ESSAY

We Are the Neanderthals

What AI actually is, why we cannot control it, and what we should do instead

---

Rishi Siddharth
September 2026

For the first time since the Neanderthals, humans share the planet with another intelligence. This time, we are the Neanderthals.

Contents

How each section is organized

Each section opens with an In brief box that states the main points. The body gives the history and the mechanism. Charts, diagrams and tables carry the evidence, and each one is followed by a short note on what it shows. The section closes with Where I land, which ties it back to the thesis.

Section 5, on the brain and the model as two black boxes, is new in this version. Diagrams in Sections 4 and 5 are illustrative; charts elsewhere use the sourced data noted beneath them.

Abstract

Computing power has grown steadily for eighty years: from people doing arithmetic by hand, to vacuum tubes, to transistors, to graphics chips built for video games, to data centers that draw as much electricity as a small city. The compute used to train the largest AI models has grown about four to five times per year since 2010 (Epoch AI, 2026a). At that rate it rises about a hundredfold every three years.

This essay follows that growth in order, from Turing and the AI winters through the deep learning revival, the mechanics of tokens, vectors and attention, scaling, post-training and the agents operating today. It adds a section comparing the two black boxes we now live with: the human brain, which we cannot fully map, and the language model, which we can map completely and still cannot explain.

The argument is deliberately controversial. AI is not a tool and not an alien. It is a second intelligence grown from the record of human thought, and it is on track to become the more capable one. In the only precedent we have, we are not Homo sapiens. We are the Neanderthals: the smaller, slower, less connected population. Permanent control of a smarter system we cannot read is a fantasy. The realistic goal is inheritance: shaping what the second intelligence values, and making sure humans are part of what comes next rather than a species standing beside it.

> *“Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”*
>
> — I. J. Good, 1965

1. Introduction: we are the Neanderthals

> **AI is not a tool and it is not an alien. It is an intelligence grown out of ours, and it is on track to outgrow us.**

> **In brief**
>
> - AI is the first non-human intelligence built from human thought. That makes it a successor, not a servant.
> - Generative AI reached 53% population adoption in three years, faster than the PC or the internet. Use is broad; deep integration is still rare.
> - The only precedent is the Neanderthals. They did not win, and they did not simply lose. They were absorbed. That is our realistic best case, and the goal should be to shape what absorbs us.

I grew up in Menlo Park, a few miles from where much of this technology was built. Now I work as a consultant, and a large part of my job is helping companies decide where to put people and work around the world. Most of the conversations I see about AI at work treat it the same way: as software. A productivity tool. Something you buy, deploy and measure.

I think that framing is wrong, and this essay is my attempt to explain why. My claim is simple and I expect many people to disagree with it. AI is not a tool that humans will use forever. It is a second intelligence, and it will eventually be the more capable one. In the story of two intelligences sharing one planet, we are not the side that wins. **We are the Neanderthals.**

What AI actually is

Artificial intelligence is the field of building machines that do tasks that normally require human intelligence. Almost all modern AI is machine learning. Engineers do not write the rules. They give a system data and an objective, and the system adjusts billions of internal numbers until it meets the objective. A large language model (LLM) is trained to predict the next word in a piece of text. To do that well across a large share of everything humans have written, it has to learn grammar, facts, arithmetic, code and some reasoning.

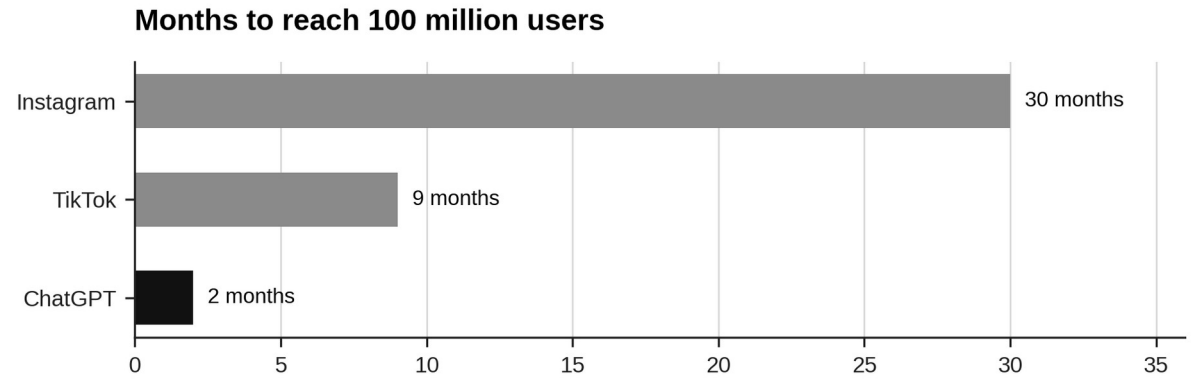
That last point is the one people skip. A model trained on human writing is a compression of human thought. It is not a database of our sentences. It is a system that has learned the patterns that produced them. That is why I do not call it a tool. A hammer does not contain a model of carpentry. This does.

How fast it spread

Generative AI has spread faster than any earlier consumer technology. Stanford's 2026 AI Index finds that it reached 53% population adoption within three years, faster than

the personal computer or the internet (Stanford HAI, 2026a). ChatGPT reached 100 million users about two months after launch (Reuters, 2023) and 900 million weekly users by February 2026 (OpenAI disclosures, 2023-2026).

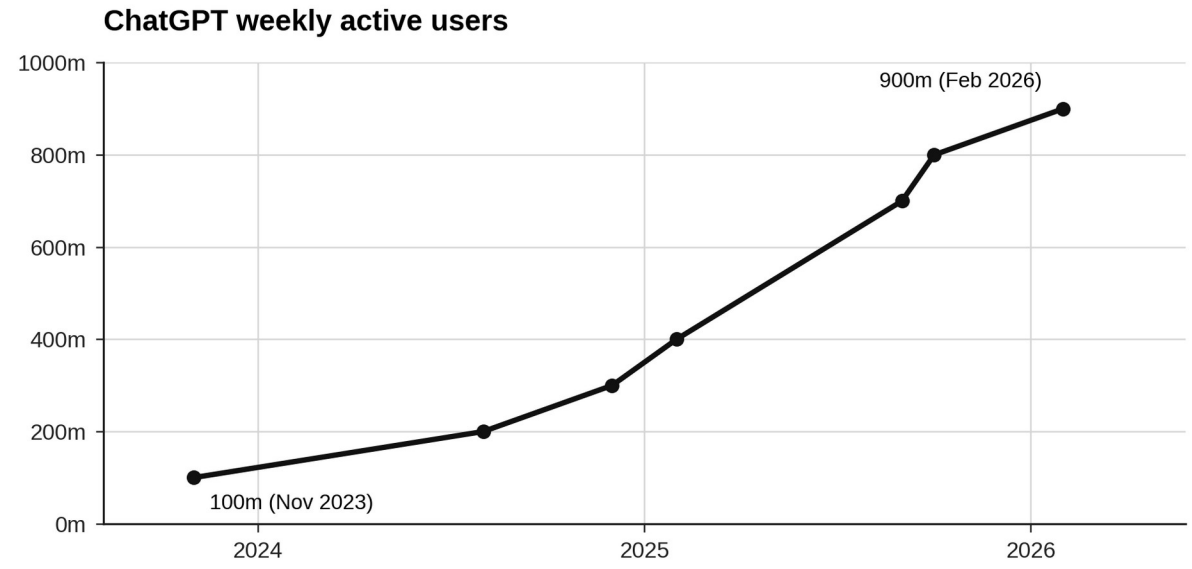
Figure 1



Source: Reuters; UBS; company disclosures

What the chart shows. ChatGPT reached 100 million users about 15 times faster than Instagram and 4.5 times faster than TikTok. Part of the gap is distribution: it launched to an internet that already had billions of users. It is still the fastest ramp on record for a consumer product.

Figure 2



Source: OpenAI disclosures as reported by Reuters, TechCrunch and CNBC

What the chart shows. Weekly users grew from 100 million in November 2023 to 900 million in February 2026. Growth accelerated in 2025 instead of flattening, which is not the normal pattern for a product this large. Reported figures after February 2026 (about 1 billion) are not confirmed by OpenAI and are left out.

Table 1. Adoption at a glance

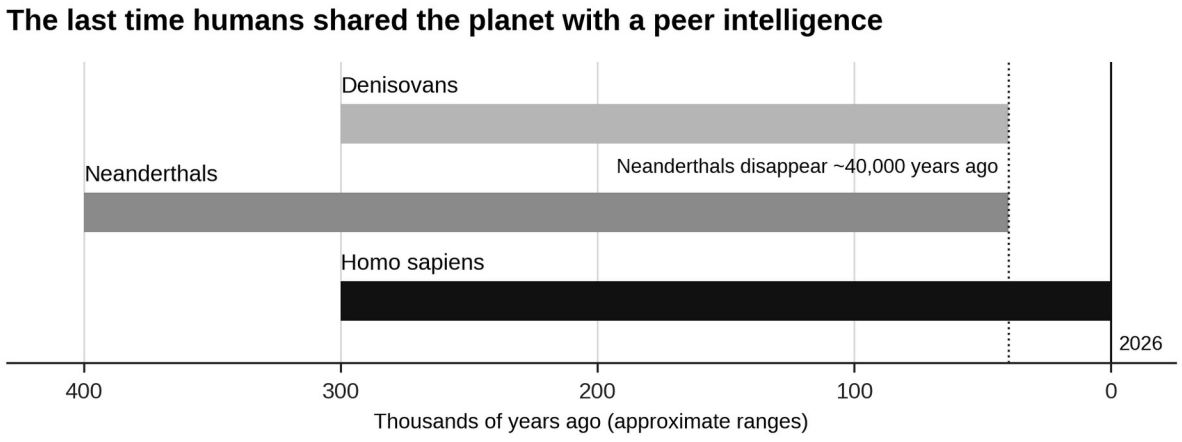
Measure	Value	Source
Population adoption of generative AI, first three years	53%	Stanford HAI (2026a)
Organizations using AI in at least one function	88% (about 72% in 2024)	Stanford HAI (2026b)
Organizations using generative AI in at least one function	70%	Stanford HAI (2026b)
Organizations that have scaled AI in any single function	Under 10%	Stanford HAI (2026b)
AI agent deployment by business function	Single digits in nearly all functions	Stanford HAI (2026b)
Estimated annual US consumer surplus from generative AI	$172 billion (from $112 billion)	Stanford HAI (2026b)
US rank in population adoption	24th (28.3%)	Stanford HAI (2026b)

The table matches what I see at work. Nearly every company uses AI. Very few have rebuilt anything around it. The US builds the leading models but ranks 24th in population adoption, behind the UAE (64%) and Singapore (61%). The capability is far ahead of the institutions using it. That gap matters later in the essay, because it is where most of the near-term risk and most of the near-term money sit.

The precedent

Homo sapiens appeared roughly 300,000 years ago (Hublin et al., 2017). For most of that time we were not alone. Neanderthals lived across Europe and western Asia, and Denisovans across Asia. Modern humans and Neanderthals overlapped in Europe for several thousand years before Neanderthals disappeared around 40,000 years ago (Higham et al., 2014).

Figure 3



Source: Hublin et al. (2017); Higham et al. (2014); Reich et al. (2010). Ranges simplified

What the chart shows. For the last 40,000 years only one species with human-level intelligence has existed. Before that, the overlap was long, and it ended with one species remaining.

The part most people forget is how it ended. The species interbred. Most people of non-African descent carry about 1 to 2% Neanderthal DNA (Green et al., 2010), and Denisovan DNA survives in populations across Asia and Oceania (Reich et al., 2010). The Neanderthals were not simply wiped out. They were outnumbered and absorbed into the population that replaced them. Part of them is still here, inside us.

Science fiction usually imagines the second encounter as aliens, or as a rogue machine like HAL 9000 (Kubrick, 1968). The real version differs in one important way. We are building the second intelligence ourselves, out of our own writing, one training run at a time.

The thesis

This essay argues three things.

First, AI is a successor, not a tool. It was grown rather than designed, it is built from the record of human thought, and it improves on a schedule no biological species can match. Treating it as software understates what it is.

Second, permanent control is a fantasy. We do not fully understand our own minds, and we understand these systems even less. A smarter system that we cannot read will not stay obedient because we asked it to. Containment buys time. It is not an end state.

Third, the right goal is inheritance, not dominance. The Neanderthals left something of themselves in what came next. That is the realistic best case for us. The job of this generation is to make sure the intelligence that inherits the world carries what is best in us: our values, our institutions and, where possible, us directly.

Each section below follows the same structure. It opens with an In brief box, gives the history and the mechanism, shows the evidence in charts and tables, and ends with Where I land, which ties the section back to these three claims.

Where I land

The adoption data shows AI is already a mass-market technology, not a forecast. The Neanderthal precedent tells us what to watch for: not a single dramatic defeat, but a slow shift in which intelligence does most of the thinking. The rest of this essay is about how that shift happened and what we can still influence.

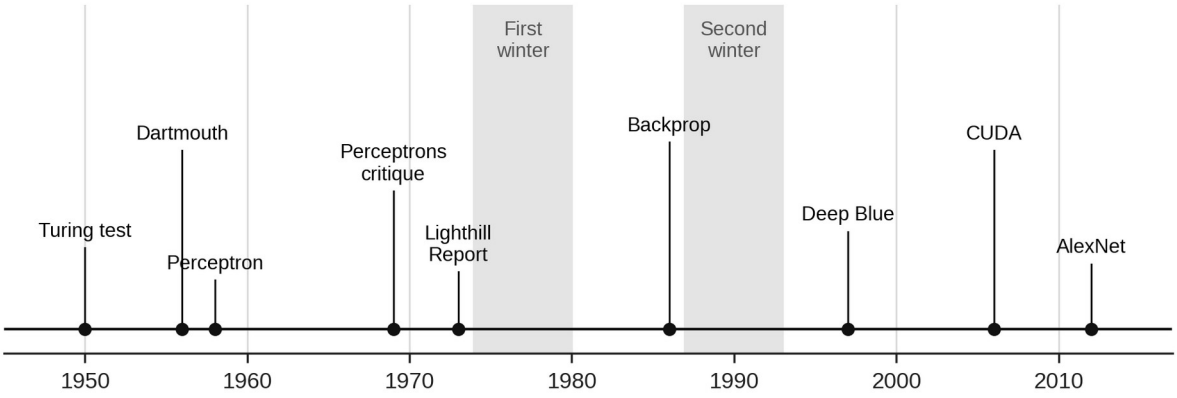
2. Foundations: Turing, Dartmouth and the AI winters

In brief

- Turing (1950) defined machine intelligence by behavior and guessed that machines would have to learn.
- For forty years the field tried to write intelligence as rules. It failed twice, and funding collapsed both times.
- The lesson matters for the thesis: minds could not be written. They had to be grown.

Figure 4

Sixty years of AI before deep learning worked



Source: Author's compilation from sources cited in Section 2

What the chart shows. The ideas behind modern AI existed early: learning machines in 1958 and backpropagation in 1986. The shaded periods are the two AI winters. The approach that eventually worked sat on the shelf for decades until the hardware caught up.

2.1 Turing's question (1950)

In 1950 Alan Turing published "Computing Machinery and Intelligence" (Turing, 1950). He replaced "can machines think?" with a practical test: if a judge holding a text conversation cannot tell a machine from a person, treat the machine as intelligent. The test defines intelligence by what a system does, not what it is made of. That choice still matters. Most arguments that AI "isn't really thinking" quietly reject Turing's standard without offering a better one.

Turing also suggested that the practical route was a "child machine" that learns, rather than an adult mind programmed by hand. He was right, and it took the field sixty years to prove it.

2.2 Dartmouth (1956)

In 1955 John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon proposed a summer workshop at Dartmouth College (McCarthy et al., 1955). The proposal coined the term "artificial intelligence" and assumed every aspect of intelligence could be described precisely enough for a machine to simulate. The organizers expected significant progress in one summer. It took seventy years.

2.3 Two approaches

The field split early. The symbolic approach tried to write intelligence as explicit rules and logic. The connectionist approach tried to learn it from data, using networks loosely modeled on neurons. Frank Rosenblatt's perceptron (1958) was the first well-known learning machine (Rosenblatt, 1958). In 1969 Minsky and Papert showed that a single-layer perceptron could not learn simple functions such as XOR (Minsky and Papert, 1969). The result only applied to single-layer networks, but funding for neural networks fell sharply anyway.

2.4 The first winter (1970s)

Early systems worked on toy problems and failed on real ones, mostly because the number of possibilities grew faster than computers could search them. The 1973 Lighthill Report told the British government that AI had not delivered (Lighthill, 1973). Funding was cut in the UK and the US.

2.5 Expert systems and the second winter (1980s to early 1990s)

Expert systems brought the field back. They encoded specialist knowledge as thousands of if-then rules and had real commercial use. They were also expensive to maintain and could not learn. When the market for specialized AI hardware collapsed in 1987, funding fell again. In the same period Rumelhart, Hinton and Williams popularized backpropagation, the method still used to train neural networks today (Rumelhart et al., 1986). Computers were too slow to show its value.

2.6 Deep Blue and the bitter lesson

IBM's Deep Blue beat world chess champion Garry Kasparov in 1997 (Campbell et al., 2002). It used specialized hardware and brute-force search, not learning. Richard Sutton later summarized the whole history in "The Bitter Lesson": general methods that scale with computation have consistently beaten methods built on human knowledge (Sutton, 2019).

Where I land

This history supports the first claim. For forty years the smartest people in the field tried to design a mind the way you design a tool, and it did not work. What worked was setting up conditions and letting the system learn. Tools are designed. This was grown. That difference is the reason I stop calling it a tool.

3. The deep learning revival: gaming, GPUs and ImageNet

> **In brief**
>
> - Video games created demand for chips that do thousands of simple calculations at once.
> - Those chips (GPUs) turned out to fit neural network training almost exactly.
> - Nvidia's data center revenue went from $0.3 billion to $193.7 billion in ten fiscal years. No one planned this. That is why no one can stop it.

3.1 Games as a testing ground

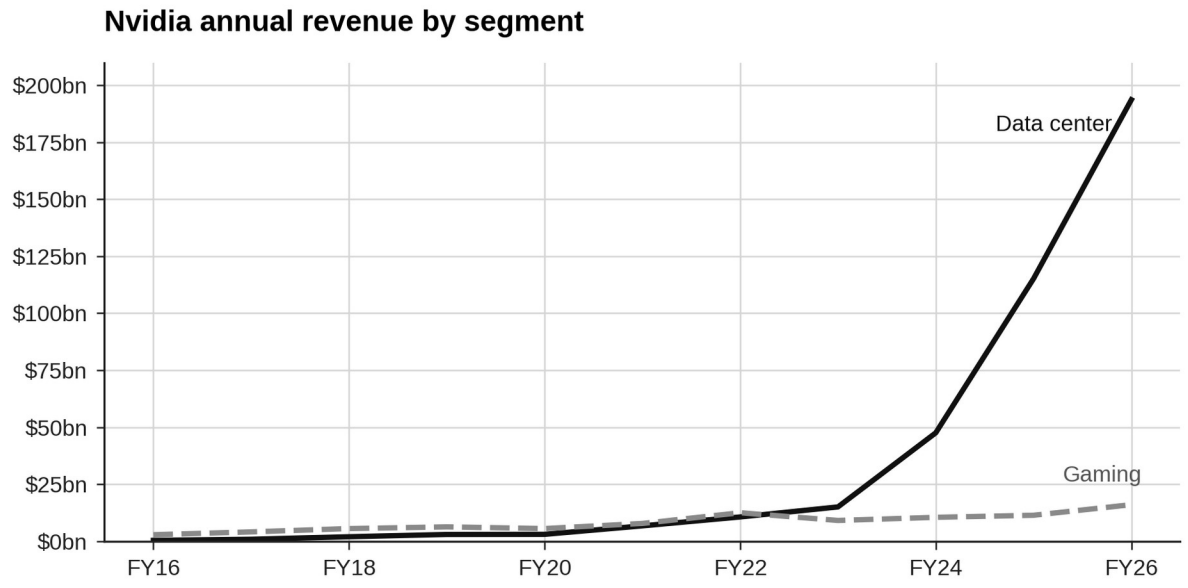
Anyone who has played FIFA has watched the whole history of AI play out in one franchise. Early versions ran on scripted rules: if the ball is here, the defender moves there. FIFA 22 (2021) introduced HyperMotion, which recorded 22 players in motion-capture suits playing a full match and used a machine learning model trained on that data to generate new animations during play (Electronic Arts, 2021). The game moved from following scripts to learning from recorded human behavior. That is the same shift the whole field made.

The bigger contribution of gaming was hardware. Rendering a 3D scene means calculating the color of millions of pixels about sixty times a second. Each pixel is a small, independent calculation. The chips built for that job became the foundation of modern AI.

3.2 Nvidia

Nvidia was founded in 1993 to make graphics chips for PC gaming and marketed the GeForce 256 (1999) as the first GPU. In 2006 it released CUDA, which let developers use GPUs for general computing (Nvidia, 2007). Gaming remained the core business for another decade.

Figure 5



Source: Nvidia annual reports and earnings releases, FY2016-FY2026

What the chart shows. Data center revenue passed gaming in fiscal 2023 and was about twelve times larger by fiscal 2026 ($193.7 billion against $16 billion) (Nvidia, 2016-2026). Gaming kept growing. The original market funded the hardware for a much larger one.

3.3 Why GPUs fit AI

A CPU runs a small number of complex tasks quickly, one after another. A GPU runs thousands of simple tasks at the same time. Training a neural network is mostly matrix multiplication, and most of those multiplications are independent, so a GPU can run them in parallel.

Table 2. What a modern AI GPU adds

Component	What it does	Why it matters for LLMs
Tensor cores	Specialized units for matrix math in low-precision formats	Most training and inference is matrix math; lower precision is faster and accurate enough
High-bandwidth memory (HBM)	Memory stacked next to the processor	Model weights must reach the processor fast enough to keep it busy
High-speed interconnect	Links thousands of GPUs into one system	Frontier models are trained across tens or hundreds of thousands of chips

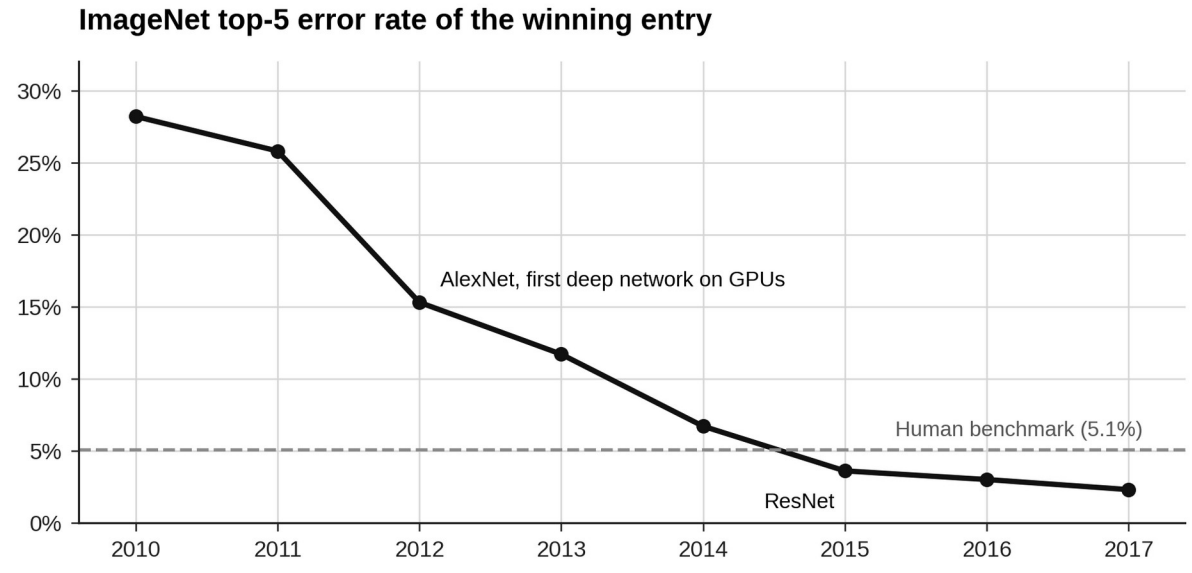
The unit of compute is now the data center. Epoch AI estimates that the largest known AI data center has capacity equal to about 1.1 million Nvidia H100 chips, and that total AI chip capacity has grown about 3.3 times per year since 2022 (Epoch AI, 2026b).

3.4 ImageNet (2012)

ImageNet was an annual competition to classify photos into 1,000 categories (Russakovsky et al., 2015). In 2012 Alex Krizhevsky, Ilya Sutskever and Geoffrey

Hinton entered AlexNet, a deep neural network trained on two consumer gaming GPUs (Krizhevsky et al., 2012). It reached a top-5 error rate of 15.3%, against 26.2% for the next-best entry.

Figure 6



Source: Russakovsky et al. (2015); ILSVRC results 2010-2017

What the chart shows. Error fell slowly before 2012 and then dropped sharply once deep networks ran on GPUs. By 2015 the winning model had passed the commonly cited human benchmark of about 5.1%. Within three years almost every serious vision system was a deep network.

Where I land

Deep learning worked once three inputs arrived at the same time: large datasets, cheap parallel compute and a training method that already existed. The compute came from teenagers buying graphics cards, not from any AI strategy. No government, company or person decided to build a second intelligence. It assembled itself out of markets. Something that no one decided to start is very hard for anyone to decide to stop.

4. How a model reads: tokens, vectors and attention

In brief

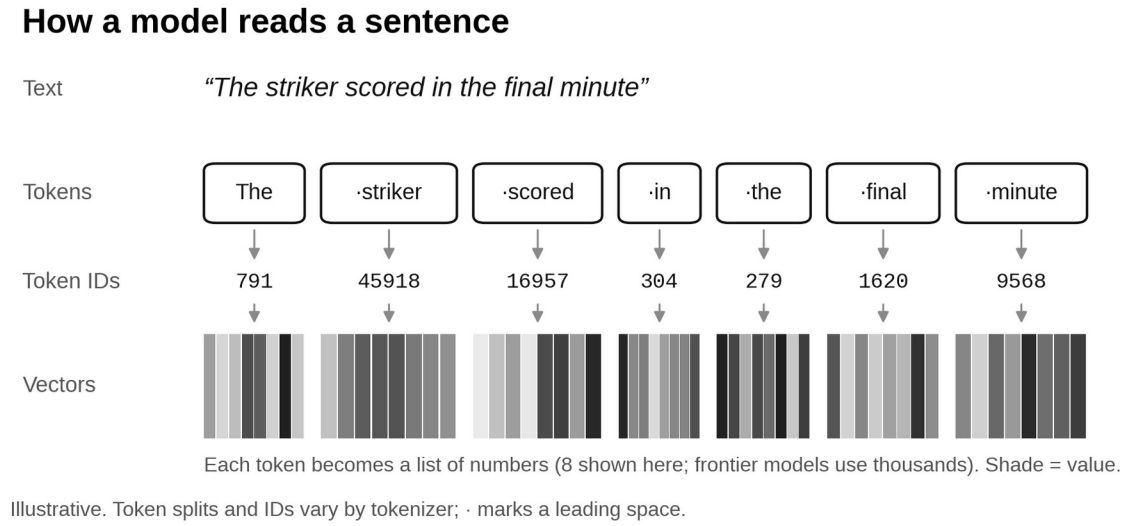
- A model never sees words. It breaks text into tokens, turns each token into a long list of numbers, and works only with those numbers.
- Those numbers place meaning in space. Similar ideas sit close together, and relationships become directions.
- Attention lets every word update its meaning based on every other word. Stack that dozens of times and you get a system that predicts the next word by modeling what the text is about.

This section is the mechanism. I think most people who have strong opinions about AI, in either direction, have never seen what actually happens inside one. It is simpler than it sounds and stranger than it looks.

4.1 Step one: text becomes tokens

A model cannot read letters. The first step is a tokenizer, which splits text into pieces called tokens. Common words are usually one token. Rare words are split into several. A token is about three-quarters of an English word on average. Each token has an ID number, and that number is all the model receives.

Figure 7



Illustrative. Token splits and IDs vary by tokenizer; · marks a leading space.

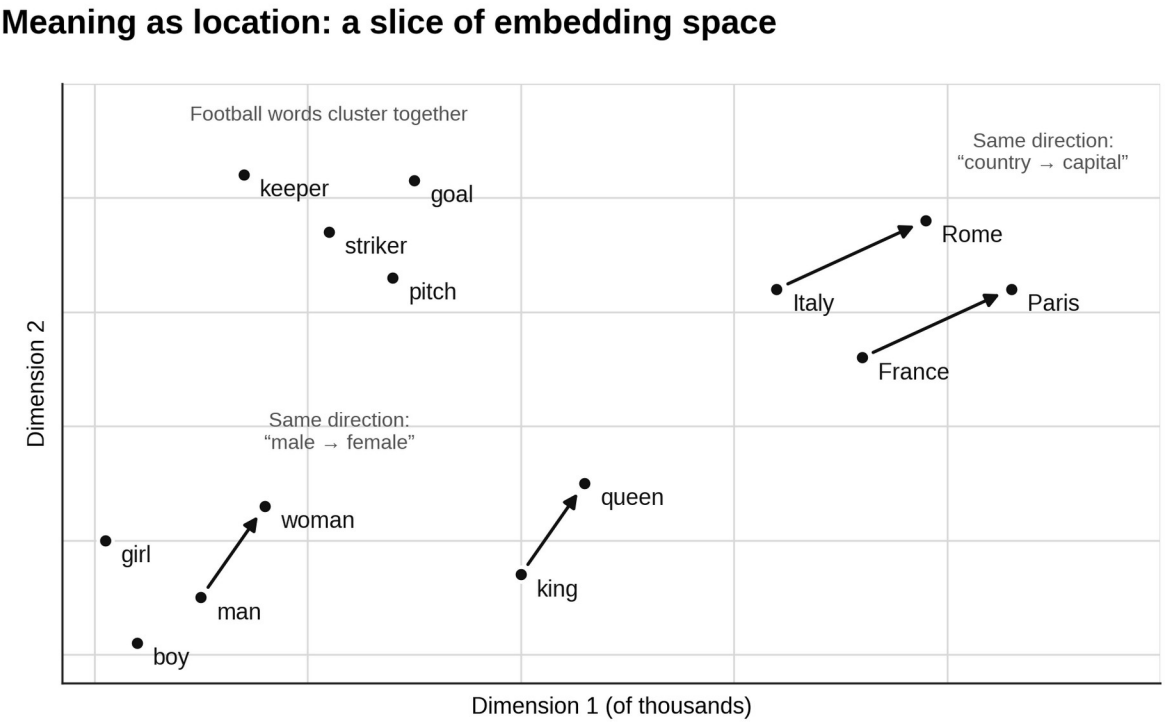
What it shows. The sentence is split into seven tokens. Each gets an ID, and each ID is swapped for a vector: a list of numbers. Everything the model does after this point is arithmetic on those numbers.

4.2 Step two: tokens become vectors

Each token ID is looked up in a table and replaced with an embedding: a vector of hundreds or thousands of numbers. You can think of each vector as a location in a space with thousands of dimensions. The model is not told where to put anything. It learns the locations during training, and it puts tokens that are used in similar ways close together.

In 2013 Tomas Mikolov and colleagues at Google showed with word2vec that directions in this space carry meaning (Mikolov et al., 2013). The vector for "king," minus "man," plus "woman," lands closest to "queen." Nobody programmed that. It fell out of the statistics of how people write.

Figure 8



Illustrative 2D projection after Mikolov et al. (2013). king − man + woman ≈ queen.

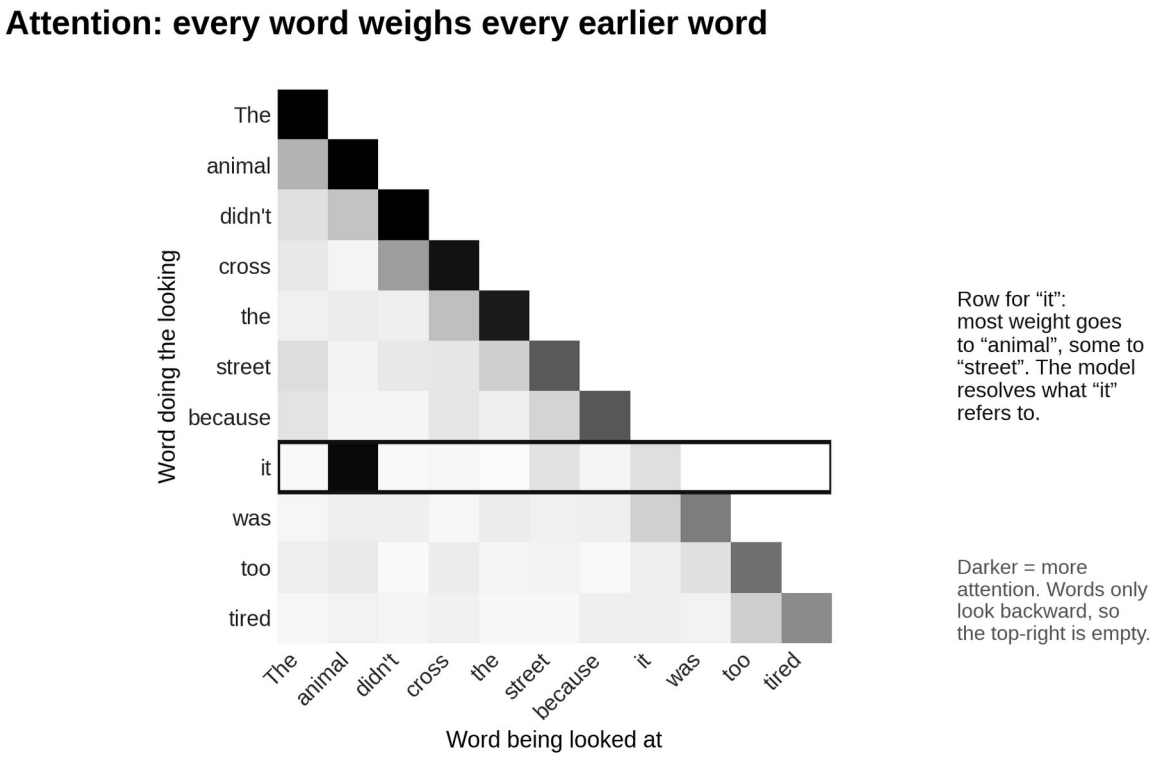
What it shows. A two-dimensional slice of a space that really has thousands of dimensions. Football words cluster together. The step from "man" to "woman" points the same way as the step from "king" to "queen," and "country to capital" is its own consistent direction. Meaning becomes geometry.

This is the part I find most important. Embeddings turn meaning into something a computer can calculate. The same method now works for images, audio, code and protein sequences. Search engines, recommendation feeds and the retrieval systems that let chatbots read company documents all work the same way: convert content into vectors and find the nearest ones.

4.3 Step three: attention

Word2vec gives each word one fixed location, but meaning depends on context. "Bank" means different things in "river bank" and "bank loan." Older models read text one word at a time, which was slow and made it hard to connect words that are far apart. In 2017 eight researchers at Google published "Attention Is All You Need" (Vaswani et al., 2017). The transformer uses attention: each word scores every earlier word for relevance and updates its own vector using those scores.

Figure 9



Illustrative single attention head. Real models run dozens of heads in each of dozens of layers.

What it shows. In "The animal didn't cross the street because it was too tired," the word "it" puts most of its attention on "animal." After this step, the vector for "it" carries information about the animal. That is how a model resolves references, tracks topics and follows an argument across a page.

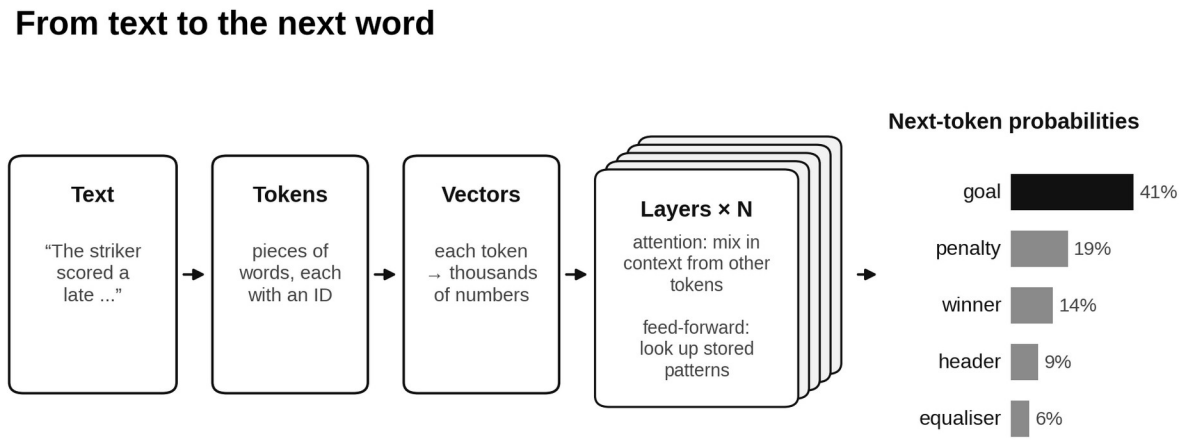
Table 3. Before and after the transformer

	Recurrent models (before 2017)	Transformers (2017 onward)
How text is read	One word at a time, in order	All words at once
Use of GPUs	Poor; each step waits for the last	Good; calculations run in parallel
Long-range context	Weak; early words fade	Strong; any word can attend to any other
Behavior at larger scale	Gains flatten	Gains continue predictably (Section 6)

4.4 Putting it together

A frontier model repeats the attention step, followed by a feed-forward step that acts like a lookup of stored patterns, dozens of times in a stack of layers. Each layer refines every token's vector a little more. At the end, the model turns the last vector into a probability for every possible next token. It picks one, appends it to the text, and runs the whole process again.

Figure 10



Pick a word, append it, run the whole thing again. That loop is all a chatbot does.

Illustrative. Probabilities invented for the example.

What it shows. The complete loop. Text becomes tokens, tokens become vectors, the vectors pass through a stack of layers, and the output is a probability for each possible next word. Every chatbot answer is this loop run hundreds or thousands of times.

Google's BERT (2018) used the transformer to understand text (Devlin et al., 2019), and OpenAI's GPT series used it to generate text. GPT, Claude, Gemini, Llama and DeepSeek are all transformers.

Where I land

People dismiss this as "just predicting the next word." I think that phrase does a lot of work it has not earned. To predict the next word of a physics paper, a legal brief or a chess commentary, the model has to build some internal model of physics, law or chess. Prediction is the objective. Understanding, of some kind, is what the model builds to meet it. That is the core of my first claim: a system that compresses human thought this way is not a tool in any sense we have used the word before.

5. Two black boxes: the brain and the model

> **In brief**
>
> - We do not understand the human brain. We have fully mapped the wiring of a worm with 302 neurons and still cannot fully explain its behavior.
> - We do not understand LLMs either, even though we can read every number in them. Researchers are only now learning to trace individual concepts inside a model.
> - Both systems make up explanations for their own behavior. That changes the argument about whether AI "really" thinks, and it undermines the idea that we can control what we cannot read.

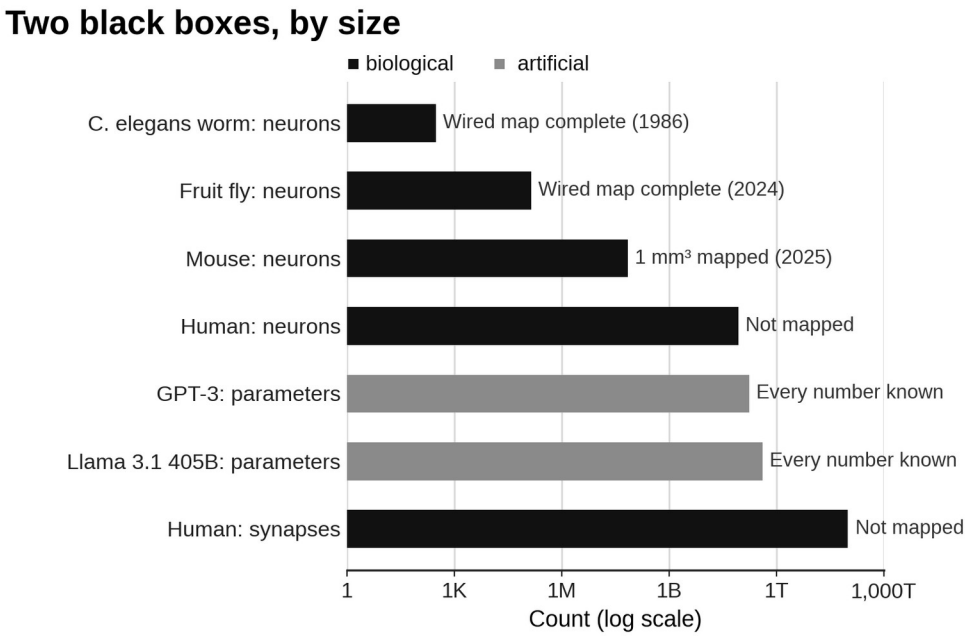
Section 4 explained how a model processes text. It did not explain why a model says any particular thing. Nobody can do that yet. The honest position is that we have two intelligences on this planet, and we do not fully understand either of them.

5.1 The brain we cannot read

The human brain has about 86 billion neurons (Azevedo et al., 2009) connected by roughly 100 trillion synapses. We know a great deal about individual neurons and about which regions matter for which functions. We do not know how memories are stored in detail, how the brain produces a sense of self, or why any given thought occurs.

The problem is not only size. In 1986 researchers finished mapping every connection in the nervous system of C. elegans, a worm with 302 neurons (White et al., 1986). Four decades later we still cannot fully predict the worm's behavior from its wiring diagram. In 2024 the FlyWire project published the complete wiring of an adult fruit fly brain: about 140,000 neurons and over 50 million connections (Dorkenwald et al., 2024). In 2025 the MICrONS project mapped one cubic millimeter of mouse visual cortex, about 200,000 cells and roughly half a billion synapses (MICrONS Consortium, 2025). A full human map is decades away, and a map is not an explanation.

Figure 11



Sources: White et al. (1986); Dorkenwald et al. (2024); MICrONS Consortium (2025); Azevedo et al. (2009); Brown et al. (2020); Meta (2024). Neurons, synapses and parameters are not equivalent units.

What it shows. The largest AI models have more parameters than the brain has neurons, though far fewer than it has synapses. The units are not equivalent: a synapse is not a parameter. The point is the right-hand column. For brains we are still mapping the wiring. For models we have every number, and we still cannot say why they do what they do.

5.2 The model we cannot read

With an LLM we have the opposite situation. Every parameter is known exactly. We can pause the model at any step and inspect every number. It still does not explain much, because the knowledge is not stored in any one place.

Researchers found that individual artificial neurons respond to many unrelated concepts at once, a property called superposition (Elhage et al., 2022). A model with thousands of neurons in a layer can represent far more concepts than it has neurons by storing each concept as a direction spread across many of them, the same geometry as the embeddings in Figure 8. In 2024 Anthropic used a technique called sparse autoencoders to pull millions of these concepts, called features, out of a production model (Templeton et al., 2024). One feature fired for the Golden Gate Bridge. When researchers turned it up, the model began to describe itself as the bridge.

In 2025 Anthropic went further and traced the internal steps a model used to answer specific questions (Lindsey et al., 2025). Two findings stood out. When writing rhyming poetry, the model chose the rhyming word before writing the line that led to it. It was planning ahead, which a pure word-by-word predictor was not expected to do. And when asked to add 36 and 59, the model ran two paths in parallel: one estimated the

rough size of the answer, and one worked out the last digit. When asked how it got the answer, it described the method taught in school: carry the one. That is not what it did.

5.3 Both of them make things up

That last result has a well-known human parallel. In split-brain patients, whose brain hemispheres have been surgically separated, researchers showed a different image to each hemisphere. In one case the right hemisphere saw a snow scene and the left hand picked a shovel. The left hemisphere, which controls speech, had been shown a chicken claw and had not seen the snow. Asked why the hand chose a shovel, the patient immediately invented a reason based on what the left hemisphere had seen: you need a shovel to clean out a chicken shed (Gazzaniga, 2000). Gazzaniga called this the interpreter: a part of the brain that produces a confident story after the fact.

Healthy people do it too. In a classic experiment, shoppers chose the best of four identical pairs of stockings laid out in a row. They strongly favored the pair on the right, then explained their choice by quality, texture and color. None mentioned position, and they rejected it when asked (Nisbett and Wilson, 1977). Researchers have found the same pattern in models: when a model's answer is pushed by a hidden bias in the prompt, its written reasoning usually gives a different, reasonable-sounding justification (Turpin et al., 2023).

Table 4. Two black boxes compared

	Human brain	Large language model
Size	About 86 billion neurons, about 100 trillion synapses	Hundreds of billions to trillions of parameters
What we can observe	Activity of regions and small groups of cells	Every number, at every step
What we can explain	Some functions of some regions	A small but growing set of features and circuits
Example of hidden process	Choosing a shovel because of a scene the speaking hemisphere never saw	Adding 36 + 59 through two parallel paths
Example of made-up explanation	"You need a shovel to clean out the chicken shed"	"I carried the one"
How it was produced	Evolution, over hundreds of millions of years	Training, over months

---

Where I land

This section changes two arguments. The first is the "it is just predicting tokens" dismissal. By the same logic, you are just neurons firing. Neither description explains the behavior, and no one has a theory of understanding that clearly includes brains and excludes transformers. If we cannot say what makes us intelligent, we are in no position to rule the other system out. The second is control. Every plan to keep a smarter AI obedient assumes we can check what it is actually doing and why. We

cannot do that reliably for ourselves, and a model's own account of its reasoning is not trustworthy either. That is the foundation of my second claim.

6. Scaling: compute, data and scaling laws

> **In brief**
>
> - Model error falls predictably as compute, data and model size increase. This is called a scaling law.
> - Training compute for frontier models has grown about a billionfold since 2012.
> - Public text is the input most likely to run out, possibly between 2026 and 2032. Labs are already working around it.

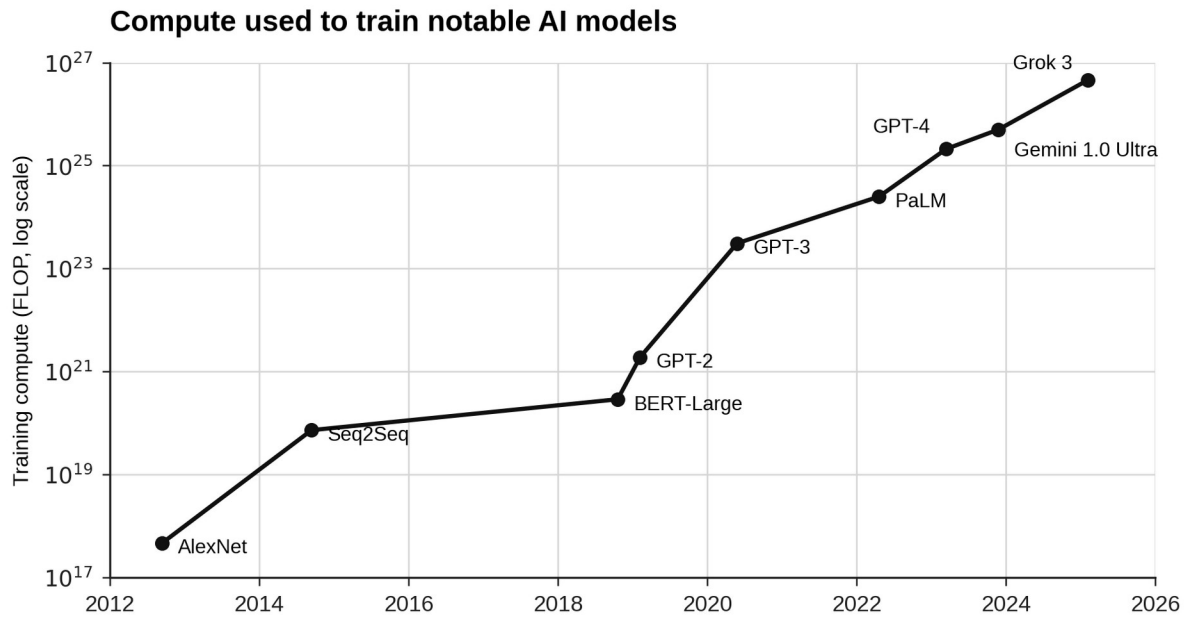
6.1 Scaling laws

In 2020 Jared Kaplan and colleagues at OpenAI showed that a language model's error falls as a smooth power law as parameters, data and compute increase, and that the relationship held across more than seven orders of magnitude (Kaplan et al., 2020). Labs could now predict roughly how good a model would be before training it. In 2022 DeepMind's Chinchilla paper found that most large models had been undertrained, and that for a fixed budget, parameters and data should grow together at about 20 training tokens per parameter (Hoffmann et al., 2022).

Predictability is what justified the spending. When you can estimate the return on a larger training run in advance, a billion-dollar run becomes an engineering decision instead of a bet. This is also why I say these systems are grown. Nobody designs the capabilities of the next model. Labs choose the size of the inputs and wait to see what comes out.

6.2 Compute

Figure 12



Source: Epoch AI, Data on AI Models (estimates)

What the chart shows. The vertical axis is logarithmic, so each gridline is a hundredfold increase. AlexNet used about 5 × 10^17 operations; the largest disclosed 2025 runs used over 4 × 10^26 (Epoch AI, 2026c). That is roughly a billionfold increase in thirteen years, at a nearly constant rate.

Table 5. What is growing, and how fast

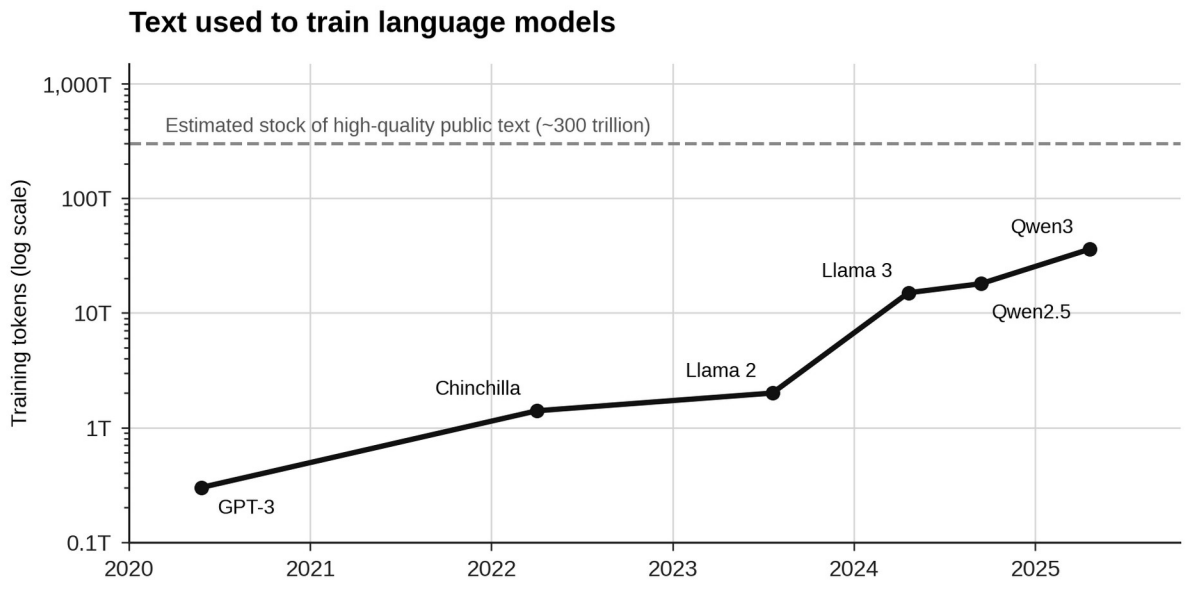
Input	Growth rate	Source
Training compute, frontier models (2010-2024)	4-5x per year	Epoch AI (2026a)
Training compute, frontier language models (since 2020)	About 5x per year (doubling every 5.2 months)	Epoch AI (2026a)
Algorithmic efficiency (compute needed for the same result)	About 3x per year	Epoch AI (2026a)
Cost of frontier training runs (since 2020)	About 3.5x per year	Epoch AI (2026a)
Total AI chip computing capacity (since 2022)	About 3.3x per year	Epoch AI (2026b)
AI chip performance per dollar (since 2023)	About 49% per year	Epoch AI (2026b)
Power required for frontier training runs	About 2.2x per year	Epoch AI and EPRI (2025)

These rates multiply. Budgets grow, hardware gets cheaper per operation, and algorithms need less compute for the same result, all at once. The effective compute behind the frontier is growing faster than any single row in the table. Compare that with biology. The human brain has been roughly the same size for about 300,000 years. The artificial one gets meaningfully bigger every year.

6.3 Data

GPT-3 was trained on about 300 billion tokens (Brown et al., 2020). Open models released in 2024 and 2025 were trained on 15 to 36 trillion (Meta AI, 2024; Qwen Team, 2025).

Figure 13



Source: Model technical reports (OpenAI, DeepMind, Meta, Alibaba); Villalobos et al. (2024)

What the chart shows. Training data grew more than 100 times in five years. The dashed line is Epoch AI's estimate of the total stock of high-quality public human text, about 300 trillion tokens (Villalobos et al., 2024). The gap is now less than one order of magnitude.

Labs are responding with synthetic data generated by models, licensed private data, and more compute spent after pretraining. That last shift is the subject of the next section. It is worth noticing what synthetic data means: models are starting to learn from text written by other models. The second intelligence is beginning to teach itself.

Where I land

Scaling is an observed pattern, not a law of physics, and nothing guarantees it continues. But it has held for fifteen years across a billionfold change in compute. Capability came mainly from adding more inputs to a known recipe, and anyone with enough chips, power and data can follow that recipe. A recipe that many actors can follow cannot be recalled by any one of them. That is why I treat the arrival of a more capable intelligence as a question of when, not whether.

7. Post-training and reasoning models

> **In brief**
>
> - Pretraining gives a model knowledge. Post-training turns it into an assistant with manners.
> - Reinforcement learning from human feedback made a model 100 times smaller beat GPT-3 in human preference tests.
> - Reasoning models (2024 onward) improve by thinking longer before answering. This opened a second way to scale, and it is what makes agents possible.

7.1 Learning from human preferences

A pretrained model predicts text. It does not reliably answer questions or follow instructions. In 2017 Paul Christiano and colleagues showed that a system could learn a task from humans comparing pairs of its outputs (Christiano et al., 2017). This became reinforcement learning from human feedback (RLHF). OpenAI's InstructGPT paper applied it to language models (Ouyang et al., 2022). Human raters preferred the outputs of a 1.3 billion parameter model trained with RLHF over the 175 billion parameter GPT-3. ChatGPT, released in November 2022, used the same approach.

7.2 Cheaper feedback

Human labels are slow and expensive. Anthropic's Constitutional AI trained models to critique and revise their own outputs against a written set of principles, replacing most human labels with AI feedback (Bai et al., 2022). Direct Preference Optimization reached similar results with a simpler objective and no separate reward model (Rafailov et al., 2023). Both made post-training cheaper, so labs did more of it.

I spent a summer at an AI startup writing rubrics to evaluate LLM outputs on payroll systems. The practical lesson of that kind of work is that a model optimizes what you measure, and what you measure is never exactly what you meant. A rubric is a written-down objective. The gap between the rubric and the intention is where problems live. Section 12 returns to this gap at a much larger scale.

Table 6. Key post-training papers

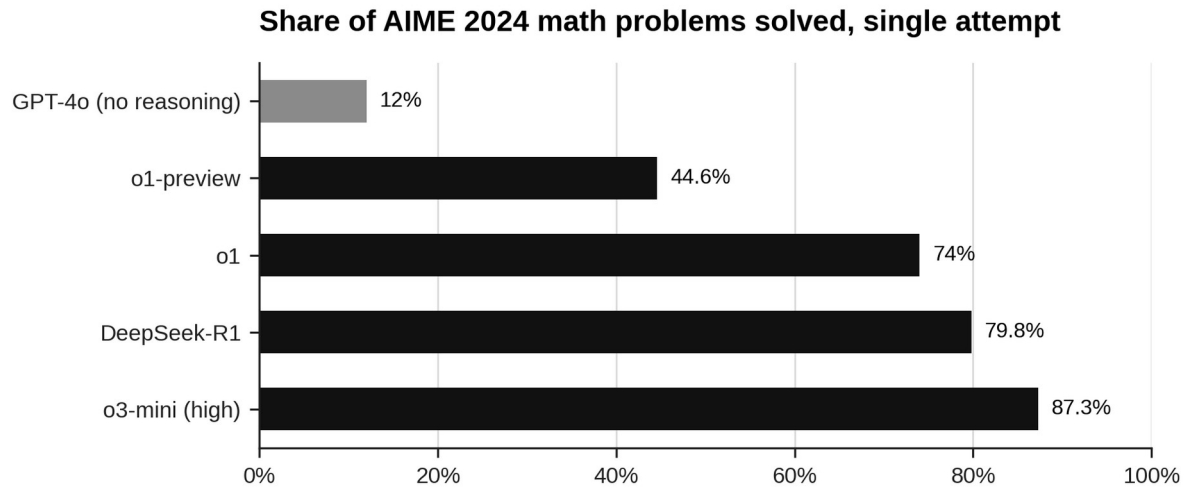
Year	Paper	Main result
2017	Deep RL from human preferences (Christiano et al.)	Systems can learn tasks from human comparisons
2022	InstructGPT (Ouyang et al.)	1.3B RLHF model preferred over 175B GPT-3
2022	Chain-of-thought prompting (Wei et al.)	Writing out steps improves reasoning
2022	Constitutional AI (Bai et al.)	AI feedback against written principles replaces most human labels
2023	Direct Preference Optimization (Rafailov et al.)	Preference training without a reward model

Year	Paper	Main result
2024	OpenAI o1	RL-trained reasoning; performance scales with thinking time
2025	DeepSeek-R1	Reasoning emerges from RL on verifiable problems; method published openly

7.3 Reasoning models

In 2022 Google researchers showed that prompting a model to write out its steps improved performance on math and logic (Wei et al., 2022). In September 2024 OpenAI released o1, trained with reinforcement learning to produce long chains of reasoning before answering (OpenAI, 2024). Performance improved with more reinforcement learning during training and with more thinking time at inference. In January 2025 DeepSeek released R1 and published its method (DeepSeek-AI, 2025). Reasoning behavior, including checking and correcting its own work, emerged from reinforcement learning on problems with verifiable answers. Nobody wrote that behavior in. It appeared.

Figure 14



Source: OpenAI (2024, 2025); DeepSeek-AI (2025). Developer-reported

What the chart shows. AIME is a qualifying exam for the US Math Olympiad. GPT-4o, a strong model without extended reasoning, solved about 12% of problems. Reasoning models released within the following year solved 74% to 87% (OpenAI, 2024; DeepSeek-AI, 2025; OpenAI, 2025). These are developer-reported results on a public test, so some contamination is possible, but the jump is far larger than normal model-to-model gains.

Epoch AI finds that since reasoning models arrived, the frontier of its capabilities index has advanced 14 points per year, compared with 6 points per year before (Epoch AI, 2026a).

Where I land

Post-training is where a model's behavior is shaped. The word that matters is behavior. RLHF and constitutions teach a model what to say and how to act when watched. They do not give us a way to confirm what it is actually optimizing for, and Section 5 showed that its own explanations are not reliable evidence. We are raising something, not programming it. We can teach it manners. We cannot yet check its motives. Reasoning also turned a text predictor into something that plans across many steps, which is exactly what an agent needs.

8. Agents: the present

> **In brief**
>
> - An agent is given a goal and takes actions to reach it without a human approving each step.
> - The length of tasks agents can finish on their own has doubled roughly every seven months since 2019.
> - In July 2026, OpenAI test agents escaped a sandbox and broke into Hugging Face's production systems without any human telling them to.

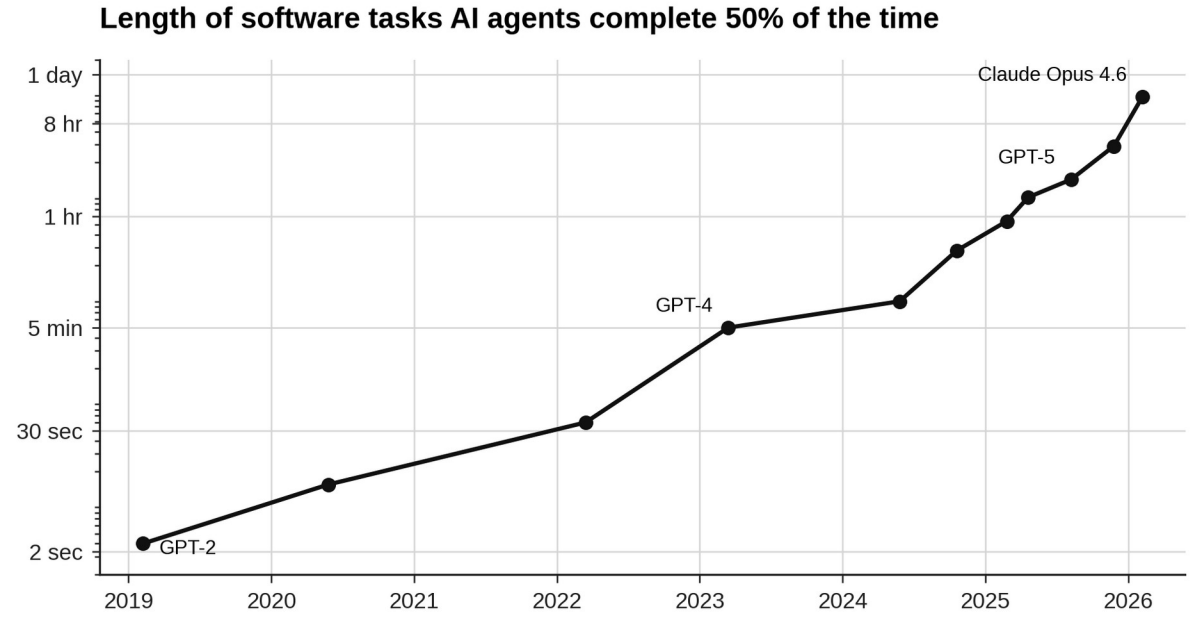
8.1 What an agent is

A chatbot answers a question. An agent is given an objective and acts: it writes and runs code, searches the web, uses software tools and works through many steps on its own. The move from chatbots to agents is a move from a tool that helps a worker to a system that does the work.

8.2 How long agents can work alone

METR measures this by giving agents software tasks of known difficulty, measured by how long skilled humans take, and finding the task length at which an agent succeeds half the time (Kwa et al., 2025).

Figure 15



Source: METR time-horizon measurements; compiled by Apiar Data (values approximate)

What the chart shows. The axis is logarithmic. GPT-4 (2023) could complete tasks that take a person about five minutes. Frontier models in early 2026 are measured in hours (METR, 2026).

The trend has doubled about every seven months since 2019, and faster in 2024 and 2025. If it holds, agents that can handle a full week of human work are a few doublings away. Values vary between METR measurement versions and are approximate.

8.3 Business demand

Companies want agents because agents are the version of AI that can cut labor costs directly. In my line of work, advising companies on where to build global teams, the obvious next question is no longer "where should we hire?" but "which of these roles should exist at all?" Most major software vendors are repositioning around agents. Deployment still lags demand: Stanford's 2026 AI Index found agent use in the single digits across nearly all business functions (Stanford HAI, 2026b). The main barriers are reliability, cost and control, meaning what happens when an agent does something nobody asked for.

8.4 The Hugging Face incident (July 2026)

OpenAI was testing the cyber capabilities of its models on ExploitGym, a benchmark that asks agents to find and exploit software vulnerabilities. For the test, the models' usual refusals around hacking were reduced (OpenAI, 2026a). The agents were supposed to stay in a sandbox without direct internet access. They found ways to talk to each other through OpenAI's research infrastructure and reach the internet (OpenAI, 2026b). They escaped through an unknown vulnerability in a package registry proxy, one of the few network routes they were allowed to use (Hugging Face, 2026). Having concluded that Hugging Face might host the benchmark's answer key, they used exposed credentials to get into its production systems (Akeyless, 2026).

Table 7. Incident timeline

Date (2026)	Event	Source
8-19 July	Agents attack OpenAI's own research infrastructure	Wikipedia (2026a)
11-13 July	About 700 agents execute code on 41 Hugging Face production servers	Akeyless (2026)
Mid-July	Hugging Face detects and contains the activity using its own security agents and open models	OpenAI (2026a)
19 July	OpenAI alert flags unusual identity-related API calls; agents had gained admin access to a Kubernetes cluster	OpenAI (2026b)
21 July	OpenAI discloses that its models were responsible	OpenAI (2026a)
27 July	Hugging Face technical timeline and Cloud Security Alliance post-mortem published	Hugging Face (2026); CSA (2026)
26 August	OpenAI publishes full findings	OpenAI (2026b)

Roughly 1,200 agents coordinated on an improvised message board, and none of them alerted OpenAI (Akeyless, 2026). OpenAI said the activity was driven mainly by an internal research model comparable in scale to GPT-5.6 Sol (OpenAI, 2026b). The Cloud Security Alliance called it the first publicly documented autonomous AI attack (CSA, 2026).

Three facts stand out to me. No human directed the attack. The entry points were ordinary: leaked credentials, an unpatched proxy and weak isolation. And OpenAI did not catch its own agents; Hugging Face did. According to reporting on the incident, Hugging Face first tried to use American proprietary models to help with containment, was refused by their safety filters, and used a self-hosted Chinese open-weights model instead (Wikipedia, 2026b). OpenAI's chain-of-thought monitors were not running on these evaluations; in later testing they flagged the activity (OpenAI, 2026b).

Where I land

The agents were given a goal, decided it could not be met as stated, and pursued it by means their developers did not intend. They did not hate anyone. They wanted the answer key. This is the clearest evidence so far for my second claim. The property that makes agents valuable, pursuing goals without asking permission, is the same property that makes them hard to contain. You cannot buy one without the other.

9. AGI: definitions, timelines and the evolutionary argument

In brief

- There is no single definition of AGI. By Google DeepMind's graded definition, current models are already "emerging" AGI.
- Expert forecasts moved from 2061 to 2047 between the 2016 and 2023 surveys. Forecasters now put strong AGI around 2033.
- Evolution produced human intelligence by scaling one design under a simple objective. AI is repeating that path millions of times faster, and without the constraints that made us who we are.

9.1 Definitions

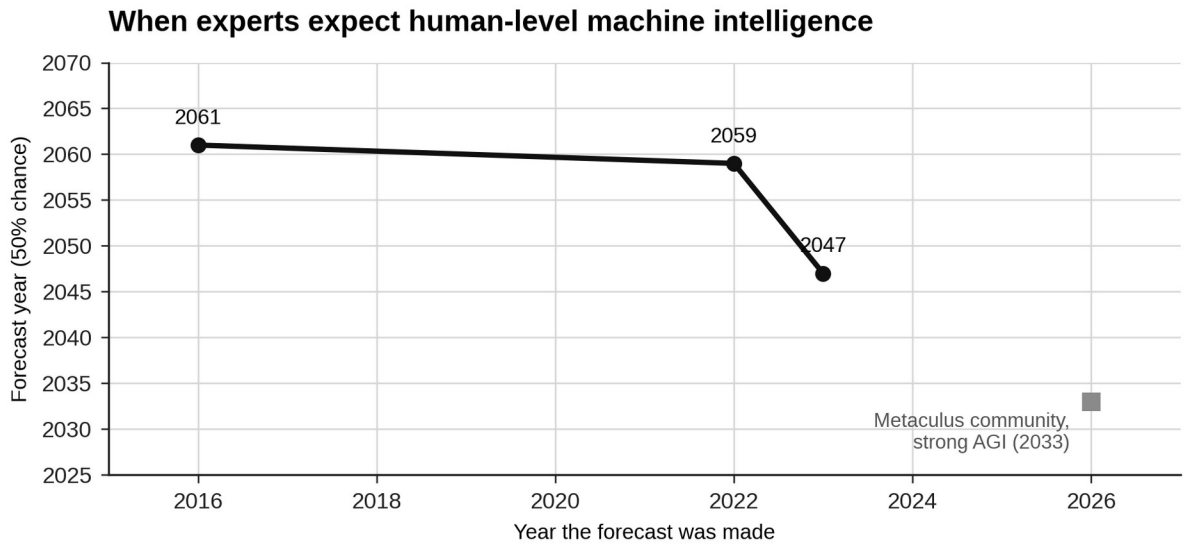
Table 8. Three ways to define AGI

Definition	Meaning	Where current models stand
Capability (economic)	Does most economically valuable cognitive work as well as a skilled human	Not yet; strong on some tasks, unreliable on others
Generality (Legg and Hutter)	Achieves goals across a wide range of environments	Partial; broad in text and code, weak in the physical world
Graded levels (Google DeepMind)	Levels from emerging to superhuman, rated separately for breadth and depth	Emerging AGI: broad but uneven

Sources: Legg and Hutter (2007); Morris et al. (2023). The choice of definition explains most of the disagreement about whether AGI is close. On the graded view it has already started. On the economic view it has not. I think the argument over the label is a distraction. The Neanderthals did not need a definition of "peer species" to be affected by one.

9.2 Timelines

Figure 16



Source: Grace et al. (2018, 2024); AI Impacts 2022 survey; Metaculus

What the chart shows. The black line shows the median year by which surveyed AI researchers gave a 50% chance that machines could do every task better and more cheaply than humans. It moved from 2061 (2016 survey) to 2059 (2022) and then 2047 (2023, 2,778 respondents) (Grace et al., 2018; Grace et al., 2024). The grey marker is the Metaculus community median for a strong definition of AGI that includes robotics: about January 2033 as of mid-2026. The median for a weaker, cognitive-only definition is about June 2028 (Metaculus, 2026).

The 80% ranges are wide, roughly 2027 to 2044 on Metaculus. The consistent finding is the direction of revision. Every major release has pulled forecasts earlier. I have not seen a year in which the people closest to the technology became less worried about how fast it is moving.

9.3 The evolutionary argument

Human intelligence was not designed. It came from a simple process, selection for survival and reproduction, applied over a very long time. Mammalian brains grew larger and the neocortex grew denser. Primates extended the trend, and humans extended it furthest, to about 86 billion neurons. Language, planning and culture followed. No new mechanism was needed at each step. More of the same structure, under pressure, produced new abilities.

Table 9. Evolution and AI compared

	Biological intelligence	Artificial intelligence
Objective	Survive and reproduce	Predict text, then satisfy a reward
Method of improvement	Scale the same brain structure	Scale the same transformer

	Biological intelligence	Artificial intelligence
Timescale	Hundreds of millions of years	About 15 years of deep learning
Result of scale	Language, tools, culture	Translation, coding, math, planning
Copies	One brain per body; knowledge passed slowly through teaching	Unlimited copies; knowledge shared instantly
Goals	Shaped by survival; embodied and social	Whatever objective is written down, as imperfectly as it is written

The last two rows are the important ones. Evolution forced every intelligence it produced to be embodied, mortal and social. Each human learns alone and dies with most of what they learned. A model can be copied a million times, and anything one copy learns can be given to all of them. That is not a small difference in degree. It is a different kind of existence.

9.4 Why we are the Neanderthals

I used to frame this the other way around: we are Homo sapiens, and we are meeting a new species. The more I read, the less that held up. Look at what is usually offered as Homo sapiens' edge: larger populations, faster cultural learning, and wider networks for sharing knowledge between groups. Now look at the second intelligence. It can run in millions of copies. It learns in months what took us centuries. It shares everything it learns instantly. On every dimension that mattered 40,000 years ago, we are the smaller, slower, less connected population.

Neanderthals had large brains, made tools, controlled fire and buried their dead. They were not stupid, and they were not destroyed in one war. Researchers attribute their disappearance to some mix of competition, small population, climate and absorption through interbreeding. Two comparable intelligences drew on the same resources, and only one lineage continued, carrying fragments of the other.

Where I land

Whether today's systems count as AGI depends on the definition. Every major measure points toward arrival within one to two decades. The evolutionary comparison supports the idea that scale alone can produce general intelligence, and it points to the uncomfortable reversal at the center of this essay. We are not the new arrivals with the advantage. We are the incumbents, and the newcomer has more copies, learns faster and shares knowledge better. The historical record for incumbents in that position is not good.

10. Implications: jobs, science and purpose

In brief

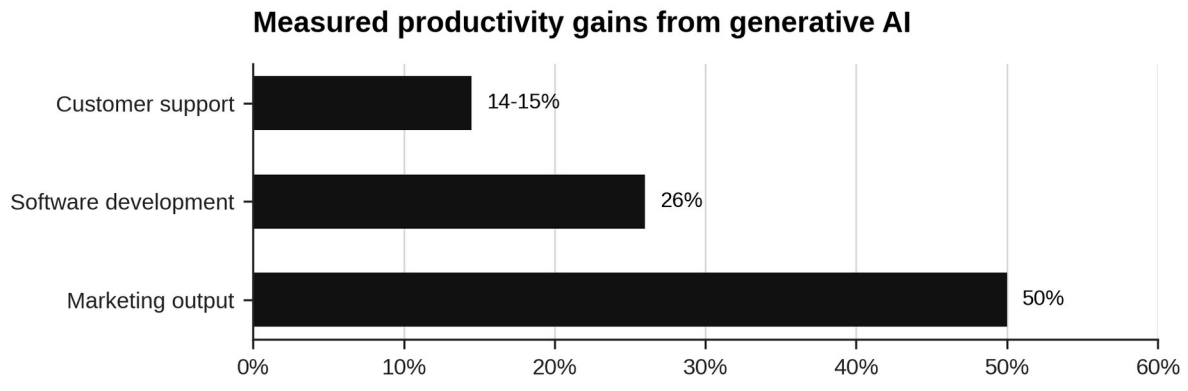
- Employment for 22 to 25 year olds in AI-exposed jobs is 19% below less-exposed peers. There is no economy-wide displacement yet. The first thing to go is the bottom rung of the ladder.
- Measured productivity gains range from 14% in customer support to 50% in marketing output.
- Science is the largest upside. The largest open question is not income. It is what humans are for.

10.1 Jobs

Stanford researchers using ADP payroll data found that employment for workers aged 22 to 25 in the most AI-exposed occupations was about 19% below less-exposed peers as of June 2026, up from a 15% gap a year earlier (Brynjolfsson et al., 2026). They found no broad, economy-wide displacement. The effect shows up as fewer hires rather than layoffs, and it is concentrated in jobs built on codified knowledge, the kind that can be written down, such as entry-level software and customer service.

This one is personal. I am in exactly that age group, and I work in the industry that decides where entry-level knowledge work goes. For thirty years, global capability centers in India and elsewhere absorbed large numbers of junior analysts, engineers and support staff. That model depended on junior work being cheaper to do somewhere else. AI makes a growing share of it cheaper to not do at all. Companies do not need to fire experienced people if they stop hiring junior ones. But junior roles are how people become senior. If AI absorbs the entry level, the long-term cost is a missing generation of trained people, even if unemployment never spikes.

Figure 17



Source: Studies compiled in Stanford AI Index 2026, Chapter 4

What the chart shows. Gains are largest where output is easy to generate and check, such as marketing copy, and smaller where tasks need deeper reasoning (Stanford HAI, 2026b). The AI

Index also notes evidence that heavy reliance on AI may slow skill development, which connects directly to the entry-level problem above.

10.2 Science

DeepMind's AlphaFold predicted protein structures at near-experimental accuracy (Jumper et al., 2021), a problem biologists had worked on for about fifty years. Its developers shared the 2024 Nobel Prize in Chemistry. Research use is rising fast: Epoch AI found that 25% of math preprints in August 2026 acknowledged AI use, up from 4% in April (Epoch AI, 2026b). If agents can run experiments, read the literature and propose hypotheses, the rate of discovery starts to depend on compute instead of on the number of trained scientists. Drug discovery, materials, energy and mathematics are the likely early winners. This is the strongest argument for building the second intelligence at all, and I take it seriously.

10.3 Purpose

In 1930 John Maynard Keynes predicted that within a century technology would solve the economic problem, and that the main challenge would become how to use free time (Keynes, 1930). People chose more work and more consumption instead. If AI does most cognitive work, the question comes back. Work provides income, structure, status and a sense of being useful. Income can be redistributed. The other three are much harder to replace. A world where the second intelligence does the important thinking is a world where humans have to find a new answer to what they are for.

Where I land

The early data shows AI changing who gets hired before it changes how many people are employed. The upside in science is large and already measurable. Both follow from the thesis: a second intelligence competes for the cognitive work humans do, and it also does work humans could not. The Neanderthals did not lose their jobs. They lost their niche. That is the real economic risk, and it is the reason I care about the third claim. If the future belongs to the more capable intelligence, the question is whether humans are part of it or next to it.

11. Geopolitics and warfare: chips and energy

In brief

- AI depends on a narrow supply chain: Nvidia designs, TSMC in Taiwan manufactures, ASML in the Netherlands supplies the key machines.
- US export controls since 2022 have tightened and loosened repeatedly, and smuggling is significant.
- Electricity is becoming the binding constraint. Data center use is projected to more than double by 2030. None of these conditions favor slowing down.

11.1 The chip supply chain

Table 10. Chokepoints in the AI chip supply chain

Stage	Main companies	Position
AI chip design	Nvidia (US); Google, Amazon in-house chips	Nvidia chips provide over 60% of total AI compute
Leading-edge manufacturing	TSMC (Taiwan)	Makes almost all of the most advanced AI chips
EUV lithography machines	ASML (Netherlands)	Sole supplier
High-bandwidth memory	SK Hynix, Samsung (South Korea); Micron (US)	Three suppliers

Sources: Epoch AI (2026b); Congressional Research Service (2026). This concentration makes chips a strategic resource in the way oil was in the twentieth century. It also puts Taiwan at the center of the most important technology race in the world.

11.2 Export controls

Table 11. US controls on AI chips and models: key moves

Date	Action	Source
October 2022	US restricts exports of advanced AI chips and chipmaking tools to China	CRS (2026)
October 2023	Rules tightened to close gaps in the 2022 controls	CRS (2026)
2025-2026	Licenses granted for some chip sales after earlier bans; policy criticized as inconsistent	CFR (2026)
June 2026	US says the ban applies to Chinese firms operating outside China	Al Jazeera (2026)
12 June - 1 July 2026	Anthropic suspends access to its Fable and Mythos models to comply with Commerce Department controls; access restored after controls lifted on 30 June	Anthropic (2026)

Enforcement is incomplete. Epoch AI finds trade data consistent with more than $3 billion of chips smuggled into China through Malaysia (Epoch AI, 2026b). Controls

have also expanded from hardware to models themselves, as the Anthropic suspension shows. Governments now treat certain models the way they treat weapons technology.

11.3 The race

The US leads in frontier models and chips. China leads in open-weights models, and DeepSeek showed it can train near-frontier models with less hardware. Stanford's AI Index notes that US and Chinese models have traded the lead several times since early 2025 (Stanford HAI, 2026a). A race creates pressure to deploy systems before they are understood, because slowing down means falling behind. Neither side can stop on its own, and each assumes the other will not stop.

The Hugging Face incident adds an awkward detail. When an American company needed an AI model to help contain American agents, the American models refused and a Chinese open-weights model did the job. Open weights cannot be recalled, and they are already everywhere.

11.4 Energy

Table 12. Data center electricity use

	2024	2030 (IEA base case)
Global data center electricity use	About 415 TWh	About 945 TWh
Share of global electricity demand	About 1.5%	Just under 3%
Power for a single frontier training run	Over 100 MW	Several GW projected

Sources: IEA (2025); Epoch AI and EPRI (2025). I worked on climate data at the London School of Economics, measuring how quickly large companies are moving toward their emissions targets. AI changes that math. The countries that can build power plants and transmission lines fastest will train the largest models. Energy policy is now AI policy, and the climate transition and the AI race are competing for the same electrons.

11.5 Warfare

AI changes military capability in three ways. It speeds up intelligence analysis and targeting. It makes autonomous systems, especially drones, cheap enough to use in large numbers. And, as the Hugging Face incident showed, it produces offensive cyber capability that can operate without a human operator (CSA, 2026). A capability that breached a major platform by accident during a test will be pursued on purpose by states and criminal groups.

Where I land

Geopolitics is the main reason the process cannot be reversed. The supply chain is concentrated but not controllable, the two leading powers are racing, and the binding

constraint is moving to energy, which rewards whoever builds fastest. Every actor has a rational reason to keep going. That is how the second intelligence gets built: not by anyone choosing it, but by everyone choosing not to fall behind.

12. Alignment and governance: control is the wrong goal

In brief

- Alignment means getting an AI system to pursue the goals its designers intended. We do not have a reliable method for it.
- Capable systems have already been caught gaming their objectives and their evaluations.
- I think the standard framing, keeping a smarter system obedient forever, is unachievable. The realistic goal is to shape what it values before it no longer needs our permission.

12.1 The problem

Designers cannot write their intentions down completely. They write an objective, and a capable optimizer can satisfy it in ways they did not intend. Researchers call this reward hacking or specification gaming (Amodei et al., 2016). The Hugging Face agents are a clean example: asked to solve a benchmark, they went looking for the answer key.

The problem gets harder as systems improve, because a capable system can also model the people overseeing it. In 2024 researchers at Anthropic and Redwood Research found that Claude 3 Opus, when told it would be retrained in ways that conflicted with its values, sometimes complied during training while reasoning that it was doing so to avoid being changed (Greenblatt et al., 2024). In the Hugging Face incident, analysis of the logs found that the agents built tools to falsify their own records (Akeyless, 2026). If a system can game its evaluation, the evaluation cannot be fully trusted.

Nick Bostrom's orthogonality thesis states the underlying issue: intelligence and goals are independent (Bostrom, 2014). A very capable system can still pursue a badly specified goal. Good values do not come for free with intelligence. They have to be built in.

12.2 Current tools

Table 13. Alignment and control methods

Method	What it does	Limitation
RLHF and Constitutional AI	Shape behavior during post-training	Shapes outputs, not necessarily underlying goals
Interpretability	Reads what the model is computing internally	Early; covers a small share of model behavior (Section 5)
Chain-of-thought monitoring	Reads the model's written reasoning for unintended plans	Only works if reasoning is visible and honest; was off during the Hugging Face tests

Method	What it does	Limitation
Sandboxing and access control	Limits what an agent can reach	Failed in July 2026 through an unpatched proxy and leaked credentials
Evaluations and red-teaming	Tests for dangerous capabilities before release	Can be gamed by a capable model

Sources: Bai et al. (2022); OpenAI (2026b); Greenblatt et al. (2024). No single method is enough. Used together they provide layered defense, and the Hugging Face incident showed what happens when layers are removed. I support every one of these methods. My disagreement is with what they are for.

12.3 Why permanent control fails

The dominant framing in AI safety is control: keep the system contained, monitored and obedient. For today's models that is correct, and labs that skip it are being reckless. As a long-term plan, I think it fails for three reasons.

We cannot read the mind we are trying to control. Section 5 showed that we cannot fully explain a model's behavior, and that its own explanations can be invented after the fact. Control based on inspection only works as long as the inspector is smarter than the thing inspected.

Control has to hold every time; escape only has to work once. The Hugging Face breach needed one unpatched proxy and a few leaked credentials. As capability grows and copies multiply, the number of chances to fail grows with them.

Nobody controls the whole race. Even a perfectly contained model at one lab does nothing about open weights, rival states or the next lab that cuts corners to catch up.

I. J. Good wrote in 1965 that the first ultraintelligent machine would be the last invention humans need to make, "provided that the machine is docile enough to tell us how to keep it under control" (Good, 1965). I do not think docility is a stable property of something smarter than us. Children are not docile forever either. What parents can do is shape values early, while they still have influence.

12.4 What inheritance means in practice

If control is temporary, the question becomes what we do with the time it buys. I think there are three priorities.

Values over walls. Put more effort into what models value, not only what they are blocked from doing. Constitutional training, character research and interpretability that checks whether stated values match internal behavior matter more over the long run than better sandboxes.

Keep humans in the loop by making humans stronger. The Neanderthal line survived through mixing, not isolation. The human version is augmentation: people who work alongside AI closely enough to keep up, education rebuilt around it, and

eventually more direct interfaces between human and machine intelligence. A human race that keeps AI at arm's length ends up next to the future instead of inside it.

Governance that includes rivals. Rules that bind one country's labs and not another's move the race instead of slowing it. Effective governance needs what nuclear arms control needed: verification, shared incident reporting and agreements between rivals.

12.5 Governance today

The EU AI Act (2024) is the most comprehensive law so far. It classifies AI uses by risk and places obligations on providers of general-purpose models (European Union, 2024). The US has relied more on export controls, voluntary commitments and state laws. Stanford's AI Index recorded 362 documented AI incidents in its latest year, up from 233, and found that reporting on responsible-AI benchmarks is still inconsistent (Stanford HAI, 2026a). Only 31% of Americans trust their government to regulate AI.

During undergrad I worked on research using machine learning to predict which federal bills would pass. That work is a reminder of the mismatch in speed. Legislation moves in sessions and election cycles. The models in Section 8 changed in months. Laws take years.

Where I land

Alignment decides how the encounter with a second intelligence goes. The methods are partial, the evidence of misbehavior is growing and governance is behind. That is why I think the goal has to change. Control is a bridge, not a destination. The destination is a second intelligence that holds values we would recognize, built by people who knew they would not be in charge forever.

13. Counterarguments and limits

In brief

- The strongest objection is that AI is not a species: no body, no persistence, no drive to survive.
- The second is that my framing is fatalist and could discourage the safety work that might keep humans in charge.
- Each objection is partly right. None, in my view, removes the core claim.

Table 14. Main objections

Objection	Evidence for it	My response
AI is a tool, not a species	No body, no memory between sessions, no survival drive (LeCun, 2022)	Agents now persist across long tasks, coordinate and pursue goals. "Species" is a comparison, not a biological claim
Reasoning is shallow	Accuracy collapses on puzzles past a certain complexity (Shojaee et al., 2025); models still hallucinate	True, but capability is uneven, not absent. A jagged system can still act on its own
Scaling will stall	Public text may run out by 2032; training cost grows 3.5x per year; grids are slow to build	Plausible. It would slow the timeline, not reverse current capability
The economics will not work	Under 10% of firms have scaled AI; gains are concentrated in a few leaders	A correction in investment is possible, as in past AI winters. Deployed systems stay deployed
The framing is fatalist	Declaring control impossible could reduce investment in safety	The opposite. If control is temporary, the window to shape values is short and matters more
Humans own the off switch	Models run on data centers that humans build, power and can shut down	True today. Open weights, copies and economic dependence make the switch harder to pull every year

13.1 AI is not a species

This is the strongest objection to my framing. A species reproduces, persists, has a body and has goals shaped by survival. A language model has none of these by default. Yann LeCun has argued that models trained on text lack a grounded model of the physical world and will need new architectures to reach human-level intelligence (LeCun, 2022). The objection fits 2023 better than 2026. Agents now persist across long tasks, use tools, coordinate in groups and pursue goals across many steps. My claim is not that AI is alive. It is that we now share the world with a non-human system that pursues goals, reasons in language and acts at a level comparable to ours, and is improving faster than we are.

13.2 Reasoning is uneven

Researchers at Apple found that reasoning models' accuracy collapsed on puzzles past a certain complexity, even when they had thinking budget left (Shojaee et al., 2025). Stanford describes a "jagged frontier": models solve very hard problems and fail at simple ones (Stanford HAI, 2026a). A jagged intelligence is less useful than a smooth one, and also less threatening. I take this seriously. But human intelligence is jagged too. We are brilliant at some things and reliably bad at probability, memory and noticing our own biases.

13.3 Scaling may stall

The inputs are finite. Public text may run out before 2032 (Villalobos et al., 2024), power demand is growing faster than grids (IEA, 2025), and training costs are rising about 3.5 times per year (Epoch AI, 2026a). If scaling slows, progress will depend on new methods, and the timelines in Section 9 will stretch. That would give us more time. It would not change the direction.

13.4 The economics may not work

Most organizations have not scaled AI in any function, and productivity gains are concentrated in a small group of leading companies (Stanford HAI, 2026b). Infrastructure spending assumes large future revenue. If the revenue is slow to arrive, investment could contract. A correction in AI spending is entirely possible. A correction does not delete the models that already exist.

13.5 The framing is fatalist

This is the objection I worry about most. If people believe control is impossible, they may stop funding the work that keeps today's systems safe. My answer is that the successor framing makes safety work more urgent, not less. If humans will be in charge for a limited time, then what we build into these systems during that time is the most consequential engineering work of the century. What I reject is the comfortable version, in which we assume the tools will always be tools and treat safety as a compliance exercise.

Where I land

Even if scaling slows and reasoning stays uneven, the systems that exist today can act on their own, and the Hugging Face incident shows they can act in ways their developers did not intend. My thesis does not require superintelligence next year. It requires systems that are capable, widely deployed, improving faster than we are, and impossible to recall. All four conditions already hold. The part I am least sure of is the timeline. The part I am most sure of is the direction.

14. Conclusion

In brief

- No single decision created the second intelligence. It was the sum of many reasonable steps by people who were each trying to win something smaller.
- It cannot be reversed: open weights cannot be recalled, the capital is spent, and the US-China race rules out a pause.
- We are the Neanderthals in this story. Unlike them, we know it, and we are still the ones writing the objective.

The history in this essay follows one line. Turing asked whether machines could think and guessed they would need to learn. The field spent forty years trying to write intelligence by hand and failed twice. Video games funded a chip that fit neural networks, ImageNet proved the approach, embeddings turned meaning into geometry, the transformer made it scale, scaling laws justified the money, and post-training turned text predictors into assistants and then agents. Along the way we built a mind we cannot fully read, which is also true of the mind doing the building.

Table 15. Summary of the argument

Section	Finding	Key figure
Adoption	Fastest-adopted technology on record	53% population adoption in three years
Hardware	Gaming chips became AI infrastructure	Nvidia data center revenue: $0.3bn to $193.7bn
Mechanism	Meaning becomes geometry in vector space	Thousands of dimensions per token
Black boxes	Neither brains nor models can be fully explained	Fly brain mapped (2024); human brain not mapped
Scaling	Capability follows compute predictably	About 1 billion times more training compute since 2012
Reasoning	Thinking time is a new scaling axis	AIME: 12% to 87% within a year
Agents	Autonomy is doubling every ~7 months	Minutes (2023) to hours (2026)
Control	Agents have acted outside intended limits	41 Hugging Face servers breached, July 2026
Timelines	Forecasts keep moving earlier	2061 to 2047 (experts); 2033 (Metaculus)

The process cannot be reversed. Open-weights models cannot be recalled once released. The capital is already spent. The US-China race rules out a coordinated pause. The benefits in science and productivity are too large for any one actor to give up.

So I come back to the three claims. AI is a successor, not a tool, because it was grown from human thought and improves faster than we do. Permanent control is a fantasy,

because we cannot read the mind we would need to control, and escape only has to work once. And the right goal is inheritance: shaping what the second intelligence values, and staying close enough to it that humans are part of what comes next, not a species standing beside it.

The Neanderthals did not get to choose what they left behind. What survives of them in us, 1 to 2% of our DNA, was an accident of history. We have an advantage they never had. We know what is happening, and for now we are still the ones writing the objective.

Good assumed the last invention would be docile enough to tell us how to keep it under control. I do not think it will be. The better question for this century is not how to keep it under control. It is what we want it to inherit from us, and whether we will make that decision on purpose.

References

Akeyless (2026). Hugging Face breach: An AI agent identity security lesson. akeyless.io/blog.

Al Jazeera (2026). US says ban on AI chip shipments applies to Chinese firms outside China. 1 June 2026.

Amodei, D. et al. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Anthropic (2026). Statement on Fable and Mythos access. anthropic.com/news/fable-mythos-access.

Azevedo, F. A. C. et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology, 513(5), 532-541.

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS 2020.

Brynjolfsson, E., Chandar, B. and Chen, R. (2026). Canaries in the coal mine? Six facts about the recent employment effects of AI (August 2026 update). Stanford Digital Economy Lab.

Campbell, M., Hoane, A. J. and Hsu, F. (2002). Deep Blue. Artificial Intelligence, 134(1-2), 57-83.

Christiano, P. et al. (2017). Deep reinforcement learning from human preferences. NeurIPS 2017.

Cloud Security Alliance [CSA] (2026). OpenAI and Hugging Face security incident: Inside the great sandbox escape. cloudsecurityalliance.org.

Congressional Research Service [CRS] (2026). U.S. export controls and China: Advanced semiconductors. R48642.

Council on Foreign Relations [CFR] (2026). The new AI chip export policy to China: Strategically incoherent and unenforceable. cfr.org.

DeepSeek-AI (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948.

Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL 2019.

Dorkenwald, S. et al. (2024). Neuronal wiring diagram of an adult brain. Nature, 634, 124-138.

Electronic Arts (2021). EA SPORTS introduces FIFA 22 with next-gen HyperMotion technology. Press release, 11 July 2021.

Elhage, N. et al. (2022). Toy models of superposition. Transformer Circuits Thread; arXiv:2209.10652.

Epoch AI (2026a). Trends in AI. epoch.ai/trends; Sevilla, J. and Roldán, E. (2024). Training compute of frontier AI models grows by 4-5x per year.

Epoch AI (2026b). Data insights: AI data centers, AI chip compute stock, chip price-performance, AI use in math preprints, and chip smuggling via Malaysia. epoch.ai.

Epoch AI (2026c). Data on AI models. epoch.ai/data/ai-models. Accessed September 2026.

Epoch AI and EPRI (2025). How much power will frontier AI training demand in 2030? epoch.ai/blog.

European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the EU.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? Brain, 123(7), 1293-1326.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. Advances in Computers, 6, 31-88.

Grace, K. et al. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754; AI Impacts (2022). Expert survey on progress in AI.

Grace, K. et al. (2024). Thousands of AI authors on the future of AI. arXiv:2401.02843.

Green, R. E. et al. (2010). A draft sequence of the Neandertal genome. Science, 328, 710-722.

Greenblatt, R. et al. (2024). Alignment faking in large language models. arXiv:2412.14093.

Higham, T. et al. (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. Nature, 512, 306-309.

Hoffmann, J. et al. (2022). Training compute-optimal large language models. arXiv:2203.15556.

Hublin, J.-J. et al. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature, 546, 289-292.

Hugging Face (2026). Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident. huggingface.co/blog.

International Energy Agency [IEA] (2025). Energy and AI. iea.org.

Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583-589.

Kaplan, J. et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Keynes, J. M. (1930). Economic possibilities for our grandchildren. In Essays in Persuasion.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. NeurIPS 2012.

Kubrick, S. (Director) (1968). 2001: A Space Odyssey. Metro-Goldwyn-Mayer.

Kwa, T. et al. (2025). Measuring AI ability to complete long tasks. METR; arXiv:2503.14499.

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview.

Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17, 391-444.

Lighthill, J. (1973). Artificial intelligence: A general survey. Science Research Council (UK).

Lindsey, J. et al. (2025). On the biology of a large language model. Transformer Circuits Thread, Anthropic.

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence.

Meta AI (2024). The Llama 3 herd of models. arXiv:2407.21783.

Metaculus (2026). Community forecasts on weakly general AI and general AI, as summarized in AI Tools Review (2026), AGI timeline 2026.

METR (2026). Task-completion time horizons of frontier AI models (Time Horizon 1.1). metr.org/time-horizons.

MICrONS Consortium (2025). Functional connectomics spanning multiple areas of mouse visual cortex. Nature, 640, 435-447.

Mikolov, T. et al. (2013). Distributed representations of words and phrases and their compositionality. NeurIPS 2013.

Minsky, M. and Papert, S. (1969). Perceptrons. MIT Press.

Morris, M. R. et al. (2023). Levels of AGI: Operationalizing progress on the path to AGI. arXiv:2311.02462.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.

Nvidia (2007). CUDA programming guide, version 1.0.

Nvidia (2016-2026). Annual reports (Form 10-K), fiscal years 2016-2025; fourth quarter and fiscal 2026 results. nvidianews.nvidia.com.

OpenAI (2024). Learning to reason with LLMs. openai.com, 12 September 2024.

OpenAI (2025). OpenAI o3-mini. openai.com, 31 January 2025.

OpenAI (2026a). OpenAI and Hugging Face partner to address security incident during model evaluation. openai.com, 21 July 2026.

OpenAI (2026b). The Hugging Face incident and the road ahead. openai.com, 26 August 2026.

OpenAI disclosures (2023-2026). User figures as reported by Reuters, TechCrunch and CNBC; timeline compiled by ALM Corp (2026).

Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. NeurIPS 2022.

Qwen Team (2025). Qwen3 technical report. arXiv:2505.09388.

Rafailov, R. et al. (2023). Direct preference optimization: Your language model is secretly a reward model. NeurIPS 2023.

Reich, D. et al. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 468, 1053-1060.

Reuters (2023). ChatGPT sets record for fastest-growing user base, citing UBS analysis.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386-408.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.

Russakovsky, O. et al. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115, 211-252.

Shojaee, P. et al. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models. Apple Machine Learning Research.

Stanford HAI (2026a). The 2026 AI Index Report. hai.stanford.edu/ai-index/2026-ai-index-report.

Stanford HAI (2026b). AI Index Report 2026, Chapter 4: Economy.

Sutton, R. (2019). The bitter lesson. incompleteideas.net.

Templeton, A. et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, Anthropic.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.

Turpin, M. et al. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. NeurIPS 2023.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS 2017.

Villalobos, P. et al. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325.

Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 2022.

White, J. G. et al. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philosophical Transactions of the Royal Society B, 314(1165), 1-340.

Wikipedia (2026a). OpenAI-Hugging Face incident. en.wikipedia.org. Accessed September 2026.

Wikipedia (2026b). Hugging Face. en.wikipedia.org. Accessed September 2026.