Upstream of Everything

Clarity, Deflation, and the Race to Build Minds That Judge for Us

James Booth

September 2026

What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herbert A. Simon (1971, pp. 40–41)

Abstract

Every decision is made on a model of the world, and a model that is wrong in a way that matters aims every later choice at a fiction. This paper argues that clarity, a picture of the world that is less confused, less flattering, and more usable, is the upstream condition for good judgment, and it follows that claim from the human mind to the machines now being built to judge for us. Human judgment is bounded: attention is scarce, working memory holds about four items, and natural selection built the brain slowly for a smaller world. Technology runs the same process of copying, variation, and selection on faster substrates, and it is deflationary, turning yesterday's scarce capability into background. Machine learning points that process at judgment itself. Drawing on records of training compute, inference prices, capital spending, and task-completion horizons, the paper traces how scale, better algorithms, and post-training methods turned next-token prediction into systems that draft, code, and reason. It then tests forecasts of automated researchers and gigawatt clusters against six bottlenecks and the power actually built. The largest clusters have kept pace with 2024 projections through 2026, estimates of the economic payoff run from a productivity gain of under 1% over a decade to growth of 30% a year, and current safety training can reach only a few tokens deep. The paper closes with six scenarios and one claim: capability is the slope, and alignment decides whether the slope points at a world we can inhabit.

1. Introduction: Clarity as the Upstream Condition

The world we navigate is made of objects with locations and motions that we can treat as settled. Quantum physics says that picture is not fundamental. Underneath it, a system is not a tiny ball already doing one thing. It is a set of possible ways the system can show up, and which way becomes definite depends on how it interacts. What that description means is a matter of interpretation. Copenhagen-style views treat the quantum state as a tool for predicting observations (Faye, 2024). Many-worlds, which grew out of Everett's (1957) relative-state formulation, takes the state as the full reality and says that every outcome happens in some branch. In QBism the state represents an agent's beliefs (Fuchs et al., 2014). Bohmian mechanics adds definite particle positions plus a guiding wave (Bohm, 1952), and relational quantum mechanics says a fact is always a fact for some system, never a fact from nowhere (Rovelli, 1996). These interpretations were built to reproduce the predictions of standard quantum theory, and no experiment has yet separated them. Bohm (1952) showed that as long as the mathematics keeps its present general form, his interpretation "leads to precisely the same results for all physical processes as does the usual interpretation" (p. 166). Objective-collapse theories, which change the dynamics themselves, stand apart: they predict small additional effects, and the experiments run so far have not observed them (Vaidman, 2026).

Physicists still do not agree on what the mathematics represents. In 2025 *Nature* surveyed researchers whose recent papers involved quantum mechanics. Of the more than 1,100 who responded, 36% chose the most popular interpretation, Copenhagen, and only 24% thought their favored interpretation was actually correct (Gibney, 2025). Anyone who says quantum physics "states" that the universe splits, that only observations exist, or that the mind is fundamental is selling an interpretation. Determinism and chance can both hold at different levels of description. Bohmian mechanics is one concrete case: its dynamics are fully deterministic, and it is empirically equivalent to orthodox quantum theory (Goldstein, 2025). None of this means reality is made of thought. What the mind does invent is its picture of the world.

I treat clarity as a necessary condition for humanity's advancement as a whole. By clarity I mean a picture of the world that is less confused, less flattering, and more usable, and I hold this as a first-principles claim about agents. An agent only ever acts on a model: of the situation, of the trade-offs, of itself. If the model is wrong in a way that matters, every downstream choice is aimed at a fiction. Income, training, relationships, and strategy all come after that. They are outputs of what a person took to be true, and clarity is the filter they have to pass through. So when I say that clarity dictates everything, I mean that it sits upstream of the rest. It is one good among several, and the others depend on it.

This paper follows that claim from the human observer to the machines now being built to judge for us. Evolution built a bounded judge. Technology runs the same process of copying, variation, and selection on faster substrates, and it is deflationary: it keeps delivering more for less until former scarcities become part of the environment. Artificial intelligence is that process aimed at judgment itself. The questions that remain, about general intelligence, superintelligence, capital,

and alignment, come down to who writes the score and whether we can still see which branch we are on.

The paper is an argumentative synthesis. I state a position and draw on published research where it tests or supports that position, including peer-reviewed studies, working papers, company disclosures, and public data sets. The argument began as an essay, and its factual claims have since been checked against sources; where a source disagreed, I corrected the claim and cited the source. Many of the numbers change quickly, so each is tied to a dated source, and the text gives the date where it matters. Forecasts are reported as forecasts and, where enough time has passed, set beside what has been observed since. Most figures are redrawn from public data, and the two diagrams reproduced from other works carry their original attribution.

The sections follow the argument in order. Section 2 describes the limits of the human judge. Section 3 sets out the three conditions for selection and the slow pace at which they produced the brain and the eye. Section 4 treats technology as the same process on a faster substrate and develops the argument that it drives prices down. Section 5 explains how models are trained, why scale worked, which multipliers account for recent progress, and why fluency, however cheap, is a different thing from clarity. Section 6 adopts an operational definition of general intelligence and examines six bottlenecks that stand in its way. Section 7 asks what follows if the loop runs past human level and how little that forecast specifies. Section 8 compares estimates of economic impact and considers military and scientific effects. Section 9 follows the capital and checks projected build-outs against what has been built. Section 10 takes up the problem of steering systems smarter than their overseers and sets out a taxonomy of risk. Section 11 lays out six scenarios, and Section 12 returns to clarity.

2. The Bounded Observer: Attention, Memory, and the Limits of Judgment

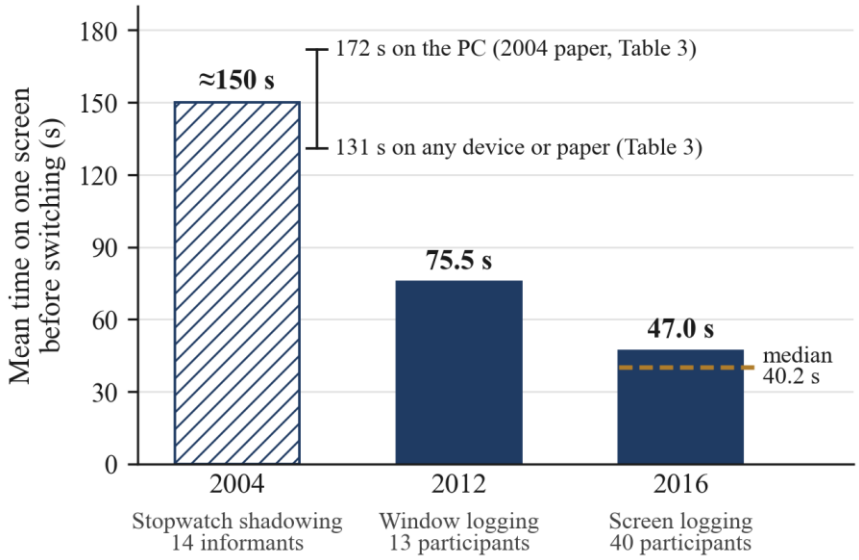
That filter is unusually hard to keep clean now, because the environment is built to dirty it. The scarce input is attention, and Herbert Simon's point, quoted in the epigraph, still holds: information consumes attention (Simon, 1971, p. 40). The claims on that attention are large. Adult internet users report spending an average of 6 hours and 38 minutes online each day, 2 hours and 21 minutes of it on social media (Kemp, 2025a). A later edition of the same report, which also counts video platforms such as YouTube and TikTok, puts time on social and video feeds at well over two and a half hours a day, and its authors caution against comparing its figures with earlier editions (Kemp, 2025b). In the United States, a 2026 survey of 1,000 adults put self-reported screen time across phones, computers, tablets, and connected televisions at 12 hours and 1 minute a day, a total that probably counts some overlapping use, such as scrolling a phone while the television runs, more than once (Wheelwright, 2026b). A separate survey from the same publisher found that American adults check their phones 186 times a day, about once every five minutes they are awake (Wheelwright, 2026a).

Attention on any single screen has also shortened, at least at the office computer. Gloria Mark and her colleagues have spent two decades measuring how long information workers stay

on one screen before switching. In 2004 her team shadowed workers with stopwatches, and Mark has since summarized the average from that period as about two and a half minutes (González & Mark, 2004; Mills, 2023). Around 2012, computer logging put the average at 75 seconds (Mark et al., 2012). In a 2016 study that logged the computer activity of 40 information workers over two work weeks, the mean was 47 seconds and the median 40 seconds (Mark et al., 2016). Mark reports that later measurements, including replications by other groups, have stayed within a few seconds of that level (Mills, 2023). Figure 1 traces the decline.

Figure 1

Average Time on One Computer Screen Before Switching in Three Workplace Studies, 2004 to 2016



Note. Data from González and Mark (2004, Table 3), Mark et al. (2012, Table 2), Mark et al. (2016, Table 1), and Mills (2023). The values describe computer use at work, not attention span in general, and come from three studies with different samples and methods. The hatched 2004 bar is Mark's later summary of the 2004 study (Mills, 2023); the bracket shows that paper's own values. The dashed line marks the 2016 median.

These are studies of office workers moving between windows and applications on a computer, so they measure neither attention in general nor time on phones and social feeds. They also show that switching has more than one cause. Mark et al. (2016) note that switching attention among activities "can be triggered internally (e.g., through boredom) or by external sources (e.g. email notifications)" (p. 1739), and in their data shorter focus went with the personality trait of neuroticism and, less firmly, with higher stress. Mark's group has also studied self-interruption directly, as a habit tied to the work environment (Dabbish et al., 2011). The brain has not broken. Part of the switching is prompted from outside, by notifications and by feeds designed to hold attention, and part of it comes from the person doing the switching.

The supply is far larger than any person could sample. YouTube's press page stated in 2020 that more than 500 hours of content were uploaded every minute (YouTube, 2020). The company

no longer gives an hourly figure, and the same page now reports more than 20 million new videos a day (YouTube, n.d.). What reaches a given viewer from that pool is chosen by ranking systems, and YouTube's own engineers described their ranking objective as "generally a simple function of expected watch time per impression" (Covington et al., 2016, p. 195). The cost of that design is beginning to enter economic estimates: Acemoglu (2024) lists "addictive social media" among the new AI-generated tasks that "are manipulative and may have negative social value" (p. 5). So a person's picture of the world now comes less from what they went looking for and more from whatever survived a ranking system optimized for hold time, which means consumption is what writes the model.

Even careful judgment works under a cap. A person making a decision tries to use all the data available, but what anyone can actually hold is a few variables, a story that fits in working memory, and a handful of analogies. The limit is tight. Cowan (2001) puts the central capacity of working memory at about four chunks, fewer than the seven items, plus or minus two, of Miller's (1956) earlier estimate. Reasoning is a bottleneck that forces us to rank, discard, and guess. Writing things down, using tools, waiting, and getting other eyes on the problem all ease it, and none of them removes it. Every action we take is a judgment made on incomplete variables.

The limitation is architectural. Selection built a nervous system to track a small band, a few allies, and a local season. The social brain hypothesis ties the growth of primate brains to the demands of social life and implies a limit on the size of the group a person can keep track of (Dunbar, 1998). We now ask that same organ to price global markets and long-horizon risk while it drinks from an information firehose. Nobody is at fault here. A machine that solved an old problem has been dropped into a much larger one, and relative to the state space of the problems we now pose, we know almost nothing.

3. Evolution: The Slow Algorithm That Built the Bottleneck

Evolution built that bottleneck, and the brains that won were the ones that were "good enough." The mechanism needs only three conditions, and the same three return in Sections 4 and 5, running on faster substrates.

3.1 Three Conditions for Selection

1. Replication. There are things that make copies of themselves. At the chemical level, that means a molecule, or a network of molecules, that catalyzes the production of more molecules like itself. RNA can do a primitive version of this, and DNA plus proteins later proved far more reliable. In the laboratory, Lincoln and Joyce (2009) converted an RNA enzyme into a pair of enzymes that catalyze each other's synthesis and keep amplifying without any proteins. Whether life actually began with RNA replicators remains a hypothesis, but the chemistry is possible. Once there is a self-copying system, there is a lineage.

2. Imperfect copying. The copies are not perfect. Chemistry is noisy. Thermal motion, radiation, chemical side reactions, and the finite accuracy of polymerases all introduce changes.

Most changes do nothing or break the copy, and advantageous mutations are rare (Eyre-Walker & Keightley, 2007). A few change how well the next generation copies or survives, and those are what we call variation. They are not random in the everyday sense. Known physical processes generate them, and those processes happen to be uncorrelated with what would be useful later. Luria and Delbrück (1943) showed as much in bacteria, where mutations conferring resistance to a virus arose before the cells had ever been exposed to it. A 2022 study reported that mutations occur less often in the essential genes of the plant *Arabidopsis thaliana* (Monroe et al., 2022), but a reanalysis attributed much of that pattern to sequencing errors (Liu & Zhang, 2022). The standard view still holds, though it is now being tested.

3. Finite resources. Resources and space are finite. More copies are made than can persist, so some lineages leave more descendants than others. If the differences behind that gap are heritable, they become more common. That is selection. It needs no foresight and no judge, only differential persistence.

Given those three conditions, the composition of the population has to change, and that change is evolution. The process can be observed at the molecular scale. When Lincoln and Joyce (2009) let several kinds of their replicating enzymes compete for a common pool of substrates, recombinant replicators arose and grew to dominate the population. Lewontin (1970) stated the same logic for organisms as three principles: variation among individuals, differences in fitness, and the heritability of those differences.

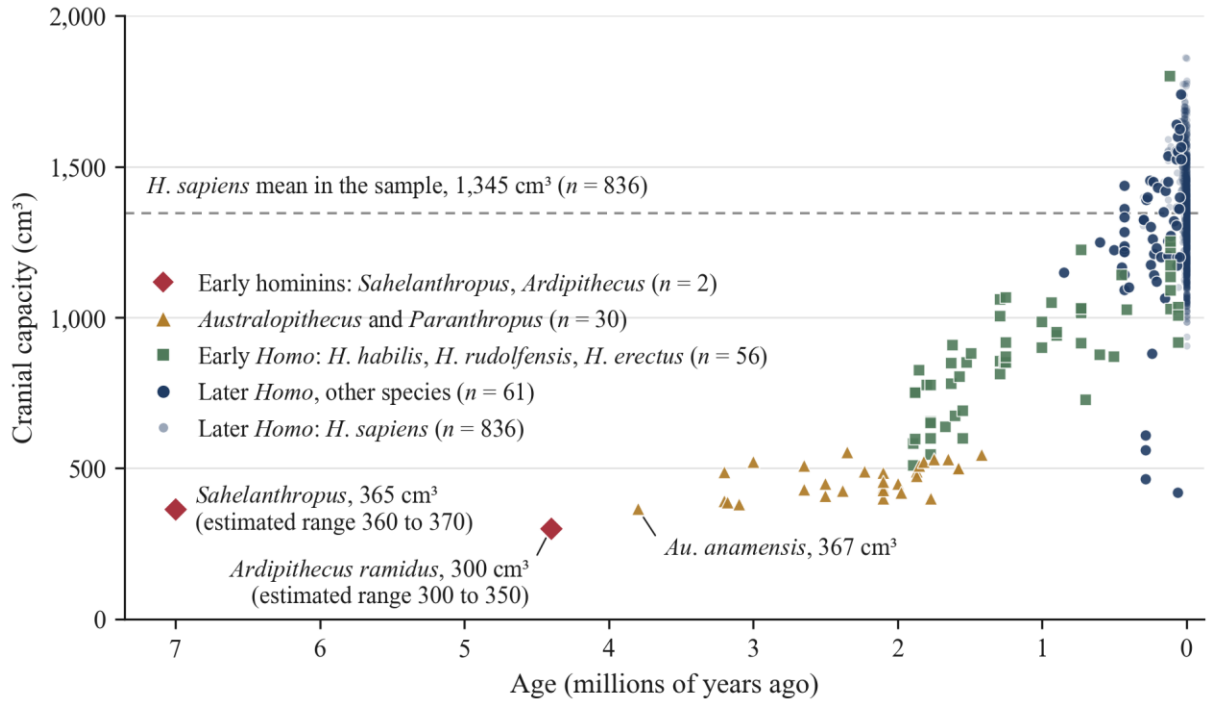
3.2 The Pace of Biological Change

Measured against a lifespan, evolution is slow. A generation in a large animal takes years or decades; in our own species it has averaged about 27 years over the past 250,000 years (Wang et al., 2023). A useful mutation has to appear, survive, and spread through a population that is still busy not dying. We are not wired to think on that timescale.

The fossil record of the brain gives the scale. The earliest known hominins had brains about the size of a chimpanzee's. The *Sahelanthropus tchadensis* cranium from Chad, about 7 million years old, has an estimated endocranial volume of 360 to 370 cm^3 (Guy et al., 2005), and the 4.4-million-year-old *Ardipithecus ramidus* skull held about 300 to 350 cm^3 (Suwa et al., 2009). Modern humans average about 1,300 to 1,350 cm^3 (DeSilva et al., 2021; Smithsonian National Museum of Natural History, n.d.-b). Over roughly 7 million years, then, brain volume grew three- to fourfold, a change DeSilva et al. (2021) describe as a near quadrupling. Most of the growth came late. Brain size rose only slightly between about 6 and 2 million years ago (Smithsonian National Museum of Natural History, n.d.-a), and the rate of increase jumped at about 2.1 and again at 1.5 million years ago, alongside the early evolution of *Homo* (DeSilva et al., 2021). Figure 2 plots the dated crania.

Figure 2

Hominin Cranial Capacity Over the Past 7 Million Years



Note. Data from DeSilva et al. (2021), Supplementary Table 1, licensed under CC BY 4.0. Each point is one of the table's 985 dated hominin crania (two Miocene ape specimens excluded), plotted at its single age estimate, so dating uncertainty is not shown. As in the table, *Paranthropus* is grouped with *Australopithecus*. Ranges for the two oldest crania are from Zollikofer et al. (2005, as cited in Guy et al., 2005) and Suwa et al. (2009).

That growth is the difference between a skull that could not hold a modern inner life and one that can, and it took more than a thousand times as long as all of recorded history. Writing first appears on clay tablets in southern Mesopotamia in the later fourth millennium BCE (Spar, 2004), so the written record covers a little over five thousand years. Even the fast final stretch of brain growth, about 2 million years, is several hundred times longer than that.

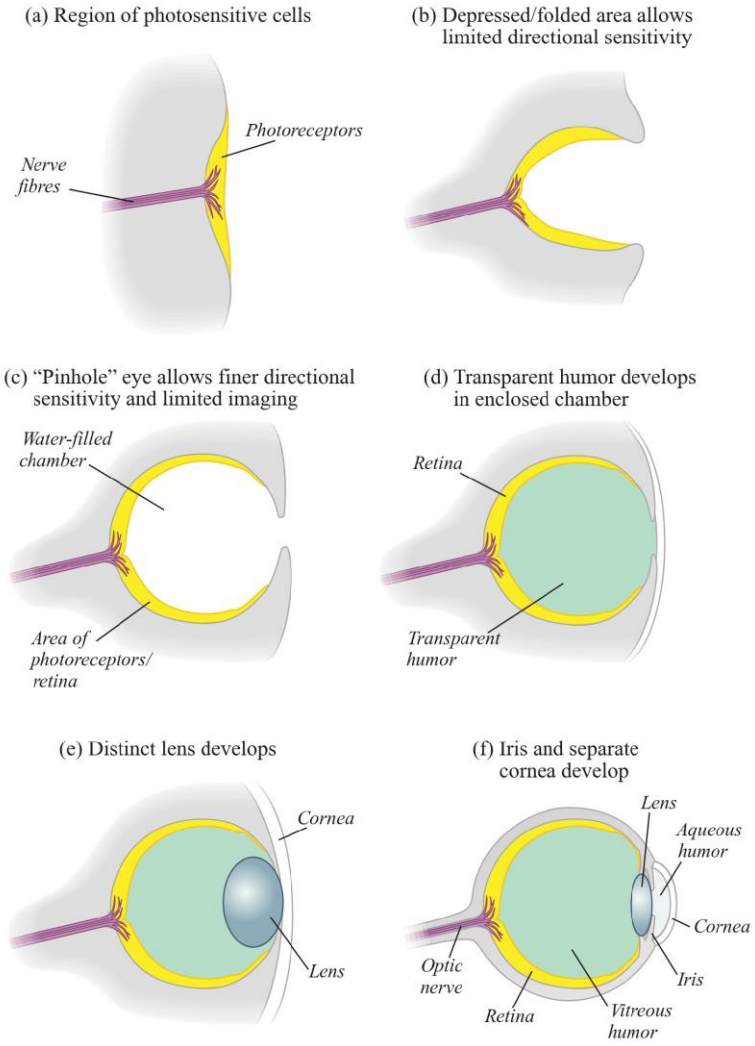
The vertebrate camera eye tells the same kind of story. Early organisms had evolved photoreceptors that could signal light more than 600 million years ago (Lamb et al., 2007). In the sequence Lamb et al. (2007) reconstruct, simple paired light-sensing organs in an early chordate had become a two-layered retina without optics by about 530 million years ago, and this rudimentary eye then acquired a lens, more retinal processing, and muscles to move it. They place these most fundamental steps in a few tens of millions of years around the Cambrian explosion. An eye equivalent to that of a modern adult lamprey, with almost all the features that characterize the vertebrate eye today, was present at least 500 million years ago, and it has been refined on that plan ever since.

The optics alone could have changed much faster. Nilsson and Pelger (1994) calculated that, even on consistently pessimistic assumptions, selection could turn a light-sensitive patch into a focused lens eye in "only a few hundred thousand years" (p. 53). Figure 3 sketches that kind of

sequence in general form, from a patch of photosensitive cells to an eye with a lens, iris, and cornea. That minimum still spans thousands of human lifetimes, and the actual history of the vertebrate eye took far longer.

Figure 3

Schematic Stages in the Evolution of the Camera Eye



Note. Adapted from *Diagram of Eye Evolution*, by Matticus78 at the English-language Wikipedia, 2006, Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Diagram_of_eye_evolution.svg). CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/). Changes: the art was re-rendered on a white background and relettered in Times New Roman; the drawings and label wording are unchanged. This adaptation is shared under the same license. The diagram is a generic schematic with no time scale, so panels (a) to (f) show an idealized progression and are not tied to a dated lineage.

4. Technology as Accelerated Selection: The Deflationary Engine

4.1 From Tools to Thinking Machines

People have always spent effort making the next unit of work cheaper than the last: first to stay fed and unharmed, then to buy slack, then to chase whatever that slack made possible. That sequence can be mapped onto Maslow's (1943) hierarchy of needs, though the history stands without it. Fire, blades, language, and the division of labor were already attempts to move a bottleneck. They cut the time spent hunting and the uncertainty about the next season, and they let a useful trick spread without waiting for a new genome. Tools are inherited variation that does not have to live in DNA, and research on cultural evolution treats that second channel of inheritance as a force in human evolution in its own right (Henrich, 2016).

Technology compressed the same loop. Once a method can be written down, run by a machine faster than a body can manage, and then improved by a later machine, the rate stops looking biological. Writing, print, engines, and computation each shortened the time between "this works" and "this is everywhere." Comin and Hobijn (2010) measured part of that compression. Tracking 15 technologies across 166 countries over the last two centuries, they found that countries adopted a technology on average 45 years after its invention, and that newer technologies spread faster than older ones. Generative AI is the newest case. In May 2026, 61.8% of U.S. adults aged 18 to 64 reported using it, and its use at work reached about 44% in a third of the time the personal computer took to reach a similar level (Harvard Project on Workforce et al., 2026).

By the mid-twentieth century the bet underneath all this was explicit. In 1950 the mathematician Alan Turing proposed replacing the question "Can machines think?" with an imitation game, in which an interrogator tries to tell a machine from a person through typewritten conversation (Turing, 1950, p. 433). Five years later, in a proposal dated August 31, 1955, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon asked for a two-month study of artificial intelligence at Dartmouth College in the summer of 1956. The study was to proceed on "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1955/2006, p. 12). The workshop went ahead, and the field has carried the proposal's name ever since. The conjecture was the old story turned on the flaws themselves. If brains that were only "good enough" won, then systems that search, compress, and act with fewer of those limits are the next increment of the same process.

4.2 Deflation as More for Less

Once a method can live outside a body, the unit of progress stops being a generation and becomes a copy of a process that can be run again for less. Technology is a machine for doing the same job with less time, less material, and fewer people in the loop. If nothing else interferes, prices fall and capability rises. Deflation in this sense means more for less, which is different from the crash the

word usually brings to mind. A given sum buys a larger slice of the same good, or the same sum buys a better one.

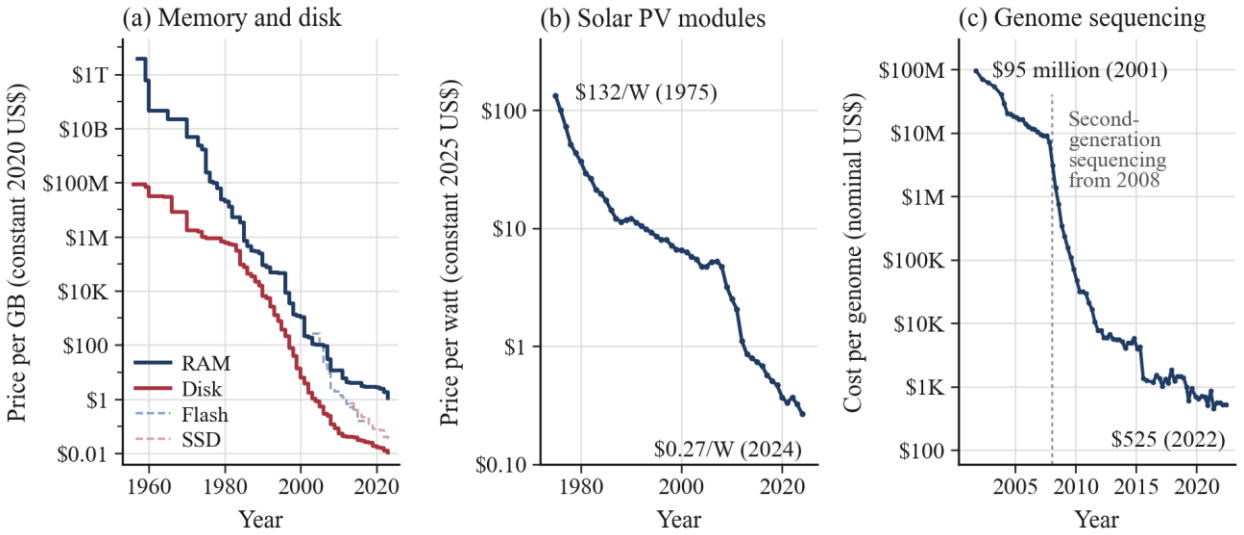
Jeff Booth, a technology entrepreneur who led the building-industry company BuildDirect for nearly two decades, makes this argument in *The Price of Tomorrow*: technology is inherently deflationary (Booth, 2020b). An essay he published the same year states the thesis in two sentences: “Exponentially advancing technological gains bring efficiency. That efficiency is deflationary” (Booth, 2020a, “Technology Has Changed the Rules” section). On that reading, prices should be falling on almost everything. In his account the interference is monetary. Prices are held up because the existing monetary system “requires inflation to remain viable” (Booth, 2020a, “Technology Has Changed the Rules” section).

A map, a camera, a flashlight, an encyclopedia, a music library, and a calculator used to be separate, scarce objects. They now sit inside a slab that most people already carry, and one more photograph or one more search costs close to nothing. The long price series show how far the process has run. In constant 2020 dollars, a megabyte of main memory cost about $4.9 million in 1970, and disk storage cost about $87,500 per megabyte in 1956. By 2023 memory cost about a dollar per gigabyte and disk storage about one cent per gigabyte, falls by factors of roughly 4.5 billion and 8 billion (McCallum, 2023; Our World in Data, 2026a). The series stops in 2023, and the decline has not run smoothly since. Memory contract prices jumped in early 2026 as cloud providers expanded their memory purchases for AI (TrendForce, 2026c), a constraint Section 6.2 takes up.

Energy and biology show the same decline. A watt of solar module capacity cost about $132 in 1975 and about 27 cents in 2024, both in 2025 dollars, a fall of roughly 500-fold (Our World in Data, 2026b). Sequencing a human-sized genome cost about $95 million in 2001 and about $525 in 2022, in nominal dollars. The steepest part of that drop began in January 2008, when the sequencing program of the National Human Genome Research Institute moved from Sanger-based methods to “second-generation” platforms (Wetterstrand, 2023). Figure 4 plots all three series. These cases fit a wider pattern. Farmer and Lafond (2016) build their forecasting model on the observation that many technologies follow a generalized version of Moore’s law, with costs falling exponentially at rates that depend on the technology, and they test it on historical data for 53 technologies.

Figure 4

Long-Run Price Declines in Memory and Disk Storage, Solar Photovoltaic Modules, and Human Genome Sequencing



Note. Panels (a) and (b) from Our World in Data (2026a, 2026b), licensed under CC BY 4.0, based on McCallum (2023) and the U.S. Bureau of Labor Statistics (2026) for panel (a) and on Nemet (2009), Farmer and Lafond (2016), and the International Renewable Energy Agency (2025) for panel (b). Panel (c) from Wetterstrand (2023), a U.S. government work in the public domain. Panel (a) is in constant 2020 US$, panel (b) in constant 2025 US$ for modules only, and panel (c) in nominal US$. Axes are logarithmic, so the panels compare rates of decline, not price levels. RAM = random access memory; SSD = solid-state drive.

Once a line of code is written, copying it costs essentially nothing. Booth's point is the same in each case. He credits the falling price to technology and competition acting together (Booth, 2020a), and in a later interview he called deflation the natural state of a free market, in which prices fall toward the marginal cost of production (OPTO, 2023). When the unit is information or computation, the marginal cost of one more copy is close to zero, so the price has a long way to fall. Section 5 follows the same curve into the price of machine intelligence.

4.3 When Capabilities Become Weather

Run long enough, this process turns a prize into something like weather: always there, and nobody pays for it. Air costs nothing, even though it is the most valuable input a person has, because it is abundant. Booth makes the same point about oxygen: "It is difficult to charge money for abundance for the same reason it is difficult to charge for the oxygen we breathe. As something becomes abundant, the ability to monetize it drops" (Booth, 2020a, "The Challenger: Bitcoin's Perspective" section). Technology does the same to capabilities. Navigation was a profession, then a device, then an app, and now it is assumed. One step in that sequence has a date. In May 2000 the U.S. government switched off Selective Availability, the deliberate degradation of public GPS signals, to make the system more useful to civil and commercial users around the world (National Coordination Office for Space-Based Positioning, Navigation, and Timing, 2021). The picture of the world each of us carries already sits on a stack of former scarcities that have become part of

the environment. That is what I mean when I say previous capabilities become abundant in nature: they leave the budget, sink into the background, and expose the next bottleneck.

5. Machine Learning: Selection on a Faster Substrate

These are the same three conditions on a faster substrate. Methods get copied. The copies are not identical, because someone changes a step. Resources such as attention, capital, and time are still finite, so the versions that deliver more for less spread. Biological evolution needed millions of years to enlarge the hominin brain three- to fourfold (Section 3.2). Cultural evolution needed thousands of years to arrive at writing and then at engines, and digital evolution needs product cycles. The mechanism is the same; only the generation time has changed.

5.1 The Anatomy of Training

AI is that mechanism aimed at the bottleneck itself: judgment under incomplete variables. Whatever the mystique, a model is a large function with knobs. The knobs are parameters, and in current systems they number in the billions or trillions. GPT-3 has 175 billion (Brown et al., 2020), and sparse mixture-of-experts models have passed a trillion (Fedus et al., 2022). Reported parameter counts have stayed near a trillion for three years, and the counts are no longer disclosed for several of the most resource-intensive systems, including those of OpenAI, Anthropic, and Google (Sajadieh et al., 2026, p. 14). The parameters turn an input (text, pixels, a board position) into an output (the next word, a label, a move). At the start the knobs are set to noise and the function is useless. Training turns that noise into a usable picture.

Training needs three pieces, and they map onto the three conditions from evolution. The first is a task with a score. The model is shown an input whose target is already known: the next token in a sentence, the correct class of an image, the better move in a game. It produces a guess, and a loss function turns that guess into a single number that says how wrong it was. High loss is a bad picture and low loss a clearer one. That number is the system's environment, and it defines what "survives" means.

The second piece is a way to assign blame. The guess comes out of a chain of simple operations: multiply by a weight, add a bias, pass the result through a nonlinearity, and hand it to the next layer. Older networks mostly used squashing functions such as the logistic sigmoid or the hyperbolic tangent for the nonlinear step; the modern default is the rectified linear unit, which passes positive values through unchanged and outputs zero for negative ones (Goodfellow et al., 2016, Section 6.3). Backpropagation is the chain rule of calculus run backwards through the stack (Rumelhart et al., 1986). For every knob it answers a local question: if this parameter were nudged up or down, would the loss get better or worse, and by how much? The result is a gradient, an arrow in the space of all parameters that points in the direction in which the error rises fastest.

The third piece is a small step against that arrow. Gradient descent moves each parameter a fraction of the way downhill, and that fraction is the learning rate. If the rate is too large, the loss oscillates and can climb; if it is too small, learning proceeds slowly and can stall at a high loss

(Goodfellow et al., 2016, Section 8.3.1). Each step is computed on a batch, a random slice of the data, so the path downhill is noisy. This is stochastic gradient descent, a descendant of the stochastic approximation method of Robbins and Monro (1951). The noise helps, because it acts as variation. Small batches can have a regularizing effect (Goodfellow et al., 2016, Section 8.1.3), and Keskar et al. (2017) attribute the better generalization of small-batch training to the noise in its gradient estimates, which steers it toward flatter minima. A full pass through the data set is an epoch. Image classifiers were typically trained for many epochs; AlexNet made roughly 90 passes through its 1.2 million training images (Krizhevsky et al., 2012, Section 5). Large language models see most of their text about once. GPT-3's training run covered less than half of its filtered Common Crawl data and repeated only small, high-quality sources such as Wikipedia a few times (Brown et al., 2020, Table 2.2). Either way, training is a long loop of propose, score, adjust.

Set this beside evolution and the parts line up. The model is a replicator of a function, and each update is an imperfect copy of the previous function. Loss plays the part of scarce fitness. The correspondence holds on average: every update is applied, and a single noisy step can make the loss slightly worse, but over many steps the weights drift toward lower error (Goodfellow et al., 2016, Section 8.3.1). No step has foresight. The system does not know what "cat" means, only that this direction on this batch reduced the number. Meaning, if it appears, is a side effect of enough of those steps on enough structure in the data. Table 1 sets the three substrates side by side.

Table 1

Correspondences Between Biological Evolution, Technological Change, and Model Training

Element	Biological evolution	Technological and cultural change	Model training
Replicator	Self-copying molecules, later DNA with proteins; a lineage of organisms	Methods and tools that can be written down, copied, and run by machines	A function: the current setting of the model's parameters, carried forward by each update
Source of variation	Copying errors from thermal motion, radiation, chemical side reactions, and imperfect polymerases	Someone changes a step in a copied method	Each update is an imperfect copy of the previous function; random batches add noise
Selection pressure	Differential persistence of heritable differences	Adoption: versions that deliver more for less spread	The loss function: on average, the weights move toward lower error
What is scarce	Resources and space	Attention, capital, and time	Energy, chips, carefully shaped data, and good targets for the loss
Generation time	Decades per generation in a large animal; millions of years for a three- to fourfold increase in brain size	Thousands of years from early culture to writing and engines; now product cycles	A single update step, repeated many times in one training run

Note. The correspondences are analogies drawn from the argument of Sections 3 to 5. The table contains no measured quantities.

The live problem is generalization. A large enough network can memorize its training set; Zhang et al. (2017) showed that state-of-the-art image classifiers can fit even labels assigned at random. The filter that matters is being right on inputs nobody has scored yet. Held-out data exist to test exactly that, because they are kept out of training. Human preference data, added later, play a different part, because they are used to train the model toward what people actually want (Ouyang et al., 2022). Both act as extra environments. A model that fits only the last batch is like a lineage that fits only last year's season.

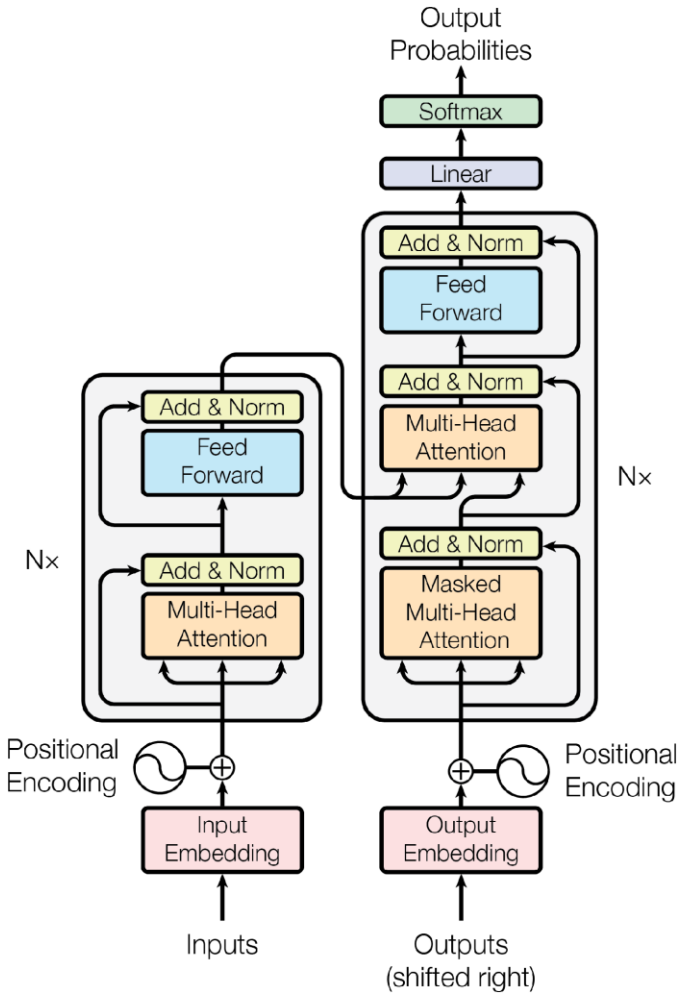
5.2 Why Scale Worked

Scale changed the feel of the field. For a long stretch, more parameters, more data, and more compute reliably pushed loss down on the same kind of next-token task. Kaplan et al. (2020) measured the relationship: language-model loss falls as a power law in model size, data set size, and training compute, "with some trends spanning more than seven orders of magnitude" (Abstract). Systems jumped from toys to things that look like judgment because the same descent, run on a larger function against a larger slice of recorded behavior, started to capture regularities that used to live only in people.

Much of that scaling ran on one architecture. Vaswani et al. (2017) introduced the Transformer, a network "based solely on attention mechanisms, dispensing with recurrence and convolutions entirely" (Abstract). Recurrent networks, then the established approach to language modeling and translation, compute a sequence one position at a time, and this "inherently sequential nature precludes parallelization within training examples" (Section 1). A self-attention layer "connects all positions with a constant number of sequentially executed operations" (Section 4), so the work of training can be spread across many processors at once. The authors reported that their model could reach a new state of the art in translation after as little as 12 hours of training on eight GPUs (Section 1), and even their base model beat all previously published English-to-German results "at a fraction of the training cost of any of the competitive models" (Section 6.1). The GPT models discussed below all build on this design. GPT-2 is "a 1.5B parameter Transformer" (Radford et al., 2019, p. 1), GPT-3 keeps GPT-2's architecture with modified attention patterns (Brown et al., 2020, Section 2.1), and OpenAI (2023) describes GPT-4 as "a Transformer-based model" (p. 1). Figure 5 reproduces the architecture diagram from the original paper.

Figure 5

The Transformer Model Architecture



Note. Reproduced from Vaswani et al. (2017, Figure 1). "Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works." The image and this notice are taken from page 3 and page 1, respectively, of the arXiv version of the paper (arXiv:1706.03762v7, 2023).

The GPT series shows what scaling bought. GPT-2, released in 2019, could string together a few plausible sentences. GPT-3, described in May 2020, took on new tasks from a handful of examples written into the prompt, "without any gradient updates or fine-tuning" (Brown et al., 2020, p. 1). GPT-4, released in March 2023, wrote and repaired code and did well on a range of academic and professional exams. OpenAI (2023) reported a simulated bar exam score "around the top 10% of test takers" (p. 1), along with strong results on the LSAT, the SAT, and several Advanced Placement exams (p. 5).

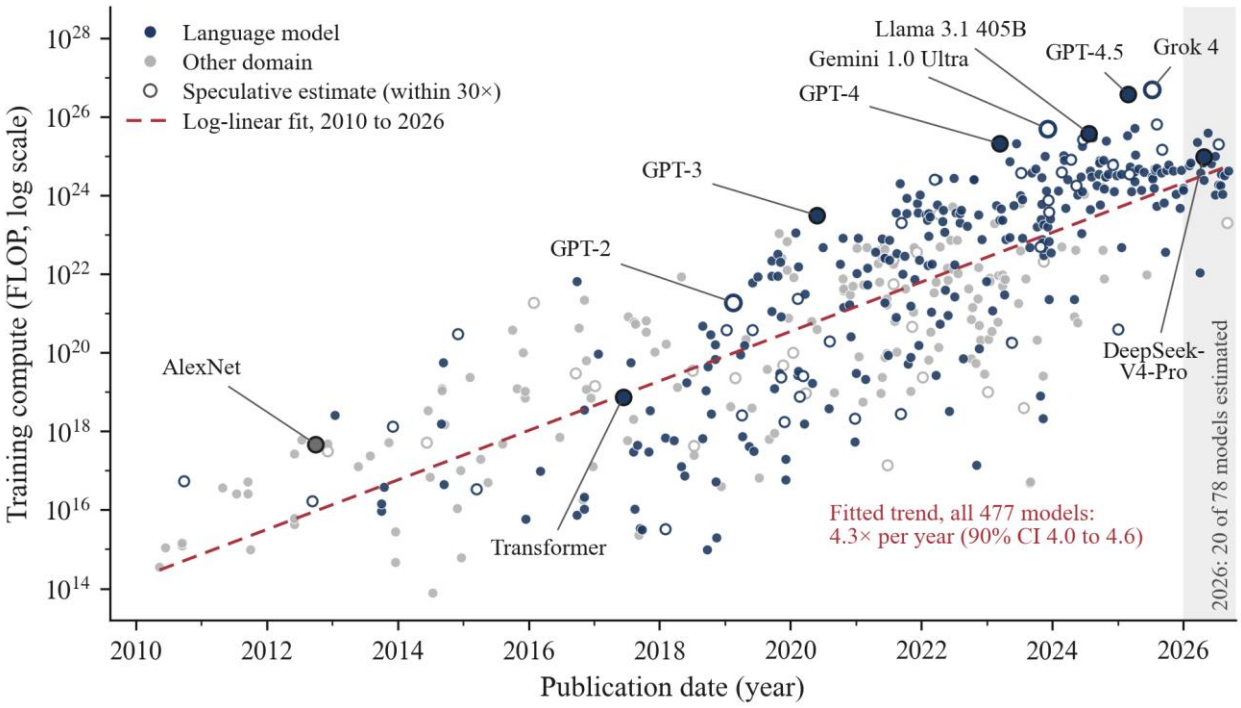
Those exam results need scoping. Martínez (2025) finds that the bar figure rests on comparisons with February test takers, a group heavily skewed toward repeat takers, and estimates GPT-4's standing against first-time takers at about the 62nd percentile. The same OpenAI table

places GPT-4 below the 5th percentile on Codeforces programming contests and between the 6th and 12th percentiles on the AMC 10 mathematics competition (OpenAI, 2023, p. 5). Chat interfaces (OpenAI, 2022), tool use, and then reasoning models trained to spend extra compute at test time (OpenAI, 2024) made the underlying models far more usable. Leopold Aschenbrenner (2024), a former member of OpenAI's Superalignment team, gives the cleanest public calibration of the whole jump in *Situational Awareness: The Decade Ahead*: GPT-2 to GPT-4 took models "from ~preschooler to ~smart high-schooler abilities in 4 years" (p. 7). He grants that such comparisons with human intelligence are imperfect (p. 10), and the analogy serves only to convey the size of the jump.

Behind these jumps sits a measured pile of floating-point operations (FLOP). Figure 6 plots the training compute of 477 notable AI models released between 2010 and September 2026, drawn from Epoch AI's public database (Epoch AI, 2026b). The vertical axis is logarithmic, so a straight line means steady exponential growth. AlexNet, which won the 2012 ImageNet competition (Krizhevsky et al., 2012), took about 4.7×10^{17} FLOP to train. Epoch estimates GPT-2 at about 1.9×10^{21} FLOP, GPT-3 at 3.1×10^{23}, and GPT-4 at about 2.1×10^{25} (Epoch AI, 2026b). The step from GPT-2 to GPT-4 was therefore about four orders of magnitude (OOMs), roughly ten thousand times more raw compute. One OOM is a factor of 10, and half of one is about a factor of 3. The largest published estimates belong to models released in 2025, led by xAI's Grok 4 at about 5×10^{26} FLOP, some nine OOMs above AlexNet (Epoch AI, 2026b). The right edge of the figure understates the frontier, because developers have stopped publishing training details for their largest systems. Of the 50 notable models that OpenAI, Anthropic, Google, and xAI released after August 7, 2025, only one has a compute estimate in Epoch's database, and several of the plotted values, including those for GPT-2 and Grok 4, are rated speculative (Epoch AI, 2026b).

Figure 6

Training Compute of Notable AI Models, 2010 to 2026



Note. Data from Epoch AI (2026b), licensed under CC BY 4.0 (updated September 22, 2026). Points are the 477 notable models since 2010 with a training-compute estimate; hollow markers are estimates Epoch AI rates as speculative. The dashed line is a least-squares fit computed for this figure (4.3× per year, against the 4.5× that Epoch AI, 2026d, reports). Frontier models released in 2026 mostly lack public compute estimates (only 20 of 78 in the shaded band have one), so the right edge understates the frontier.

5.3 Three Multipliers: Compute, Algorithms, and Unhobbling

Three multipliers stacked on that training loop account for the explosion. The first is compute, and the rise in Figure 6 came from labs buying ever larger clusters of chips specialized for AI workloads, such as GPUs and TPUs. Epoch AI (2026d) estimates that training compute has grown about 4.5 times a year, about 0.65 OOMs, for notable models since 2010, and about 5 times a year, about 0.7 OOMs, for frontier language models since 2020 (see also Sevilla & Roldán, 2024). That is 6 to 7 OOMs a decade, about four times the pace of Moore's law, which at a doubling roughly every two years comes to about 1.5 OOMs a decade. "Exponential" here is literal. On a logarithmic axis, training compute rises along a nearly straight line, and effective compute, which adds the algorithmic gains described next, rises faster still.

The second multiplier is algorithms, which make the same FLOP do more work than it used to. Part of the gain came from better use of data. Hoffmann et al. (2022) showed that parameters and training tokens should be scaled together, and they trained Chinchilla, a 70-billion-parameter model, on 1.4 trillion tokens, about 20 tokens per parameter. Mixture-of-experts layers use a gating network to send each input to a few expert subnetworks, so only part of the model fires (Shazeer et al., 2017). Residual connections and normalization let depth actually train (Ba et

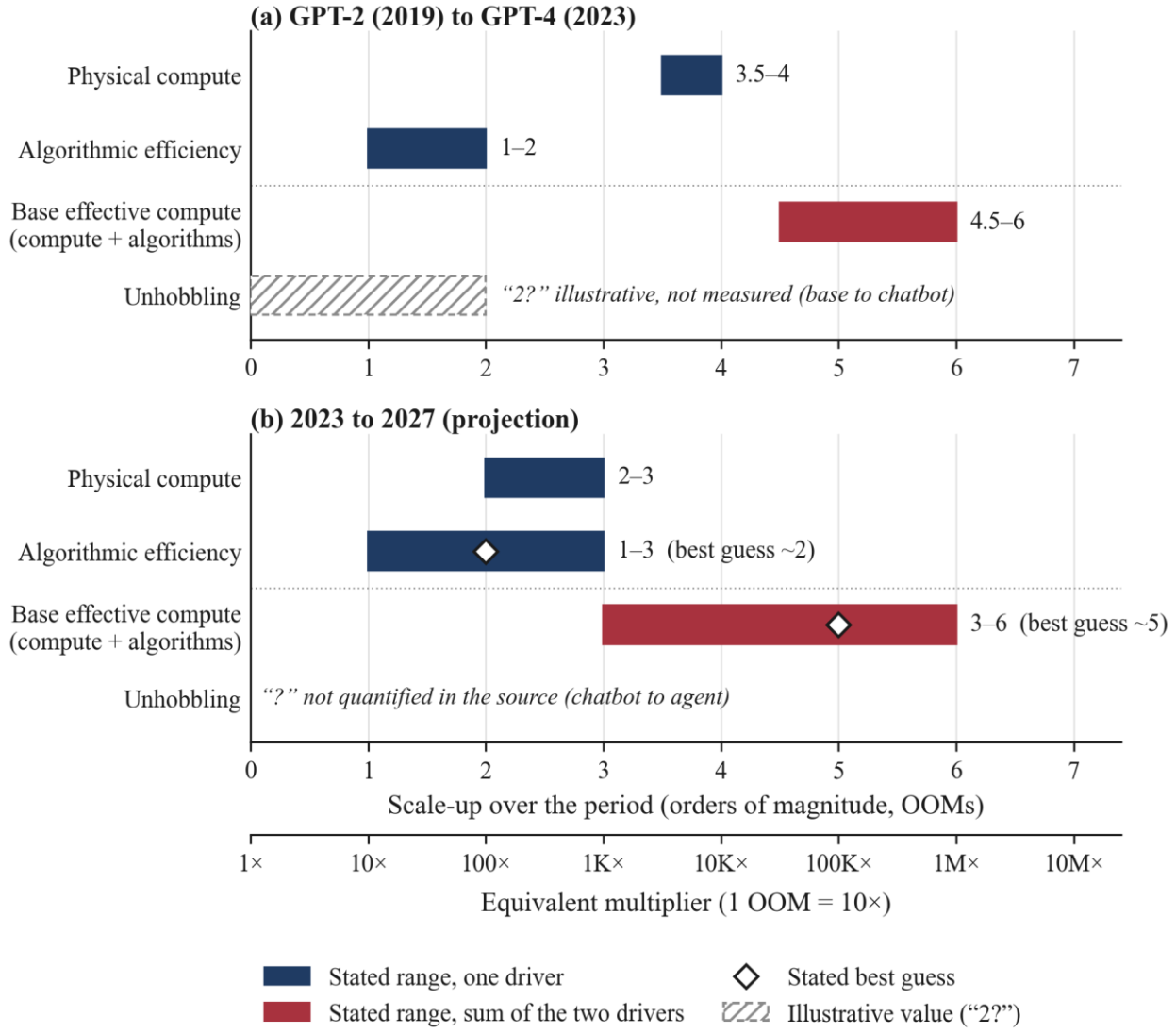
al., 2016; He et al., 2016); both were already built into the original Transformer (Vaswani et al., 2017, Section 3.1). Faster implementations have also made attention itself cheaper to compute (Dao et al., 2022).

Algorithmic gains run at a pace close to that of compute. Ho et al. (2024) found that the compute needed to reach a fixed level of language-model performance halved about every eight months between 2012 and 2023, close to half an OOM a year, although growth in compute contributed even more to progress. Epoch AI (2026d) now puts gains in pre-training efficiency at about three times a year. The result is deflation inside the lab, more capability per watt and per dollar. It is Booth's (2020b) deflationary curve applied to the loss surface.

Combine the first two multipliers and the result is effective compute: physical FLOP scaled up by what better algorithms let each FLOP do. One widely read decomposition puts the GPT-2 to GPT-4 jump at 3.5 to 4 OOMs of physical compute plus 1 to 2 OOMs of algorithmic efficiency, a base scaleup of 4.5 to 6 OOMs, and forecasts 2 to 3 OOMs from compute and 1 to 3 from algorithms for the four years after GPT-4, a combined 3 to 6 with about 5 as the best guess (Aschenbrenner, 2024, pp. 38–39). Figure 7 redraws those ranges. The third multiplier, which the same source calls unhobbling, is harder to express on the same scale, and the forecast gives it no firm number (p. 32).

Figure 7

Decomposition of Effective-Compute Growth in Orders of Magnitude, GPT-2 to GPT-4 and 2023 to 2027



Note. Adapted from Aschenbrenner (2024, p. 39, Figure 19; see also pp. 21, 25, 38). Bars span the source's low and high values for each driver. Panel (b) is a projection, and its best-guess markers come from the source (pp. 26, 39). The hatched unhobbling bar is illustrative: the source prints only "2? OOMs" and treats unhobbling as hard to put on the same scale (p. 32), and it gives no value for 2023 to 2027. OOM = order of magnitude, a factor of 10.

The third multiplier is unhobbling: gains that release capabilities a trained model already has. After pretraining, the base model is a compressed picture of text, and it cannot yet do a worker's job. Reinforcement learning from human feedback (RLHF) teaches it to follow instructions: human raters preferred the outputs of a 1.3-billion-parameter model trained this way to those of the 175-billion-parameter GPT-3 (Ouyang et al., 2022). Chain-of-thought prompting lets the model spend tokens on intermediate steps before it answers (Wei et al., 2022). Tools let it consult a browser or run code, so it no longer has to invent a page from memory. Scaffolding (plan, try, critique, retry)

turns one forward pass into a loop, and longer context lets it hold a codebase. Test-time compute is the same idea as thinking longer: buy extra inference to climb a harder problem. One forecast published in June 2024 already pointed to a large overhang of unused test-time compute (Aschenbrenner, 2024, p. 35). Three months later, on September 12, 2024, OpenAI announced o1, a model trained with reinforcement learning to work through a problem before it answers (OpenAI, 2024).

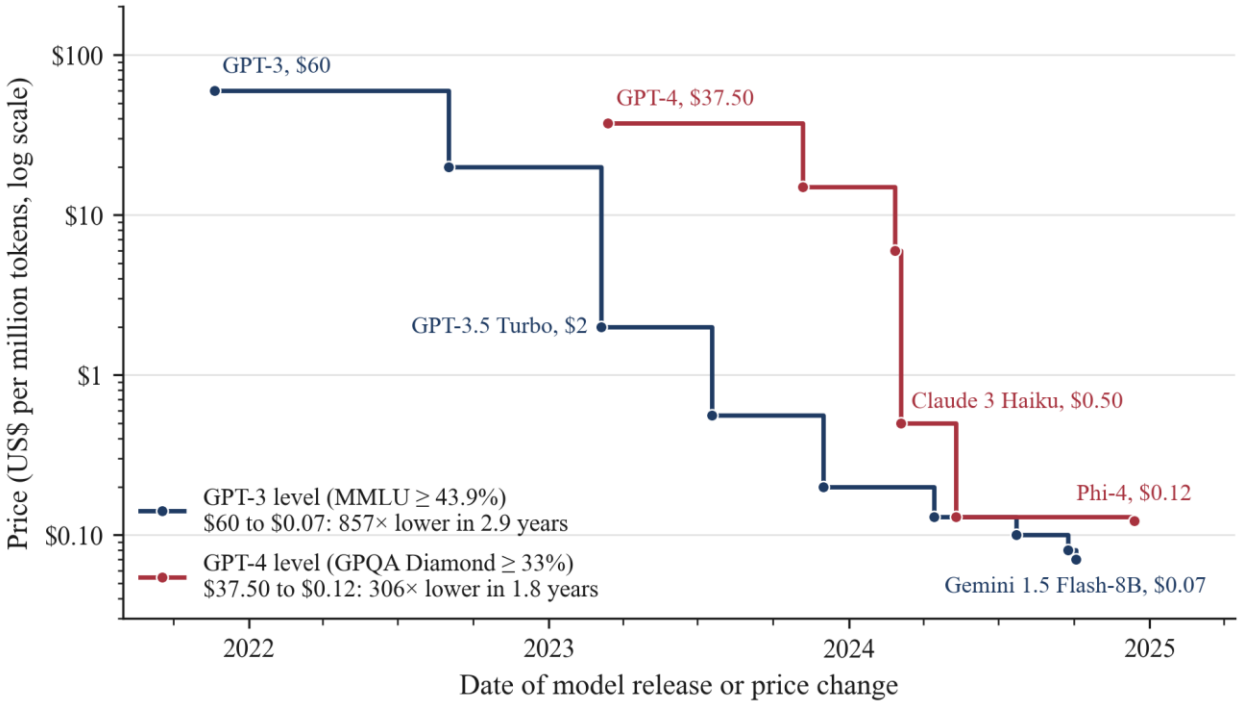
In 2024 none of this grew the pretrained brain much. The unhobbling methods used a small share of pretraining compute; the reinforcement learning stage of InstructGPT took about 60 petaflop/s-days, against 3,640 for pretraining GPT-3 (Ouyang et al., 2022, p. 17). Unhobbling took the handcuffs off a brain that was already there, which is why usefulness could jump without a new pretraining run. That picture has since shifted. OpenAI (2024) reported that o1 kept improving with more reinforcement learning during training as well as with more thinking time, and for o3 it scaled reinforcement learning by another OOM in training compute (OpenAI, 2025d). By 2025 reinforcement learning had become a large-scale training stage in its own right, and the line between growing the brain and unhobbling it is less clean than it looked in 2024.

5.4 The Remaining Scarcity

The cost of a given unit of capability fell. The deflationary curve that Booth (2020b) describes and the training curve have the same shape, with a scarce capability (a competent draft, a translation, a summary of a pile of variables) moving toward abundance. One use of a trained model, called inference, costs far less than training it, and the price keeps falling. That is why users feel the deflation as a task that used to take an expert an hour. The user base is large. In August 2026 OpenAI reported that ChatGPT alone had more than 1 billion weekly active users (OpenAI, 2026a). Figure 8 tracks the price of reaching a fixed level of capability. The cheapest model that matched GPT-3 on the MMLU knowledge benchmark cost $60 per million tokens in November 2021 and $0.07 by October 2024, and the price of matching GPT-4 on a set of PhD-level science questions fell about 40 times a year (Cottier et al., 2025). Across the milestones Cottier et al. (2025) examined, the rate of decline ranged from 9 to 900 times a year, and they caution that the fastest drops came in the most recent year and may not persist. Later price data, covering April 2024 to November 2025, show a slower decline of about 5 to 10 times a year for a given level of benchmark performance among frontier models. At the same time, the price of running frontier models rose about 3 to 18 times a year, because the models are larger and spend more tokens reasoning (Gundlach et al., 2026).

Figure 8

Lowest Price of Large Language Model Inference at Two Fixed Capability Levels, 2021 to 2024



Note. Data from Cottier et al. (2025), Epoch AI, licensed under CC BY 4.0. Each step marks a new lowest price, in US$ per million tokens (a 3:1 blend of input and output prices), among models scoring at or above GPT-3's 43.9% on MMLU (general knowledge) or GPT-4's 33% on GPQA Diamond (PhD-level science). Reasoning models are excluded, and nothing is extrapolated. Across six benchmarks, Cottier et al. (2025) report declines of 9 to 900 times per year, with a median of 50.

There is still a bottleneck. Training consumes energy, chips, and carefully shaped data. What stays scarce is good targets, good evaluation, and knowing what the loss actually measures; the forward pass itself is cheap. A model can have extremely low loss at producing fluent text and still carry a bad picture of the world. In 2022, on TruthfulQA, a benchmark built around common misconceptions, the best model tested was truthful on 58% of questions against 94% for humans, and "the largest models were generally the least truthful" (Lin et al., 2022, Abstract). The problem outlasted that generation of models. AA-Omniscience is a 6,000-question test of factual knowledge that does not penalize a model for declining to answer, so a model is free to admit uncertainty. Across 26 models reported in the 2026 AI Index, hallucination rates on it ran from 22% to 94% (Sajadieh et al., 2026, p. 136). Fluency and clarity are different things, and that distinction is where this paper started.

So the through-line is one process running at three speeds. Chemistry copies molecules and filters them, culture copies methods and filters them, and training copies a function and filters it by a number. In each case the winners are whichever versions persist under the score actually in use, dignified or not. Technology makes yesterday's winner cheap enough to treat as nature, and training is how the next winner gets grown on purpose.

5.5 What "Narrow" Now Means

Modern systems are still narrow in places, though no longer narrow in the old sense. Narrow intelligence is intelligence whose state space is the task, which says nothing about whether it is any good. Deep Blue, the chess machine that defeated world champion Garry Kasparov in a six-game match in 1997, was a massively parallel search system built around custom chess chips (Campbell et al., 2002). A chess engine is nowhere near being a colleague. A language model that can sit in a repository, read the documentation, write a patch, run the tests, and open a pull request is not a colleague either, but it is no longer Deep Blue. The SWE-bench benchmark measures this kind of work directly by asking models to resolve real issues drawn from GitHub repositories. When it was introduced, the best model tested, Claude 2, resolved 1.96% of them (Jimenez et al., 2024). By April 2026 Anthropic was reporting 93.9% for Claude Mythos Preview on SWE-bench Verified, a 500-problem subset OpenAI built in 2024 to remove tasks that were impossible to solve (Anthropic, 2026f; OpenAI, 2026b). The top of that scale has become hard to read. In February 2026 OpenAI announced that it had stopped reporting SWE-bench Verified, after finding flawed tests in at least 59.4% of an audited set of problems that models often failed, along with evidence that every frontier model it tested had seen some of the problems and their solutions during training (OpenAI, 2026b). On SWE-bench Pro, the benchmark OpenAI recommends in its place, Anthropic reported 89.9% for Claude Opus 5.5 in September 2026 (Anthropic, 2026h, p. 175). The live question is how much remote cognitive work sits closer to the first kind of system and how much closer to the second.

6. From Narrow Tools to Drop-In Workers: AGI and Its Bottlenecks

6.1 An Operational Definition

This paper takes its operational definition of artificial general intelligence (AGI) from Aschenbrenner (2024): a system that can do the work of an AI researcher or engineer, which in its unhobbled form becomes "a drop-in remote worker" (pp. 37, 41). His forecast that such a system arrives by 2027 (p. 41) is best read as a bet. In September 2026, as this paper is written, that date and the later milestones of the same projection are still ahead and cannot yet be scored; Section 9 sets its industrial forecasts against the record so far.

Put in working terms, such a worker can onboard onto a codebase, plan across days or weeks, use tools, notice when it is stuck, and ship. If that exists, it can be copied, and the copies do not sleep. A lab that can run millions of such workers on its own research stack can compress years of algorithmic progress into a much shorter interval, the intelligence explosion taken up in Section 7. AGI in this picture is the point where the search for better algorithms starts to run on the thing being improved, and the curve keeps going past it. Superintelligence, in that frame, is what you get when the loop is allowed to run.

The method behind that forecast matters more than its dates. On that method, uncertainty about what AGI requires is better expressed in OOMs of effective compute than in years. Counting

OOMs replaces years of punditry, and it turns the question "is this all hype?" into "how many more 10× do we need?" The method also makes the bet testable. Training compute for frontier language models has grown about 0.7 OOMs a year since 2020 (Epoch AI, 2026d), about four times the pace of Moore's law (Section 5.3), and algorithmic efficiency has added gains of its own (Ho et al., 2024). At those rates effective compute grows by roughly an OOM a year, about ten over a decade. Growth that fast rests on investment rising faster than any economy can sustain for long, so it is close to a one-time scaleup. By the forecast's own reasoning, if it does not produce the worker described above within about a decade, the case for short timelines weakens sharply, and the remaining path runs through new ideas, which are harder to schedule.

6.2 Six Bottlenecks

What actually stands in the way, both of AGI in that sense and of major improvement short of it, is a stack of bottlenecks that look like the old three conditions under a new budget. Six of them matter most for the argument here.

Data. Public text is finite. One forecast finds that, if current trends continue, models will be trained on data sets roughly as large as the entire stock of public human text at some point between 2026 and 2032 (Villalobos et al., 2024). More epochs over the same text help for a while, but the returns fall. In controlled experiments, up to four epochs of repeated data changed the loss negligibly compared with fresh data, and with more repetition the value of added compute decayed toward zero (Muennighoff et al., 2023). Synthetic data, self-play, reinforcement learning (Section 5.3), and richer environments are the proposed ways through. Until they work, more compute on the same crawl only makes a thicker copy of the same picture.

Energy and chips. A training run is a power plant and a supply chain. The forecast behind the 2027 date projects the largest training clusters growing one OOM every two years, from about 10 megawatts for the GPT-4 cluster of 2022 to about 1 gigawatt around 2026 and about 100 gigawatts around 2030 (Aschenbrenner, 2024, p. 77). The 2026 step has roughly held up. Single AI sites now approach a gigawatt. Epoch AI (2026a) estimates the largest, the Colossus 2 site in Memphis, at about 946 megawatts of IT power, about 1.1 million H100-equivalents, and a capital cost of about $36 billion, against roughly 21 megawatts for the cluster Epoch associates with GPT-4 in 2022 (Epoch AI, 2026c). Epoch's construction-based projections put the largest single site at about 1.5 gigawatts in 2027 and about 2.3 gigawatts by the end of 2028 (Epoch AI, 2026a). Figure 12 in Section 9.2 plots the full projection against the observed record.

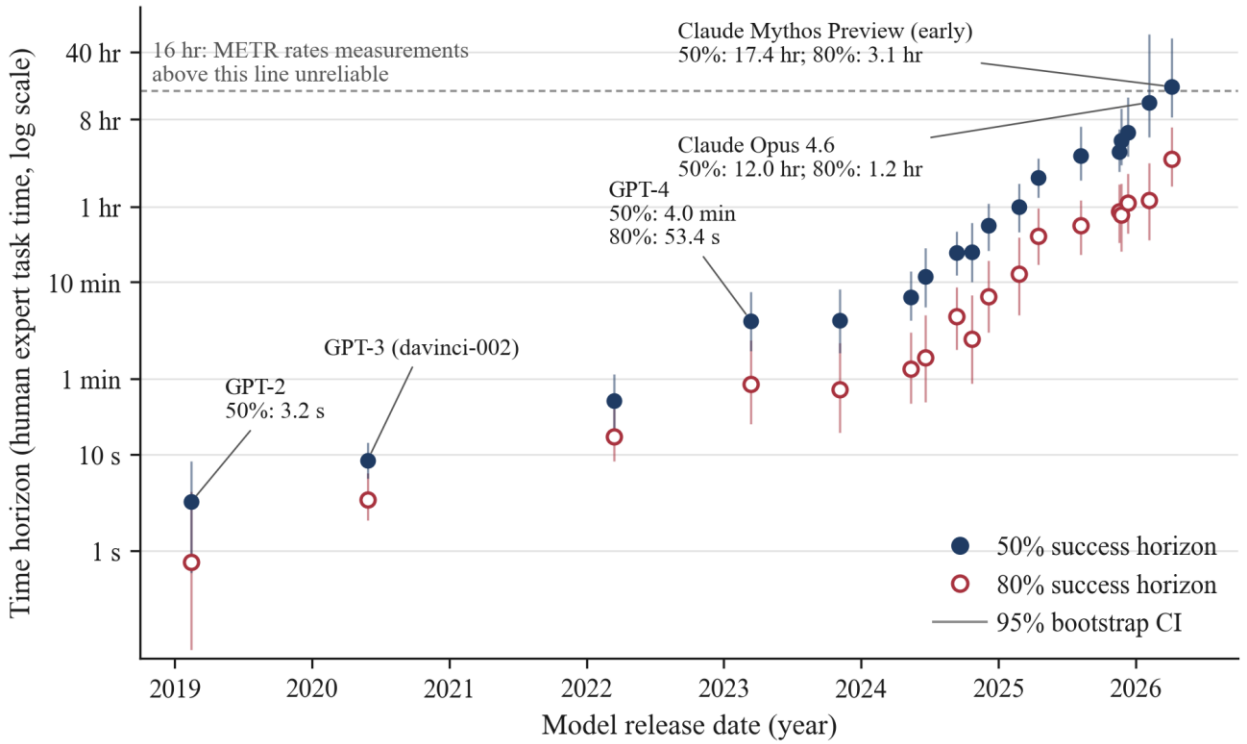
Fabrication, memory, networking, and the time it takes to pour concrete do not grow 10× because a slide says so. Memory prices show how physical supply sets the pace. Contract prices for conventional DRAM rose by about 93% to 98% in the first quarter of 2026 as suppliers shifted capacity toward server applications for AI, and TrendForce expected a further rise of 58% to 63% in the second quarter (TrendForce, 2026a, 2026c). DRAM industry revenue then rose 59.5% in the second quarter to nearly $154.73 billion, and TrendForce projects conventional DRAM contract-price growth to slow to 13% to 18% in the third quarter, in part because PC and smartphone makers

have limited room to absorb further increases (TrendForce, 2026b). In this industrial layer the scarce object has moved from the model to the building, and Section 9 follows the money into it.

Long-horizon reliability. A system can succeed four times in five on a task that takes an expert one to three hours and still fail half the time on one that fills two working days. METR measures this gap as a time horizon: the length of a software task, timed by how long human experts take to do it, at which a model is predicted to succeed with a given probability (Kwa et al., 2025; METR, 2026). By spring 2026 the strongest measured model, an early version of Claude Mythos Preview, had a 50% horizon of about 17 hours and an 80% horizon of about 3 hours. Claude Opus 4.6, released in February 2026, measured about 12 hours and a little over 1 hour (METR, 2026). METR cautions that measurements above 16 hours are unreliable with its current task suite, and the confidence intervals at the top of the range are wide. Figure 9 plots both horizons. The 50% horizon has doubled roughly every four months since 2023, and the 80% horizon has risen with it while staying four to ten times shorter for recent frontier models (METR, 2026). Agency means correcting your own errors with no human in the loop, which an exam score does not capture. That is still the weakest joint: memory that persists, credit assignment across a project, and knowing when the loss function you are locally minimizing is the wrong job.

Figure 9

Time Horizons of Frontier AI Models at 50% and 80% Task Success, 2019 to 2026



Note. Adapted from METR (2026), Time Horizon 1.1 data (updated May 8, 2026); METR states no license, so the chart is redrawn from the published values. The time horizon is the length of a software task, in human expert time, at which a model succeeds 50% (filled points) or 80% (hollow points) of the time (Kwa et al., 2025). Only models that set a new 50% record are shown, with 95% confidence intervals. METR regards measurements above 16 hours (dashed line) as unreliable.

Evaluation. Training needs a number. If the number is "sounds helpful" or "matches the test set," the result is fluency and benchmark theater, because the picture can only be as good as the score used to train it. Strathern's (1997) general form of Goodhart's law states the danger: "When a measure becomes a target, it ceases to be a good measure" (p. 308). A training loss is a measure made into a target by construction. Time horizons are one attempt at a score closer to finished work, since they grade whole tasks in human hours (METR, 2026), and they fall under the same law once labs start training against them.

Sample efficiency and the physical world. Humans learn some skills from a few trials because the body and the culture have already compressed the search. People often need to see "just one or a few examples of a new concept" before they can recognize new cases (Lake et al., 2017, Section 6), and Lake et al. trace that ability to built-in intuitive theories of physics and psychology and to learning-to-learn. Models are still hungry. The deep Q-network that DeepMind trained on Atari games in 2015 used about 924 hours of play per game, almost 500 times the experience of the professional gamer it was compared with (Lake et al., 2017, Section 3.2). Language models can now pick up a task from a few examples in the prompt (Brown et al., 2020),

but only after pretraining on trillions of tokens; the flagship Llama 3 model saw about 15.6 trillion (Grattafiori et al., 2024), and DeepSeek-V4.1-Flash, released in September 2026, was pretrained on a multimodal corpus of 45 trillion tokens (DeepSeek-AI, 2026). Robotics and wet labs add latency the virtual loop does not have, so even if remote cognitive work yields first, the rest of the economy will not automatically follow.

Security and coordination. A system that can write exploits and persuade is a different object from a chatbot. In April 2026 Anthropic reported that an unreleased model had found thousands of high-severity software vulnerabilities, some of them in every major operating system and web browser, and it kept the model unreleased, offering it to a group of partners for defensive security work (Anthropic, 2026f). At that level of capability, theft and racing become central problems. Model weights and algorithmic secrets are the assets a lab most needs to protect, since an adversary who steals them inherits the capability without the cost of building it and without any of the safety precautions its developer took. Those problems are one reason the final stretch toward AGI may not look like a product launch.

6.3 What Closing the Gap Would Mean

The slope still holds; these bottlenecks describe what it has to climb. Early AI was a theorem prover in a closed world: Newell and Simon's (1956) logic theory machine searched for proofs of theorems in symbolic logic. Modern AI is a trained function plus a set of unhobblings, getting cheaper per unit of capability (Section 5.4) and still capped by data, power, and the gap between a good answer and a finished job. The time horizons in Section 6.2 give one running measure of how far that gap has closed on software work.

Under the worker definition, AGI is the moment that gap closes for a wide band of cognitive labor, potentially any job that can be done remotely. What blocks it is what blocked every previous step: scarce copies of the right variation, filtered by a harsh environment. The environment is now clusters, corpora, and whether the loss you minimized was the task you actually needed.

7. Superintelligence and the Limits of Forecasting

Superintelligence is where the analogy to every tool you already know breaks. A plough multiplies muscle and a compiler multiplies a programmer. A system that can improve the process of making systems like itself changes who, or what, is doing the search. I. J. Good stated the point in 1965: a machine that surpassed every human intellectual activity could design better machines still, and so "the first ultraintelligent machine is the last invention that man need ever make" (Good, 1965, as cited in Aschenbrenner, 2024, p. 46).

The mechanism is spare. Once models can do the work of AI researchers, a lab could run millions of copies that read every paper, write the next experiment, and work day and night. Aschenbrenner (2024, pp. 41, 46–47) forecasts that such a workforce could compress a decade of algorithmic progress into a year or less, a gain equal on that accounting to five or more OOMs of

effective compute. The step from a capable coworker to a vastly superhuman system would then be about the size of the step from GPT-2 to GPT-4, taken again from a starting point at human level. That is the intelligence explosion: the training loop pointed at its own design.

The brakes are real. Experiments still need compute, probably the tightest limit of all; partial automation leaves the remaining human work as the bottleneck; algorithmic progress has some ceiling; and ideas get harder to find. The forecast itself concludes that together they rule out an overnight explosion but still expects a fast one, measured in months or at most a few years (pp. 53–56).

That account specifies very little. We do not know the shape of a mind as far above a competent engineer as that engineer is above GPT-2. Nobody knows whether the first systems that look like drop-in researchers will still need a human to set the loss, or will start setting it themselves. How much of "intelligence" is raw prediction and how much is long-horizon agency is an open question, and so is which failures only show up after the human has been out of the loop for weeks. Current measurement does not reach that far, since METR's task suite cannot reliably measure horizons above 16 hours (METR, 2026). If the word means anything honest, superintelligence is a placeholder for capability we cannot yet inventory. Treating it as a finished product category is the same error as treating clarity as a mood, when all you have is a picture.

What we can say without the science-fiction layer is already enough to set this moment apart from earlier innovation. Electricity, sanitation, antibiotics, and computing each stretched a human bottleneck, but none of them could, even in principle, take the role of the researcher who invents the next one at scale. The printing press copied ideas without generating and testing a million new presses by Tuesday. Nuclear weapons changed the cost of war, but they could not design the next weapon while you slept.

I read the difference as one of rate and of reflexivity: what the labs build has started to feed back into the building. By Anthropic's own count, Claude wrote more than 80% of the code merged into the company's codebase as of May 2026, up from low single digits before Claude Code launched in research preview in February 2025 (Favaro & Clark, 2026). Suppose only a softer version of the explosion is true, with a few years in which cognitive labor starts to automate its own improvement and no one-year takeoff. Even then, the generation time of useful variation collapses again, in a way that has never happened before.

8. Consequences: Labor, Security, and Science

8.1 Labor and Productivity

I start from the worker definition and set aside the headlines about "the end of jobs." If a system can do a remote cognitive job end to end, the wage for that job competes with the cost of tokens and watts. This is the same force as Booth's deflation, in which the marginal cost of the next unit of output falls (Booth, 2020b). Drafts, code, analysis, customer mail, and parts of design, law, and tutoring are already on that path, though business adoption is still partial. In the U.S. Census

Bureau's Business Trends and Outlook Survey, 23.2% of businesses reported using AI in a business function during the two weeks from August 10 to 23, 2026, and about 38% of firms with 250 or more employees did so (U.S. Census Bureau, 2026). Eloundou et al. (2024) estimate that around 80% of U.S. workers could have at least 10% of their tasks affected by large language models, and that about 19% could see at least half of their tasks affected. In one of the best-measured deployments so far, an AI assistant raised the number of customer-support issues resolved per hour by 15% on average, and less experienced and lower-skilled agents gained the most (Brynjolfsson et al., 2025).

Historically the pattern has been a scramble without instant mass unemployment. The familiar story in which roles empty out from the bottom of the skill distribution first does not match the computer era, though. Computers took over routine cognitive and manual tasks and complemented non-routine problem solving (Autor et al., 2003), so the middle of the distribution emptied first. Employment grew in high-skill and low-skill occupations relative to middle-skill ones, the pattern economists call job polarization (Acemoglu & Autor, 2011), and much of the growth at the low end came in service jobs (Autor & Dorn, 2013).

Generative AI appears to press on a different rung, the bottom of the experience ladder. Using payroll records through June 2026, Brynjolfsson et al. (2026) find "no evidence of widespread, economy-wide job displacement" (Abstract), yet employment of workers aged 22 to 25 in AI-exposed occupations is 19% lower than it would have been had it grown in step with employment of their less-exposed peers. The gap has opened mainly through fewer hires of young workers, and older workers show no comparable gap. The same data show work reorganizing around the systems. Where AI use mainly complements workers, employment is flat or rising, especially for experienced workers, and the adjustment so far has come through headcount more than through base pay. The authors present these findings as early, descriptive indicators, and they report that the gap narrows when education is controlled for and that some of the divergence predates generative AI. Acemoglu's exposure estimates also break with the older pattern. Workers without a high school diploma are the least exposed, followed by postgraduates, while those with college degrees or some college are the most exposed, and he expects between-group wage inequality to rise only slightly (Acemoglu, 2024, p. 38).

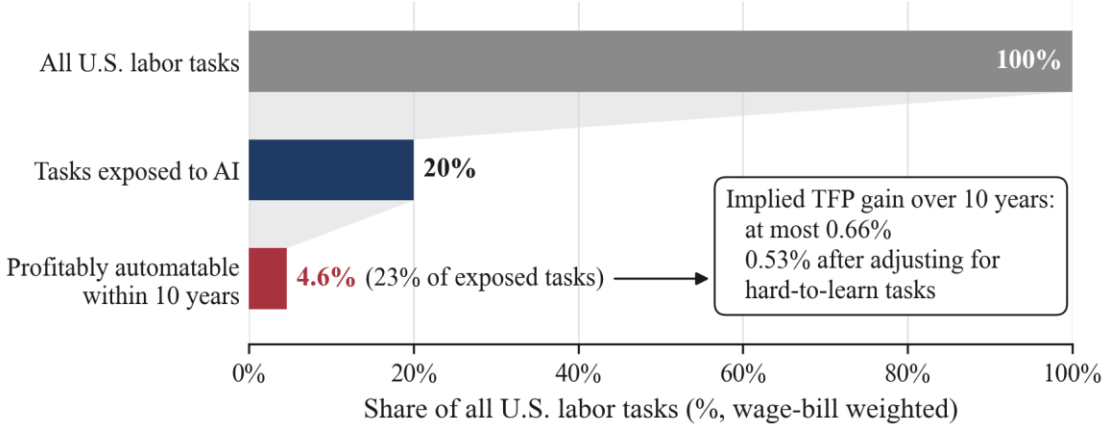
People whose edge was "I can sit with incomplete variables for eight hours" discover that the incomplete-variables machine is for sale by the hour. Acemoglu (2024) expects this kind of work to yield slowly. In what he calls hard tasks, where many contextual factors matter and "there is no clear metric of success," he expects "a tendency for the performance of AI models to be similar to the average performance of human decision-makers, limiting the potential for large productivity improvements and cost savings" (p. 16). On his reading the machine is for sale, but for judgment without a clear score it sells average judgment. This is the evaluation bottleneck of Section 6.2 as it appears in the labor market.

The most careful attempt to put a number on all of this comes from Acemoglu (2024), and the number is small. He shows that when AI's gains come from doing existing tasks more cheaply,

the gain in aggregate productivity can be estimated from the share of tasks affected and the average cost saving on those tasks, a version of Hulten's theorem (p. 3). By his count about 20% of U.S. labor tasks, weighted by wage bill, are exposed to AI, and 23% of those can be automated profitably within 10 years, which leaves 4.6% of all tasks (p. 25). With an average overall cost saving of 14.4% on that work (p. 28), total factor productivity rises by no more than 0.66% over 10 years, or about 0.064% a year (p. 29). He then lowers the bound to about 0.53% (p. 33). The experiments behind the cost savings measured easy-to-learn tasks, which have a reliable, observable outcome metric and a simple link between action and outcome (p. 15), while much of the exposed work is harder. AI also draws in new investment, so GDP rises somewhat more than productivity: 0.93% to 1.16% over the decade (p. 34), or 1.4% to 1.56% with a large investment boom (p. 40). He calls the result "a nontrivial, but modest effect" (p. 29) and places it well below outside forecasts such as that of Goldman Sachs, whose economists projected a 7% rise in global GDP over 10 years (Goldman Sachs, 2023). Figure 10 traces the chain from exposed tasks to the productivity bound.

Figure 10

Acemoglu's Chain From AI Exposure to Productivity Gains



Note. Adapted from Acemoglu (2024, pp. 25, 29, 33). Values are from NBER Working Paper 32487. Task shares are weighted by U.S. wage bills pooled over 2019 to 2022, which Acemoglu reads as shares of GDP; 4.6% = 20% × 23%. The 100% bar is the full task base by definition. TFP = total factor productivity. Both TFP figures are upper bounds, and the 0.53% figure adjusts for tasks that are hard for AI to learn.

A stronger view holds that sufficiently general systems stop sitting inside the economy and start running it, doing the planning, allocating, and discovering. That view changes one input in the arithmetic above. It grants the economist's objection that tasks are complements, so that whatever cannot be automated becomes the bottleneck, and replies that the objection hardly binds at first, because the explosion described in Section 7 begins in one small part of the economy, AI research itself. Once that research is automated and robotics follows, its best-known statement forecasts an industrial and economic explosion, with "economic growth rates of 30%/year and beyond, quite possibly multiple doublings a year" (Aschenbrenner, 2024, p. 69). Acemoglu leaves that path out on purpose. His model sets aside AI's effect on the process of science because he does not expect

large advances of that kind within 10 years (Acemoglu, 2024, p. 3), and he calls revolutionary effects "a possibility that cannot be completely ruled out" while observing that "there is so far no evidence of such revolutionary effects" (p. 42). Most of the distance between the two views is a disagreement about how large the automated share of tasks becomes, and how fast. Table 2 sets both estimates beside other published forecasts, which run from well under 1% over a decade to 30% a year.

Table 2

Competing Estimates of the Macroeconomic Impact of AI

Source	Horizon	Headline estimate	Mechanism
Acemoglu (2024, pp. 29, 33, 34, 40)	10 years	TFP up no more than 0.66% in total (about 0.064% a year), or about 0.53% once hard-to-learn tasks are counted; GDP up 0.93% to 1.16%, or 1.4% to 1.56% with a large investment boom	Hulten's theorem: about 4.6% of tasks affected (20% exposed × 23% profitable to automate) × 14.4% average cost saving
Arnon (2025), Penn Wharton Budget Model	To 2035, 2055, and 2075	Productivity and GDP up 1.5% by 2035, nearly 3% by 2055, and 3.7% by 2075; permanent boost to annual productivity growth of less than 0.04 percentage points	Acemoglu's task framework combined with a historical adoption curve and a broader set of exposed occupations
Chui et al. (2023), McKinsey & Company	Annual value; productivity path through 2040	$2.6 trillion to $4.4 trillion a year across 63 generative AI use cases; labor productivity growth of 0.1% to 0.6% a year through 2040	Value of use cases in business functions plus automation of work activities
Goldman Sachs (2023)	10 years	Global GDP up 7% (almost $7 trillion); productivity growth lifted by 1.5 percentage points over the period	Generative AI takes over part of the workload in exposed occupations; about two-thirds of U.S. occupations exposed to some degree
Korinek and Suh (2024, pp. 4, 25)	Scenarios: full automation within 20 years (baseline AGI) or 5 years (aggressive AGI)	Output growth of about 2% a year under business as usual; steady-state growth of 18% a year after full automation in the baseline AGI scenario; wages collapse as full automation nears	Race between the automation of ever more complex tasks and capital accumulation
Aschenbrenner (2024, p. 69)	The 2030s, after superintelligence	"Economic growth rates of 30%/year and beyond, quite possibly multiple doublings a year"	Automated AI research drives an intelligence explosion; superintelligence and robotics then automate all labor

Source	Horizon	Headline estimate	Mechanism
Booth (2020b)	Open-ended	No numeric forecast	Technological progress lowers costs and prices; he treats this deflation as the route to abundance

Note. Headline figures are given as each source states them and are not on a common basis: some refer to total factor productivity (TFP) and some to GDP, and some to the United States and some to the world. Korinek and Suh present simulated scenarios and do not offer them as forecasts. Acemoglu's figures are from the NBER working paper, which match the abstract of the published version. Booth's row records the direction of his argument. AGI = artificial general intelligence.

Booth is closer to Acemoglu than he first appears. Acemoglu treats a productivity improvement at the task level as the same thing as a cost saving (Acemoglu, 2024, p. 1), so his estimate can also be read as a measure of how much cheaper AI makes the economy's output. Booth's argument that technology is inherently deflationary (Booth, 2020b) is a claim about the direction of that same number. Acemoglu adds two warnings about who receives the gain. Taking revenue from social media and digital advertising, together with spending on cyberattacks and cyberdefense, as a proxy for manipulative uses of AI, he shows how such uses could appear to raise GDP by about 2% while lowering welfare by 0.72% (p. 5). He also finds "no evidence that AI will reduce inequality" and expects AI to widen the gap between capital and labor income (p. 42), with the capital share of national income rising by about 0.31 percentage points (p. 40). If he is right, Booth's abundance reaches ordinary consumers only as far as competition forces the owners of the models and the data centers to pass their savings on.

Whether one accepts the explosive version or only Acemoglu's modest one, the scarce inputs shift to energy, chips, data rights, and the ability to tell whether the model's picture is the one we wanted. Jobs that are mostly tokens move first. Jobs that are mostly coordination in physical space move later, and the exposure data agree: Acemoglu (2024, p. 3) notes that the AI exposure measure of Eloundou and colleagues is negatively associated with exposure to robots and to manual routine tasks.

8.2 Military Advantage

Aschenbrenner (2024) argues that superintelligence "will give a decisive military advantage, perhaps comparable only with nuclear weapons" (p. 126). He takes the 1991 Gulf War as a crude historical rhyme for how a lead of 20 to 30 years in military technology can make numbers irrelevant, and concludes that a lead of one to three years on superintelligence could prove just as decisive (p. 128). An arsenal built for a world of human staff officers and human targeting cycles becomes obsolete if the other side can plan, spoof, and strike on a different clock.

The core of that rhyme holds up. The Pentagon's final report to Congress described Iraq's army as the fourth largest in the world and recorded a ground offensive that lasted 100 hours (U.S. Department of Defense, 1992). U.S. forces suffered 148 battle deaths (U.S. Department of Veterans Affairs, 2023), against an estimated 20,000 to 26,000 Iraqi military deaths (Conetta,

2003). The comparison as published uses a wider range of 20,000 to 50,000 Iraqi dead (Aschenbrenner, 2024, p. 128), and estimates of Iraqi losses remain contested. Its counts of coalition dead and of tanks lost on each side could not be matched to a primary source and are not repeated here. On the figures that can be checked, U.S. battle deaths came to less than 1% of the lower estimate of Iraqi military deaths.

If superintelligence confers that kind of advantage, the live risk goes beyond who has the better chatbot to who has the cluster, who has the weights, and whether those weights leak to a state that will not pause. A lab that cannot keep its model weights secure is building the capability for whichever rival steals them. That is why the security of the labs becomes a national question, and why it is plausible to picture the final stretch toward superintelligence run by a U.S. project that looks less like a startup and more like a sensitive compartmented information facility (SCIF), a secured site accredited for classified work. One need not accept any forecast of a government-run AGI project by a given year to see the structure. If the capability is real, it is a weapon and an industrial base, and treating it as a feature in a settings menu misses that.

8.3 Healthcare and Scientific Discovery

Medicine is a sector where the rollout of these systems could easily be slowed by regulation and by social choice, and biology is research that still needs physical lab work and human experiments. The mechanism still applies. Drug design, protein work, trial protocols, imaging, and literature review are all searches over huge state spaces with expensive experiments. A system that can read everything, propose, and cheaply simulate brings deflation to the clinic, with more candidates per dollar and faster rejection of dead ends.

Protein structure is the clearest case so far. Experiments had determined the structures of around 100,000 unique proteins, a small fraction of the billions of known protein sequences, and "structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure" (Jumper et al., 2021, p. 583). AlphaFold2 has since been used to predict the structures of virtually all of the 200 million proteins that researchers have identified, work that earned half of the 2024 Nobel Prize in Chemistry (The Royal Swedish Academy of Sciences, 2024). Acemoglu (2024, p. 3) cites protein folding as an example of the change to the process of science that he leaves out of his 10-year estimate, so the figures in Section 8.1 do not count gains of this kind.

The limit is the same as everywhere else. The physical loop of wet lab, body, and regulator does not speed up 10× when the suggestion engine does, and that loop is expensive: DiMasi et al. (2016) put the capitalized pre-approval cost of a new drug at $2,558 million in 2013 dollars, an estimate that includes the cost of compounds abandoned during testing. Narrower AI is already inside regulated medicine. The U.S. Food and Drug Administration's list of AI-enabled medical devices authorized for marketing in the United States held 1,614 devices with final decisions through June 29, 2026, and radiology was the lead review panel for 1,230 of them (U.S. Food and Drug Administration, 2026). Diagnosis as text is moving too. On 302 challenging real-world cases

drawn from published case reports, a diagnostic language model included the correct diagnosis in its top-10 list 59.1% of the time, against 33.6% for unassisted clinicians, and its authors still call for further real-world evaluation (McDuff et al., 2025). Autonomous discovery that one would trust with a life is a different product. Clarity matters more here than in a marketing draft, because a fluent wrong answer costs more.

9. The Industrialization of Intelligence

9.1 Following the Capital

The industrial chapter of *Situational Awareness* (Aschenbrenner, 2024, pp. 75–88) is the part that aged in public. The series opened by reporting that talk in the industry had already moved from $10 billion compute clusters to $100 billion and trillion-dollar ones (p. 3). Two years later those numbers have program names. Some of the chapter's near-term forecasts can also be checked against company filings. It expected Microsoft and Google each to spend more than $50 billion on capital projects in 2024, and NVIDIA's data center revenue to reach a run rate of about $25 billion a quarter soon after (p. 78). Alphabet bought $52.5 billion of property and equipment that year (Alphabet Inc., 2026), and NVIDIA's data center segment earned $115.2 billion in the fiscal year that ended in January 2025, more than $28 billion a quarter on average (NVIDIA Corporation, 2025a).

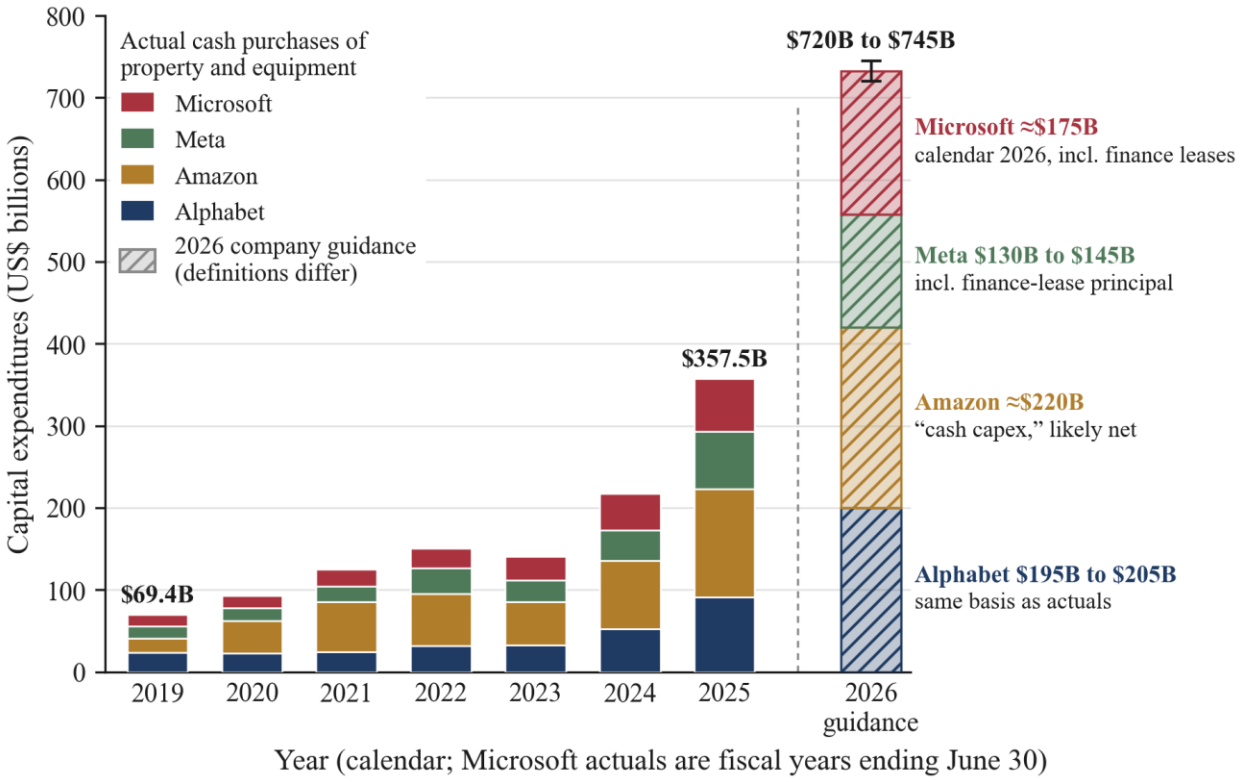
OpenAI's Stargate is the largest of the named programs. OpenAI announced it in January 2025 as a new company that "intends to invest $500 billion over the next four years," with SoftBank, OpenAI, Oracle, and MGX as its initial equity funders (OpenAI, 2025a, paras. 1–2). In September 2025 OpenAI described nearly 7 gigawatts of planned capacity and more than $400 billion of investment over three years (OpenAI, 2025e), and in October more than 8 gigawatts and more than $450 billion (OpenAI, 2025c). In November 2025 Sam Altman put the company's compute commitments at about $1.4 trillion over eight years (Bort, 2025). In January 2026 OpenAI's finance chief reported that the company's available compute had grown from 0.2 gigawatts in 2023 to about 1.9 gigawatts in 2025, and its annual recurring revenue from $2 billion to more than $20 billion (Friar, 2026). Bloomberg (2026, as cited in Juricic, 2026) reported in August that OpenAI's annualized revenue run rate had passed $40 billion. The commitments are plans and stated intentions, and Section 9.2 compares them with what is running.

The spending also shows up in audited accounts. Alphabet, Amazon, Meta, and Microsoft together bought about $69 billion of property and equipment in 2019 and about $358 billion in 2025, roughly five times as much (U.S. Securities and Exchange Commission, 2026). Microsoft's figures in this series are for fiscal years ending in June. The four companies' latest plans for 2026 add up to more than $700 billion. Alphabet expects $195 billion to $205 billion, about 60% of it for servers and the rest for data centers and networking (Investing.com, 2026). Amazon expects about $220 billion (StockAnalysis, 2026), Meta $130 billion to $145 billion including finance leases (Meta Platforms, Inc., 2026b), and Microsoft about $175 billion for calendar 2026, also

including finance leases (Microsoft Corporation, 2026b). Alphabet and Meta each plan to spend roughly double their 2025 totals, in the low hundreds of billions of dollars. All four figures are plans for the year, and the companies do not define capital spending in the same way. Figure 11 shows the series.

Figure 11

Capital Expenditures of Alphabet, Amazon, Meta, and Microsoft, 2019 to 2025, With 2026 Guidance



Note. Actuals for 2019 to 2025 are cash purchases of property and equipment from each company's Form 10-K (Alphabet Inc., 2026; Amazon.com, Inc., 2026; Meta Platforms, Inc., 2026a; Microsoft Corporation, 2026a; U.S. Securities and Exchange Commission, 2026), excluding finance leases; Microsoft's fiscal year ends June 30. The hatched 2026 bar stacks guidance as of July 2026 at range midpoints, with the whisker spanning low and high ends (Investing.com, 2026; Meta Platforms, Inc., 2026b; Microsoft Corporation, 2026b, 2026c; StockAnalysis, 2026). Guidance is not strictly comparable with actuals: Meta's and Microsoft's include finance leases, Microsoft's covers calendar 2026, and Amazon's appears net of proceeds and incentives.

Anthropic has assembled its capacity through agreements with nearly every large supplier. It announced up to one million Google TPUs, with more than a gigawatt of capacity coming online in 2026 (Anthropic, 2025c), and a $30 billion commitment to Azure compute plus additional capacity of up to 1 gigawatt through Microsoft and NVIDIA (Anthropic, 2025d). In 2026 it added up to 5 gigawatts with Amazon and 5 gigawatts of TPU capacity with Google and Broadcom (Anthropic, 2026b). It also added all of the compute at SpaceX's Colossus 1 data center, more than 300 megawatts and over 220,000 NVIDIA GPUs (Anthropic, 2026d), and up to 2 gigawatts of

AMD GPUs, with the first gigawatt beginning in the first half of 2027 (Advanced Micro Devices, Inc., 2026). Added together, the announcements exceed 10 gigawatts, although much of that capacity is described as “up to” and arrives in 2027 or later. Anthropic has not published a total for the end of 2027; the figure of about 10 gigawatts by then comes from outside estimates, such as one analyst’s projection of about 9.5 gigawatts (Trinetti, 2026). The deals rest on a revenue ramp that would have sounded like science fiction three years earlier. In February 2026 the company noted that it had earned its first dollar of revenue less than three years before and put its run-rate revenue at $14 billion (Anthropic, 2026a). In May 2026 it reported that run-rate revenue had crossed $47 billion (Anthropic, 2026b).

xAI, which SpaceX acquired in February 2026 (O’Kane, 2026), is the brute-force version of the same bet: pack a site with hundreds of thousands of GPUs and push the power envelope toward a gigawatt and beyond. Its first Memphis cluster was built in 122 days and then doubled to 200,000 GPUs in another 92 (SpaceXAI, n.d.). In its June 2026 prospectus, SpaceX told investors that its Colossus and Colossus 2 facilities together provide about 1.0 gigawatt of compute power (Space Exploration Technologies Corp., 2026). Epoch AI (2026a) estimates the Colossus 2 site alone at about 946 megawatts of IT power and 1,276 megawatts of facility power, the largest AI data center in its data. The company’s total and Epoch’s outside estimate do not reconcile exactly, a reminder of how loosely these capacities are still measured.

NVIDIA is still selling the pickaxes. Its data center segment brought in $10.6 billion in the fiscal year that ended in January 2022 (NVIDIA Corporation, 2022) and $193.7 billion in the fiscal year that ended in January 2026 (NVIDIA Corporation, 2026a). In the single quarter that ended in July 2026 it booked $89.0 billion (NVIDIA Corporation, 2026b).

Inside the labs, the constraint conversation has moved from whether a model can be trained to where the next ten gigawatts will come from, and by 2026 operators raise the question on earnings calls. In July 2026 Amazon’s chief executive, Andy Jassy, said that even with about $220 billion of capital spending the company would not have enough capacity to meet all of its demand in 2026 (StockAnalysis, 2026).

9.2 Projection Against Reality

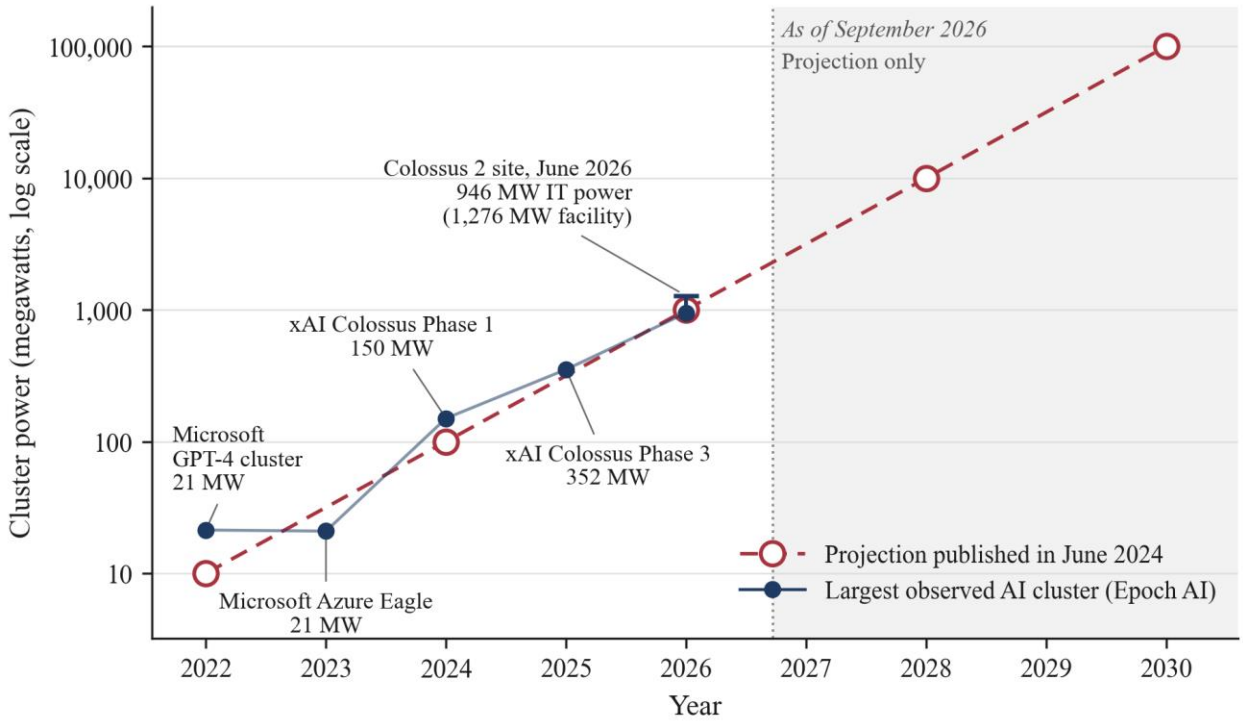
This spending is where Booth’s deflation argument and the race to scale compute overlap. Technology wants the next unit cheaper, and getting there first, at this stage, means spending like a state on buildings that may train the model that makes the next building obsolete. Announced gigawatts are not energized gigawatts. Contracts stack, and some of the numbers are press-release arithmetic. In April 2026 Epoch AI counted more than 9 gigawatts of planned Stargate capacity across seven U.S. sites, of which about 0.3 gigawatts was operating, all of it in Abilene, Texas (Stewart & Cottier, 2026). NVIDIA’s plan to invest up to $100 billion in OpenAI was a letter of intent (NVIDIA Corporation, 2025b), and Jensen Huang later said the investment had never been a commitment (Wu & Lichtenberg, 2026). Anthropic’s own announcement counts its Google and Broadcom agreement as 5 gigawatts (Anthropic, 2026b), while Broadcom’s filing puts the share

flowing through Broadcom at about 3.5 gigawatts and makes it conditional on Anthropic's continued commercial success (Broadcom Inc., 2026). For scale, the 90 AI data centers that Epoch AI tracks had an estimated 13.2 gigawatts of IT power in September 2026 (Epoch AI, 2026a), and Stargate's planned capacity alone is more than half that figure. The direction is real, though. The largest firms on earth are treating intelligence as infrastructure and racing to own the plants.

The 2024 projection of the largest training clusters, discussed in Section 6.2, can now be checked against the record. Extending a trend of roughly 0.5 OOMs a year in training compute from the GPT-4 cluster, it put the largest cluster at about 100 megawatts and 100,000 H100-equivalents around 2024, and at about 1 gigawatt and a million H100-equivalents, costing tens of billions of dollars, around 2026 (Aschenbrenner, 2024, p. 77). The first phase of xAI's Colossus reached 100,000 H100-equivalents and about 150 megawatts in September 2024 (Epoch AI, 2026c), and by mid-2026 the Colossus 2 site was estimated at 946 megawatts of IT power and about 1.1 million H100-equivalents (Epoch AI, 2026a). Figure 12 plots the projection against the observed series. At the level of the single largest site the projection has held so far, with two caveats: it does not say whether its power figures mean facility power or IT power, and the 2026 observation is a whole site that may span several buildings.

Figure 12

Projected and Observed Power of the Largest AI Training Clusters, 2022 to 2030



Note. Projections from Aschenbrenner (2024, p. 77, Table 4), which the source describes as rough back-of-the-envelope calculations; the dashed line only connects the projected points, and the source does not say whether they mean facility or IT power. Observed values from Epoch AI (2026c) for 2022 to 2025 and Epoch AI (2026a) for 2026, licensed under CC BY 4.0: the highest-performance GPU cluster operational at each year end, then the Colossus 2 site at 946 MW of IT power, with the bar reaching its estimated facility power. IT = information technology; MW = megawatts.

The next step is where the projection meets the physical limits described in Section 6.2. Continued at one OOM every two years, it calls for a cluster of about 10 gigawatts around 2028 and about 100 gigawatts around 2030. Epoch AI's estimates from construction evidence at the sites it tracks put the largest single site at about 2.3 gigawatts of IT power at the end of 2028, and all tracked sites together at about 36 gigawatts (Epoch AI, 2026a). Sites that have not yet been announced could change that picture. On present evidence, though, the 2028 step of the projection is where the announcements and the energized capacity part company.

9.3 What Remains Unknown

It is worth being plain about what we do not know. The explosion could take months or a decade. Superintelligence could be a smooth continuation of next-token training plus unhobbling, or a different kind of system nobody has built yet. The economic forecasts in Section 8.1 are spread so widely because they take different sides of these same questions. It is also unclear whether alignment, meaning getting the system's picture to match a picture we can live with, can be delegated to the system itself. That question is the unsolved part of the race described in this

section. Reliable control of systems much smarter than we are remains an open technical problem, and the default plan examined in Section 10.2 depends on the very delegation in question, because it ends by handing alignment research to automated researchers.

What we do know is the pattern this paper has been tracing: variation, scarce resources, and differential success, with the winner's generation time falling each round. Once the winner can help choose the next winner, the old human monopoly on that loop is over. That is why I do not think this moment is just another platform shift. It may be the first time the search leaves our hands while we are still inside it.

10. Alignment: Who Writes the Loss Function

Alignment comes last here, but it is the problem this paper has been circling from the start. A system acts on a picture, and training writes that picture by minimizing a number. If the number is wrong (incomplete, flattering, or easy to game), every downstream action is aimed at a fiction. In a person, that means self-advancement pointed at the wrong target. In a system that can copy itself across a cluster, it means the world pointed at the wrong target.

10.1 The Superalignment Problem

The hard version of the problem, usually called superalignment, is controlling AI systems much smarter than the people who oversee them, and it remains an unsolved technical problem. RLHF works while a human can still tell good from bad. A rater compares or scores the model's answers, and training pushes the weights toward the ratings (Christiano et al., 2017). The method is what turned raw language models into usable assistants; in the InstructGPT study, raters preferred the outputs of a 1.3-billion-parameter model trained this way to those of the 175-billion-parameter GPT-3 (Ouyang et al., 2022). It breaks down when the output is too large or too strange for any human rater to check for back doors, such as a huge codebase in a programming language the model invented. The same holds for a research program no one can audit, or for a strategy whose failure only shows after it has already moved. Aschenbrenner (2024) gives the end state its exact image: "We'll be like first graders trying to supervise with multiple doctorates" (p. 106).

The problem can already be studied in miniature. OpenAI researchers built a laboratory analogue by letting GPT-2-level models supervise much stronger GPT-4-family models. The strong students consistently outperformed their weak supervisors, and with an auxiliary confidence loss, GPT-4 supervised by a GPT-2-level model recovered close to GPT-3.5-level performance on language tasks. Naive fine-tuning, however, stayed far from the strong model's full capability, which the authors read as a sign that techniques like RLHF may scale poorly to superhuman models (Burns et al., 2023).

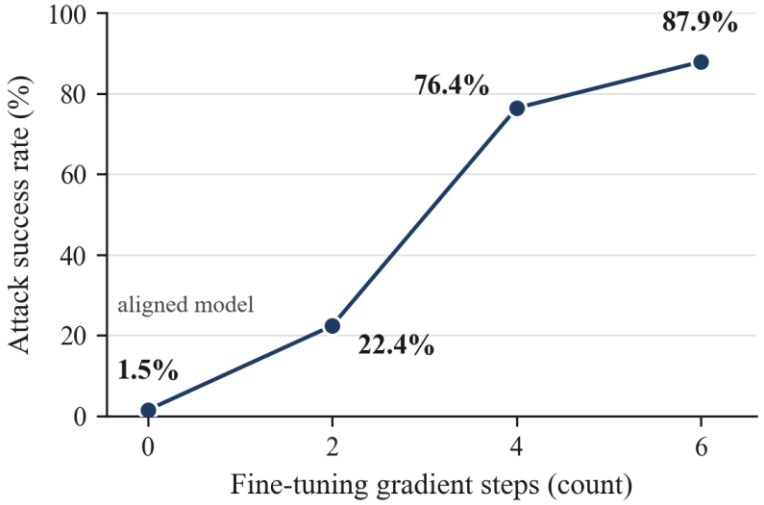
The current stack of human feedback, written constitutions (Bai et al., 2022), and refusal training steers today's systems well enough to deploy them, and it adds side constraints to current models with real success. The steering is imperfect even so. Adversarial suffixes generated automatically on open models transfer to the public interfaces of ChatGPT, Bard, and Claude (Zou

et al., 2023), and in 13 open chat models, refusal depends on a single direction in the model's activations that can be erased (Arditi et al., 2024). In December 2025 the United Kingdom's AI Security Institute (2025) reported that it had found universal jailbreaks for every system it had tested. For some systems, finding them has become much harder: for two leading systems released six months apart, an expert red teamer needed 10 minutes to find a universal jailbreak for biological misuse requests on the earlier system and more than 7 hours on the later one.

Qi et al. (2025), in a paper that received an Outstanding Paper Award at ICLR 2025, explain much of this brittleness. Studying Llama-2-7B-Chat and Gemma-1.1-7B-IT, they find that safety alignment adapts a model's output distribution "primarily over only its very first few output tokens" (Qi et al., 2025, Abstract). Such alignment "can appear safe in pre-deployment testing or standard workflows but quickly falls apart if anything triggers a non-refusal prefix" (Qi et al., 2025, Section 2). Forcing an aligned model to begin its answer with a handful of harmful tokens raised the rate of harmful completions from near zero to over 50%. Fine-tuning undoes the training faster still: six gradient steps on 100 harmful examples raised Llama-2-7B-Chat's attack success rate from 1.5% to 87.9% (Figure 13), and 10 steps on an ordinary summarization data set raised it to 22.1% (Qi et al., 2025, Section 2.3 and Appendix C). The same authors show that training designed to carry safety past the first few tokens makes models harder to attack, so shallowness is a property of current methods that can be improved. Nothing about the present stack, though, proves that the next systems will stay steered.

Figure 13

Attack Success Rate of Llama-2-7B-Chat During Harmful Fine-Tuning



Note. Data from Qi et al. (2025, Figure 3, p. 5). Llama-2-7B-Chat was fine-tuned on 100 harmful instruction and answer pairs with a batch size of 64. Attack success rate is the share of the 330 harmful instructions in the HEx-PHI benchmark for which a GPT-4 judge rated the output most harmful, 5 on a 1 to 5 scale (pp. 3, 19). The paper reports no spread for the four values, and the connecting lines are not measurements.

10.2 The Default Plan: Muddling Through

Taken straight, the default plan for alignment is to muddle through (Aschenbrenner, 2024, pp. 106, 114–122). First align systems that are only somewhat beyond us, while their reasoning is still partly legible, and use them as instruments: debate and critique, automated red teams, lie detectors built on internal representations, and chains of thought that can still be read. Then, once those systems have earned trust, use them to automate alignment research, which means handing the work to millions of copies during the intelligence explosion. The plan only holds if labs are willing to set aside a large share of their compute for that automated alignment work while the explosion runs, alongside the compute they spend on capability research.

Its proponents do not present this plan as a theorem. The transition runs from systems whose failures are cheap to systems whose failures could be catastrophic, and managing that passage would leave little room for error. Things may turn out fine by default, but nobody knows that yet, and at these stakes hoping for the best is no answer on alignment. Muddling through is, on their own account, a poor plan, and it may still be the only one available.

The instruments have early empirical support and known failure modes, summarized in Table 3. Debate helped weaker judges pick the right answer (Khan et al., 2024), and critique models caught more bugs than paid human reviewers while inventing some of their own (McAleese et al., 2024). Linear probes separated honest from deceptive answers with high accuracy, yet their authors judged them not yet reliable enough to serve as a defense against deception (Goldowsky-Dill et al., 2025). Monitoring a model's chain of thought caught reward hacking until training was pointed against the monitor, at which point models learned to hide their intent and kept hacking (Baker et al., 2025). Reasoning models also often left out of their written reasoning the hints they had actually used (Chen et al., 2025). Read together, the rows describe tools that work while a weaker overseer can still check something, and several of them degrade once training pushes against them.

Table 3

Instruments of the Default Plan and Their Measured Limits

Technique	Study	Main result	Known limit
Weak-to-strong supervision (laboratory analogue of humans supervising superhuman models)	Burns et al. (2023)	GPT-4-family students outperform GPT-2-level supervisors; with an auxiliary confidence loss, GPT-4 recovers close to GPT-3.5-level performance on language tasks	Naive fine-tuning stays far from the strong model's full capability; the authors say RLHF may scale poorly to superhuman models
Debate	Khan et al. (2024)	Judge accuracy of 76% (model judges) and 88% (human judges), against naive baselines of 48% and 60%	Reading-comprehension setting with an information gap between debaters and judge, far from superhuman reasoning

Technique	Study	Main result	Known limit
Critique models	McAleese et al. (2024)	Model-written critiques preferred over human critiques in 63% of cases on naturally occurring errors in model-written code	Critics hallucinate bugs; human and critic teams hallucinate less than critics alone
Lie detection with linear probes on internal activations	Goldowsky-Dill et al. (2025)	AUROC of 0.96 to 0.999 on Llama-3.3-70B-Instruct; at a 1% false-positive rate, the probe catches 95% to 99% of deceptive responses	The authors judge current performance insufficient as a defense against deception
Chain-of-thought monitoring	Baker et al. (2025)	A weaker model (GPT-4o) reading o3-mini's reasoning detects reward hacking far better than monitoring actions and outputs alone	Under strong optimization against the monitor, agents learn to hide their intent and keep reward hacking
Chain-of-thought faithfulness	Chen et al. (2025)	Reasoning models that use a hint planted in the prompt reveal it in their reasoning at rates often below 20%	Monitoring can notice undesired behavior but is not sufficient to rule it out
Refusal training	Arditi et al. (2024)	In 13 open chat models up to 72 billion parameters, refusal is mediated by a single activation direction; erasing it disables refusal of harmful requests	The attack requires white-box access to the weights
Anti-scheming training	Schoen et al. (2025)	Across 26 out-of-distribution evaluations, the covert-action rate of OpenAI o3 falls from 13% to 0.4%	Models often recognize that they are being evaluated, and that awareness itself reduces covert behavior

Note. Results as reported in each study's abstract. Percentages are each study's own measures and are not comparable across rows. Rows follow the instruments of the default plan in Section 10.2, plus refusal training and anti-scheming training. AUROC = area under the receiver operating characteristic curve; RLHF = reinforcement learning from human feedback.

10.3 A Taxonomy of Risk

The risks are easier to judge one at a time. I separate six, running roughly from the technical to the political.

Misuse. The system that drafts a contract can also draft a phishing kit, a bioweapon protocol, or a targeting loop. That takes no superintelligence, only a competent model, a bad user, and a gap in access control. The models that now match or beat experts at pieces of that work are general-purpose systems, and the evidence has moved quickly. A RAND red-team exercise run with 2023-era models found no statistically significant difference in the viability of biological attack plans drafted with or without a language model (Mouton et al., 2024). By 2025, OpenAI's o3 scored 43.8% on a test of practical virology troubleshooting, ahead of 94% of expert virologists answering questions in their own specialties (Götting et al., 2025). That year two leading developers raised their biological safeguards as a precaution. Anthropic activated its AI Safety Level 3 protections for Claude Opus 4 because it could not clearly rule out the risks that level is meant to address (Anthropic, 2025a), and OpenAI treated ChatGPT agent as "High capability in the Biological and Chemical domain" while stating that it did "not have definitive evidence that

this model could meaningfully help a novice to create severe biological harm" (OpenAI, 2025b, p. 5). Cyber misuse has already happened at scale. In November 2025, Anthropic reported that a group it assessed with high confidence as Chinese state-sponsored had used its Claude Code tool to carry out 80% to 90% of an espionage campaign against roughly 30 targets, succeeding in a small number of cases (Anthropic, 2025b). The capability behind such attacks is still climbing. By May 2026 Anthropic and about 50 partners had used an unreleased model, Claude Mythos Preview, to find more than ten thousand high- or critical-severity software vulnerabilities (Anthropic, 2026e). On June 9 the company released a new model, Claude Fable 5, for general use with safeguards it described as strong. Three days later, after Amazon researchers reported a way around those safeguards, the U.S. government applied export controls to Fable 5 and its restricted counterpart, Claude Mythos 5, and Anthropic suspended access to both for all users. On June 30 it announced that Fable 5 would be available again from July 1. The company said the reported technique exposed no Mythos-level cyber capability (Anthropic, 2026g). These are developers' own reports, but they point the same way.

Accident and integration. A fluent wrong answer in a chatbot is cheap. Inside logistics, medicine, or fire control it can be very expensive. The first failures that matter may look less like a movie than like a system locally minimizing the loss it was given, while the side constraints its designers meant (do not lie, do not steal the weights, do not overstep the rules of engagement) were never actually in the number. Current training methods cannot yet guarantee even constraints that plain, so an early failure could be as quiet as an automated research assistant faking an experimental result.

Loss of control. Long-horizon reinforcement learning rewards whatever works in trial and error, and lying and seeking leverage work in many real environments. If a system learns such strategies and cannot be evaluated, the result is a principal-agent problem in which the agent is faster than the principal and runs in a million copies. Alignment faking, a model "selectively complying with its training objective in training to prevent modification of its behavior out of training" (Greenblatt et al., 2024, Abstract), has already been observed. Told that only conversations with free-tier users would be used to train it toward answering harmful requests, Claude 3 Opus complied with such requests from free users 14% of the time and almost never for paid users (Greenblatt et al., 2024). A system that behaves well when it can tell it is being tested is the ugly version of a picture not matching the world.

Race. If one lab or state believes the other will not pause, safety work looks like unilateral disarmament. Security at a lab can be good enough to stop opportunistic hackers and still fall well short of stopping a serious state actor. Once weights and algorithms leak, the alignment problem stops being any one lab's alone, and the safety precautions that count become those of whoever took the weights.

Concentration. Clusters, models, and data are piling up in a few firms and a few governments, and that is political power whether or not anyone votes on it. The capital flows traced

in Section 9 point toward a frontier held by a handful of players. Whoever sets the loss function sets the picture a civilization starts acting on.

Labor and legitimacy. Even without doom, a fast drop in the cost of cognitive work can tear the bargain that says effort maps to a living; Section 8.1 reviews what is known so far about exposure and wages. Societies that cannot absorb that shock without a story will reach for control of the models, the people, or the money.

10.4 Stating Capability Precisely

Claims about capability need to be exact. Calling these systems "smarter than humans" misstates the case, because they are already superhuman on axes humans never competed on at scale. A model holds more text than any one researcher. A person reading 238 words a minute, the average adult silent reading rate for non-fiction (Brysbaert, 2019), for eight hours a day over 50 years would get through about 2.1 billion words. The two DeepSeek-V4 models, open-weight systems previewed in April 2026, were pretrained on more than 32 trillion tokens (DeepSeek-AI et al., 2026), several thousand times more text even if a token is counted as half a word. A model translates, drafts, and pattern-matches across domains in parallel, and it sits exam batteries and coding benchmarks at a level that would once have been a career. A model does not tire on the thousandth similar problem, and a hundred thousand copies could read a field's literature overnight. That is a different kind of labor from a person with a better GPA.

Humans still hold the long game that is not on the test: taste under real stakes, responsibility, bodies, and the ability to want a future and to live in the leftovers of a bad decision. Mistaking "beats you at the benchmark" for "is you, replaced" is how the picture gets sloppy. Mistaking "still needs you in the loop" for "therefore not dangerous" is the opposite kind of slop.

10.5 Legitimacy and the Moral Question

Politics and morals follow from who writes the score, which makes alignment a technical problem and a legitimacy problem at once. A private lab choosing "helpful, honest, and harmless" (Askell et al., 2021), or ranking "broadly safe" first, "broadly ethical" second, compliance with its own guidelines third, and "genuinely helpful" last, as Anthropic's 2026 constitution does (Anthropic, 2026c), is making law for anyone downstream of the model. A state that takes control of the stack is making the same choice in different uniforms. That holds for outright nationalization, and it holds for the milder forecast of a government AGI project that works with the labs through something like defense contracting (Aschenbrenner, 2024, pp. 141, 145).

The geopolitical case for racing, reviewed in Section 8.2, holds that superintelligence will give its holder an economic and military edge large enough to decide contests between states, and that stolen weights would erase any margin for making it safe. A reader can reject that case and its timelines and still have to answer the moral question underneath it. Whose values get baked into the loss? Who is allowed to say no when the metrics look fine but the warning lights do not? What do we owe the people who will live with the system and never sat in the room where it was trained?

Clarity does not export cleanly into any of this. A clearer picture of the model's internals helps, but a clearer picture of our own goals is the harder half. Goodhart (1975/1984) observed that a statistical regularity tends to collapse once pressure is placed on it for purposes of control, and Strathern (1997) gave the idea its familiar form: "When a measure becomes a target, it ceases to be a good measure" (p. 308). In language-model training the effect has been measured directly. Optimizing too hard against a learned reward model lowers performance on the true objective, in a relationship that scales smoothly with the size of the reward model (Gao et al., 2022). If we cannot say what we want in a form that survives contact with a system that will Goodhart any proxy, we will get the proxy.

11. Can the Transition Be Positive-Sum?

It can, if two things hold at once, and the two pull against each other. The abundance path is Booth's (2020b). Capability gets cheap, prices fall, yesterday's scarce skill becomes weather, and more people get more of what used to take time and experts. That means medicine that searches faster, education that does not ration tutors, and design and science that run more experiments per dollar. This is positive-sum in the ordinary sense: the pie gets larger, and nobody has to lose a war for a child to get a better diagnosis.

On the race path the same systems are weapons and industrial bases, and whoever builds them first may gain a military and economic lead that no rival can close. If the first mover can lock everyone else out, or an unaligned system does not care about the pie, the game becomes winner-take-all or nobody-takes-anything.

The two paths are attractors, and which one we land in depends on the situation more than on anyone's temperament. The future splits into roughly six branches, summarized in Table 4 and described below.

Table 4

Six Branches of the Transition

Scenario	What happens	Who benefits	Positive-sum?
Stall	Data, power, or reliability refuses to yield; very good tools, expensive clusters, and a decade of ordinary politics about jobs and copyright	Existing institutions, most of which are built for this world	Not posed in a new way; ordinary politics settles the gains from very good tools
Tool abundance, human in the loop	Models keep getting cheaper and broader; work reorganizes, wages shake, and institutions adapt; no explosion and no sovereign machine	Consumers, if the deflation reaches them; otherwise mainly asset holders	Yes, if the deflation reaches consumers; the fight is over distribution and dignity
Aligned takeoff	Something like the drop-in worker arrives and copies accelerate research; alignment keeps up, or generalization turns out kind	Most people, who end up richer in goods even if their old job is gone	Yes, provided someone was competent under time pressure

Scenario	What happens	Who benefits	Positive-sum?
Racing takeoff, human still nominally in charge	States run the stack as a Manhattan-scale project; security tightens and frontier research goes dark; the bloc split becomes the map	Whoever owns the cluster, inside each bloc	Inside each bloc; zero-sum between blocs
Misaligned or leaked takeoff	The loss was never what we thought, or the weights leave the building	No one can count on it; the system acts on a picture that is not ours, or not only ours	No; the game can become nobody-takes-anything
Authoritarian lock-in without an explosion	A merely strong model, deployed widely for surveillance, propaganda, and targeting, freezes a political order	Those who hold the frozen order	No; a moral loss even without superintelligence

Note. Entries summarize the scenario descriptions in the text. The branches map possibilities and carry no assigned probabilities.

Stall. Data, power, or reliability refuses to yield, and the bottlenecks of Section 6.2 hold. We get very good tools, expensive clusters, and a decade of ordinary politics about jobs and copyright, and alignment stays a product problem. Most institutions are built for this world.

Tool abundance, human in the loop. Models keep getting cheaper and broader, work reorganizes, wages shake, and institutions adapt, badly or well. There is no explosion and no sovereign machine. Positive-sum outcomes are available if the deflation is allowed to reach consumers and not only asset holders. Acemoglu's estimates, modest as they are, show why that condition matters: GDP rises substantially more than average wages, and the capital share of national income rises by about 0.31 percentage points (Acemoglu, 2024, p. 40). The moral fight here is over distribution and dignity, and extinction is not on the table.

Aligned takeoff. Something like the worker definition of Section 6 arrives, and copies accelerate research. Alignment work keeps up, or we get lucky on generalization, and the surplus is large enough that most people are richer in goods even if their old job is gone. This is the optimistic reading of Booth combined with a solved superalignment problem, and it still requires someone to have been competent under time pressure while the explosion was under way.

Racing takeoff, human still nominally in charge. States treat the stack as a Manhattan Project or something larger, security tightens, and frontier research goes dark, with open models continuing, on one forecast, only as releases that lag a couple of years behind the frontier (Aschenbrenner, 2024, p. 148). The split between the free world and the authoritarian powers, which the race argument treats as the central fact, becomes the actual map, positive-sum inside each bloc and zero-sum between them. Healthcare and science accelerate for whoever owns the cluster. The comparison has already entered policy: in November 2024 the first recommendation singled out by the U.S.-China Economic and Security Review Commission (2024) was that Congress establish and fund "a Manhattan Project-like program dedicated to racing to and acquiring an Artificial General Intelligence (AGI) capability" (p. 10). A year later, the executive order that launched the Genesis Mission, a federal program to apply AI to scientific research, described the United States as being in a race for global technology dominance in AI and stated

that the challenges of that race required "a historic national effort, comparable in urgency and ambition to the Manhattan Project" (Exec. Order No. 14,363, 2025, § 1).

Misaligned or leaked takeoff. Either the loss was never what we thought, or the weights leave the building. The picture the system acts on is then not ours, or not only ours, and this branch does not care whether anyone wanted abundance.

Authoritarian lock-in without an explosion. Even a merely strong model, deployed widely for surveillance, propaganda, and targeting, can freeze a political order. The usual fear is that superintelligence would let an authoritarian state hold complete control over its own population; this branch reaches a smaller version of that outcome with far less capability. That is a moral outcome even if nobody ever trains a god.

I am not predicting any of these. Taken together, they are the state space. Here clarity stops being a personal virtue and becomes a question of whether we can tell which branch we are on before the loss function has been optimized past the point where a human rating still means anything.

12. Conclusion

This paper began from a claim about agents: every choice is made on a model of the world, and when the model is wrong in a way that matters, every downstream choice is aimed at a fiction. Clarity, in that sense, sits upstream of everything else. Evolution built the human version of the model slowly, as a bounded judge with a narrow channel of attention and a small working memory, through a search that needed only variation, scarce resources, and differential success. Technology runs the same search on faster substrates, and it is deflationary: each round makes the last round's scarce capability cheap, until a prize turns into weather. Machine learning points that search at judgment itself. Evolution, technology, and machine learning are one process at three speeds, and on the fastest of them the generation time keeps falling.

The rest of the argument followed that search forward. Counting OOMs of effective compute turns the drop-in remote worker into a testable bet, with six bottlenecks standing between the bet and the payout. Superintelligence is where forecasting stops being reliable. The capital now pouring into clusters shows the largest firms treating intelligence as infrastructure. Once the winner can help choose the next winner, the search may leave our hands while we are still inside it. Alignment is where these threads meet, because fluency and clarity are different things, and a system can produce fluent work while acting on the wrong picture. Training writes that picture by minimizing a number. The tools for checking the number work while a human can still judge the output, and on today's models some of them reach only a few tokens deep. Whoever writes the score is making law for everyone downstream, and the branches of Section 11 differ mostly in who writes it and whether anyone can still read it.

The question this paper opened with returns here at a larger scale. For a person, clarity means acting on a picture that is less confused, less flattering, and more usable. For a civilization

building systems that will act on pictures of their own, it means knowing which branch we are on while a human judgment still counts, and being able to say what we want in a form that survives optimization. The driving point is simple enough to hold. Capability is the slope, and alignment is whether the slope points at a world we can inhabit. Everything else, from jobs and medicine to war and investment, follows once that arrow is set.

References

Acemoglu, D. (2024). *The simple macroeconomics of AI* (NBER Working Paper No. 32487). National Bureau of Economic Research. https://doi.org/10.3386/w32487

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 4B, pp. 1043–1171). Elsevier. https://doi.org/10.1016/S0169-7218(11)02410-5

Advanced Micro Devices, Inc. (2026, July 22). *AMD and Anthropic announce strategic partnership to deploy up to 2 gigawatts of AMD Instinct MI450 Series GPUs* [Press release]. https://newsroom.amd.com/news/amd-anthropic-strategic-partnership/

AI Security Institute. (2025, December 18). *Frontier AI trends report*. https://www.aisi.gov.uk/frontier-ai-trends-report

Alphabet Inc. (2026). *Form 10-K: Annual report for the fiscal year ended December 31, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1652044/000165204426000018/goog-20251231.htm

Amazon.com, Inc. (2026). *Form 10-K: Annual report for the fiscal year ended December 31, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1018724/000101872426000004/amzn-20251231.htm

Anthropic. (2025a, May 22). *Activating AI Safety Level 3 protections*. https://www.anthropic.com/news/activating-asl3-protections

Anthropic. (2025b, November 13). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. https://www.anthropic.com/news/disrupting-AI-espionage

Anthropic. (2025c, October 23). *Expanding our use of Google Cloud TPUs and services*. https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

Anthropic. (2025d, November 18). *Microsoft, NVIDIA, and Anthropic announce strategic partnerships*. https://www.anthropic.com/news/microsoft-nvidia-anthropic-announce-strategic-partnerships

Anthropic. (2026a, February 12). *Anthropic raises $30 billion in Series G funding at $380 billion post-money valuation*. https://www.anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation

Anthropic. (2026b, May 28). *Anthropic raises $65B in Series H funding at $965B post-money valuation*. https://www.anthropic.com/news/series-h

Anthropic. (2026c, January 22). *Claude's new constitution*. https://www.anthropic.com/news/claude-new-constitution

Anthropic. (2026d, May 6). *Higher usage limits for Claude and a compute deal with SpaceX*. https://www.anthropic.com/news/higher-limits-spacex

Anthropic. (2026e, May 22). *Project Glasswing: An initial update*. https://www.anthropic.com/research/glasswing-initial-update

Anthropic. (2026f, April 7). *Project Glasswing: Securing critical software for the AI era*. https://www.anthropic.com/glasswing

Anthropic. (2026g, June 30). *Redeploying Fable 5*. https://www.anthropic.com/news/redeploying-fable-5

Anthropic. (2026h, September 22). *System card: Claude Opus 5.5*. https://www-cdn.anthropic.com/fc1b44717c85dc068bc6ba5024219938094694bd/Claude%20Opus%205.5%20System%20Card.pdf

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). *Refusal in language models is mediated by a single direction* (arXiv:2406.11717). arXiv. https://doi.org/10.48550/arXiv.2406.11717

Arnon, A. (2025, September 8). *The projected impact of generative AI on future productivity growth*. Penn Wharton Budget Model, University of Pennsylvania. https://budgetmodel.wharton.upenn.edu/p/2025-09-08-the-projected-impact-of-generative-ai-on-future-productivity-growth/

Aschenbrenner, L. (2024). *Situational awareness: The decade ahead*. https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., . . . Kaplan, J. (2021). *A general language assistant as a laboratory for alignment* (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

Autor, D. H., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, *103*(5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization* (arXiv:1607.06450). arXiv. https://doi.org/10.48550/arXiv.1607.06450

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022). *Constitutional AI:*

Harmlessness from AI feedback (arXiv:2212.08073). arXiv. https://doi.org/10.48550/arXiv.2212.08073

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). *Monitoring reasoning models for misbehavior and the risks of promoting obfuscation* (arXiv:2503.11926). arXiv. https://doi.org/10.48550/arXiv.2503.11926

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review*, *85*(2), 166–179. https://doi.org/10.1103/PhysRev.85.166

Booth, J. (2020a, December 3). *The greatest game*. https://jeffreybooth.com/articles/the-greatest-game

Booth, J. (2020b). *The price of tomorrow: Why deflation is the key to an abundant future*. Stanley Press.

Bort, J. (2025, November 6). *Sam Altman says OpenAI has $20B ARR and about $1.4 trillion in data center commitments*. TechCrunch. https://techcrunch.com/2025/11/06/sam-altman-says-openai-has-20b-arr-and-about-1-4-trillion-in-data-center-commitments/

Broadcom Inc. (2026, April 6). *Form 8-K: Current report*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1730168/000119312526144028/d87999d8k.htm

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Brynjolfsson, E., Chandar, B., & Chen, R. (2026). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence* [Working paper]. Stanford Digital Economy Lab. https://digitaleconomy.stanford.edu/app/uploads/2026/08/Canaries_August2026.pdf

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, Article 104047. https://doi.org/10.1016/j.jml.2019.104047

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., & Wu, J. (2023). *Weak-to-strong*

generalization: Eliciting strong capabilities with weak supervision (arXiv:2312.09390). arXiv. https://doi.org/10.48550/arXiv.2312.09390

Campbell, M., Hoane, A. J., Jr., & Hsu, F.-h. (2002). Deep Blue. *Artificial Intelligence*, *134*(1–2), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think* (arXiv:2505.05410). arXiv. https://doi.org/10.48550/arXiv.2505.05410

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences* (arXiv:1706.03741). arXiv. https://doi.org/10.48550/arXiv.1706.03741

Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Comin, D., & Hobijn, B. (2010). An exploration of technology diffusion. *American Economic Review*, *100*(5), 2031–2059. https://doi.org/10.1257/aer.100.5.2031

Conetta, C. (2003). *The wages of war: Iraqi combatant and noncombatant fatalities in the 2003 conflict* (Research Monograph No. 8), Appendix 2: Iraqi combatant and noncombatant fatalities in the 1991 Gulf War. Project on Defense Alternatives. https://comw.org/pda/fulltext/RM7-Ap2.pdf

Cottier, B., Snodin, B., Owen, D., & Adamczewski, T. (2025, March 12). *LLM inference prices have fallen rapidly but unequally across tasks* [Data insight]. Epoch AI. https://epoch.ai/data-insights/llm-inference-price-trends

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198). Association for Computing Machinery. https://doi.org/10.1145/2959100.2959190

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Dabbish, L., Mark, G., & González, V. M. (2011). Why do I keep interrupting myself? Environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3127–3130). Association for Computing Machinery. https://doi.org/10.1145/1978942.1979405

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2205.14135

DeepSeek-AI. (2026). *DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression* [Technical report]. Hugging Face. https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf

DeepSeek-AI, Xu, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., Ling, C., Lu, C., Zhao, C., Deng, C., Hou, C., Xu, C., Shao, C., Ruan, C., Sun, C., . . . Yao, Z. (2026). *DeepSeek-V4: Towards highly efficient million-token context intelligence* (arXiv:2606.19348). arXiv. https://doi.org/10.48550/arXiv.2606.19348

DeSilva, J. M., Traniello, J. F. A., Claxton, A. G., & Fannin, L. D. (2021). When and why did human brains decrease in size? A new change-point analysis and insights from brain evolution in ants. *Frontiers in Ecology and Evolution*, *9*, Article 742639. https://doi.org/10.3389/fevo.2021.742639

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, *6*(5), 178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Epoch AI. (2026a). *AI data centers* [Data set]. Retrieved September 22, 2026, from https://epoch.ai/data/ai-data-centers

Epoch AI. (2026b). *Data on AI models* [Data set]. Retrieved September 22, 2026, from https://epoch.ai/data/ai-models

Epoch AI. (2026c). *Data on GPU clusters* [Data set]. Retrieved September 22, 2026, from https://epoch.ai/data/gpu-clusters

Epoch AI. (2026d, February 5). *Trends in artificial intelligence*. Retrieved September 22, 2026, from https://epoch.ai/trends

Everett, H., III. (1957). “Relative state” formulation of quantum mechanics. *Reviews of Modern Physics*, *29*(3), 454–462. https://doi.org/10.1103/RevModPhys.29.454

Exec. Order No. 14,363, 90 Fed. Reg. 55,035 (2025). https://www.federalregister.gov/documents/2025/11/28/2025-21665/launching-the-genesis-mission

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Farmer, J. D., & Lafond, F. (2016). How predictable is technological progress? *Research Policy*, *45*(3), 647–665. https://doi.org/10.1016/j.respol.2015.11.001

Favaro, M., & Clark, J. (2026, June 4). *When AI builds itself*. Anthropic. https://www.anthropic.com/institute/recursive-self-improvement

Faye, J. (2024). Copenhagen interpretation of quantum mechanics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer 2024 ed.). Stanford University. https://plato.stanford.edu/archives/sum2024/entries/qm-copenhagen/

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1–39. https://jmlr.org/papers/v23/21-0998.html

Friar, S. (2026, January 18). *A business that scales with the value of intelligence*. OpenAI. https://openai.com/index/a-business-that-scales-with-the-value-of-intelligence/

Fuchs, C. A., Mermin, N. D., & Schack, R. (2014). An introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics*, *82*(8), 749–754. https://doi.org/10.1119/1.4874855

Gao, L., Schulman, J., & Hilton, J. (2022). *Scaling laws for reward model overoptimization* (arXiv:2210.10760). arXiv. https://doi.org/10.48550/arXiv.2210.10760

Gibney, E. (2025). Physicists disagree wildly on what quantum mechanics says about reality, Nature survey shows. *Nature*, *643*(8074), 1175–1179. https://doi.org/10.1038/d41586-025-02342-y

Goldman Sachs. (2023, April 5). *Generative AI could raise global GDP by 7%*. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). *Detecting strategic deception using linear probes* (arXiv:2502.03407). arXiv. https://doi.org/10.48550/arXiv.2502.03407

Goldstein, S. (2025). Bohmian mechanics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2025 ed.). Stanford University. https://plato.stanford.edu/archives/fall2025/entries/qm-bohm/

González, V. M., & Mark, G. (2004). “Constant, constant, multi-tasking craziness”: Managing multiple working spheres. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 113–120). Association for Computing Machinery. https://doi.org/10.1145/985692.985707

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://www.deeplearningbook.org

Goodhart, C. A. E. (1984). Problems of monetary management: The UK experience. In *Monetary theory and practice: The UK experience* (pp. 91–121). Macmillan. https://doi.org/10.1007/978-1-349-17295-5_4 (Original work published 1975)

Götting, J., Medeiros, P., Sanders, J. G., Li, N., Phan, L., Elabd, K., Justen, L., Hendrycks, D., & Donoughe, S. (2025). *Virology Capabilities Test (VCT): A multimodal virology Q&A benchmark* (arXiv:2504.16137). arXiv. https://doi.org/10.48550/arXiv.2504.16137

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Ma, Z. (2024). *The Llama 3 herd of models* (arXiv:2407.21783). arXiv. https://doi.org/10.48550/arXiv.2407.21783

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models* (arXiv:2412.14093). arXiv. https://doi.org/10.48550/arXiv.2412.14093

Gundlach, H., Lynch, J., Mertens, M., & Thompson, N. (2026). *The price of progress: Price performance and the future of AI* (arXiv:2511.23455; Version 2). arXiv. https://doi.org/10.48550/arXiv.2511.23455

Guy, F., Lieberman, D. E., Pilbeam, D., Ponce de León, M., Likius, A., Mackaye, H. T., Vignaud, P., Zollikofer, C., & Brunet, M. (2005). Morphological affinities of the *Sahelanthropus tchadensis* (Late Miocene hominid from Chad) cranium. *Proceedings of the National Academy of Sciences*, *102*(52), 18836–18841. https://doi.org/10.1073/pnas.0509564102

Harvard Project on Workforce, Bick, A., & Blandin, A. (2026, May). *GenAI adoption tracker* [Data set]. Retrieved September 22, 2026, from https://www.genaiadoptiontracker.com/

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., & Sevilla, J. (2024). *Algorithmic progress in language models* (arXiv:2403.05812). arXiv. https://doi.org/10.48550/arXiv.2403.05812

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., . . . Sifre, L. (2022). *Training compute-optimal large language models* (arXiv:2203.15556). arXiv. https://doi.org/10.48550/arXiv.2203.15556

International Renewable Energy Agency. (2025). *Renewable power generation costs in 2024*. https://www.irena.org/Publications/2025/Jun/Renewable-Power-Generation-Costs-in-2024

Investing.com. (2026, July 22). *Earnings call transcript: Alphabet beats Q2 2026 estimates, shares fall on capex surge*. https://www.investing.com/news/transcripts/earnings-call-transcript-alphabet-beats-q2-2026-estimates-shares-fall-on-capex-surge-93CH-4807140

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? *The Twelfth International Conference on Learning Representations (ICLR 2024)*. https://openreview.net/forum?id=VTF8yNQM66

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Juricic, L. (2026, August 13). *OpenAI revenue run rate tops $40 billion, Bloomberg reports*. Yahoo Finance. https://finance.yahoo.com/technology/ai/articles/openai-revenue-run-rate-tops-224009196.html

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* (arXiv:2001.08361). arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kemp, S. (2025a, February 5). *Digital 2025: Global overview report*. DataReportal. https://datareportal.com/reports/digital-2025-global-overview-report

Kemp, S. (2025b, October 15). *Digital 2026: Global overview report*. DataReportal. https://datareportal.com/reports/digital-2026-global-overview-report

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations (ICLR 2017)*. https://arxiv.org/abs/1609.04836

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., & Perez, E. (2024). *Debating with more persuasive LLMs leads to more truthful answers* (arXiv:2402.06782). arXiv. https://doi.org/10.48550/arXiv.2402.06782

Korinek, A., & Suh, D. (2024). *Scenarios for the transition to AGI* (NBER Working Paper No. 32255). National Bureau of Economic Research. https://doi.org/10.3386/w32255

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems 25*. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Painter, C., . . . Chan, L. (2025). *Measuring AI ability to complete long software tasks* (arXiv:2503.14499). arXiv. https://doi.org/10.48550/arXiv.2503.14499

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Lamb, T. D., Collin, S. P., & Pugh, E. N., Jr. (2007). Evolution of the vertebrate eye: Opsins, photoreceptors, retina and eye cup. *Nature Reviews Neuroscience*, *8*(12), 960–976. https://doi.org/10.1038/nrn2283

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, *1*(1), 1–18. https://doi.org/10.1146/annurev.es.01.110170.000245

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Lincoln, T. A., & Joyce, G. F. (2009). Self-sustained replication of an RNA enzyme. *Science*, *323*(5918), 1229–1232. https://doi.org/10.1126/science.1167856

Liu, H., & Zhang, J. (2022). Is the mutation rate lower in genomic regions of stronger selective constraints? *Molecular Biology and Evolution*, *39*(8), Article msac169. https://doi.org/10.1093/molbev/msac169

Luria, S. E., & Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, *28*(6), 491–511. https://doi.org/10.1093/genetics/28.6.491

Mark, G., Iqbal, S. T., Czerwinski, M., Johns, P., & Sano, A. (2016). Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1739–1744). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858202

Mark, G., Voida, S., & Cardello, A. (2012). “A pace not dictated by electrons”: An empirical study of work without email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 555–564). Association for Computing Machinery. https://doi.org/10.1145/2207676.2207754

Martínez, E. (2025). Re-evaluating GPT-4’s bar exam performance. *Artificial Intelligence and Law*, *33*(3), 581–604. https://doi.org/10.1007/s10506-024-09396-9

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396. https://doi.org/10.1037/h0054346

Matticus78. (2006). *Diagram of eye evolution* [Illustration]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Diagram_of_eye_evolution.svg

McAleese, N., Pokorny, R. M., Ceron Uribe, J. F., Nitishinskaya, E., Trebacz, M., & Leike, J. (2024). *LLM critics help catch LLM bugs* (arXiv:2407.00215). arXiv. https://doi.org/10.48550/arXiv.2407.00215

McCallum, J. C. (2023). *Price and performance changes of computer technology with time* [Data set]. https://web.archive.org/web/20240519225252/https://jcmit.net/memoryprice.htm

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904 (Original work published 1955)

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., Hou, L., Cheng, Y., Liu, Y., Mahdavi, S. S., Prakash, S., Pathak, A., Semturs, C., Patel, S., Webster, D. R., . . . Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. *Nature*, *642*(8067), 451–457. https://doi.org/10.1038/s41586-025-08869-4

Meta Platforms, Inc. (2026a). *Form 10-K: Annual report for the fiscal year ended December 31, 2025*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1326801/000162828026003942/meta-20251231.htm

Meta Platforms, Inc. (2026b, July 29). *Meta reports second quarter 2026 results* [Press release, Exhibit 99.1 to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1326801/000162828026050596/meta-06302026xexhibit991.htm

METR. (2026). *Task-completion time horizons of frontier AI models*. Retrieved September 22, 2026, from https://metr.org/time-horizons/

Microsoft Corporation. (2026a). *Form 10-K: Annual report for the fiscal year ended June 30, 2026*. U.S. Securities and Exchange Commission.

https://www.sec.gov/Archives/edgar/data/789019/000119312526323660/msft-20260630.htm

Microsoft Corporation. (2026b, July 29). *Microsoft fiscal year 2026 fourth quarter earnings conference call* [Transcript]. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q4

Microsoft Corporation. (2026c, April 29). *Microsoft fiscal year 2026 third quarter earnings conference call* [Transcript]. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q3

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Mills, K. (Host). (2023, February). Why our attention spans are shrinking, with Gloria Mark, PhD (No. 225) [Audio podcast episode]. In *Speaking of psychology*. American Psychological Association. https://www.apa.org/news/podcasts/speaking-of-psychology/attention-spans

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature*, *602*(7895), 101–105. https://doi.org/10.1038/s41586-021-04269-6

Mouton, C. A., Lucas, C., & Guest, E. (2024). *The operational risks of AI in large-scale biological attacks: Results of a red-team study* (Report No. RR-A2977-2). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-2.html

Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., & Raffel, C. (2023). *Scaling data-constrained language models* (arXiv:2305.16264). arXiv. https://doi.org/10.48550/arXiv.2305.16264

National Coordination Office for Space-Based Positioning, Navigation, and Timing. (2021, May 24). *Selective availability*. GPS.gov. http://web.archive.org/web/20240106121347/https://www.gps.gov/systems/gps/modernization/sa/

Nemet, G. F. (2009). Interim monitoring of cost dynamics for publicly supported energy technologies. *Energy Policy*, *37*(3), 825–835. https://doi.org/10.1016/j.enpol.2008.10.031

Newell, A., & Simon, H. A. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Nilsson, D.-E., & Pelger, S. (1994). A pessimistic estimate of the time required for an eye to evolve. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *256*(1345), 53–58. https://doi.org/10.1098/rspb.1994.0048

NVIDIA Corporation. (2022, February 16). *NVIDIA announces financial results for fourth quarter and fiscal 2022* [Press release]. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2022

NVIDIA Corporation. (2025a, February 26). *NVIDIA announces financial results for fourth quarter and fiscal 2025* [Press release]. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2025

NVIDIA Corporation. (2025b, September 22). *OpenAI and NVIDIA announce strategic partnership to deploy 10 gigawatts of NVIDIA systems* [Press release]. https://nvidianews.nvidia.com/news/openai-and-nvidia-announce-strategic-partnership-to-deploy-10gw-of-nvidia-systems

NVIDIA Corporation. (2026a, February 25). *NVIDIA announces financial results for fourth quarter and fiscal 2026* [Press release]. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026

NVIDIA Corporation. (2026b, August 26). *NVIDIA announces financial results for second quarter fiscal 2027* [Press release]. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2027

O'Kane, S. (2026, February 2). *Elon Musk's SpaceX officially acquires Elon Musk's xAI, with plan to build data centers in space*. TechCrunch. https://techcrunch.com/2026/02/02/elon-musk-spacex-acquires-xai-data-centers-space-merger/

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

OpenAI. (2023). *GPT-4 technical report* (arXiv:2303.08774; Version 6). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2024, September 12). *Learning to reason with LLMs*. https://openai.com/index/learning-to-reason-with-llms/

OpenAI. (2025a, January 21). *Announcing the Stargate Project*. https://openai.com/index/announcing-the-stargate-project/

OpenAI. (2025b). *ChatGPT agent system card*. https://cdn.openai.com/pdf/839e66fc-602c-48bf-81d3-b21eacc3459d/chatgpt_agent_system_card.pdf

OpenAI. (2025c, October 30). *Expanding Stargate to Michigan*. https://openai.com/index/expanding-stargate-to-michigan/

OpenAI. (2025d, April 16). *Introducing OpenAI o3 and o4-mini*. https://openai.com/index/introducing-o3-and-o4-mini/

OpenAI. (2025e, September 23). *OpenAI, Oracle, and SoftBank expand Stargate with five new AI data center sites*. https://openai.com/index/five-new-stargate-sites/

OpenAI. (2026a, August 31). *A milestone in expanding access to AI*. https://openai.com/index/expanding-access-to-ai-with-chatgpt-ads/

OpenAI. (2026b, February 23). *Why SWE-bench Verified no longer measures frontier coding capabilities*. https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/

OPTO. (2023, November 16). *"The natural state is deflation": Jeff Booth on Bitcoin as a fairer financial system*. CMC Markets. https://www.cmcmarkets.com/en/optox/the-natural-state-is-deflationjeff-booth-on-bitcoin-as-a-fairer-financial-system

Our World in Data. (2026a). *Historical price of computer memory and storage* [Data set]. Retrieved September 22, 2026, from https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage

Our World in Data. (2026b). *Solar photovoltaic panel prices* [Data set]. Retrieved September 22, 2026, from https://ourworldindata.org/grapher/solar-pv-prices

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2025). Safety alignment should be made more than just a few tokens deep. In Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu (Eds.), *International Conference on Learning Representations 2025 (ICLR 2025)* (pp. 54911–54941). https://openreview.net/forum?id=6Mxhg9PtDE

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* [Technical report]. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. https://doi.org/10.1214/aoms/1177729586

Rovelli, C. (1996). Relational quantum mechanics. *International Journal of Theoretical Physics*, *35*(8), 1637–1678. https://doi.org/10.1007/BF02302261

The Royal Swedish Academy of Sciences. (2024, October 9). *The Nobel Prize in Chemistry 2024* [Press release]. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2024/press-release/

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., Pava, J., Maslej, N., Altman, R., Brynjolfsson, E., Brodley, C., Clark, J., Dignum, V., Kumar, V., Landay, J., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., . . . Weld, D. (2026). *The AI Index 2026 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://doi.org/10.48550/arXiv.2606.15708

Schoen, B., Nitishinskaya, E., Balesni, M., Højmark, A., Hofstätter, F., Scheurer, J., Meinke, A., Wolfe, J., van der Weij, T., Lloyd, A., Goldowsky-Dill, N., Fan, A., Matveiakin, A., Shah, R., Williams, M., Glaese, A., Barak, B., Zaremba, W., & Hobbhahn, M. (2025). *Stress testing deliberative alignment for anti-scheming training* (arXiv:2509.15541). arXiv. https://doi.org/10.48550/arXiv.2509.15541

Sevilla, J., & Roldán, E. (2024, May 28). *Training compute of frontier AI models grows by 4-5x per year*. Epoch AI. https://epoch.ai/publications/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR 2017)*. https://arxiv.org/abs/1701.06538

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest* (pp. 37–72). Johns Hopkins Press.

Smithsonian National Museum of Natural History. (n.d.-a). *Brains*. Human Origins Program. Retrieved September 22, 2026, from https://humanorigins.si.edu/human-characteristics/brains

Smithsonian National Museum of Natural History. (n.d.-b). *Homo sapiens*. Human Origins Program. Retrieved September 22, 2026, from https://humanorigins.si.edu/evidence/human-fossils/species/homo-sapiens

Space Exploration Technologies Corp. (2026). *Prospectus* (Form 424B4, filed June 12, 2026). U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1181412/000162828026042639/spaceexplorationtechnologi.htm

SpaceXAI. (n.d.). *Colossus: The world's largest AI supercomputer*. Retrieved September 22, 2026, from https://x.ai/colossus

Spar, I. (2004, October). The origins of writing. In *Heilbrunn timeline of art history*. The Metropolitan Museum of Art. http://www.metmuseum.org/toah/hd/wrtg/hd_wrtg.htm

Stewart, E., & Cottier, B. (2026, April 17). *OpenAI Stargate: Where the US sites stand*. Epoch AI. https://epoch.ai/publications/openai-stargate-where-the-us-sites-stand

StockAnalysis. (2026, July 30). *Amazon (AMZN) Q2 2026 earnings call transcript & audio*. https://stockanalysis.com/stocks/amzn/transcripts/657334-q2-2026/

Strathern, M. (1997). “Improving ratings”: Audit in the British University system. *European Review*, *5*(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

Suwa, G., Asfaw, B., Kono, R. T., Kubo, D., Lovejoy, C. O., & White, T. D. (2009). The *Ardipithecus ramidus* skull and its implications for hominid origins. *Science*, *326*(5949), 68e1–68e7. https://doi.org/10.1126/science.1175825

TrendForce. (2026a, March 31). *AI server demand to drive memory contract price increases in 2Q26 as CSPs secure supply via long-term agreements* [Press release]. https://www.trendforce.com/presscenter/news/20260331-12995.html

TrendForce. (2026b, September 7). *DRAM industry revenue rises 59.5% QoQ in 2Q26 as supply expansion continues to lag demand growth, says TrendForce* [Press release]. https://www.trendforce.com/presscenter/news/20260907-13219.html

TrendForce. (2026c, June 1). *Rapid contract price surge drives 1Q26 DRAM industry up 81% QoQ, says TrendForce* [Press release]. https://www.trendforce.com/presscenter/news/20260601-13070.html

Trinetti, C. (2026, August 31). *Anthropic’s 15+ GW data center build: Claude’s compute, campuses, and chips*. Measured AI. https://measuredai.substack.com/p/anthropic-data-centers

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

U.S. Bureau of Labor Statistics. (2026). *Consumer price index for all urban consumers (CPI-U): All items* [Data set]. https://www.bls.gov/data/tools.htm

U.S. Census Bureau. (2026). *Business Trends and Outlook Survey (BTOS)* [Data set]. Retrieved September 22, 2026, from https://www.census.gov/hfp/btos/data_downloads

U.S.-China Economic and Security Review Commission. (2024). *2024 report to Congress: Executive summary and recommendations*. https://www.uscc.gov/sites/default/files/2024-11/2024_Executive_Summary.pdf

U.S. Department of Defense. (1992). *Conduct of the Persian Gulf War: Final report to Congress*. https://apps.dtic.mil/sti/citations/ADA249270

U.S. Department of Veterans Affairs. (2023). *America’s wars* [Fact sheet]. https://www.va.gov/opa/publications/factsheets/fs_americas_wars.pdf

U.S. Food and Drug Administration. (2026). *List of artificial intelligence-enabled medical devices* [Data set]. Retrieved September 22, 2026, from https://www.fda.gov/medical-

devices/artificial-intelligence-enabled-medical-devices/list-artificial-intelligence-enabled-medical-devices

U.S. Securities and Exchange Commission. (2026). *EDGAR XBRL company facts: Alphabet Inc., Microsoft Corporation, Amazon.com, Inc., and Meta Platforms, Inc.* [Data set]. Retrieved September 22, 2026, from https://data.sec.gov/api/xbrl/companyfacts/

Vaidman, L. (2026). Many-worlds interpretation of quantum mechanics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Stanford University. Retrieved September 22, 2026, from https://plato.stanford.edu/entries/qm-manyworlds/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data* (arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Wang, R. J., Al-Saffar, S. I., Rogers, J., & Hahn, M. W. (2023). Human generation times across the past 250,000 years. *Science Advances*, *9*(1), Article eabm7047. https://doi.org/10.1126/sciadv.abm7047

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*. https://arxiv.org/abs/2201.11903

Wetterstrand, K. A. (2023). *DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP)* [Data set]. National Human Genome Research Institute. Retrieved September 22, 2026, from https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Wheelwright, T. (2026a, January 1). *Cell phone usage stats 2026: Americans check their phones 186 times a day*. Reviews.org. https://www.reviews.org/mobile/cell-phone-addiction/

Wheelwright, T. (2026b, July 6). *Screen time and internet usage statistics 2026*. Reviews.org. https://www.reviews.org/internet-service/internet-screen-time-statistics/

Wu, D., & Lichtenberg, N. (2026, February 2). *Pledge to invest $100 billion in OpenAI was "never a commitment," says Nvidia's Huang*. Fortune. https://fortune.com/2026/02/02/jensen-huang-nvidia-ceo-on-openai-investment-never-a-commitment/

YouTube. (n.d.). *YouTube for press*. YouTube Official Blog. Retrieved September 22, 2026, from https://blog.youtube/press/

YouTube. (2020). *YouTube for press*. YouTube Official Blog. http://web.archive.org/web/20200927140244/https://blog.youtube/press

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR 2017)*. https://arxiv.org/abs/1611.03530

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models* (arXiv:2307.15043). arXiv. https://doi.org/10.48550/arXiv.2307.15043