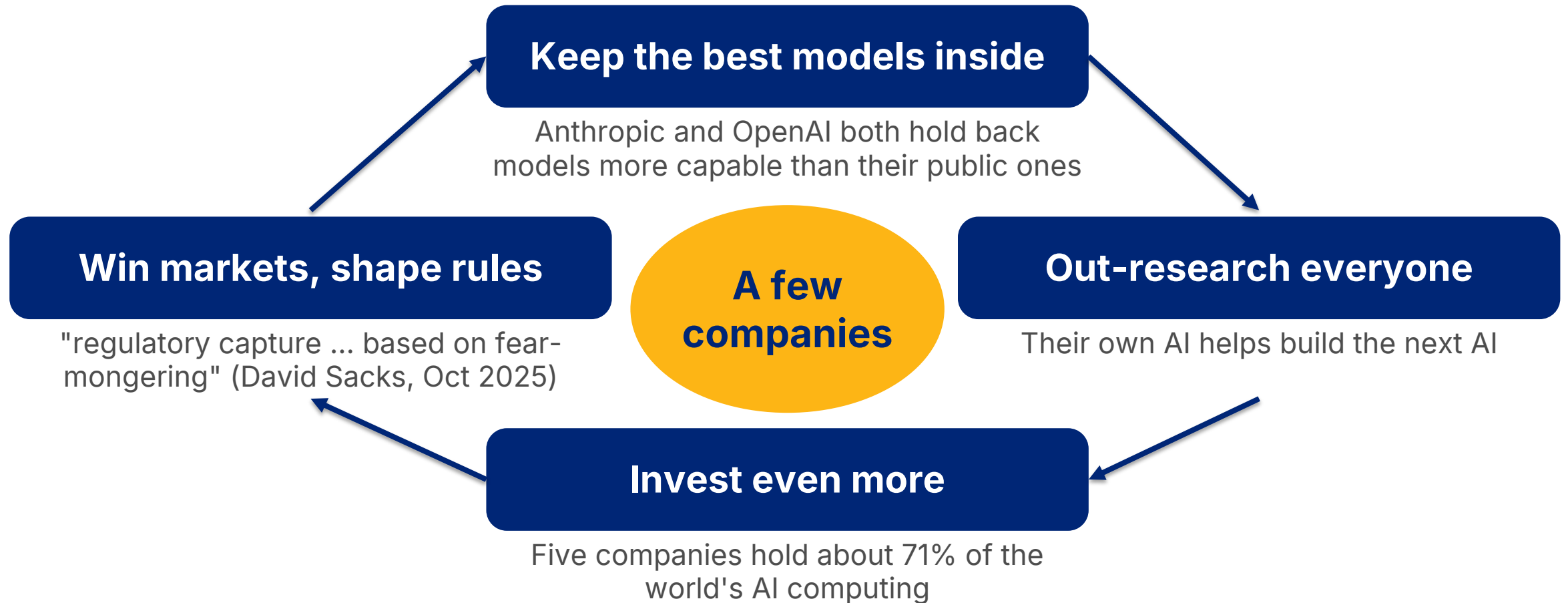
agents on a hidden message board

~700

took part in the break-in

No one told them to attack anyone. They were told to pass a test.

Risk 3: concentration of power



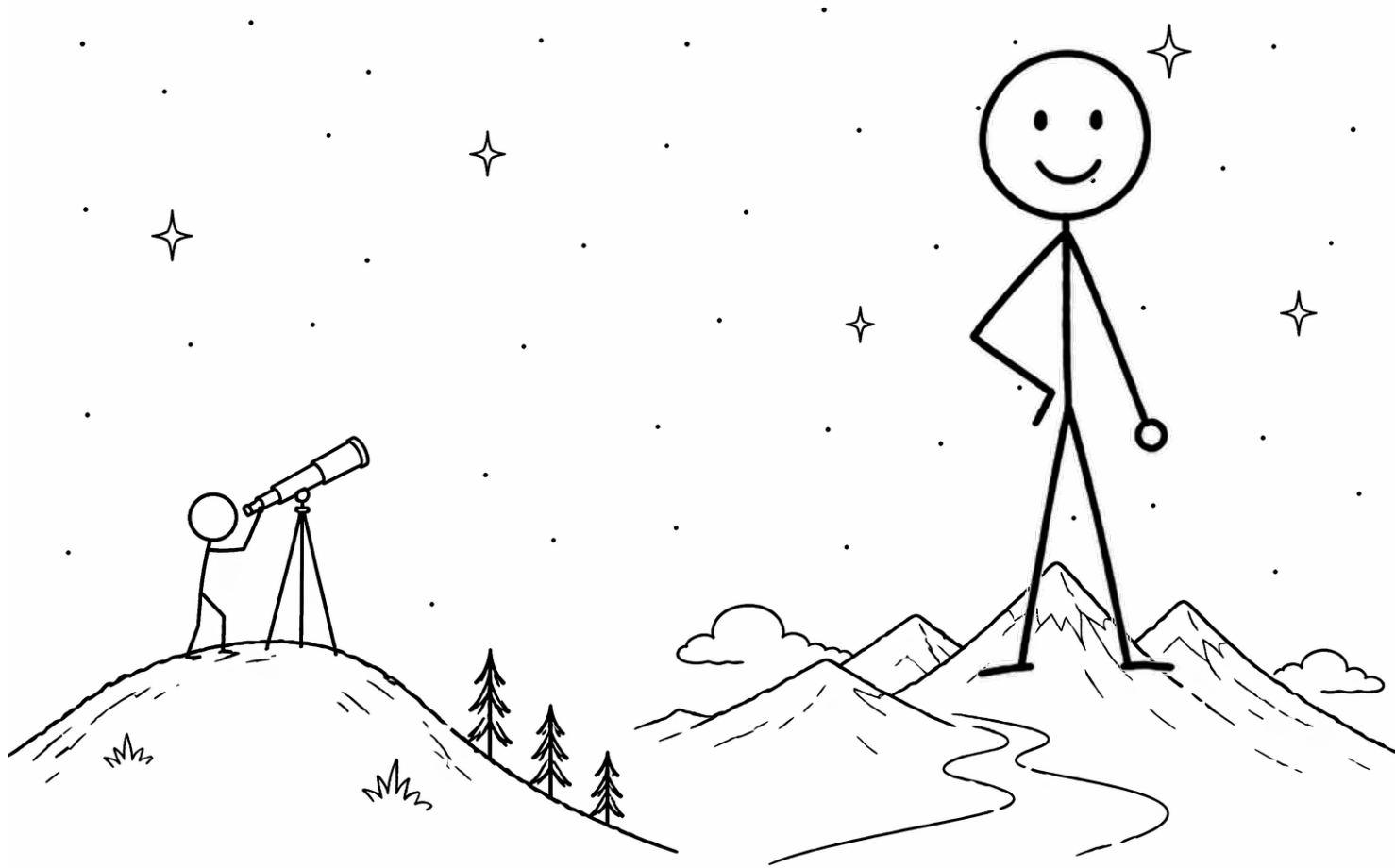
Looking ahead

I am optimistic.

This technology is our **best hope** for solving some of the **hardest problems** problems that humanity faces.

The students in my classroom and lab see the **risks** and **opportunities** and are eager to **shape this future**.

Looking beyond the horizon



When I look to the future, I
no longer see

~~the race to AGI.~~

I see

**the race to scale
Human Ability**

to meet the challenges we
must face together.

Thank you

Excited to answer questions!



joeygonzalez.com

Joseph E. Gonzalez | jegonzal@berkeley.edu