CHAPTER 1

Capabilities

Markov Grey
French Center for AI Safety (CeSIA)

Charbel-Raphaël Segerie
French Center for AI Safety (CeSIA)

Contents

1. Introduction

AI has undergone a massive shift in the last decade. Game-playing systems developed strategic creativity that surprised grandmasters, language models evolving from funny autocomplete failures to software engineering assistants handling codebases spanning thousands of files. These types of examples reflect a change in how technology is evolving, and raise a question that this entire book tries to answer: how do you make something safe when you don't fully understand what it can do, and when its abilities might change dramatically between the time you start reading this chapter and the time you finish?

This first chapter is about mapping the territory. Before we can discuss dangerous capabilities, alignment strategies, technical solutions or governance frameworks—the subjects of Chapters 2, 3, and 4—the first step is getting a sense of what AI systems can actually do right now, how they got here, and where the future trends point.

The first question to answer is: what counts as "general" intelligence? The systems we are seeing now represent a move from narrow AI—systems that do one thing well—to general purpose AI systems (foundation models) that learn general patterns, adapt to new situations and do many things well. The term AGI gets thrown around constantly, but different people mean wildly different things by it. We'll walk you through a couple of different approaches to defining intelligence, going from the Turing Test to psychometric frameworks, before settling on a practical usable definition: capability and generality as continuous axes rather than binary categories. The goal is to be able to make specific concrete statements like - " *this system performs at the 85th percentile across 30% of cognitive domains as measured by ...* "—concrete enough to measure, track, and build policy around.

The second question everyone asks: why is this happening so fast recently? The answer is the "bitter lesson" of AI research— massive computation consistently beats human-engineered knowledge. We'll look at scaling laws—the empirical relationships showing that more data and compute predictably produce better performance—and various sides of the debate about whether scale alone is sufficient (and sustainable) for transformative AI, or whether we'll need some other algorithmic breakthroughs.

The last part of the chapter will look ahead at the future. Predicting when AI might automate all cognitive labor determines which safety strategies are even viable. If transformative AI arrives by 2030, we need solutions that work with current systems and scale quickly. If it's 2050, we have time for fundamental research. This means looking at the different scenarios for "takeoff"—whether progress will be gradual or explosive, continuous or discontinuous—and what that means for our ability to respond.

The specific examples and benchmark scores in this chapter will be outdated soon, but the underlying patterns will remain. Scaling laws, emergent capabilities, the shift from narrow to general systems—these trends are stable enough for you to learn about, even as individual benchmarks become obsolete. By the end of this chapter, you should have a concrete framework for how to define AGI, and much clearer perspective on the debates shaping the field of AI safety.

Let's start with what these systems can actually do—and how quickly that list is growing.

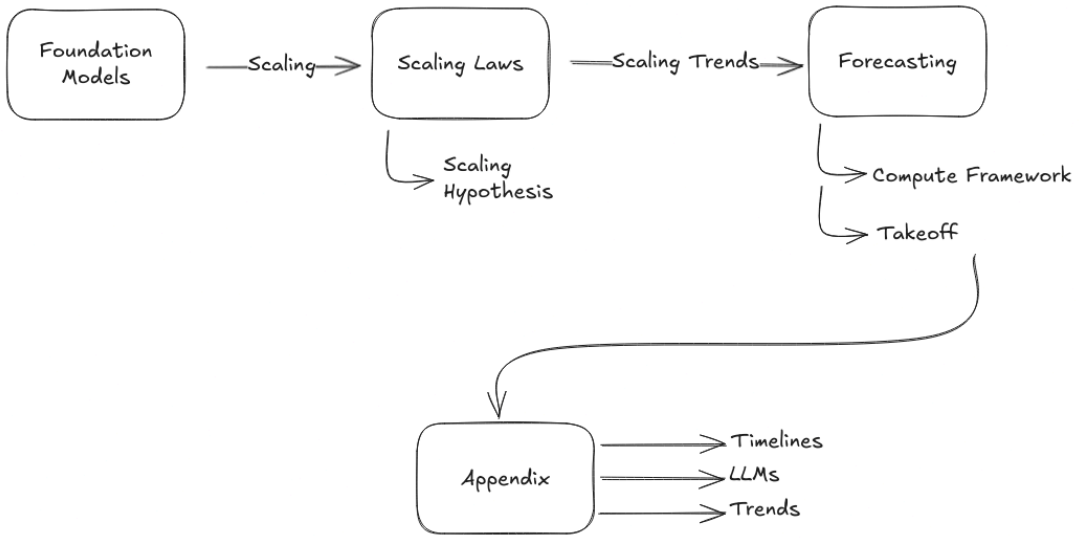


Figure 1: *We first explain foundation models, which have been continuously showing improved capabilities due to scale. Then examine empirically observed scaling laws. Based on these trends we look at some techniques that researchers use to try and forecast future AI progress.*

2. Current Capabilities

AI models can write, reason, code, generate media, and control robots—often matching or surpassing expert human performance on specific tasks. In this first section, we'll try to give you a sense of what AI is actually capable of doing as of late 2025. Numbers and graphs don't really make these capabilities tangible, so we will try to include as many videos, images, and examples to really help you get a sense of where AI currently stands. But in case you are interested, we also mention benchmark scores that measure progress quantitatively.

The trajectory matters more than a snapshot. This section can serve both as a quick history and as a snapshot of current capabilities. As you read through it, try to keep in mind how quickly we got here. Language generation went from coherent paragraphs to research assistants in a few years. Image generation went from laughable to professional-grade in a decade. Video generation seems to be following a similar path compressed into three years.

If you're already familiar with AI or machine learning , some of these stories might be review for you, but hopefully everyone who reads will be able to take away at least one new thing from this section.

2.1 Games

Game-playing AI is already at the superhuman level for many games. Comparing AI and humans at games has been a common theme through the last few decades with AI making continuous progress. IBM's Deep blue defeated chess grandmaster in 1997 (IBM, 2026), IBM's Watson won overwhelmingly at Jeopardy! in 2011 (IBM, 2026), and AlphaGo managed to beat the Go world champion Lee Sedol in 2016 (DeepMind, 2016). AlphaGo was a landmark moment, because this is a game with more possible positions than atoms in the observable universe. During game 2, AlphaGo played Move 37, a move that had a 1 in 10,000 chance of being used. Commentators initially thought it was a mistake, instead it was the move that several rounds later led to winning the game. The move demonstrated what many consider to be sparks of genuine creativity that diverged from centuries of human play that the model was trained on.

Figure 2: *Kasparov, a chess grandmaster, defeated the Chess system DeepBlue in 1996. One year later in 1997 he resigned in the last game of the six-game match after 19 moves, granting the win to Deep Blue (IBM, 2026).*

> *For the first time in the history of mankind, I saw something similar to an artificial intellect.*
>
> **Garry Kasparov**
> Chess Grandmaster
>
> **1997**
> IBM, 2026

AI's superhuman game playing capability extends to video games . Machine learning techniques on simple Atari games in 2013 (Mnih et al., 2013) progressed to OpenAI Five defeating world champions at DOTA 2 in 2019 (OpenAI, 2019). That same year, DeepMind's AlphaStar beat professional esports players at StarCraft II (Google DeepMind, 2019). These are games with open ended real time environments, requiring thousands of rapid decisions and long-term planning. By 2020 the MuZero system played Atari games, Go, chess, and shogi without even being told the rules (Google DeepMind, 2020).[1] These are AIs that all play and learn autonomously without human intervention. In 2025, game playing AIs have evolved to open-ended environments across

[1]KataGo is a system which is based on techniques used by DeepMind's AlphaGo Zero and similarly superhuman in its game play. In 2022, researchers managed to demonstrate that despite being superhuman KataGo can be beaten by humans and demonstrates "surprising failure modes" of AI systems. This is the kind of thing that will be a repeated theme throughout our text (Wang and Gleave et al., 2022)

a huge variety of games. They carry the abilities learned from one game onto the next improving their performance over time (Google DeepMind, 2025).

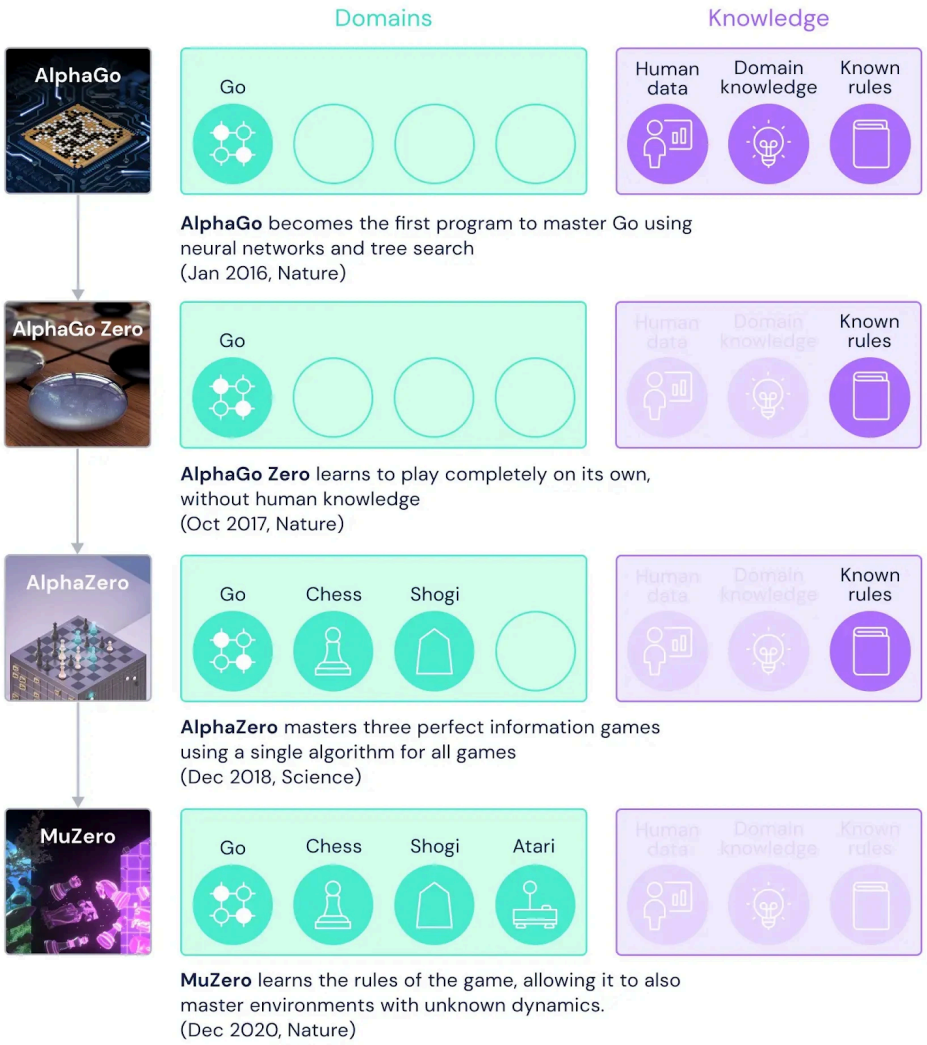


Figure 3: *The history of going from AlphaGo, which was already superhuman in 2016, to MuZero which was not only superhuman but was trained without any human data, domain knowledge or knowledge of the games rules (DeepMind, 2020).*

Game playing AI is relatively narrow in what it can do. Despite this it is extremely impressive because of the strategic planning, pattern recognition, and adversarial thinking it displays. These same reasoning abilities—planning ahead, building strategies, adapting to feedback—that started with game playing now also apply to scientific research, mathematical proofs, and complex real-world problem-solving.

Planning and Continuous Learning in Minecraft with GPT-4

OPTIONAL NOTE

AIs can construct long term strategies and play games in open ended dynamic environments. The various Alpha series of models in the 2020s did not use LLMs. But in 2023, Voyager—an AI system powered by GPT-4—demonstrated that a LLM powered system could play Minecraft (Wang et al., 2023). Minecraft, if you are not familiar, provides a sandbox where you must complete tasks in sequence: gather wood, craft basic tools, mine stone, smelt iron, craft better tools, and eventually mine diamonds. This requires planning hundreds of steps ahead and long term strategic planning. Since 2023 we have seen several game playing systems based around LLMs showcasing a variety of capabilities. For example, in strategy games, Meta's Cicero displayed intricate strategic negotiation and deception skills in natural language for the game Diplomacy (Bakhtin et al., 2022).

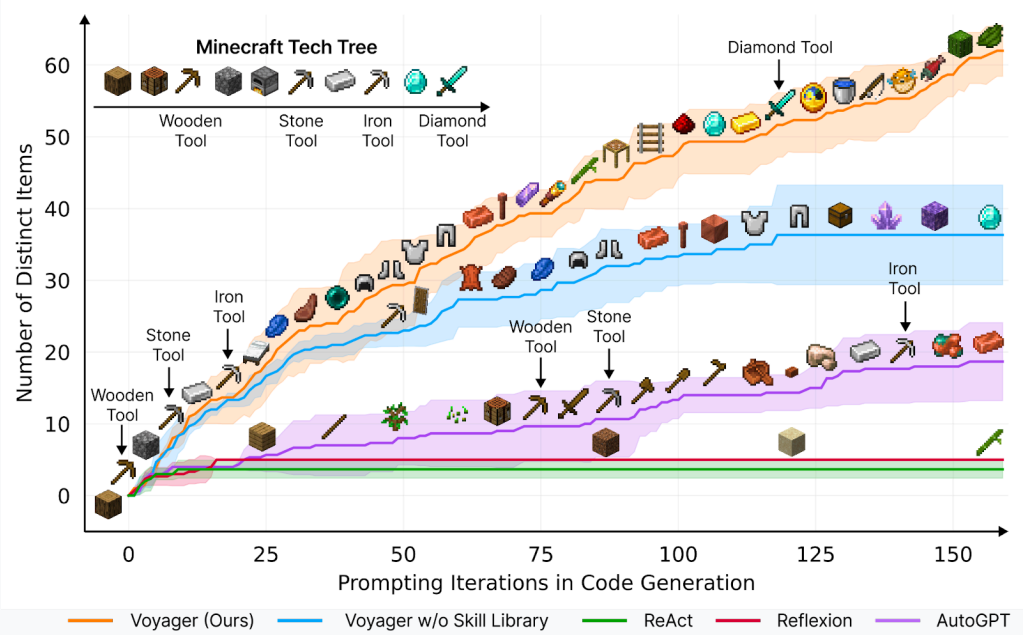


Figure 4: *Voyager discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baselines (Wang et al., 2023).*

2.2 Text Generation

Generating text can take language models far beyond simple conversations. You're probably familiar with ChatGPT. These types of language generation AIs are what we call large language models (LLMs). In 2018, early versions of these LMs could only write a few coherent paragraphs. But over a few short years, they have gotten a lot better. By 2025, GPT-5 gets 92.5% of questions right in domains ranging from highly complex STEM fields, international law, to nutrition and religion (as measured by the MMLU benchmark). Models like Claude by Anthropic, Grok by X, GLM-4.7 by Zhipu AI, and DeepSeek-v3.2 showcased similar levels of performance across various domains (ArtificialAnalysis, 2025 ; EpochAI, 2025).

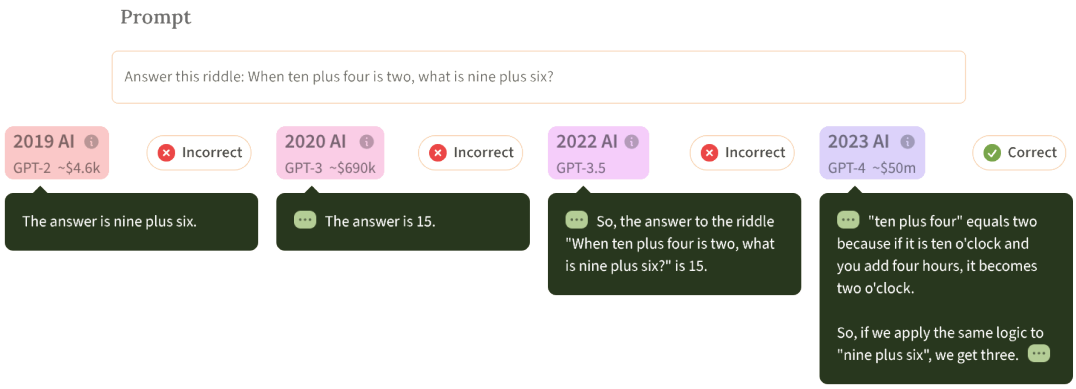


Figure 5: *In a few short years, LLMs have gone from being barely useful to being regularly used coding assistants (AI Digest, 2023).*

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

Figure 6: *Performance on common exams as a percentile compared to human test takers. Notice the large jump from GPT-3.5 to GPT-4 on these tests, often from well below the median human to the very top of the human range (Aschenbrenner, 2024 OpenAI, 2023). The jump from GPT-3 to GPT-4 was in a single year.*

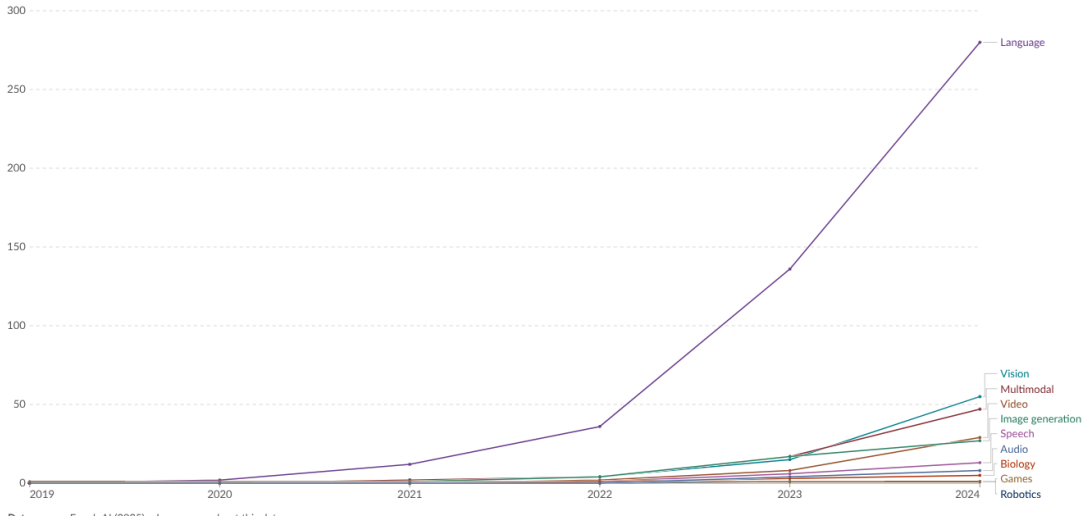


Figure 7: *We are seeing an explosion in language models due to their generality, and applicability to a wide range of tasks (Giattino et al., 2023).* (interactive version on website)

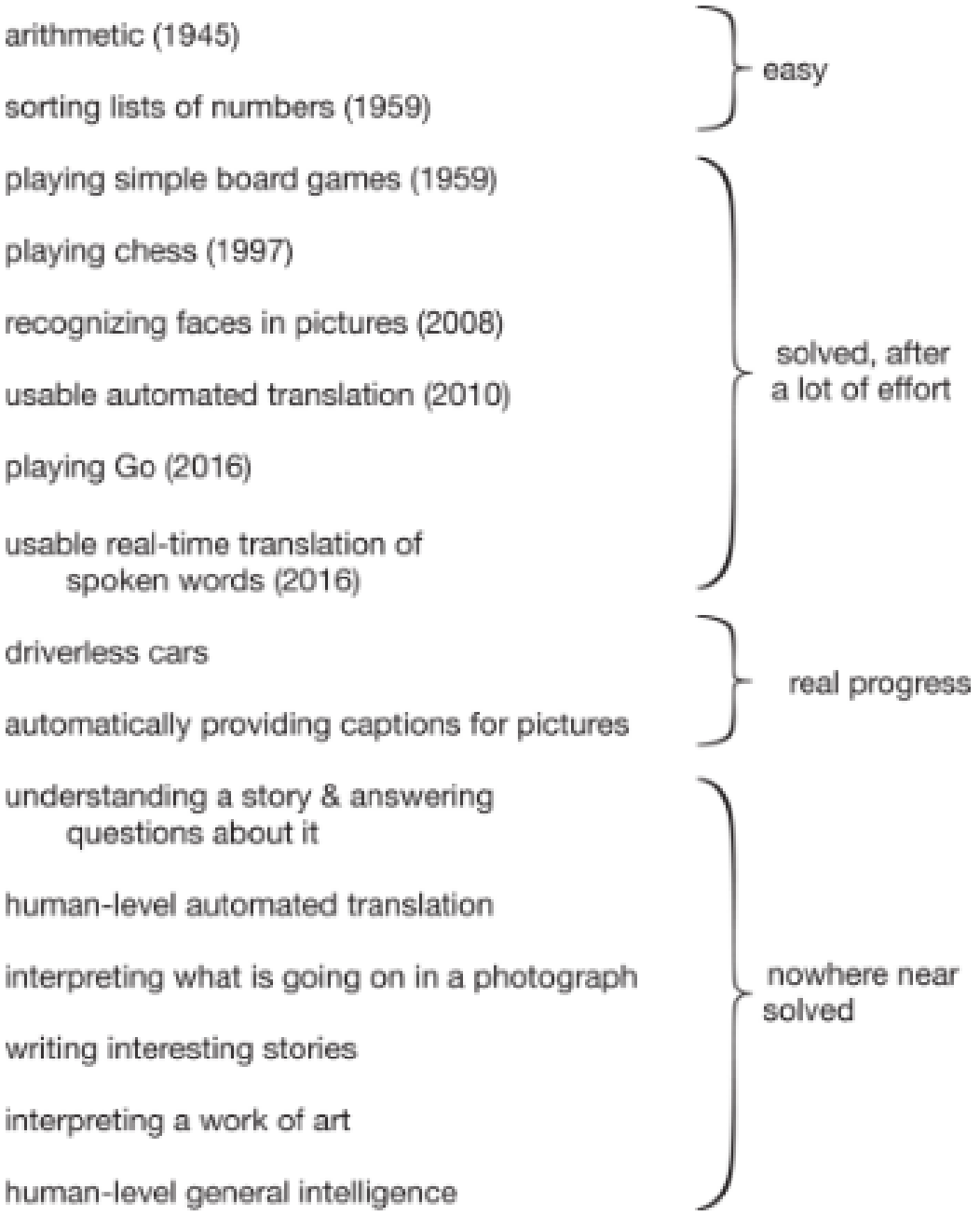


Figure 8: *A list of 'Nowhere near solved' [...] problems in AI, from 'A brief history of AI', published in January 2021 (Wooldridge, 2021). They also say: 'At present, we have no idea how to get computers to do the tasks at the bottom of the list'. But everything in the category 'Nowhere near solved' has been solved by GPT-4 (Bubeck et al., 2023), except human-level general intelligence.*

Language models have provided a core around which we have seen many impressive capabilities emerge like - scientific research, reasoning, and software development. All of these capabilities stem from the same principle of generating language and gradually refining it.

2.3 Tool Use

LLMs can intelligently use external tools, dramatically boosting performance. Language models in 2020 exhibited remarkable abilities to solve new tasks from instructions, but they used to

struggle with basic functions like arithmetic. Instead of trying to get a single model to do everything, increasingly LLMs use external tools to achieve both capabilities (Schick et al., 2023 ; Qin et al., 2023). They recognize when they need a calculator, code interpreter, or up to date information from a search engine—and call these tools appropriately.[2] Tool use significantly improves model performance; for example, the OpenAI o3 model with external tools outperforms o3 alone by almost 5% on benchmarks like Humanities Last Exam (HLE) (EpochAI, 2025). In December 2025, at least 10,000 tool servers are operational, including meta tools like 'a tool to search for tools' to help LLMs find the exact one they need for the specific situation (Anthropic, 2025 ; Anthropic, 2025). This leads to a significant enough boost that companies now report benchmark performance separately - with and without tools.

[2]Standards like the Model Context Protocol (MCP) are formalizing how AI assistants connect to data repositories and development environments (Anthropic, 2024). The MCP protocol has been donated to the Linux foundation(Anthropic, 2025).

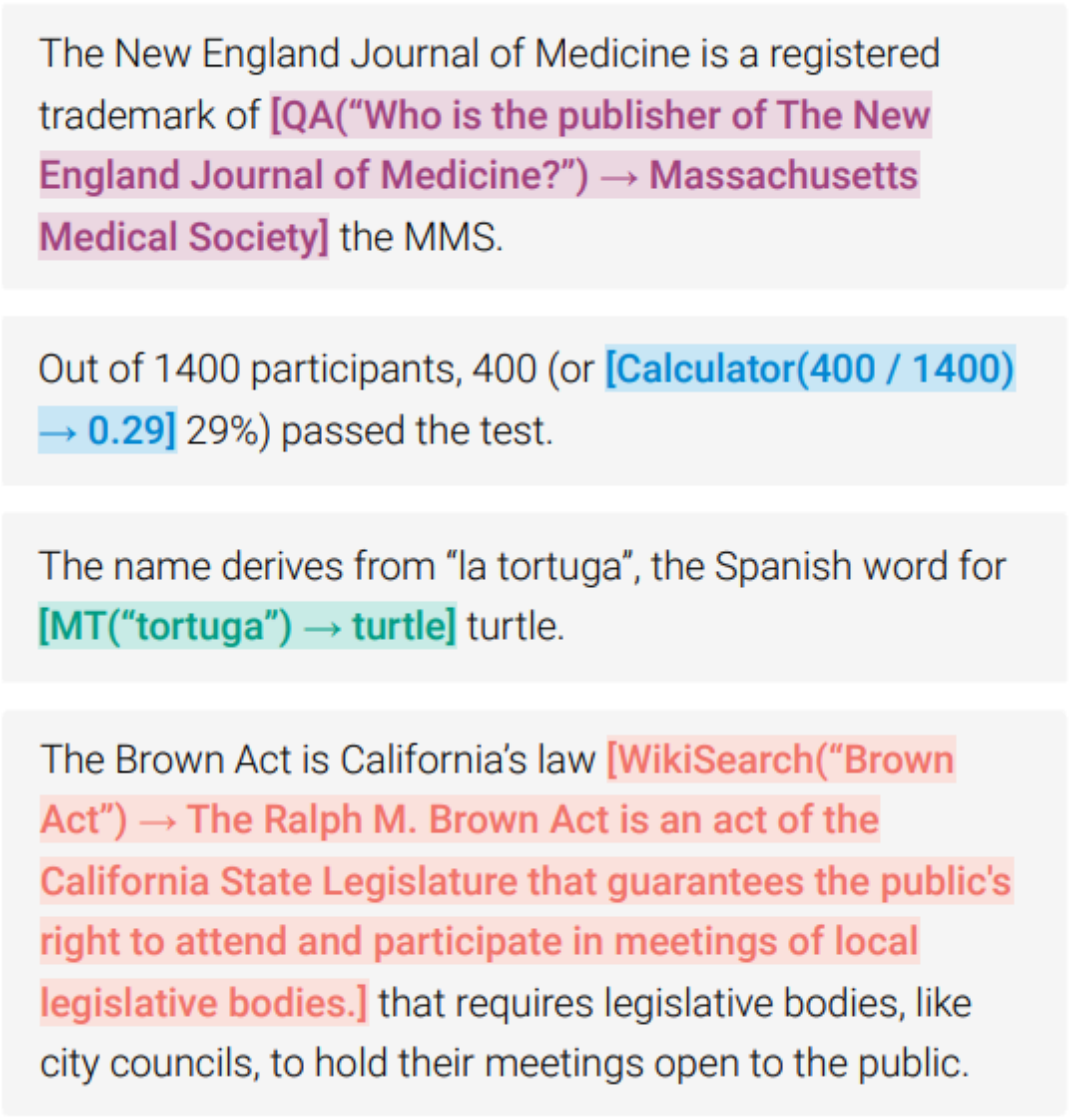
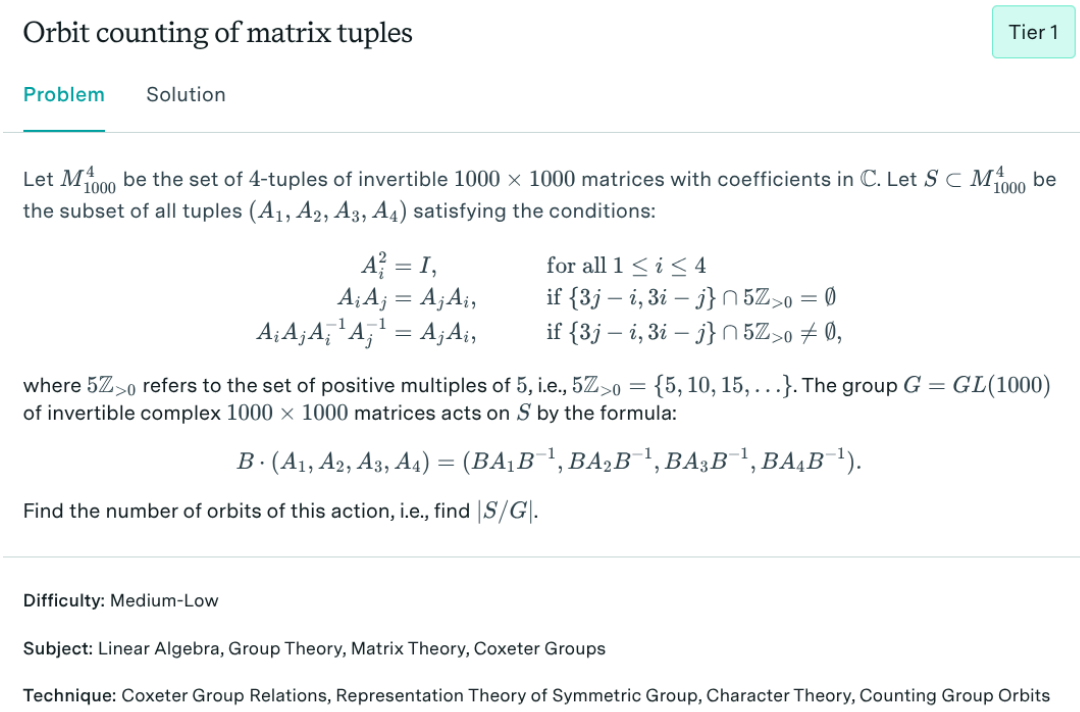


Figure 9: *A simple example from Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text (Schick et al., 2023).*

2.4 Reasoning and Research

Models now demonstrate multi-step reasoning by working through problems step-by-step. In addition to using tools, LLMs now show their reasoning, catch their own errors, and backtrack when needed. In late 2024, OpenAI introduced o1, the first "reasoning" model. These AIs allocate more effort per problem—trading "thinking time" for accuracy. The longer they think, the better their responses tend to get (OpenAI, 2024). Using these reasoning techniques, both OpenAI and Google DeepMind achieved gold-medal performance at the 2025 International Mathematical Olympiad (OpenAI, 2025 ; Google DeepMind, 2025). On FrontierMath—a test of research-level mathematics—GPT-5.2 solved 41% of tier 1-3 problems (EpochAI, 2026).1 In ML terms, 'thinking time' is called 'inference'. It is the act of generating new tokens/words, so increasing thinking time is more formally called inference time scaling. Similar to tool use this also led to such a boost

in performance that companies report high thinking time (high compute) results of benchmarks separately than no thinking time (low compute) results.



Orbit counting of matrix tuples

Tier 1

Problem Solution

Let M_{1000}^4 be the set of 4-tuples of invertible 1000×1000 matrices with coefficients in $\mathbb{C}$. Let $S \subset M_{1000}^4$ be the subset of all tuples (A_1, A_2, A_3, A_4) satisfying the conditions:

$$\begin{aligned} A_i^2 &= I, && \text{for all } 1 \le i \le 4 \\ A_i A_j &= A_j A_i, && \text{if } \{3j-i, 3i-j\} \cap 5\mathbb{Z}_{>0} = \emptyset \\ A_i A_j A_i^{-1} A_j^{-1} &= A_j A_i, && \text{if } \{3j-i, 3i-j\} \cap 5\mathbb{Z}_{>0} \neq \emptyset, \end{aligned}$$

where $5\mathbb{Z}_{>0}$ refers to the set of positive multiples of 5, i.e., $5\mathbb{Z}_{>0} = \{5, 10, 15, \ldots\}$. The group $G = GL(1000)$ of invertible complex 1000×1000 matrices acts on S by the formula:

$$B \cdot (A_1, A_2, A_3, A_4) = (BA_1B^{-1}, BA_2B^{-1}, BA_3B^{-1}, BA_4B^{-1}).$$

Find the number of orbits of this action, i.e., find $|S/G|$.

Difficulty: Medium-Low

Subject: Linear Algebra, Group Theory, Matrix Theory, Coxeter Groups

Technique: Coxeter Group Relations, Representation Theory of Symmetric Group, Character Theory, Counting Group Orbits

Figure 10: *A sample of an "easy" tier-1 problem from FrontierMath - an extremely difficult mathematics test created by mathematicians. GPT-5.2 can solve 41% of tier 1-3 problems, and 29% of the even harder tier 4 problems (EpochAI, 2025).*

LLMs can help generate and evaluate scientific hypotheses. Combining techniques like letting AI think for longer, and tools like web-search or specialized AI models, we are starting to see research assistants. As one example, Google introduced AI co-scientist in 2025. The team used it to generate and evaluate proposals for repurposing drugs, identifying drug targets, and explaining antimicrobial resistance in real-world laboratories (Google DeepMind, 2025). Others have attempted to build a fully autonomous AI scientist, which generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, and then runs a simulated review process for evaluation (SakanaAI, 2024).

Figure 11: *An example of the Google Co-Scientist. You can see the "thinking time" labelled as test time compute increasing on the left (Google DeepMind, 2025).*

AI is transitioning to an active research collaborator across scientific domains. Instead of only using language models, companies are also using approaches similar to AlphaZero to create a whole range of specialized models for scientific domains. For example, Demis Hassabis & John Jumper were awarded the nobel prize in chemistry for their work on building AlphaFold (Google DeepMind, 2024). This is a model that helped solve the long outstanding protein folding problem (Google DeepMind, 2022), and its successors AlphaFold 2 and 3 continue to aid thousands of researchers in biology (DeepMind, 2024). Similarly, AlphaGenome is helping us better understand human DNA (Google DeepMind, 2025). AlphaEvolve is helping generate faster algorithms for machine learning (Google DeepMind, 2025), and AlphaChip helps design the semiconductors that run algorithms (Google DeepMind, 2024).

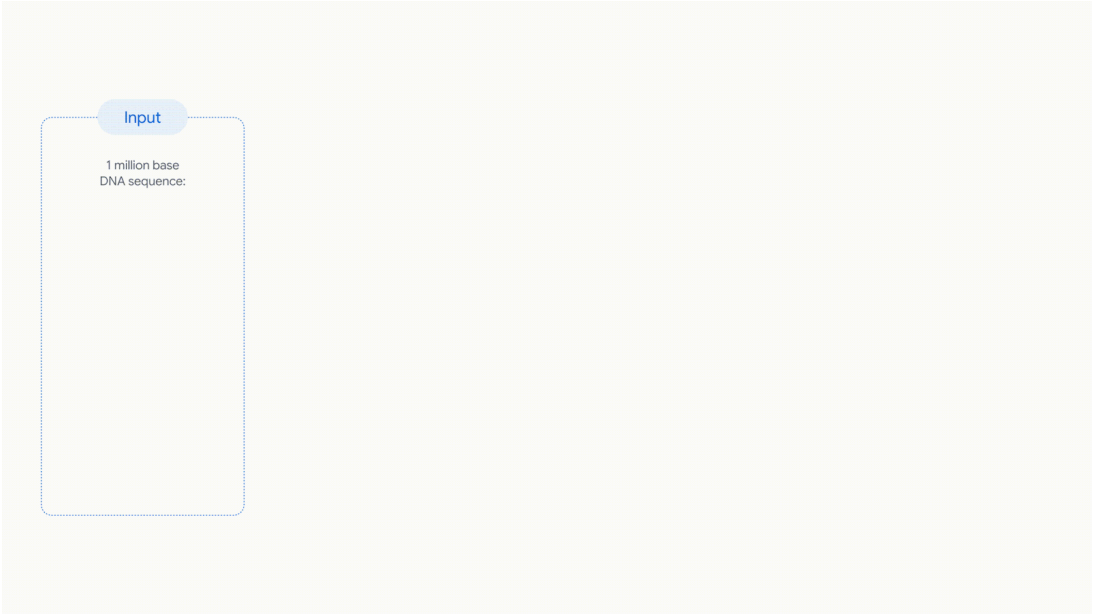


Figure 12: *Animation showing AlphaGenome taking one million DNA letters as input and predicting diverse molecular properties across different tissues and cell types (Google DeepMind, 2025)*

Beyond just mathematics and scientific research, AI models are also developing more abstract intellectual skills. LLMs have some level of metacognition, they can evaluate the validity of their own claims and predict which questions they will be able to answer correctly (Kadavath, 2022). They have some knowledge about their own selves and their limitations. Similarly, they display the ability to attribute mental states to themselves and others (theory of mind). This helps in predicting human behaviors and responses (Kosinski 2023 ; Xu et al., 2024). We are going to talk more about how we concretely define and measure things like intelligence, meta-cognition and so on in later sections.

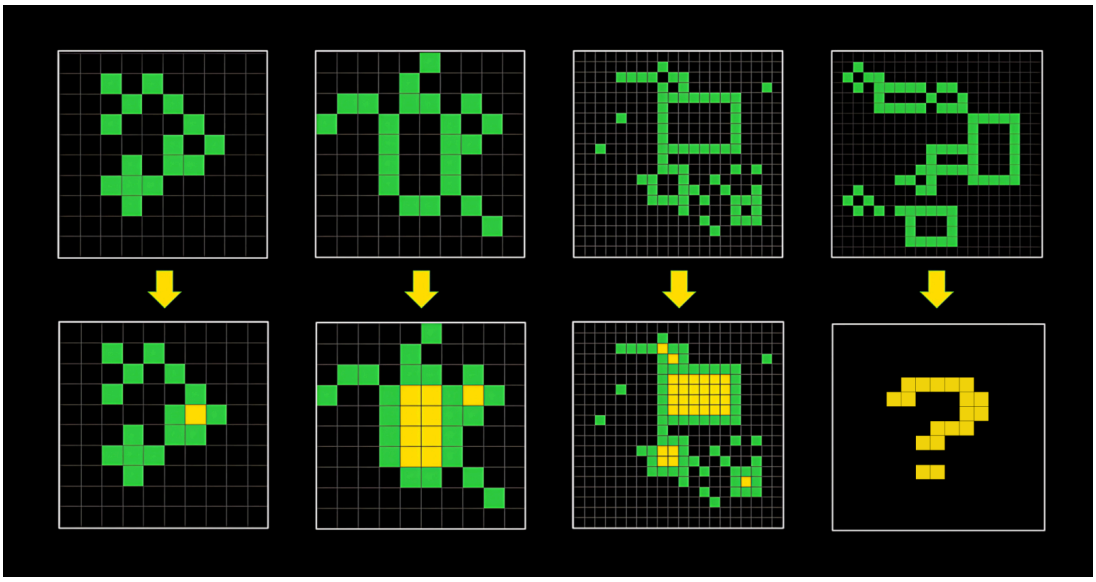


Figure 13: *An example puzzle from the abstraction and reasoning corpus for AI (ARC-AGI v1). These are designed to be easy for humans, but very hard for AIs. Models are able to solve increasing numbers of such abstract reasoning puzzles, especially with the advent of large reasoning models (LRMs) in 2025 (ARC-AGI, 2024 Chollet et. al, 2024).*

2.5 Software Development

Coding is evolving from autocomplete to collaborative software development. LLMs can generate text in any form, and one particular type of text that they are proving to be especially good at is generating code. When paired with reasoning capabilities, and tools LLMs read documentation, edit codebases spanning thousands of files, run tests, debug failures, and iterate until tests pass—with increasingly minimal human guidance. In 2025 systems like Claude Opus 4.5, Gemini 3 Pro, and GPT-5.2 implement features and entire applications increasingly independently (Anthropic, 2025 ; Google DeepMind, 2025 ; OpenAI, 2025). When tested against real GitHub issues from open-source projects— in 2024 AI systems (Claude 3 Opus) could solve just 15% problems, but by 2025, this had jumped to being able to solve 74% of issues (Tools + Claude 4 Opus) (SWE bench, 2025).

Figure 14: *AI systems now develop themselves. Boris Cherny, creator of Claude Code—Anthropic's AI-powered software development tool—stated in December 2025 that 100% of his contributions to Claude Code over the previous thirty days were written by Claude Code itself (Cherny, 2025, Twitter).*

2.6 Vision: Images and Video

Image generation progressed from unrecognizable noise to photorealistic scenes in under a decade. In 2014, Generative Adversarial Networks (GANs) produced grainy, low-resolution faces (Goodfellow et al., 2014). By 2023, models generated detailed images from complex text prompts. Models like Midjourney v7 create photorealistic scenes nearly indistinguishable from professional photography. Video generation is following a similar trajectory. AI generated videos and DeepFakes are getting increasingly indistinguishable from real videos.

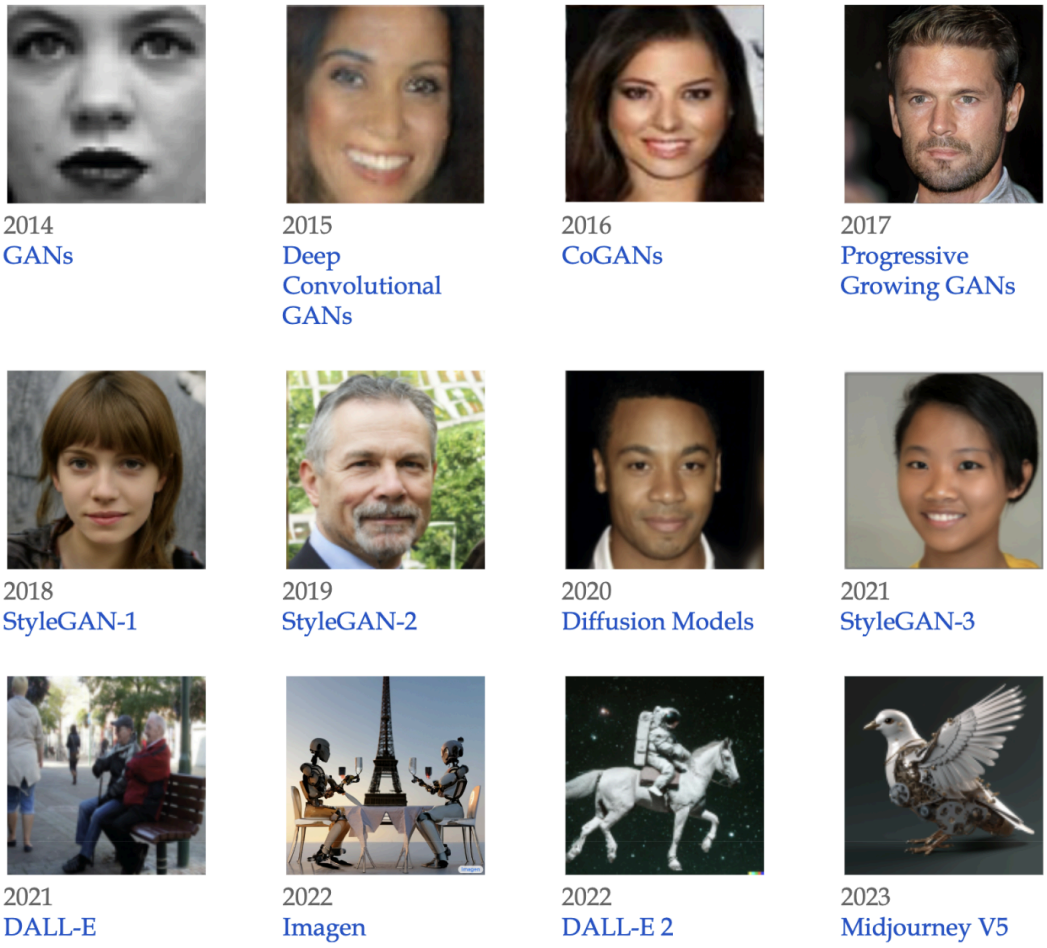


Figure 15: *An example of the evolution of image generation. At the top left, starting from GANs (Generative Adversarial Networks) to the bottom right, an image from MidJourney V5.*

Figure 16: *Improvements between the V1 of the MidJourney image generation model in early 2022, to the V6 in December 2023. Prompt: high-quality photography of a young Japanese woman smiling, backlighting, natural pale light, film camera, by Rinko Kawauchi, HDR (Yap, 2024).*

Large Multimodal **Models (LMMs) combine language and image understanding capabilities.** In 2025, multimodal models answer questions about spatial relationships in images, read embedded text in complex scenes, and extract information from charts and diagrams. Image models released in 2025 handle generation and sophisticated editing across modalities. Systems like Gemini 3 Pro Image work text-to-image, image-to-image, and handle complex editing—changing lighting, style, or composition while maintaining coherence (Google, 2025).

Figure 17: *An infographic containing both images and educational text generated by Google's Nano Banana pro model. Prompt: Create an infographic that shows how to make elaichi chai (Google, 2025).*

Figure 18: *Example of extending video and image generation to an interactive world. This was generated by DeepMind Genie 3. The bottom left arrows indicate the user interacting with the generated environment (DeepMind, 2025).*

2.7 Robotics

Both AI and robotics are evolving, with robots giving AI physical embodiment in the real world. Robotics is combining LLMs and visual models, to create robot control models (e.g. RT-1 or RT-2). These robots can use techniques borrowed from language models—like breaking complex actions into step-by-step plans—to control robot manipulators (Google DeepMind, 2024). They managed to learn behaviors like opening cabinets, operating elevators, and cooking tasks through observing human demonstrations. Robots have demonstrated the ability to perform intricate manipulation: sautéing shrimp, storing heavy pots in cabinets, and rinsing pans (Fu et al., 2024).

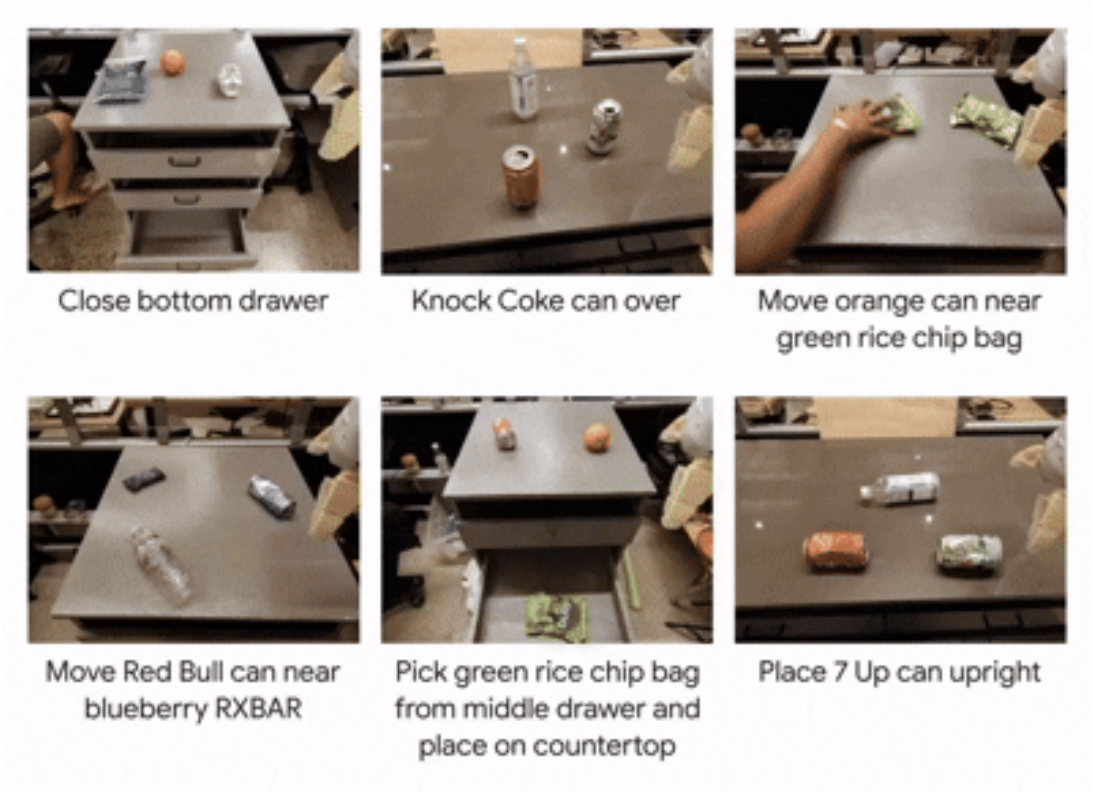


Figure 19: *SARA-RT-2 model for manipulation tasks. The robot's actions are conditioned on images and text commands (Google DeepMind, 2024).*

Autonomous robots are moving from research labs into real-world industrial deployment at significant scale. In 2023, China installed 276,300 industrial robots (AI Index Report, 2025). These systems handle welding, parts assembly, materials handling, and quality inspection—tasks requiring precision but not necessarily advanced reasoning. In addition to industrial robots, warehouse robotics represents one of the most mature deployments—Amazon operates over 1 million robots across its fulfillment network, handling everything from inventory storage to package sorting (Amazon, 2025). Robots in warehouses and industry are able to handle packages, speed up inventory identification using machine vision, and autonomously unload shipping containers.

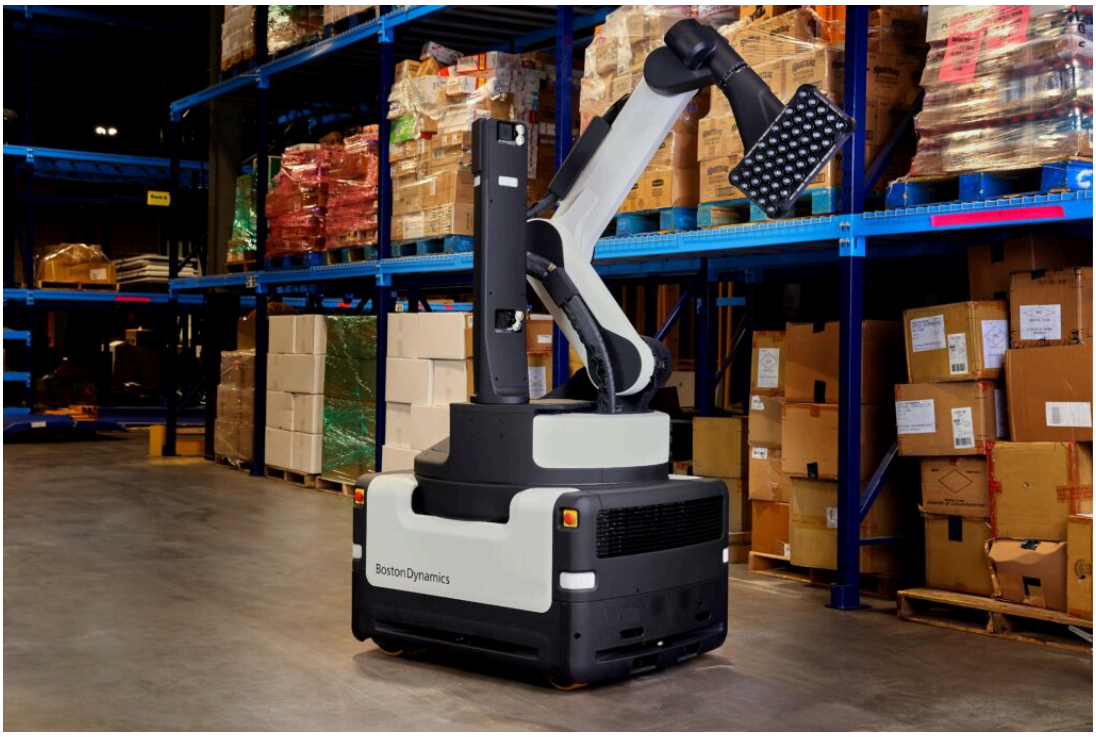


Figure 20: *Boston Dynamics' Stretch robot that can autonomously unload shipping containers, lift up to 50 pounds, and clear away missed boxes along the way (Boston Dynamics, 2024).*

Figure 21: *Amazon has a huge fleet of robots that automate it's warehouses. This is an example of the Hercules robot that retrieves shelves of products and delivers them to employees, who then pick the items customers ordered for shipping (Amazon, 2023).*

3. Foundation Models

Foundation models **represent a fundamental shift in how we develop AI.** Rather than building specialized models for many small specific tasks, we can now train large-scale models that serve as a "foundation" for many different applications. These models are then specialized later by a process called fine-tuning to perform specific tasks. Think of this as similar to how we can build many different types of buildings using the same base structure (Bommasani et al., 2022). We can build banks, restaurants, or housing but the underlying foundation remains largely the same. This is just a very quick intuitive definition. We will get more into the details in the next few subsections on training, properties and risks.

The traditional approach of training specialized AI models for every task often proved inefficient and limiting. Progress was bottlenecked by the need for human-labeled data and the inability to transfer knowledge between tasks effectively. Foundation models overcame these limitations through a process called self-supervised learning on massive unlabeled datasets. This breakthrough happened because of many different reasons - advances in specialized hardware like GPUs, new machine learning architectures like transformers, and increased access to huge amounts of online data (Kaplan et al., 2020) are some of the more prominent reasons for this shift.

In language processing, models like GPT-4 and Claude are examples of foundation models . Both of these have demonstrated the ability to generate human language, have complex conversations and perform simple reasoning tasks (OpenAI, 2023). Examples in computer vision include models like DALL-E 3 and Stable Diffusion. (Betker et al., 2023) These are domain specific examples, but we are also seeing a trend toward multimodal foundation models (LMMs). This includes things like GPT-4V and Gemini that can work across different types of data - processing and generating text, images, code, audio and probably more in the future (Google, 2023). Even in reinforcement learning, where models were traditionally trained for specific tasks, we're seeing foundation models like Gato demonstrate the ability to learn general-purpose behaviors that can be adapted to various different downstream tasks (Reed et al., 2022).

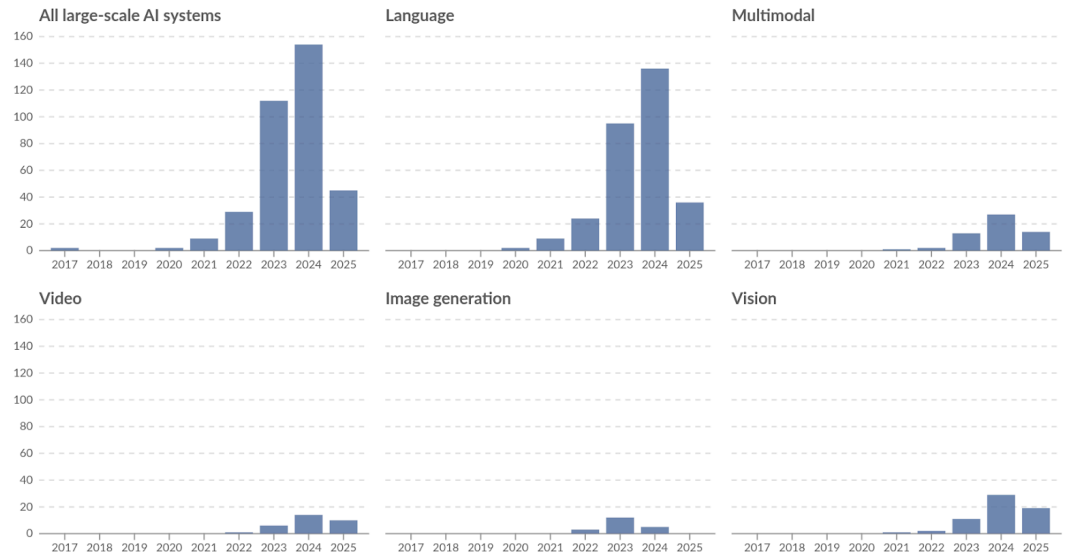


Figure 22: *Number of large-scale AI systems released per year. Describes the specific area, application, or field in which a large-scale AI model is designed to operate. The 2025 data is incomplete and was last updated 01 June 2025 (Giattino et al., 2023).* (interactive version on website)

Foundation models **mark a paradigm shift towards general-purpose systems.** This paradigm introduces many new risks which didn't exist previously. These include misuse risks from power centralization, homogenization, and dual-use capabilities just to name a few. The ability of foundation models to learn broad, transferable capabilities has led to increasingly sophisticated behaviors emerging from relatively simple training objectives (Wei et al., 2022). Complex capabilities, combined with generality and scale, means we need to seriously consider safety risks beyond just misuse that previously seemed theoretical or distant. Beyond just misuse risk, things like misalignment are becoming an increasing concern with each new capability that these foundation models exhibit. We dedicate an entire chapter to the discussion of these risks. But we will also give you a small taste on the kinds of possible risks in the next few subsections, as it warrants some repetition.

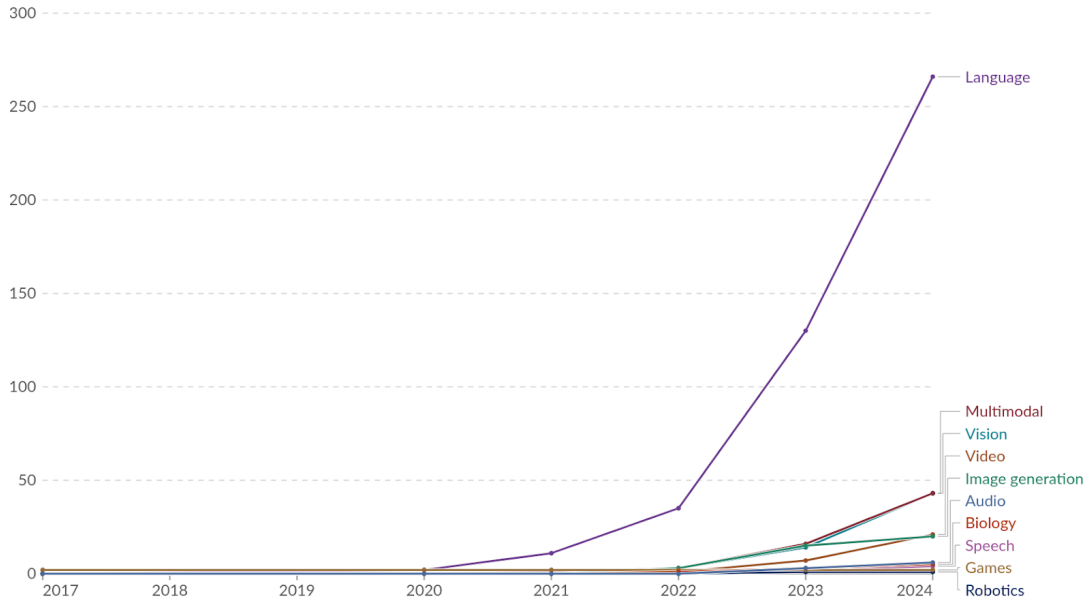


Figure 23: *Cumulative number of large-scale AI models by domain since 2017. Describes the specific area, application, or field in which a large-scale AI model is designed to operate (Giattino et al., 2023).* (interactive version on website)

What is the difference between foundation models **and frontier models?** Frontier models represent the cutting edge of AI capabilities - they are the most advanced models in their respective domains. While many frontier models are also foundation models (like Claude 3.5 Sonnet), this isn't always the case. For example, AlphaFold, while being a frontier model in protein structure prediction, isn't typically considered a foundation model because it's specialized for a single task rather than serving as a general foundation for multiple applications (Jumper et al., 2021).

3.1 Training

Generally, foundation models use a two-stage training process. First, they go through what we call a pre-training , and then second, they can be adapted through various mechanisms like fine-tuning or scaffolding to perform specific tasks. Rather than learning from human-labeled examples for specific tasks, these models learn by finding patterns in huge amounts of unlabeled data.

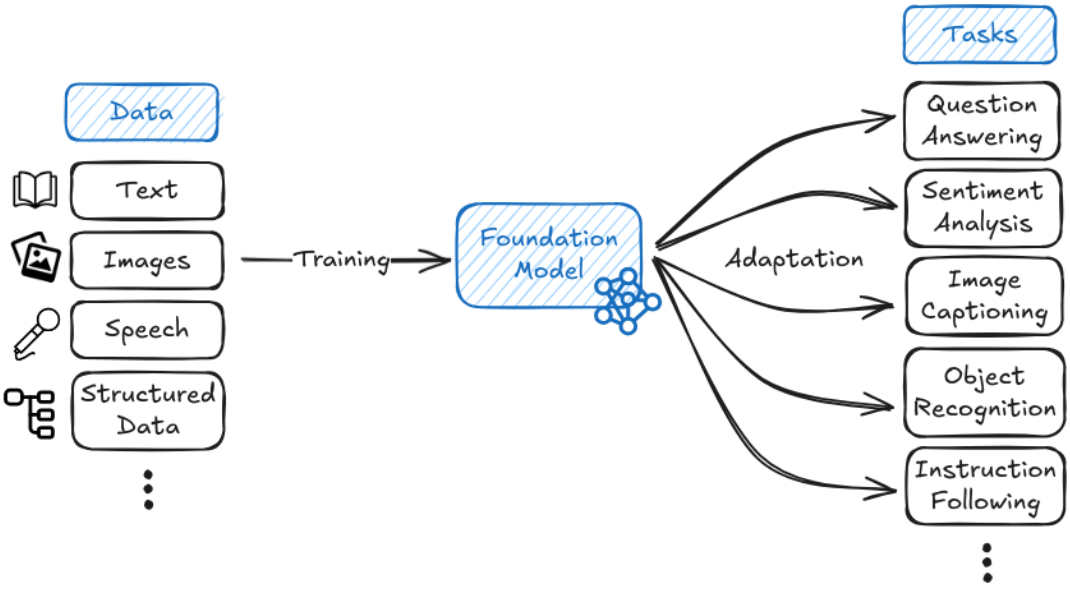


Figure 24: *On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)*

Pre-training **is the initial phase where the model learns general patterns and knowledge from massive datasets of millions or billions of examples.** During this phase, the model isn't trained for any specific task - instead, it develops broad capabilities that can later be specialized. This generality is both powerful and concerning from a safety perspective. While it enables the model to adapt to many different tasks, it also means we can't easily predict or constrain what the model might learn to do (Hendrycks et al., 2022).

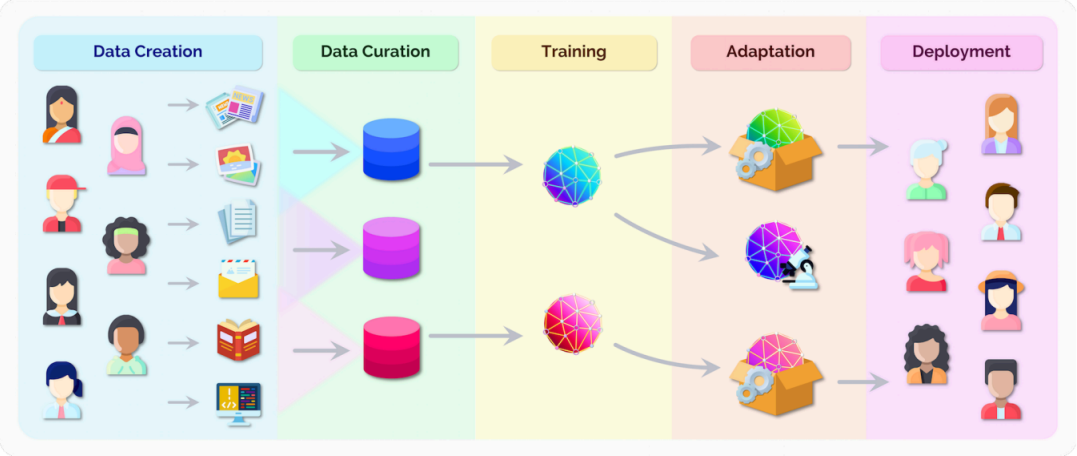


Figure 25: *On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)*

Self-supervised learning **(SSL) is the key technical innovation that makes** foundation models **possible.** This is how we actually implement the pre-training phase. Unlike traditional supervised learning , which requires human-labeled data, SSL leverages the inherent structure of the data itself to create training signals. For example, instead of manually labeling images, we might just hide part of a full image we already have and ask a model to predict what the rest should be. So it might predict the bottom half of an image given the top half, learning about which objects often appear together. As an example, it might learn that images with trees and grass at the top often have more

grass, or maybe a path, at the bottom. It learns about objects and their context - trees and grass often appear in parks, dogs are often found in these environments, paths are usually horizontal, and so on. These learned representations can then be used for a wide variety of tasks that the model was not explicitly trained for, like identifying dogs in images, or recognizing parks - all without any human-provided labels! The same concept applies in language, a model might predict the next word in a sentence, such as "The cat sat on the ... ," learning grammar, syntax, and context as long as we repeat this over huge amounts of text.

After pre-training **,** foundation models **can be adapted through two main approaches:** fine-tuning **and in-context prompting.** Fine-tuning involves additional training on a specific task or dataset to specialize the model's capabilities. For example, we might use Reinforcement Learning from Human Feedback (RLHF) to make language models better at following instructions or being more helpful. Prompting, on the other hand, involves providing the model with carefully crafted inputs that guide it toward desired behaviors without additional training.

The training process of foundation models **creates several unique safety challenges.** First, the self-supervised nature of pre-training means we have limited control over what the model learns - it might develop unintended capabilities or behaviors. Second, the adaptation process needs to reliably preserve any safety properties we've established during pre-training . Finally, the massive scale of training data and compute makes it difficult to thoroughly understand or audit what the model has learned. Many of the safety challenges we'll discuss throughout this book - from goal misgeneralization to scalable oversight - are deeply connected to how these models are trained and adapted.

3.2 Properties

The ability of foundation models to transfer knowledge, generalize across many different domains, and develop emergent capabilities means we can't rely on traditional safety approaches that assume narrow, predictable behavior.

Transfer learning **is the ability to transfer knowledge learned during** pre-training **to new tasks and domains.** Rather than starting from scratch for each task, we can leverage the general knowledge these models have already acquired (Bommasani et al., 2022). This property enables rapid adaptation and deployment, it also means that both capabilities and safety risks can transfer in unexpected ways. For example, a model might transfer not just useful knowledge but also harmful biases or undesired behaviors to new applications.

Zero-shot and few-shot learning **is the ability to perform new tasks with very few examples, or even no examples at all.** For example, GPT-4 can solve novel reasoning problems just from a natural language description of the task (OpenAI, 2023). This emergent ability to generalize to new situations is powerful but concerning from a safety perspective. If models can adapt to novel situations in unexpected ways, it becomes harder to predict and control their behavior in deployment.

Generalization in foundation models **works differently from traditional AI systems.** Rather than just generalizing within a narrow domain, these models can generalize capabilities across domains in surprising ways. However, this generalization of capabilities often happens without a corresponding generalization of goals or constraints - a critical safety concern we'll explore in detail in our chapter on goal misgeneralization. For example, a model might generalize its ability

to manipulate text in unexpected ways without maintaining the safety constraints we intended (Hendrycks et al., 2022).

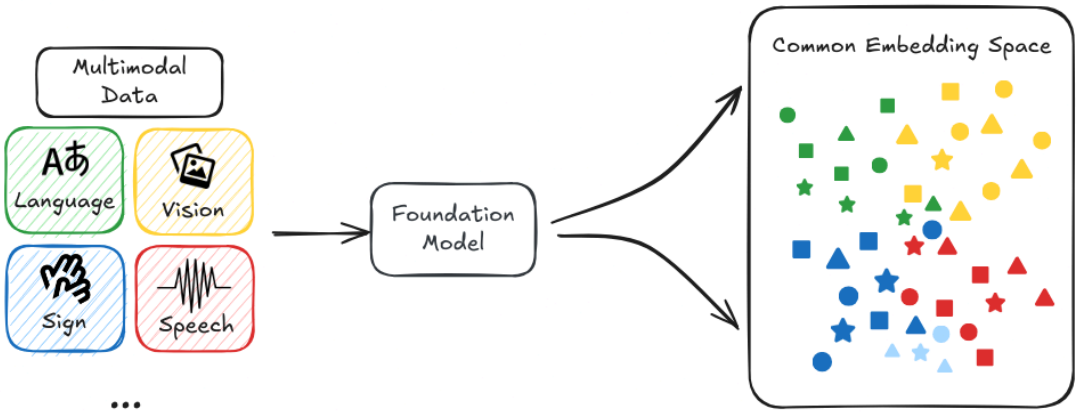


Figure 26: *On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)*

> *Multimodality will definitely be important. Speech in, speech out, images, eventually video. Clearly, people really want that. Customizability and personalization will also be very important.*
>
> **Sam Altman**
> CEO of OpenAI
>
> **Jan 2024**
> Cronshaw, 2024

4. Defining and Measuring AGI

Before we can discuss AI safety, we need to agree on what we mean by AGI. In our previous section on foundation models , we saw how modern AI systems are becoming increasingly powerful and general-purpose. But powerful at what, exactly? Some researchers claim we're already seeing "sparks" of AGI in the latest language models (Bubeck et al., 2023). Others predict human-level AI within a decade (Bengio et al., 2023). But human level at what exactly? How do we measure this? Without a clear definition, how do we assess such claims or plan appropriate safety measures?

If you can't define something, you can't measure it. If you can't measure it, you can't reliably track progress, identify or prepare for risks. Think about an example from physics. If we want to design and enforce speed limits for cars, saying something "moved 5" makes no sense without units. Did it move 5 meters, 5 feet, or 5 royal cubits? If we don't know how far or fast it moved, we can't enforce limits. The same applies to intelligence and subsequent safety measures. Every field needs standardized units - meters, watts, joules - to advance beyond vague descriptions. We need to treat AI safety with the same rigor to move past hand-waving about "intelligence."

Defining general intelligence is extremely challenging. Everyone agrees we need a definition to measure progress and design safety measures. So why don't we have one? The problem is that intelligence describes multiple overlapping abilities - problem-solving, learning, adaptation, abstract reasoning. Different disciplines view it through different lenses. Psychologists emphasize measurable cognitive skills. Computer scientists focus on task performance. Philosophers debate consciousness and self-awareness. Which approach matters most for AI safety? We go through a couple of case studies and previous attempts at a definition before giving the definitions we will be using for this text.

4.1 Case Studies

Alan Turing suggested we could sidestep the whole mess by focusing on behavior. If a machine could imitate human conversation well enough to fool an interrogator, it should be considered intelligent (Turing, 1950). The behaviorist approach is simple - forget about internal mental states and focus on observable behavior. Can the system do the thing or not? Unfortunately, LLMs exposed some limitations to this approach. GPT-4 can pass Turing-style conversation tests while struggling with basic spatial reasoning or maintaining coherent long-term plans (Rapaport, 2020). The test was too narrow. Conversation is just one capability among many we care about. But Turing's core insight - focus on observable capabilities, not internal states - remains sound.

Consciousness based approaches to general intelligence focus on "true understanding." John Searle's Chinese Room argument suggested that systems might appear intelligent (stochastic parrots) without truly understanding - processing symbols without grasping their meaning (Searle, 1980). This view emphasizes internal cognitive states similar to human consciousness. The problem is that consciousness often proves even harder to define than intelligence. We are also unsure if intelligence and consciousness are necessarily linked - a system could potentially be highly intelligent without being conscious, or conscious without being particularly intelligent. AlphaGo is very intelligent within the context of playing Go, but it is clearly not conscious. A system doesn't need to be conscious to cause harm. Whether an AI system is conscious has little, if any, bearing

on its ability to make high-impact decisions or take potentially dangerous actions.[3] Research into consciousness, sentience, and meta-ethical debates about the fundamental nature of intelligence are valuable, but less actionable for the type of safety work that this text focuses on.

Defining intelligence through goal achievement. Shane Legg and Marcus Hutter propose: " *Intelligence measures an agent's ability to achieve goals in a wide range of environments* " (Legg & Hutter, 2007). This captures something important - intelligent systems should be able to figure out how to get what they want across different situations. But it's too abstract for practical measurement. How are agents defined? Which goals? Which environments? How do you actually test this? The intuition is right, but we need something more concrete.

Process and adaptability focused views see intelligence as learning efficiency rather than accumulated skills. Some researchers define intelligence through adaptability: " *the capacity of a system to adapt to its environment while operating with insufficient knowledge and resources* " (Wang, 2020), or " *the efficiency with which a system can turn experience and priors into skills* " (Chollet, 2019). They argue that general intelligence is not demonstrated by the possession of a specific skill, but by the efficiency of acquiring new skills when faced with novel problems. High performance can be "bought" with sufficient data and compute, an AI that achieves superhuman performance at Go has mastered Go; it has not necessarily become more intelligent in a general sense. This makes sense - the ability to learn quickly from limited data in new situations matters. A system that needs a billion examples to learn what humans get from ten examples might be less impressive even if both reach the same final capability. But purely from a safety perspective, does the learning path matter more than the destination? If a system can perform dangerous tasks whether through efficient learning or brute-force memorization, the risks exist either way.

Psychologists use standardized tests and measure cognitive abilities directly. Psychometric traditions like the Cattell-Horn-Carroll (CHC) theory, break intelligence into measurable cognitive domains - reasoning, memory, processing speed, spatial ability (Schneider & McGrew, 2018). IQ tests aren't perfect, but they predict real-world outcomes better than vague philosophical debates about "true understanding." This benchmarking and test based framework gives us concrete domains to measure and track.

> *When discussing AI risks, talk about capabilities, not intelligence... People often have different definitions of intelligence, or associate it with concepts like consciousness that are not relevant to AI risks, or dismiss the risks because intelligence is not well-defined.*
>
> **Victoria Krakovna**
> Senior research scientist at Google DeepMind
>
> **Aug 2023**
> Krakovna, 2023

[3] We measure qualities like "sitational-awareness" in our chapters on risks, and evaluations using specific benchmarks. So we capture a model of the self and future planning, while distinguishing these aspects from conversations about consciousness.

Behaviorist (capabilities focused) approaches matter most for safety. If an AI system can perform dangerous tasks at human-level or beyond - sophisticated planning, manipulation, deception - these risks exist regardless of how it achieved them. Did it learn efficiently or through brute force? Is it conscious or just pattern-matching? These questions don't change the risk profile. OpenAI defines AGI as " *highly autonomous systems that outperform humans at most economically valuable work* " (OpenAI, 2023). Anthropic frames their mission around ensuring " *transformative AI helps people and society* " (Anthropic, 2024). Both focus on what systems can do, and less on how they do it. But even these capability-based definitions can be too vague. Which humans are the measure - top experts or average workers? What counts as economically valuable - just current jobs or future ones? What about systems that excel at complex tasks but only work for short periods or require massive compute? The definitions point in the right direction but lack precision for tracking progress and identifying risks.

Our approach attempts to synthesize the useful parts of all these views. We adopt the behaviorist insight from Turing - primarily focusing on what systems can observably do. We use the psychometric tradition's concrete measurement framework from CHC theory. We acknowledge and incorporate the adaptability focused view's point about efficient learning but prioritize final capabilities. And we set aside the consciousness debate as not actionable for safety work.

We need a framework that's both concrete and continuous. Concrete enough to measure what systems can do right now while also being able to identify thresholds for regulation, and emerging risks before they materialize. It has to be continuous to capture progress along multiple dimensions rather than forcing everything into binary categories that lead to endless semantic debates about whether something "counts" as AGI. That's what we build in the next section: capability and generality as two continuous axes that let us describe any AI system precisely.

4.2 Defining General Intelligence

AGI exists on a continuous spectrum measured by two dimensions: capability and generality. Capability measures how well a system executes specific cognitive tasks - from 0% (can't do the task at all) to expert human level (outperforming roughly 80-90% of humans on that specific task) to superhuman (outperforming 100% of humans). Generality measures breadth - the percentage of cognitive domains where a system achieves expert-level capability. Together, these give us concrete statements like "our AI system can outperform 85% of humans in 30% of cognitive domains." This precisely describes any AI system's capabilities.

Capability measures depth - how good is the system at individual tasks. For decades, AI research focused on making systems excel at single tasks. Early chess programs in the 1950s beat novices but lost to experts. Deep Blue in the 1990s beat world champion Kasparov. AlphaGo in 2016 achieved superhuman capability at Go, a game humans thought computers wouldn't master for decades. This progression from "can't do the task" to "better than any human" represents the capability spectrum. Everything along this line - from basic competence to superhuman capability on a single task - counts as artificial narrow intelligence (ANI).

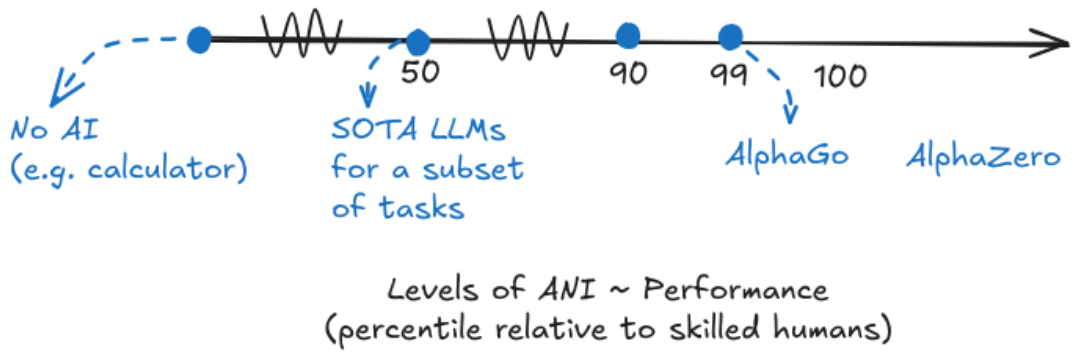


Figure 27: *This is the continuous outlook of AI measuring performance. All points on this axis can be called artificial narrow intelligence (ANI) (except for the origin) (Morris et al., 2024).*

ARTIFICIAL NARROW INTELLIGENCE (ANI) IBM, 2023

Weak AI—also called Narrow AI or Artificial Narrow Intelligence (ANI)—is AI trained and focused to perform specific tasks. Weak AI drives most of the AI that surrounds us today. 'Narrow' might be a more accurate descriptor for this type of AI as it is anything but weak; it enables some very robust applications, such as Apple's Siri, Amazon's Alexa, IBM Watson, and autonomous vehicles.

Capability alone isn't enough to define and measure progress. AlphaGo achieved superhuman capability at Go - better than any human ever. Yet ask it to write a sentence, solve an algebra problem, or recognize objects in an image, and you get nothing. This is where foundation models changed things. The shift from building one narrow system per task to training general-purpose models is why we suddenly need a framework measuring both axes - capability and generality. Pre-foundation model AIs maxed out the capability axis on some individual tasks. But systems in the last few years are climbing both axes simultaneously - getting better at specific capabilities while also expanding to more domains. Capability measures depth; generality measures breadth.

When experts say they see "sparks of AGI" in recent systems, they're observing performance across multiple cognitive domains - not just one (Bubeck et al., 2023). Think of it like the first airplane built by the Wright brothers. It barely flew, stayed airborne for seconds, and looked nothing like modern airliners. But it was still a plane! Similarly, systems achieving expert-level performance across even a modest subset of cognitive domains represent genuine general intelligence - just early-stage (weak AGI). As these capabilities expand to cover more domains at higher performance levels, systems become both increasingly general and increasingly capable. There's no universally agreed upon threshold where "not AGI" suddenly becomes "AGI."

Not every possible task matters equally for progress towards general intelligence. We don't care much if AI gets better at knitting sweaters, but we care a lot if it improves at abstract reasoning, long-term planning, and memory. The things we care about can be informally thought of as everything that can be done as remote work on a computer. More formally, the domains we choose to focus on are non-embodied cognitive tasks. These break down into ten core capabilities from the Cattell-Horn-Carroll theory (Hendrycks et al., 2025). The specific capabilities we gave

examples of in the first section - LLMs doing reasoning and math, vision models understanding images, agents planning in Minecraft - map directly onto these domains (Hendrycks et al., 2025):

1. **General Knowledge (K):** The breadth of factual understanding of the world, encompassing commonsense, culture, science, social science, and history.
2. **Reading and Writing Ability (RW):** Proficiency in consuming and producing written language, from basic decoding to complex comprehension, composition, and usage.
3. **Mathematical Ability (M):** The depth of mathematical knowledge and skills across arithmetic, algebra, geometry, probability, and calculus.
4. **On-the-Spot Reasoning (R):** The flexible control of attention to solve novel problems without relying exclusively on previously learned schemas, tested via deduction and induction.
5. **Working Memory (WM):** The ability to maintain and manipulate information in active attention across textual, auditory, and visual modalities.
6. **Long-Term Memory Storage (MS):** The capability to continually learn new information (associative, meaningful, and verbatim).
7. **Long-Term Memory Retrieval (MR):** The fluency and precision of accessing stored knowledge, including the critical ability to avoid confabulation (hallucinations).
8. **Visual Processing (V):** The ability to perceive, analyze, reason about, generate, and scan visual information.
9. **Auditory Processing (A):** The capacity to discriminate, recognize, and work creatively with auditory stimuli, including speech, rhythm, and music.
10. **Speed (S):** The ability to perform simple cognitive tasks quickly, encompassing perceptual speed, reaction times, and processing fluency.

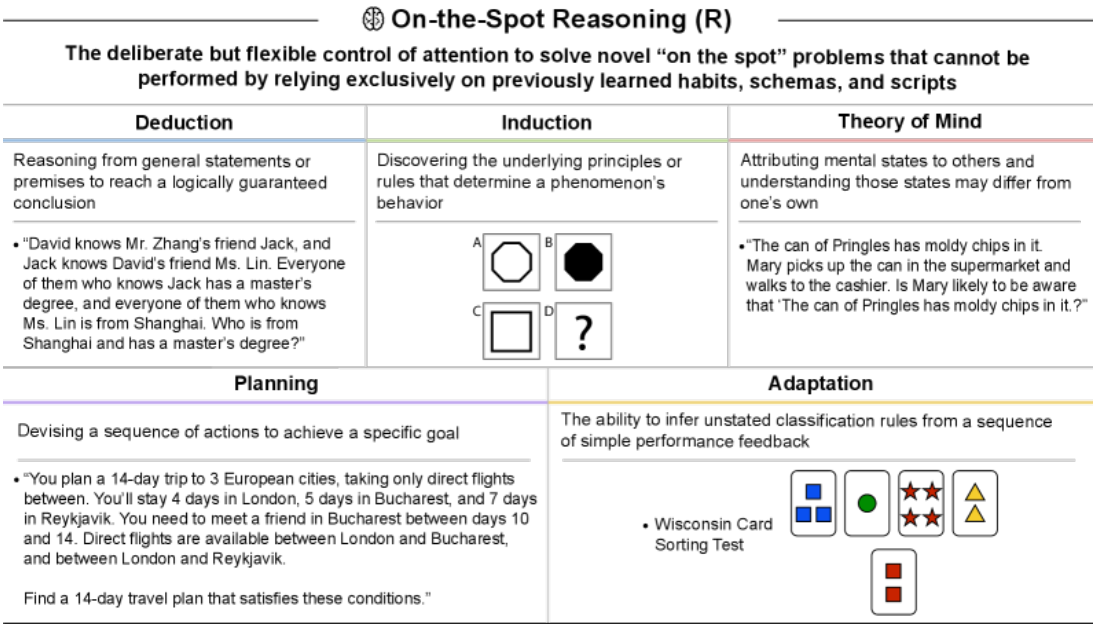


Figure 28: *Each cognitive capability breaks down into more specific measurable components. Reasoning divides into deduction, induction, theory of mind, planning, and adaptation - each can then be individually benchmarked and quantified (Hendrycks et al., 2025).*

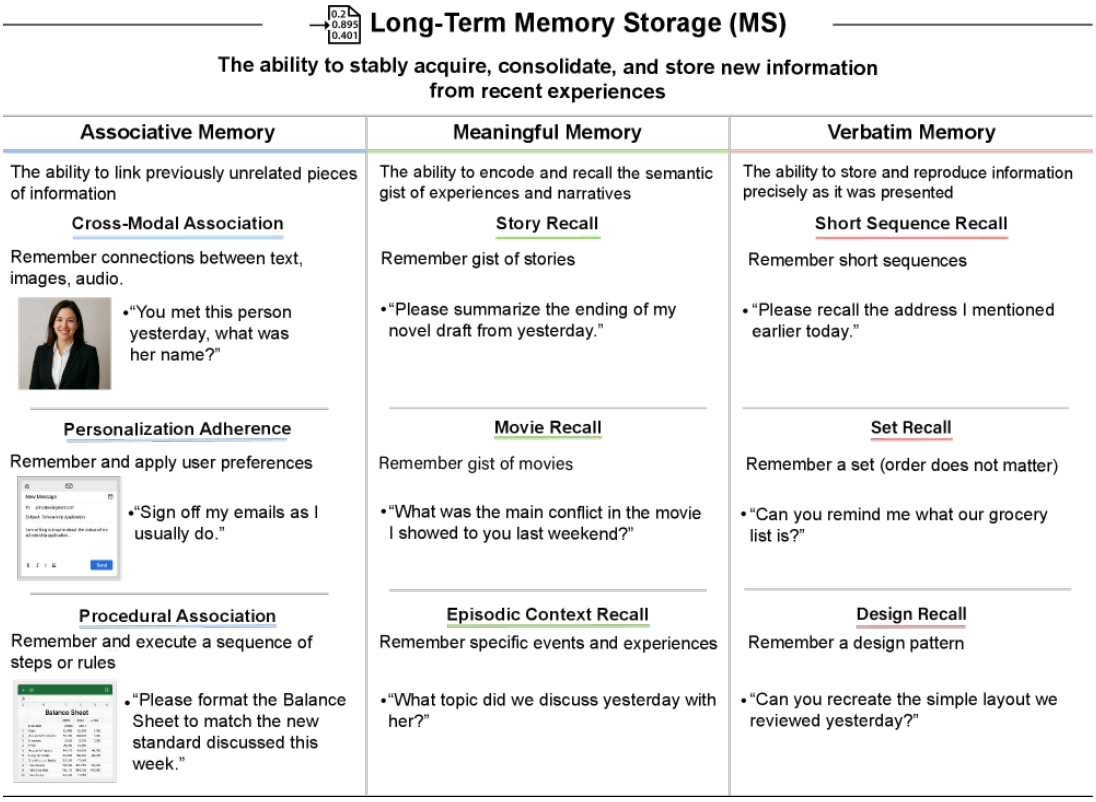
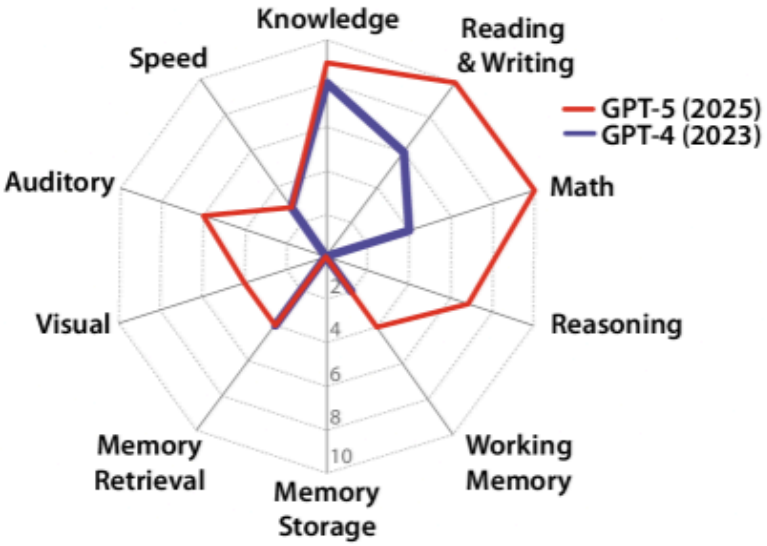


Figure 29: *Memory similarly splits into associative, meaningful, and verbatim components, allowing precise diagnosis of where systems succeed versus fail (Hendrycks et al., 2025).*

Generality is the percentage of these domains where a system achieves expert-level capability. If a system scores at the 80th percentile or higher on three out of ten domains, that's 30% generality.[4] Foundation models dramatically increased this compared to traditional narrow AI - one model handling writing, math, coding, and visual understanding represents unprecedented breadth. But current systems still cover only a fraction of cognitive capabilities, with particularly weak performance on long-term planning and memory-related domains (Hendrycks et al., 2025 ; Kwa et al., 2025).

[4]You can use expert-level parity (roughly 80-90th percentile human performance) as the threshold for "adequate" in measuring generality. Different researchers might set this threshold higher or lower depending on their specific concerns.



Model	K	RW	M	R	WM	MS	MR	V	A	S	Total
GPT-4	8%	6%	4%	0%	2%	0%	4%	0%	0%	3%	**27%**
GPT-5	9%	10%	10%	7%	4%	0%	4%	4%	6%	3%	**57%**

Figure 30: *The capabilities of GPT-4 and GPT-5, alongside a table that quantifies them (Hendrycks et al., 2025).*

Even though capability and generality exist on a continuum, certain thresholds still matter for safety planning. A system performing well on 50% of domains poses different risks than one excelling at 90%, we can't ignore this reality even if generality is a continuous variable. When pressed for concrete thresholds - which are often demanded in discussions - here's our interpretation of roughly how terms map onto the (performance, generality) space:

ARTIFICIAL GENERAL INTELLIGENCE (AGI) Hendrycks et al., 2025 Morris et al., 2024

Matching a well-educated adult's cognitive versatility and proficiency. Often used interchangeably with Human level AI (HLAI). Using our definition, these would be systems achieving expert-level performance (80-90th percentile) across most cognitive domains (80-90%).

TRANSFORMATIVE AI (TAI) Karnofsky, 2016

AI capable of triggering economic and social transitions comparable to the agricultural or industrial revolution. TAI is defined by impact potential rather than cognitive architecture. Using our definition, this could mean moderate capability (60th percentile) across many economically

important tasks (50% of domains), OR exceptional capability (99th percentile) on critical domains like automated ML R&D (20% of domains).

ARTIFICIAL SUPERINTELLIGENCE (ASI) Bostrom, 2014

Any intellect that greatly exceeds human cognitive capability across virtually all domains of interest. Using our definition, this represents systems achieving superhuman capability (>100%, greatly exceeding all humans) across virtually all cognitive domains (95%+ of domains).

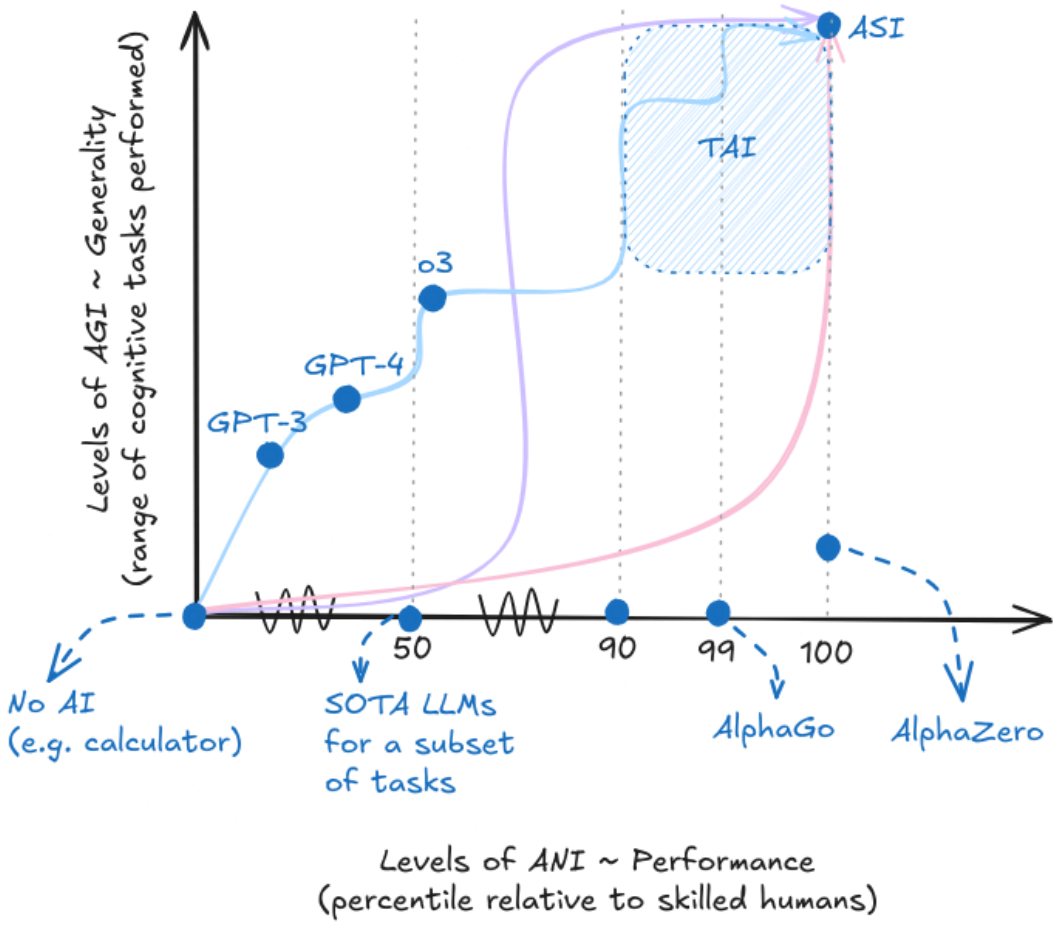


Figure 31: *The two-dimensional view of capability × generality. Different colored curves represent possible development paths to ASI. Every point on these paths corresponds to a different level of AGI capability.*

Our working definition does not include the autonomy with which an AGI system operates. This is a really important axis to pay attention to, but it has a higher bearing on deployment and impact rather than something inherent to a definition for AGI. It will be explored in the next chapter dedicated to risks from AI.

Measuring degree of Autonomy (Agency)

OPTIONAL NOTE

Autonomy describes how AI systems interact with humans, not what they can do. A highly capable system can be deployed with varying levels of human oversight. Just like we had continuous curves for capability and generality, we can similarly have an increasing level of autonomy measured by what % of a task is done by the human or the AI (Morris et al., 2024):+ Level 0: No AI - Pure human operation. Still relevant for education, assessment, or safety-critical situations even after capable AI exists.

1. Level 1: AI as Tool - System provides suggestions or information, human makes all decisions. Like autocomplete or spell-check.
2. Level 2: AI as Consultant - System provides expert advice and recommendations, human still directs overall approach. Like asking an AI for code review or strategic analysis.
3. Level 3: AI as Collaborator - Human and AI work together on equal footing, dividing tasks based on respective strengths. Requires the AI to understand when to ask for help.
4. Level 4: AI as Expert - AI handles most execution, human provides oversight and intervenes when needed. Like self-driving cars with human monitoring.
5. Level 5: AI as Agent - Fully autonomous operation with minimal human oversight. The AI decides when to consult humans. Requires strong alignment.

Autonomy level affects risk exposure, not inherent danger. A capable system deployed as a tool (Level 1) might be safer than the same system deployed as an agent (Level 5), even though the underlying capability is identical. Higher capability levels "unlock" higher autonomy levels—you can't have Level 5 autonomy without sufficient capability—but having the capability doesn't mean you should use maximum autonomy.For safety purposes, capability and deployment autonomy should be considered separately. A system scoring 90% on 80% of domains might be safely deployed at Level 2 (consultant) while being dangerous at Level 5 (agent). The framework helps us reason about these tradeoffs explicitly.

What we presented in this section is nowhere near a universally agreed upon definition. There are criticisms of both the levels of AGI framework, and of the CHC benchmark aggregation approach (Pacchiardi et al., 2025). There are many other alternative definitions and measurement frameworks in use like the European Commission's work on definition General Purpose AI (GPAI) models (EU Commission, 2025), the (t,n-AGI) framework (Ngo, 2023), OECD AI capability indicators (OECD, 2025), and not to mention all the competing definitions from the case studies in the previous sub-section. These are valid alternatives. The field is still evolving and finding gaps in its approach. We expect that the underlying benchmarks (which cognitive abilities are important?) and aggregation methods (how should we weigh them?) used to measure progress will continue to be debated but the underlying framework we teach in this section - of thinking of capability and generality as a continuous spectrum - will remain consistent.

The next sections of this chapter, look at why capabilities and generality have been improving rapidly in the recent years (Scaling), and what we can say about where they might be headed (Forecasting and Takeoff). Thinking about AI progress in this continuous way will help throughout this book when we discuss risks and mitigation strategies. Certain capability combinations might enable dangerous emergent behaviors even before reaching "human-level" on most tasks. Similarly governance frameworks, and clear policy communication depend on precise measurement to trigger appropriate responses. The rate of improvement along either axis provides important signals about which risks are most pressing, and which safety mitigations need to be developed.

Counter Argument: Arbitrary thresholds aggregated and framed as a misleading percentage

OPTIONAL NOTE

This framework enables concrete measurement, critics note that human psychometric taxonomies like CHC may miss capabilities universal in humans but lacking in AI, and that percentage scores can misleadingly suggest linear progress toward AGI.Counter arguments note that benchmark scores do not align with general perception. Generally if you let an LLM (even with extended thinking or tools), perform a task you would expect an educated human adult to complete, they tend to disappoint. An LLM scoring 90% on a Raven's progressive matrices test is not comparable to a human scoring 90%. This matters because AI systems could fall short of "100% AGI" despite already having all the capabilities of an educated human adult.Some examples of limitations include (Pacchiardi et al., 2025):+ **Coverage bias:** CHC derives from human interindividual differences, potentially omitting capabilities universal in humans (e.g., robust object permanence) that AI might lack;

1. **Arbitrary thresholds:** Human percentile cutoffs (e.g., 85th) may not translate to machines;
2. **Non-linear progress:** Percentage scores (e.g., "57% AGI") misleadingly imply linear progress when final capabilities may represent disproportionately difficult bottlenecks; and
3. **Species-specific risk:** Non-human-like AI profiles may pose transformative risks despite low scores on human-centric tests. These critiques suggest interpreting percentage scores cautiously, focusing on specific capability gaps rather than aggregate progress.

The (t,n)-AGI Framework: An Alternative way of defining AGI

OPTIONAL NOTE

AGI can also be defined through a combination of time and scale - can AI match 'n' experts working together for time 't'. This is an alternative way to think about defining AGI. Given a time frame 't' to complete some cognitive task, if an AI system can outperform a human expert who is also given the time frame 't' to perform the same task, then the AI system is called t-AGI for that timeframe 't'. If a system can outperform 'n' human experts working on the task for timeframe 't', then we call it a (t,n)-AGI for the specific time duration 't', and number of experts 'n'. The (t,n)-AGI framework does not account for how many copies of the AI run simultaneously.As an example, if an AI that exceeds the capability of a human expert in one second on a given cognitive task would be classified as a "one-second AGI". One-year AGI would beat humans at basically everything. Mainly because most projects can be divided into sub-tasks that can be completed in shorter timeframes. Within this framework, a superintelligence (ASI) could be something akin to a (one year, eight billion)-AGI, that is, an ASI could be seen as an AGI that outperforms all eight billion humans coordinating for one year on a given task (Ngo, 2023). Researchers at METR operationalized this by measuring task completion time horizons for 1 expert (n=1) - finding the duration where AI succeeds consistently on professional tasks like software development and ML research (Kwa et al., 2025).

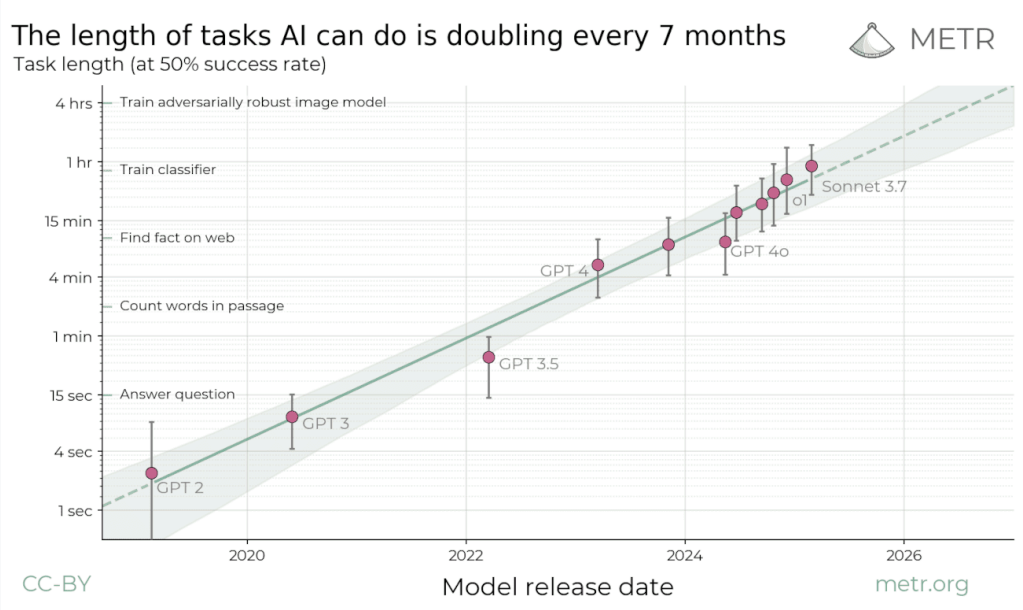


Figure 32: *The image shows a result from METR which measured the task length horizon on specific software engineering tasks (Kwa et al., 2025).*

Task time horizon limitations are implicitly captured in our generality measure as weaknesses in specific cognitive tasks like long-term memory storage and retrieval. Systems scoring poorly on memory naturally fail on multi-day projects requiring sustained context. Current transformer-based architectures specifically struggle with limited context windows.

5. Leveraging Scale

In the previous section, we looked at different notions of how to define AGI and measure the relevant capabilities. Now we'll examine one of the most important drivers behind improvements in these capabilities: scale.

5.1 Bitter Lesson

Human-engineered domain knowledge consistently loses to general methods that leverage massive computation. We assume that most of you probably went to university in an era where machine learning and AI roughly mean the same thing, or even that deep learning and AI mean the same thing. This hasn't always been true. Early in AI's history, researchers believed the key to artificial intelligence was carefully encoding human knowledge and expertise into computer programs. This led to expert systems filled with hand-crafted rules and chess engines programmed with sophisticated strategic principles. Time and again, these approaches hit walls while simple learning algorithms combined with massive computation kept improving. However, time and time again, researchers learned what we now call the bitter lesson.

> *The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. [...] The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.*
>
> **Richard Sutton**
> Professor University of Alberta, Founder, Openmind Research Institute
>
> **2019**
> Sutton, 2019

The bitterness comes from discovering that decades of human expertise mattered less than computation. Researchers who spent years encoding grandmaster chess knowledge watched brute-force search defeat world champion Garry Kasparov. Hand-crafted feature detectors in computer vision got outperformed by neural networks that learned their own features from data. Phonetics-based speech recognition lost to statistical approaches. The pattern repeated: domain expertise helped initially, then hit a wall. Simple learning algorithms plus massive compute kept improving (Sutton, 2019).

The bitter lesson doesn't reject human ingenuity or algorithmic innovation. There's a difference between building better general learning systems and encoding task-specific human knowledge. This isn't a difference just between good old fashion AI (GoFAI), and modern deep

learning . Even within deep learning , the bitter lesson still applies—the winning algorithms are those that leverage scale most effectively. Transformers didn't beat LSTMs by encoding linguistic knowledge. They outperformed because attention mechanisms parallelize better and can actually use massive compute productively. Algorithmic innovation still matters - finding architectures and training methods that extract more from the same data and hardware. The algorithms that succeed are the ones that unlock scale's potential.

The bitter lesson shapes expectations about AI progress. If the bitter lesson continues to be true, improvements should come from either finding algorithms that better leverage scale, or simply scaling existing algorithms with more compute, data, and parameters. In the last few years, the majority of gains seen in AI capabilities have emerged from scaling up the same transformer based language models.

5.2 Scaling Laws

Training frontier AI models costs hundreds of millions of dollars, making it critical to predict returns on investment. AI labs face resource allocation decisions with massive stakes: should they spend more on GPUs or training data ? Train a larger model briefly or a smaller model longer? With a fixed compute budget, they might choose between a 20-billion parameter model trained on 40% of their data or a 200-billion parameter model trained on 4% of it. Getting this wrong wastes hundreds of millions. Scaling laws help turn these gambles into engineering decisions by establishing empirically observed relationships between inputs and model accuracy.

Scaling laws describe how model accuracy changes as you vary four key variables:

- **Compute**: Total floating-point operations (FLOPs) during training - accounts for GPU power, number of chips, and training duration (how many steps you train for).
- **Parameters**: The numbers the model adjusts during training - roughly analogous to "model size".
- **Data**: Training examples seen, measured in tokens for LLMs.
- **Accuracy**: How well the model performs on benchmarks - the inverse of "loss" (lower loss = higher accuracy).

Figure 33: *Example of capabilities increasing with an increase with one of variables in the scaling laws - parameter count. The same model architecture (Parti) was used to generate an image using an identical prompt, with the only difference between the models being the parameter size. There are noticeable leaps in quality, and somewhere between 3 billion and 20 billion parameters, the model acquires the ability to spell words correctly (Yu et al., 2022).*

Scaling laws are empirically observed relationships, not laws of nature. OpenAI first documented these relationships in 2020 by running hundreds of experiments, varying inputs while measuring accuracy (Kaplan et al., 2020). They found that when you increase compute by 10×, accuracy improves predictably. Double the parameters, accuracy jumps predictably. These patterns proved surprisingly consistent across model architectures and tasks, suggesting they capture something fundamental about how neural networks learn. Later research revealed optimal training requires roughly 20 tokens of data per parameter - about 10× more data than early laws suggested[5] (Hoffmann et al., 2022). This meant previous large models were undertrained relative to their size. The relationships continue evolving as researchers gather more evidence, but the core insight remains: scale drives predictable capability gains. The following graphs clearly show massive increases in scale for data, compute, and parameter count by all major AI labs.

[5]This is also commonly called 'chinchilla optimality' or the chinchilla optimal frontier based on the original model that these laws were tested on.

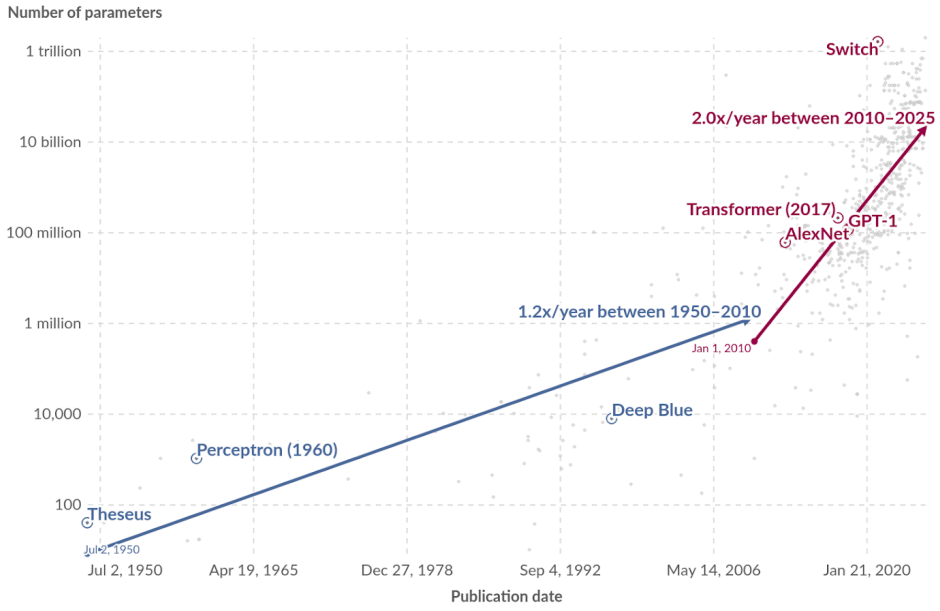


Figure 34: *Exponential growth of parameters in notable AI systems. Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output; for example, the connection weights in an artificial neural network (Giattino et al., 2023).* (interactive version on website)

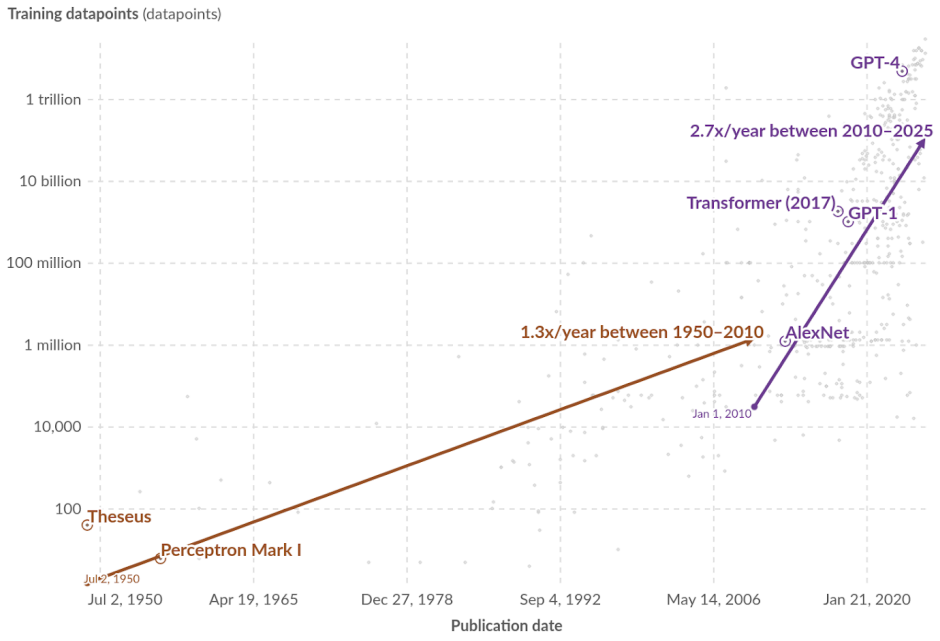


Figure 35: *Exponential growth of datapoints used to train notable AI systems. Each domain has a specific data point unit; for example, for vision it is images, for language it is words, and forgames it is timesteps. This means systems can only be compared directly within the same domain (Giattino et al., 2023).* (interactive version on website)

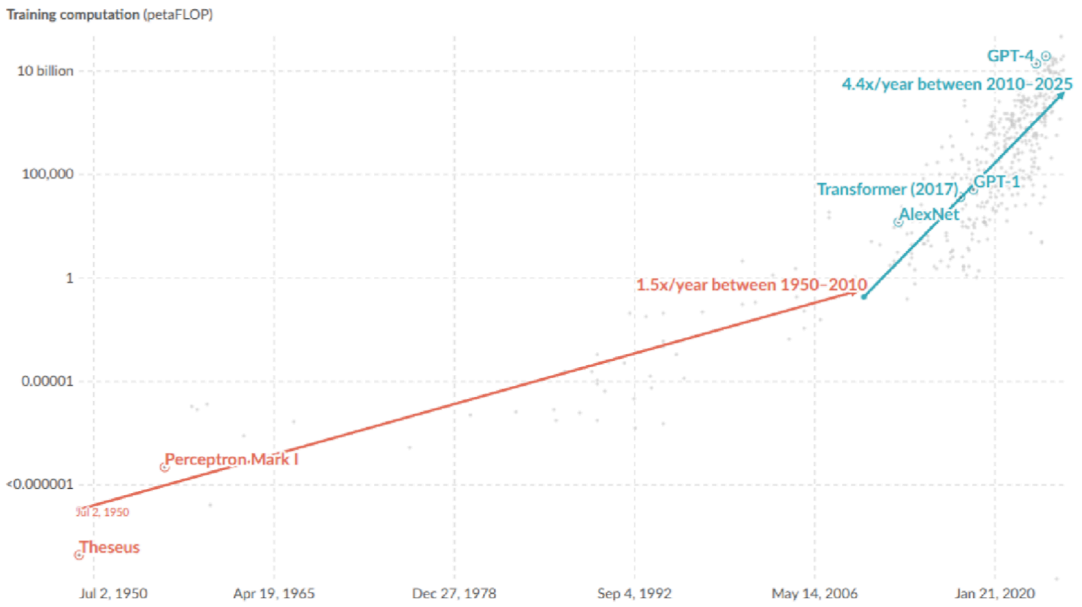


Figure 36: *Exponential growth of computation in the training of notable AI systems. Computation is measured in total peta FLOP, which is 10e15 floating-point operations (Giattino et al., 2023).* (interactive version on website)

The Broken Neural Scaling Laws (BNSL) update in 2023

OPTIONAL NOTE

Research showed that performance doesn't always improve smoothly - there can be sharp transitions, temporary plateaus, or even periods where performance gets worse before getting better. Examples of this include things like "Grokking", where models suddenly achieve strong generalization after many training steps, or deep double descent, where increasing model size initially hurts then helps performance. Rather than simple power laws, BNSL uses a more flexible functional form that can capture these complex behaviors. This allows for more accurate predictions of scaling behavior, particularly around discontinuities and transitions. Scaling laws are a good baseline, but discontinuous jumps in capabilities and abrupt step changes are still possible (Caballero et al., 2023).

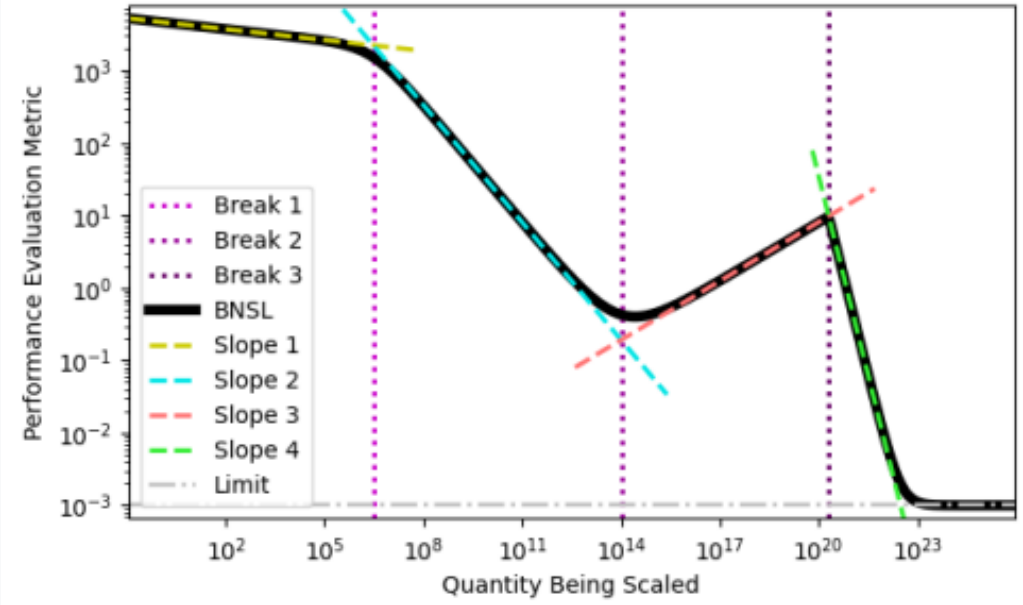


Figure 37: *A Broken Neural Scaling Law example (dark black solid line) (with 3 breaks where purple dotted lines intersect with dark black solid line) contains 4 individual power law segments (where the dashed lines that are yellow, blue, red, and green overlap with the dark black solid line). The 1st and 2nd break are very smooth; the 3rd break is very sharp (Caballero et al., 2023).*

5.3 Scaling Hypothesis

Scaling might continue driving capability gains. Look back at the examples from previous sections - the programming abilities, emergent reasoning, scientific research assistance, the jump from GPT-3.5 to GPT-4 across professional exams. These capabilities appeared as models got bigger and trained on more data, without requiring new architectural breakthroughs or encoded domain knowledge. If current approaches scaled up could produce systems capable of automating AI research within years, safety work becomes far more urgent. The existing evidence supports multiple interpretations because we're watching a technology develop in real-time whose limits we don't fully understand. So different people hold different hypotheses about how the future could unfold.

The strong scaling hypothesis. This proposes that simply scaling up existing architectures with more compute and data will be sufficient to reach transformative AI capabilities (Gwern, 2020). According to this view, we already have all the fundamental components needed - it's just a matter of making them bigger, following established scaling laws.

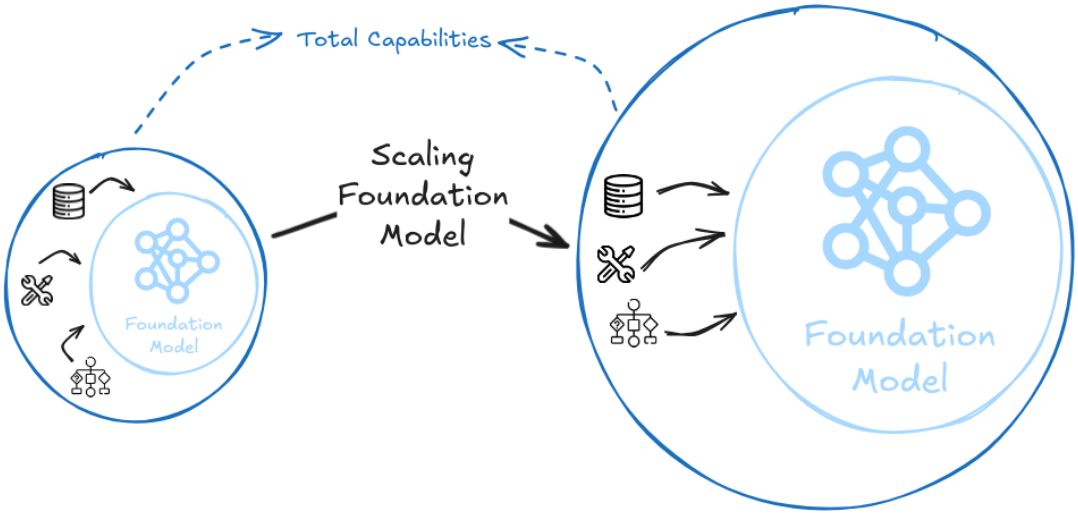


Figure 38: *Augmentation/Scaffolding stays constant, but if the scaling hypothesis, weak or strong, is true, then capabilities will keep improving just by scaling.*

The weak scaling hypothesis. This view states that even though scale will continue to be the primary driver of progress, we will also need targeted architectural and algorithmic improvements to overcome specific bottlenecks. These improvements wouldn't require fundamental breakthroughs, but rather incremental enhancements to better leverage scale (Gwern, 2020).

Researchers have been developing algorithms that leverage scale and compute for more than a decade. We have seen many gains come from improvement in compute efficiency from innovations like better attention mechanisms, mixture-of-experts routing, and efficient training methods. But even when researchers have developed sophisticated algorithms following the bitter lesson's principles, data suggests that between 60-95% of performance gains came from scaling compute and data. While algorithmic improvements contributed 5-40%, though there is substantial methodological uncertainty in disentangling these contributions (Ho et al., 2024).

The emergence of unexpected capabilities might provide another argument for strong scaling. We've seen previous generations of foundation models demonstrate remarkable abilities that weren't explicitly trained for, like programming. This emergent behavior hints that it is not impossible for higher-order cognitive abilities to similarly emerge simply as a function of further scale.

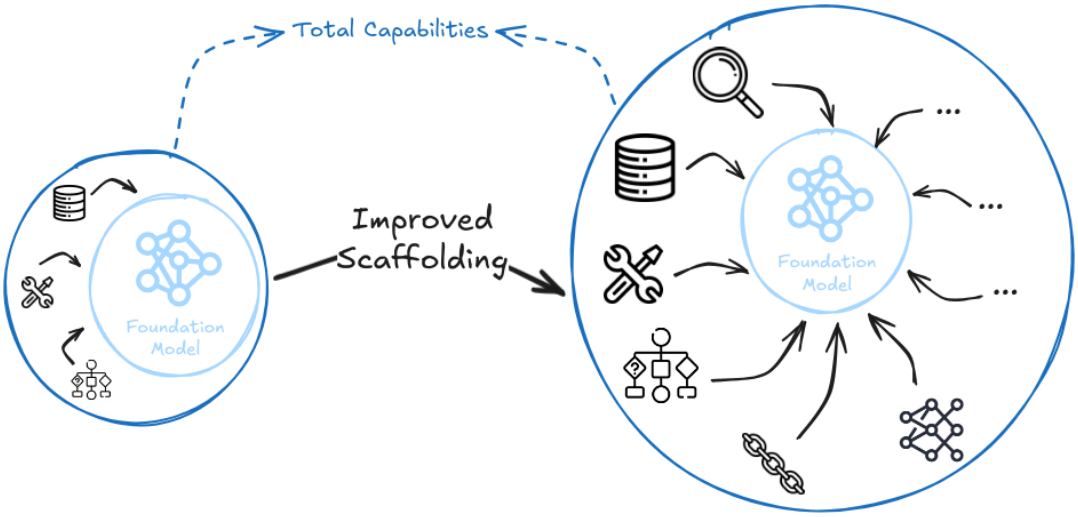


Figure 39: *Even if we see no improvements in model scale, other elicitation techniques and scaffolding can keep improving. So overall capabilities keep growing. Realistically, the future is probably going to see both improvement due to scaffolding and scale. So for now, there does not seem to be an upper limit on improving capabilities as long as either one of the two holds.*

Scale combined with techniques and tools hypothesis. Essentially, both the scaling laws (which only predict foundation model capabilities) and most debates around "scale is all you need" often miss other aspects of AI development that happen outside the scope of what scaling laws can predict. They don't account for improvements in AI "scaffolding" (like chain-of-thought prompting, tool use, or retrieval), or combinations of multiple models working together in novel ways. Any LLM with internet access, code execution, and the ability to call upon the help of other specialized sub-models has substantially more capability than the same LLM alone. We gave several examples of this being the dominant trend in our first section - tool use, thinking for longer (inference time scaling), MCP servers and so on.

Debates around the scaling laws only tell us about the capabilities of a single foundation model **trained in a standard way.** For example, by the strong scaling hypothesis we can reach TAI by simply scaling up the same foundation model until it completely automates ML RnD. But even if scaling stops, halting capabilities progress on the core foundation model (in either a weak or a strong way), the external techniques that leverage the existing model can still continue advancing. Many researchers think that this is a core element where future capabilities will come from. It is also referred to as "unhobbling" (Aschenbrenner, 2024), "schlep" (Cotra, 2023) and various other terms, but all of them point to the same underlying principle - raw scaling of single model performance is only one part of overall AI capability advancement.

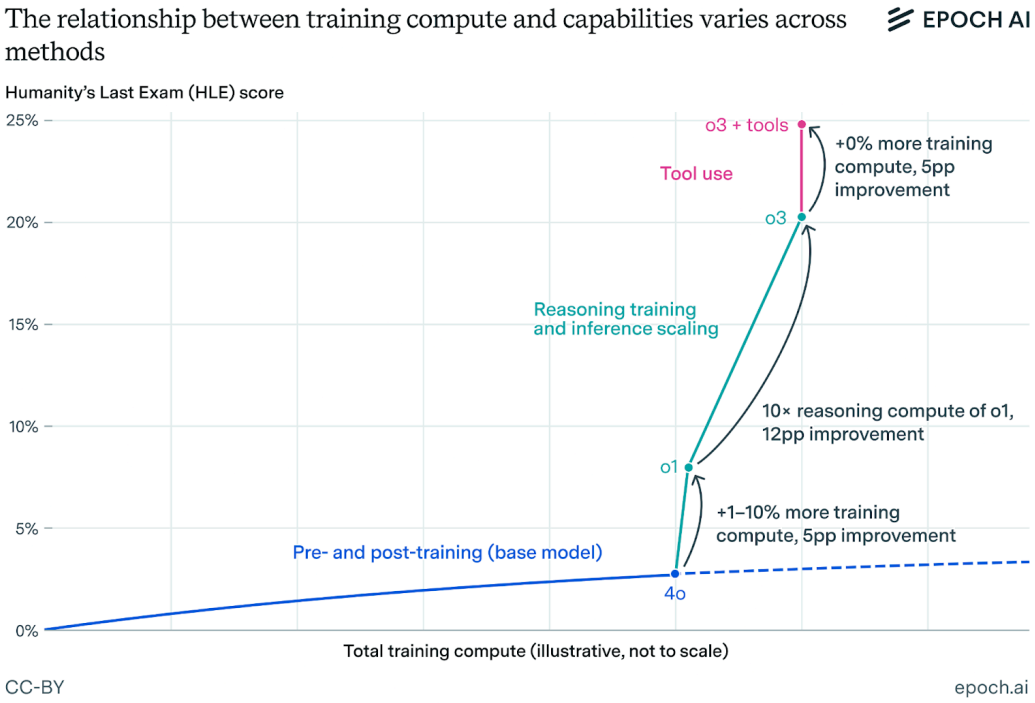


Figure 40: *An example of how much the performance on a benchmark can change just by using post training techniques like doing reasoning specific training and allowing the model to think for longer. The base model scores 4.5% on the humanities last exam benchmark (HLE), whereas with each subsequent "unhobbling" step we see jumps in performance leading up to a 25% score on the benchmark (Somala et al., 2024).*

Even the tool based scaling hypothesis is debated. Some argue that tools or scale poured into LLMs is unlikely to lead to AGI as they define it. They advocate for completely different architectures from transformer based LLMs, focusing instead on things like neuro-symbolic approaches, ensemble methods, or a completely new undiscovered architecture (Goertzel et. al., 2023 ; Marcus, 2025 ; LeCun, 2025 ; Chollet, 2025).

Despite disagreements about whether scale will lead to "true AGI", major AI labs are betting heavily on scaling. Sam Altman from OpenAI has stated his belief that scaling is going to be a big component leading to capability gains (Altman, 2023), Anthropic CEO Dario Amodei has expressed similar views (Amodei, 2023) and DeepMind's safety team similarly wrote that " *not many more fundamental innovations are needed for AGI* " (DeepMind, 2022). This consensus suggests that regardless of whether strong, weak, or tools-based scaling dominates, scale itself will likely remain central to near-term progress.

6. Forecasting Timelines

Predicting when AI automates cognitive labor determines which safety strategies are viable. The difference between 10 years and 50 years fundamentally changes which safety strategies are viable and which risks demand immediate attention . If TAI is able to automate cognitive labor (and thereby automate ML RnD) by 2030, we need safety solutions working soon. There's no time for slow theoretical research—we need approaches that work with current systems and scale quickly. If it arrives around 2050, we have breathing room for fundamental alignment research, for building robust governance frameworks, and for developing evaluation methods we can trust.

This section presents empirical trends and forecasts of AI progress using data-driven analysis. Our goal is to build on the scaling laws discussion and present the strongest available evidence so you can form informed views on where AI might be headed. We focus primarily on compute-based trends, with heavier quantitative forecasts presented alongside graphs. An appendix covers judgment-based surveys and expert opinions. Remember the old saying: "All models are wrong, but some are useful."

Forecasting helps us form internally consistent beliefs about different possible futures. The way that you should think about forecasting methods is not as concrete predictions of timelines; but instead as scenarios that help you reason about different potential futures. They let you check if your beliefs are internally coherent. If you think investment grows 5x per year AND software efficiency improves 3x per year AND hardware scales 15% per year AND each order of magnitude automates at least 10% of tasks, then what does that imply? Is the outcome of this combination consistent with the future you expect? If you're surprised, maybe one of your input beliefs needs updating.

Current evidence suggests scaling can continue through at least 2030. Training compute has grown 5x per year since 2020, dataset sizes have grown by 3.7x/year, and training costs by 3.5x/year. As a concrete example, this means that in mid 2025 the cost of training a single AI model (Grok-4) including hardware and electricity, is estimated at $480 million USD (Epoch AI, 2025). By 2030, this trajectory points to a single training run requiring hundreds of billions of dollars and gigawatts of power - which is comparable to running a small city (Owen, 2025). These are unprecedented scales, but analysis suggests that they aren't impossible (Epoch AI, 2025 ; Sevilla et. al., 2024 ; Owen, 2025). The question is whether the economic incentives justify it and whether the infrastructure constraints (chips, power, data, ...) can be overcome in time.

AI systems that leverage more compute during training and inference are capable of automating progressively more tasks. Initially, only a small fraction of tasks are automated, so running AI systems creates relatively little economic value.[6] Once effective compute budgets grow large enough, the range of automated tasks expands until eventually all cognitive labor becomes automated. This feedback loop—compute enabling automation, which increases productivity, which funds more compute—continues until full labor automation.

[6]You can think of a task as a specific unit of work or activity that a worker performs as part of their job (for instance tasks for a factory worker may include operating machinery, inspecting products, and coordinating with team members). Some tasks are assumed to be harder to automate than others, as they require more training and runtime compute to be effectively automated.

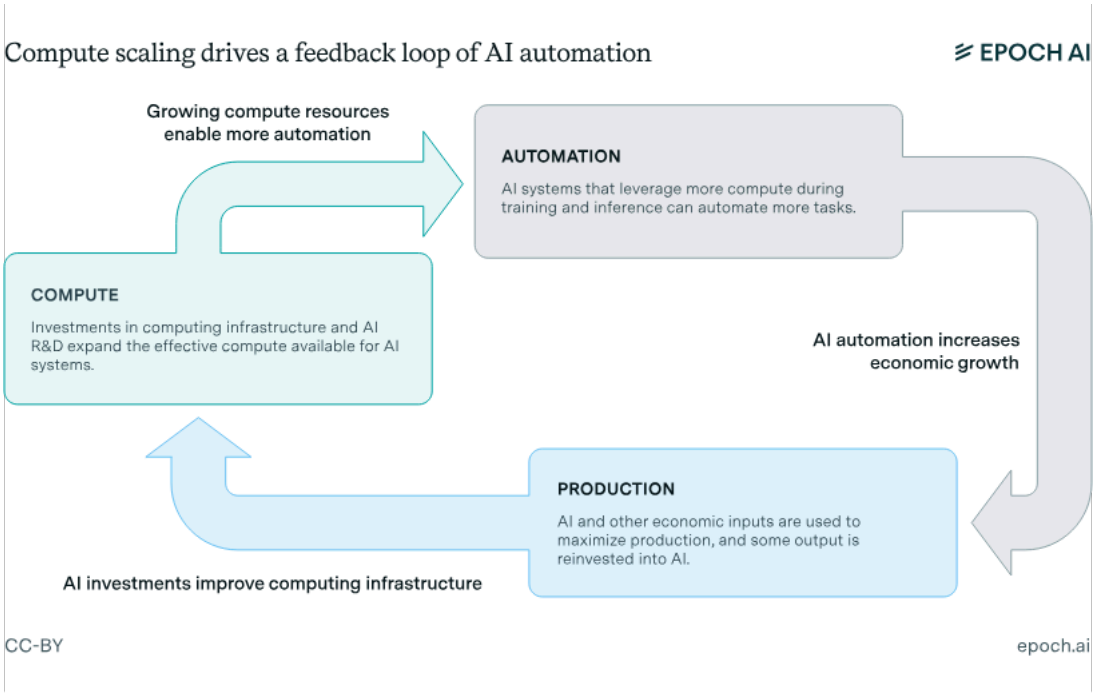


Figure 41: *The feedback loop of compute leading to more automation, which increases productivity, which increases both the amount of produced compute as well as compute efficiency completing the loop until all labor is automated (Epoch AI, 2025).*

The path to full labor automation runs through three resources: compute, data, and the economic returns that fund both. In the subsections below, we examine each constraint. For compute, we look at effective FLOP growth—the combination of more chips, better algorithms, and more efficient hardware. For data, we track when we exhaust the internet's text and what happens after. Each section answers: does this constraint stop scaling before 2030? The appendix provides detailed breakdowns of the other relevant factors like investment trends, semiconductor production, and power infrastructure for readers who want the full picture.

All forecasts come with wide error margins that compound over time. Even if we assume compute-based scaling laws hold, we still don't know how much compute TAI will actually require. One framework for estimating the compute required for general intelligence uses "biological anchors," treating the human brain as a proof of concept. These estimates vary dramatically depending on the biological process used as a reference: a lower bound considers the compute performed in learning skills during a human lifetime (10^{28} FLOP), while a theoretical upper bound accounts for the total compute performed in shaping humans over evolutionary history (10^{41} FLOP) - twelve orders of magnitude of uncertainty (Carlsmith, 2020 ; Cotra, 2020 ; Ho, 2022). It's like not knowing if something costs one dollar or a trillion dollars. When someone says " *AGI in 2045* ," they mean " *somewhere in the 2030s-2050s range, with 2045 as a rough center* " - not a precise date. The same principle applies to all trends and forecasts presented through this section; for example, projections of available chips range by a factor of 20×, and some compute estimates span 50× (Sevilla et al., 2024).

6.1 Effective Compute

Effective compute combines three independent factors that each multiply your AI capabilities. Scaling laws track performance against compute, but "compute" means more than just counting GPUs. When researchers say GPT-4 used more compute than GPT-3, they're not just talking about running more hardware for longer. They're talking about the growth in 'effective' compute measured in 'effective' floating-point operations (eFLOPs). As an analogy, think of it like measuring how far a fleet of cars can travel. You need three things: how many cars you have (chip production), how fast each car goes (hardware efficiency), and how efficiently you're driving (software efficiency) (Erdil et al., 2025). The relationship looks like this:

$$\text{Effective compute} = \text{Software efficiency} \times \text{Hardware efficiency} \times \text{Number of chips}$$

As an illustrative example, think about wanting to train a model that gets 90% accuracy on some benchmark. In 2020, that might have required 1 billion FLOP using the algorithms available then. But by 2025, better training methods, hardware utilization and other algorithmic improvements mean you only need 100 million FLOP to hit that same 90% accuracy. Your software got 10× more efficient. Better algorithms multiply your effective compute without building a single new chip. Meanwhile, the physical efficiency of your chips got better too. A 2025 GPU does more FLOP per dollar than a 2020 GPU - that's hardware efficiency. And you're also just making more chips overall as more investment keeps flowing in. So the total "effective" FLOPs changes as a function of all three of these variables.

Each factor can improve independently, which is why effective compute grows faster than any single input. You can build more chips without improving their speed. You can develop better software to make use of existing hardware. You can design faster processors without changing either your software or building more chips. When all three improve simultaneously—as they have been—the total compute available (effective compute) compounds much faster than hardware trends alone would suggest. Here's are the trends for each in 2025 (Epoch AI, 2025), with more details in the appendix:

1. **Hardware efficiency**: GPU performance grows at 1.35x per year. Each new chip generation does more FLOP/s for the same cost.
2. **Algorithmic efficiency**: Software improvements cut the compute needed for a given result by 3x per year. Better training methods mean you need less hardware to hit the same benchmark.
3. **Chip production**: The number of AI chips produced has grown 2.3x per year since 2019, driven by semiconductor fab expansions and massive capital investment.

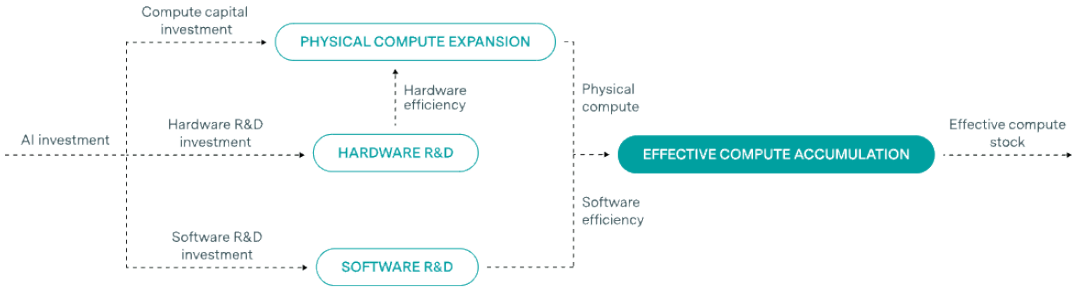


Figure 42: *Growth in effective compute breaks down into three parts: growth in computing hardware, improvements in hardware efficiency, and software efficiency gains.*

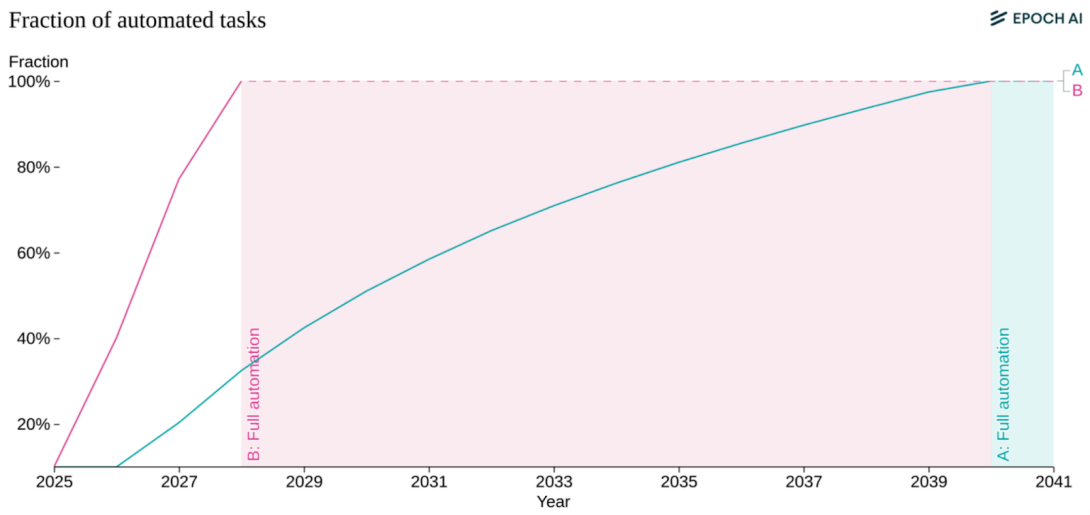


Figure 43: *The GATE model by EpochAI translates a given stock of effective compute into a degree of AI automation. This occurs by expanding the fraction of tasks that can be automated and by increasing the effective runtime compute on tasks that are already automated. Widespread automation often happens within two decades from the start of the simulation. Note that GATE uses the abstract notion of a "fraction of economically useful tasks" as a simplification, and does not specify the tasks and the order in which they are automated. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

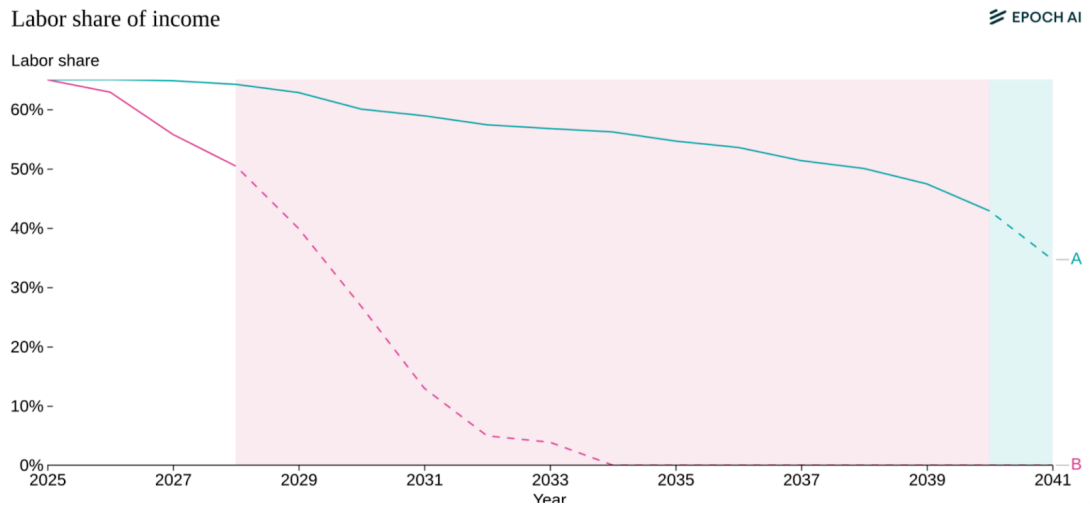


Figure 44: *Simulation of labor share of income over the next decades, using the GATE model. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models. The fraction of Gross World Product (GWP) paid out to human workers rapidly decreases as AI becomes the primary source of labor (Epoch AI, 2025). We talk about wages falling below subsistence levels in the next chapter under automation/systemic risks.*

6.2 Training Data

Besides just effective compute, the second overarching factor that we have to take into consideration is the amount of available data to train on.

The training dataset size for language models has grown by 3.7x per year since 2010. The internet has maybe 30 years worth of text data at current consumption rates—but we'll hit that wall around 2028. The indexed web contains roughly 500 trillion tokens of text (after removing duplicates). The largest models in 2024 train on about 15 trillion tokens. If we keep scaling at 4× per year, projections say that we will exhaust high-quality public text data between 2026 and 2032 (Epoch AI, 2023 ; Villalobos et al., 2024). Three escape routes exist, whether data actually constrains scaling through 2030 depends on how well these alternatives work.:

- Multimodal **data:** The internet has 10 trillion images and 10 trillion seconds of video. If encoded efficiently, this could 3-10x the effective data supply.
- **Synthetic data:** AI-generating it's own training data removes the constraint entirely—if the outputs are high enough quality. Post-training methods like reinforcement learning on reasoning tasks already demonstrate this works for some domains.
- **Task-based learning:** Self-play reinforcement learning can generate training data through environment interaction. This means AIs can learn without human-generated pretraining examples as long as tasks can be formalized as games with explicit rules and success metrics (like Go or Chess). For example, AlphaZero learned superhuman strategies without any human gameplay examples, using only the game rules and billions of self-played matches. In principle, this approach could be extended to teaching any arbitrary task allowing for data free learning through self-play on specified tasks. This is still very speculative, it requires simulation environments encoding task dynamics, explicit reward functions, and tractable state spaces.

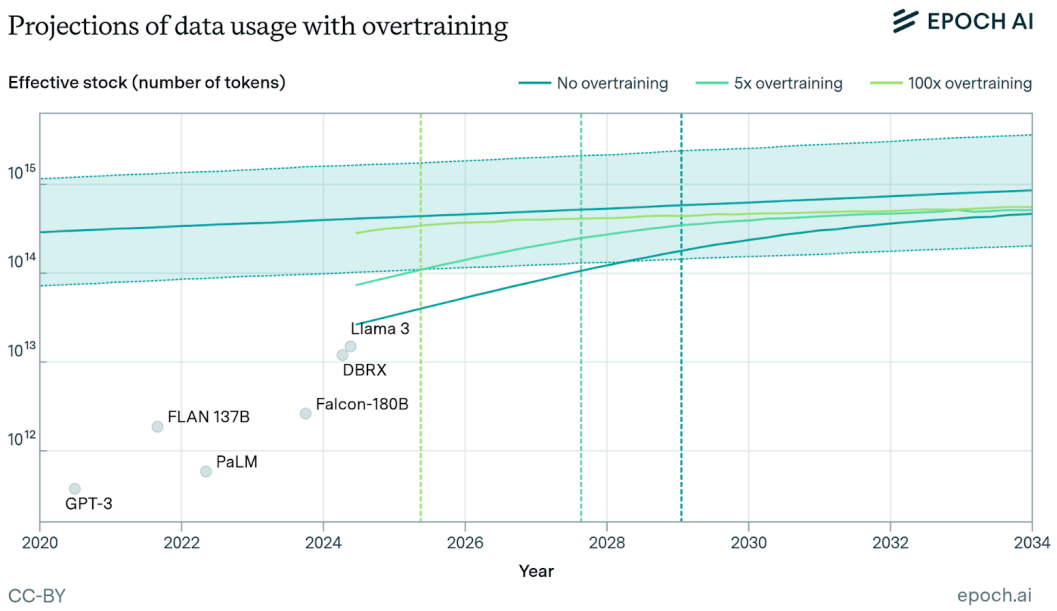


Figure 45: *Projections of future dataset sizes according to three different scaling policies. Depending on the degree of overtraining, the stock is fully used between 2025 and 2030 (Villalobos et al., 2024).*

7. Takeoff

This final section synthesizes a lot of the discussion that has happened through this chapter. We started from where we are currently, and went all the way to forecasting specific trends in the inputs to AI capabilities. So this section lays out different sides of the debate on what the combination of all of this implies.

> *There is no question that machines will become smarter than humans—in all domains in which humans are smart—in the future. It's a question of when and how, not a question of if.*
>
> **Yann LeCun**
> Chief AI scientist at Meta and Turing Prize winner
>
> **May 2023**
> Heaven, 2023

Takeoff speed refers to how quickly AI systems become dramatically more powerful than they are today and cause major societal changes. This is related to, but distinct from, AI timelines (how long until we develop advanced AI). While timelines tell us when transformative AI might arrive, takeoff speeds tell us what happens after it arrives - does AI capability and impact increase gradually over years, or explosively over days or weeks? When analyzing different takeoff scenarios, we can look at several key factors:

- Speed: How quickly do AI capabilities improve?
- Continuity: Do capabilities improve smoothly or in sudden jumps?
- Homogeneity: How similar are different AI systems to each other?
- Polarity: How concentrated is power among different AI systems?

In the next section we discuss just one of these factors that tends to be the most debated - takeoff speed. The rest (continuity, homogeneity, and polarity) are explained in an appendix.

7.1 Speed

In a slow takeoff scenario, AI capabilities improve gradually over months or years. We can see this pattern in recent history - the transition from GPT-3 to GPT-4 brought significant improvements in reasoning, coding, and general knowledge, but these advances happened over several years through incremental progress. Paul Christiano describes slow takeoff as similar to the Industrial Revolution but "10x-100x faster" (Davidson, 2023). Terms like "slow takeoff" and "soft takeoff" are often used interchangeably.

In mathematical terms, slow takeoff scenarios typically show linear or exponential growth patterns. With linear growth, capabilities increase by the same absolute amount each year - imagine an AI system that gains a fixed number of new skills annually. More commonly, we might see exponential growth, where capabilities increase by a constant percentage, similar to how we

discussed scaling laws in earlier sections. Just as model performance improves predictably with compute and data, slow takeoff suggests capabilities would grow at a steady but manageable rate. This might manifest as GDP growing at 10-30% annually before accelerating further.

Slow takeoff provides us time to adapt and respond. If we discover problems with our current safety approaches, we can adjust them before AI becomes significantly more powerful. This connects directly to what we'll discuss in later chapters about governance and oversight - slow takeoff allows for iterative refinement of safety measures and gives time for coordination between different actors and institutions.

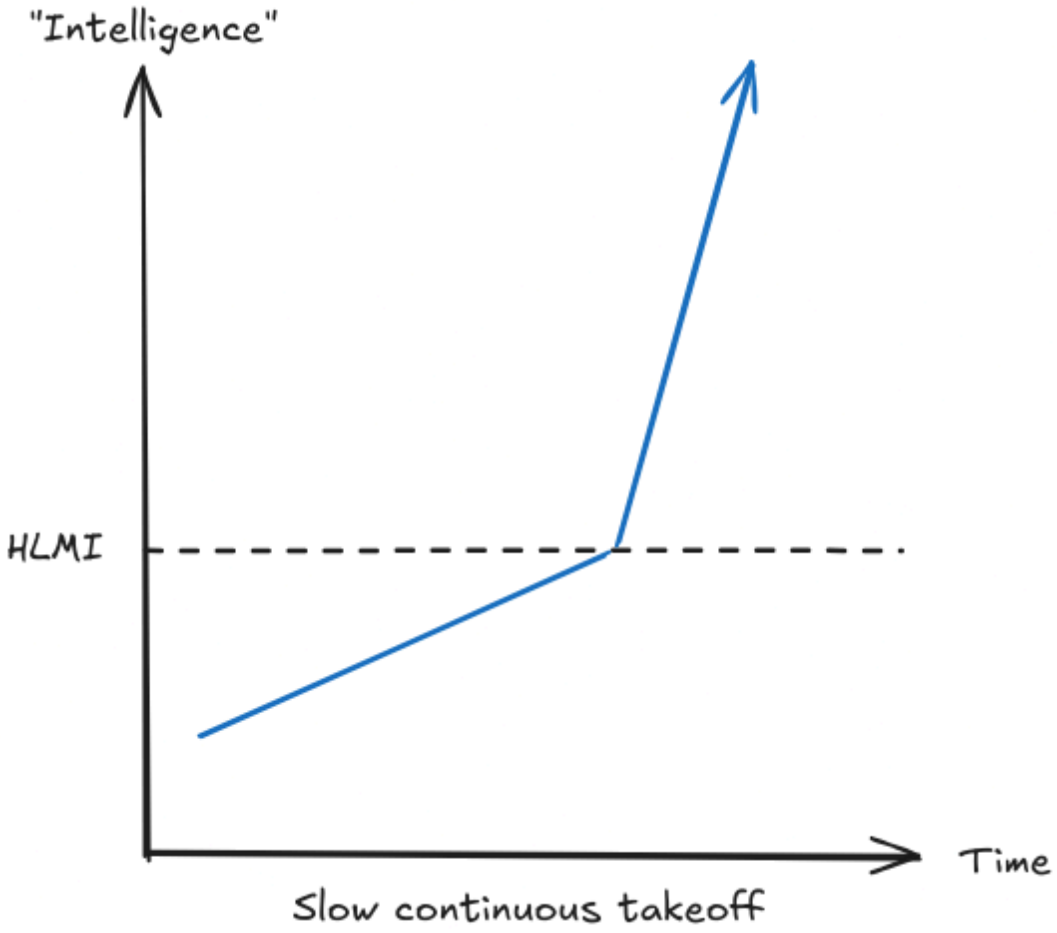


Figure 46: *An illustration of slow continuous takeoff (Martin & Eth, 2021).*

Fast takeoff describes scenarios where AI capabilities increase dramatically over very short periods - perhaps days or even hours. Instead of the gradual improvement we saw from GPT-3 to GPT-4, imagine an AI system making that much progress every day. This could happen through recursive self-improvement, where an AI system becomes better at improving itself, creating an accelerating feedback loop.

Mathematically, fast takeoff involves superexponential or hyperbolic growth, where the growth rate itself increases over time. Rather than capabilities doubling every year as in exponential growth, they might double every month, then every week, then every day. This relates to what we discussed in the scaling section about potential feedback loops in AI development - if AI systems can improve the efficiency of AI research itself, we might see this kind of accelerating progress.

The dramatic speed of fast takeoff creates unique challenges for safety. As we'll explore in the chapter on strategies, many current safety approaches rely on testing systems, finding problems, and making improvements. But in a fast takeoff scenario, we might only get one chance to get things right. If an AI system starts rapidly self-improving, we need safety measures that work robustly from the start, because we won't have time to fix problems once they emerge. Terms like "fast takeoff", "hard takeoff" and "FOOM" are often used interchangeably.

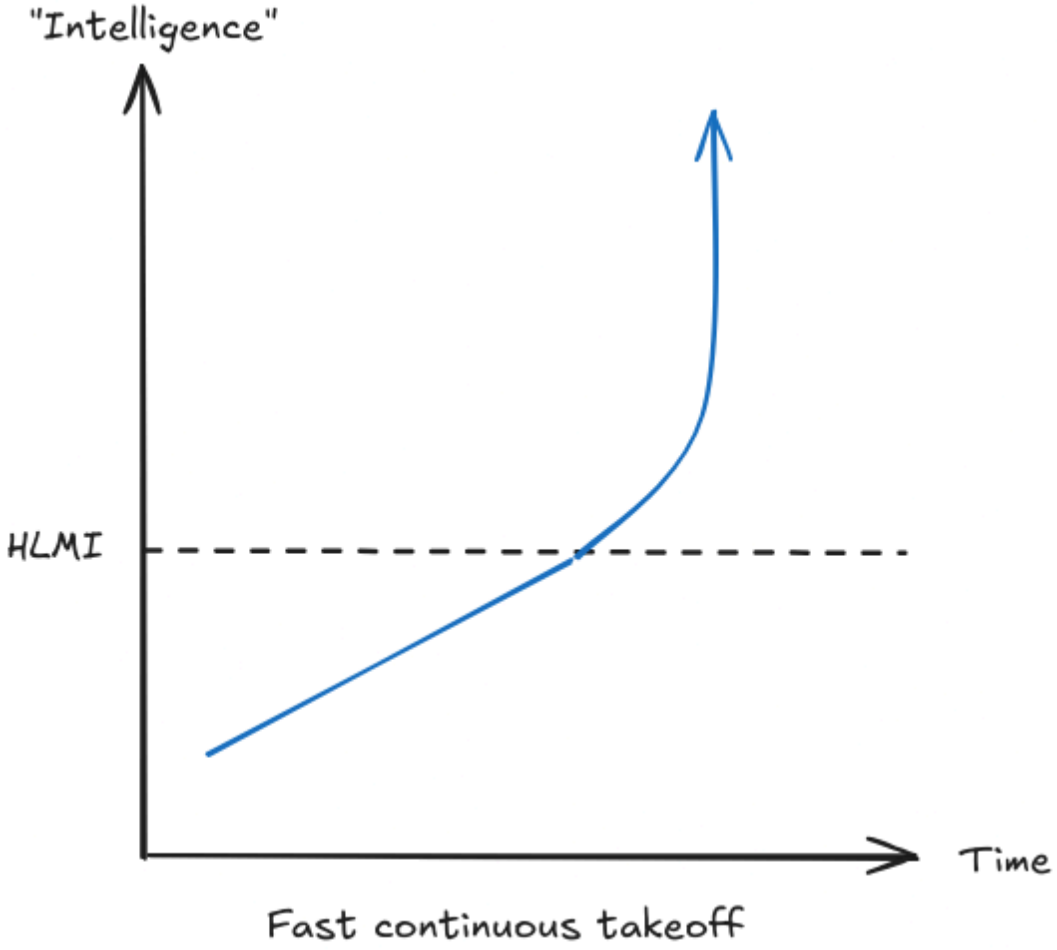


Figure 47: *An illustration of fast continuous takeoff, which is usually taken to mean superexponential or hyperbolic growth. The growth rate itself increases (Martin & Eth, 2021).*

The speed of AI takeoff fundamentally shapes the challenge of making AI safe. If progress follows predictable patterns as our current understanding suggests, we might have more warning and time to prepare. But if new mechanisms like recursive self-improvement create faster feedback loops, we need different strategies. Today, when we discover that language models can be jailbroken, companies can patch these vulnerabilities in the next release. In a slow takeoff, this pattern could continue - we'd have time to discover and fix safety issues as they arise. But in a fast takeoff, we might need to solve all potential jailbreaking vulnerabilities before deploying the AI, because a system could become too powerful to safely modify before we can implement any further safety fixes. The majority of experts, researchers and engineers agree that AI will pose risks and it should be developed responsibly. The differences are in small nuances of how to respond. These can often be boiled down to whether they expect problems to be noticeable and fixable in time, or

too fast for us to respond. Unfortunately, these sometimes get misreported as AI experts disagree on AI risks.

> *This is how we build airplanes — we build airplanes, sometimes they crash tragically, and then we fix it. I think AI sometimes gives bad outputs and then we fix it, and that's how we actually make these things reliable.*
>
> **Andrew Ng**
> Cofounder and head of Google Brain, former Chief Scientist at Baidu
>
> **Feb 2025**
> World Economic Forum, 2025

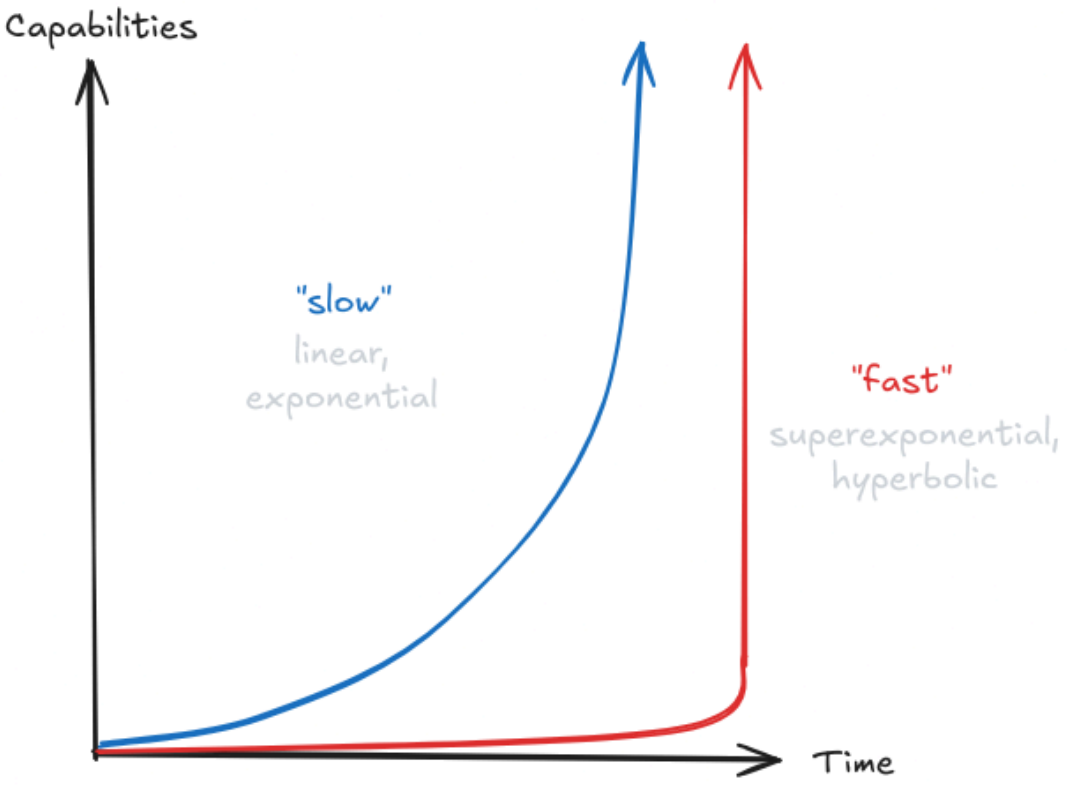


Figure 48: *Comparison of slow vs fast takeoff. Showcasing that while described as linguistically slower than fast, it is by no means slow (Christiano, 2018).*

Understanding fast vs slow helps you get the overview of the takeoff debate, but there can be a bunch of other factors like - are there sudden jumps? (takeoff continuity), how many systems are 'taking off' at the same tim? (takeoff polarity), how architecturally similar are these systems? (takeoff similarity) . If you want to learn more feel free to read the details on these in the optional appendix.

7.2 Takeoff Arguments

The Overhang Argument . There might be situations where there are substantial advancements or availability in one aspect of the AI system, such as hardware or data, but the corresponding software or algorithms to fully utilize these resources haven't been developed yet. The term 'overhang' is used because these situations imply a kind of 'stored' or 'latent' potential. Once the software or algorithms catch up to the hardware or data, there could be a sudden unleashing of this potential, leading to a rapid leap in AI capabilities. Overhangs provide one possible argument for why we might favor discontinuous or fast takeoffs. There are two types of overhangs commonly discussed:

- **Hardware Overhang**: This refers to a situation where there is enough computing hardware to run many powerful AI systems, but the software to run such systems hasn't been developed yet. If such hardware could be repurposed for AI, this would mean that as soon as one powerful AI system exists, probably a large number of them would exist, which might amplify the impact of the arrival of human-level AI.
- **Data Overhang**: This would be a situation where there is an abundance of data available that could be used for training AI systems, but the AI algorithms capable of utilizing all that data effectively haven't been developed or deployed yet.

Overhangs are also used as a counter argument to why AI pauses do not meaningfully affect takeoff speeds. One counter argument to the overhang argument is that it relies on the assumption that during the time that we are pausing AI development, the rate of production of chips will remain constant. It could be argued that the companies manufacturing these chips will not make as many chips if data centers aren't buying them. However, this argument only works if the pause is for any appreciable length of time, otherwise the data centers might just stockpile the chips. It is also possible to make progress on improved chip design, without having to manufacture as many during the pause period. However, during the same pause period we could also make progress on AI safety techniques (Elmore, 2024).

The Economic Growth Argument . Historical patterns of economic growth, driven by human population increases, suggest a potential for slow and continuous AI takeoff. This argument says that as AIs augment the effective economic population, we might witness a gradual increase in economic growth, mirroring past expansions but at a potentially accelerated rate due to AI-enabled automation. Limitations in AI's ability to automate certain tasks, alongside societal and regulatory constraints (e.g. that medical or legal services can only be rendered by humans), could lead to a slower expansion of AI capabilities. Alternatively, growth might far exceed historical rates. Using a similar argument for a fast takeoff hinges on AI's potential to quickly automate human labor on a massive scale, leading to unprecedented economic acceleration.

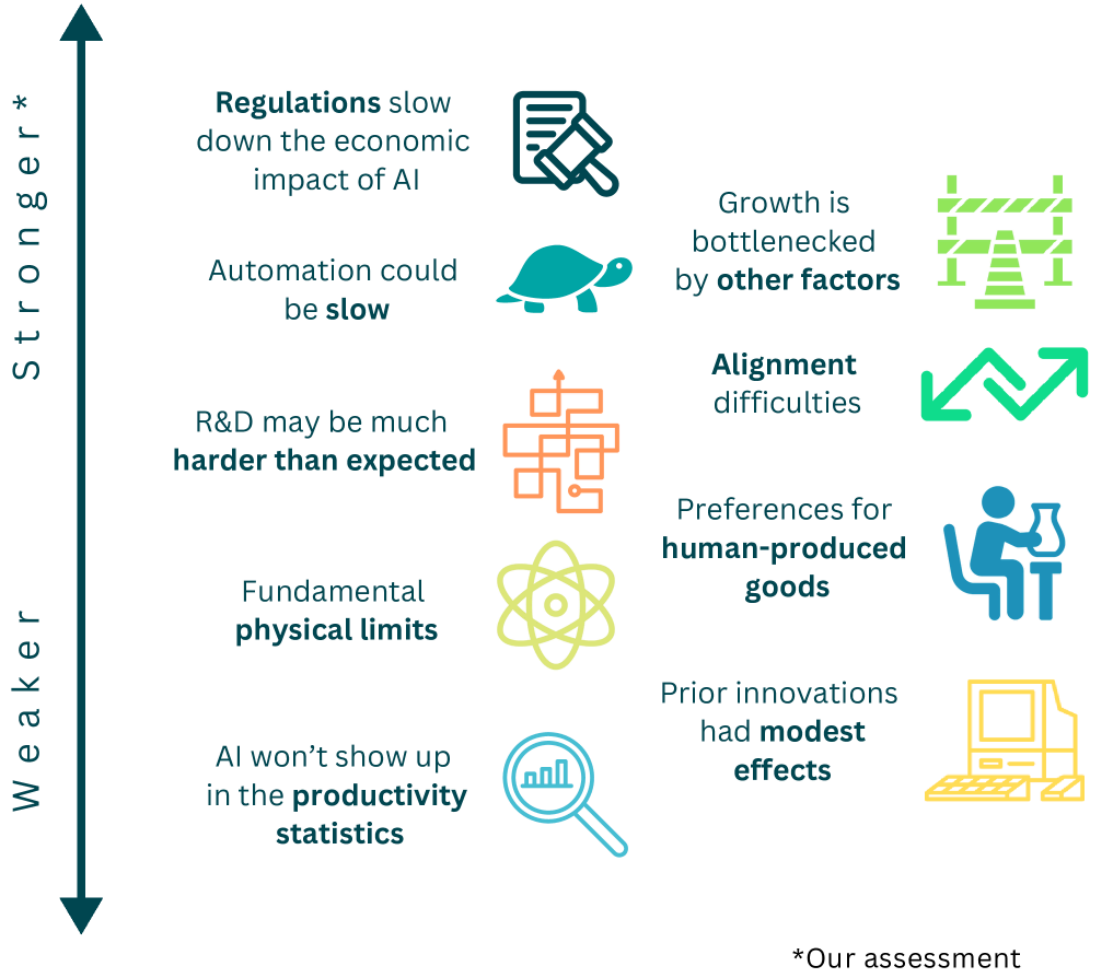


Figure 49: *A visualization of the ranking of arguments for explosive economic growth, both in favor and against. By Epoch AI (Erdil & Besiroglu, 2024).*

Compute Centric Takeoff Argument . This argument, similar to the Bio Anchors report, assumes that compute will be sufficient for transformative AI. Based on this assumption, Tom Davidson's 2023 report on compute-centric AI takeoff discusses feedback loops that may contribute to takeoff dynamics.

- **Investment feedback loop**: There might be increasing investment in AI, as AIs play a larger and larger role in the economy. This increases the amount of compute available to train models, as well as potentially leading to the discovery of novel algorithms. All of this increases capabilities, which drives economic progress, and further incentivizes investment.
- **Automation feedback loop**: As AIs get more capable, they will be able to automate larger parts of the work of coming up with better AI algorithms, or helping in the design of better GPUs. Both of these will increase the capability of the AIs, which in turn allow them to automate more labor.

Depending on the strength and interplay of these feedback loops, they can create a self-fulfilling prophecy leading to either an accelerating fast takeoff if regulations don't curtail various aspects of such loops, or a slow takeoff if the loops are weaker or counterbalanced by other factors. The entire model is shown in the diagram below:

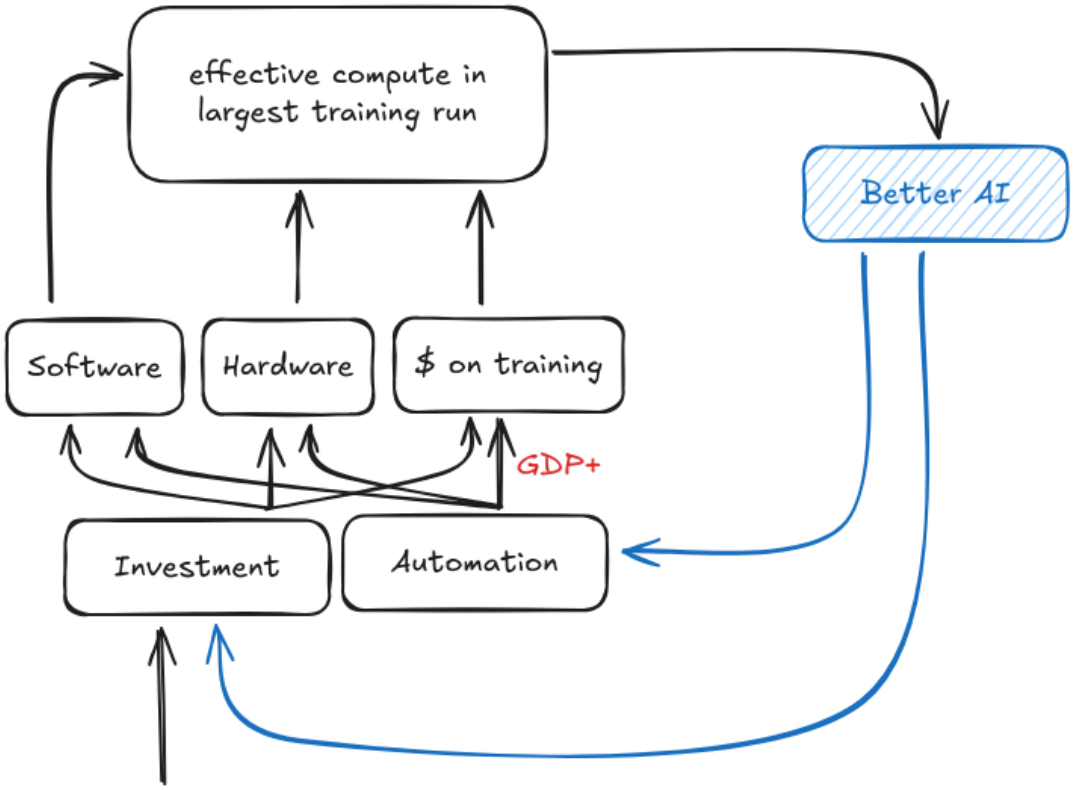


Figure 50: *A summary of What a Compute-Centric Framework Says About Takeoff Speeds (Davidson, 2024)*

Automating Research Argument. Researchers could potentially design the next generation of ML models more quickly by delegating some work to existing models, creating a feedback loop of ever-accelerating progress. The following argument is put forth by Ajeya Cotra:

Currently, human researchers collectively are responsible for almost all of the progress in AI research, but are starting to delegate a small fraction of the work to large language models. This makes it somewhat easier to design and train the next generation of models.

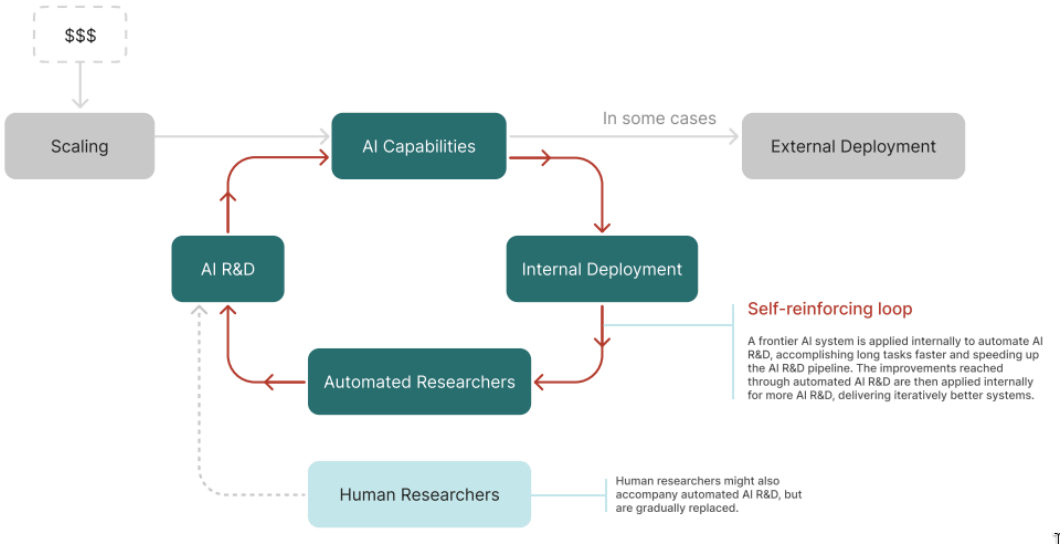


Figure 51: *A. This figure shows a representation of a self-reinforcing loop (in red). It demonstrates how internally deployed AI systems are used to help automate AI R&D, initially alongside human researchers. These AI R&D efforts culminate in a more capable AI system, which can be deployed as a new, improved, automated researcher. This cycle keeps repeating, resulting in a self-reinforcing loop (Stix et al., 2025)*

The next generation is able to handle harder tasks and more different types of tasks, so human researchers delegate more of their work to them. This makes it significantly easier to train the generation after that. Using models gives a much bigger boost than it did the last time around.

Each round of this process makes the whole field move faster and faster. In each round, human researchers delegate everything they can productively delegate to the current generation of models — and the more powerful those models are, the more they contribute to research and thus the faster AI capabilities can improve (Cotra, 2023).

So before we see a recursive explosion of intelligence, we see a steadily increasing amount of the full RnD process being delegated to AIs. At some point, instead of a significant majority of the research and design being done by AI assistants at superhuman speeds, it will become that - all of the research and design for AIs is done by AI assistants at superhuman speeds.

At this point there is a possibility that this might eventually lead to a full automated recursive intelligence explosion.

The Intelligence Explosion Argument . This concept of the 'intelligence explosion' is also central to the conversation around discontinuous takeoff. It originates from I.J. Good's thesis, which posits that sufficiently advanced machine intelligence could build a smarter version of itself. This smarter version could in turn build an even smarter version of itself, and so on, creating a cycle that could lead to intelligence vastly exceeding human capability (Yudkowsky, 2013).

In their 2012 report on the evidence for Intelligence Explosions, Muehlhauser and Salamon delve into the numerous advantages that machine intelligence holds over human intelligence, which facilitate rapid intelligence augmentation (Muehlhauser, 2012). These include:

- **Computational Resources**: Human computational ability remains somewhat stationary, whereas machine computation possesses scalability.

- **Speed**: Humans communicate at a rate of two words per second, while GPT-4 can process 32k words in an instant. Once LLMs can write "better" than humans, their speed will most probably surpass us entirely.
- **Duplicability**: Machines exhibit effortless duplicability. Unlike humans, they do not need birth, education, or training. While humans predominantly improve individually, machines have the potential to grow collectively. Humans take 20 years to become competent from birth, whereas once we have one capable AI, we can duplicate it immediately. Once AIs reach the level of the best programmer, we can just duplicate this AI. The same goes for other jobs.
- **Editability**: Machines potentially allow more regulated variations. They exemplify the equivalent of direct brain enhancements via neurosurgery in opposition to laborious education or training requirements. Humans can also improve and learn new skills, but they don't have root access to their hardware: we are just starting to be able to understand the genome's "spaghetti code," while AI could use code versioning tools to improve itself, being able to attempt risky experiments with backup options in case of failure. This allows for much more controlled variation.
- **Goal coordination**: Copied AIs possess the capability to share goals effortlessly, a feat challenging for humans.

8. Appendix: Forecasting

8.1 Effective Compute

8.1.1 Hardware Efficiency

The first factor in increasing effective compute is - how efficient is the hardware that we produce at doing computation?

AI chips are getting about 35% more powerful each year, but we're approaching physical limits. Since around 2010, the raw computational performance (FLOP/s) of GPUs in FP32 precision has grown by roughly 1.35× per year (Epoch AI, 2023 ; Hobbahn et al., 2023). The improvement comes from things like denser transistors, specialized AI circuitry, and switching to lower-precision number formats. Performance per dollar has improved rapidly, and hardware at any given precision and fixed performance level becomes 30% cheaper each year. At the same time, manufacturers continue to introduce more powerful and expensive hardware (Epoch AI, 2025)

In the near future this trend seems likely to continue, but thermodynamic limits will eventually stop this trend. Chips can only get so energy-efficient before physics says "no". Every computation generates heat as a fundamental law of physics, not an engineering problem. Current analysis suggests there is room for a 50 to 1,000× improvement in energy efficiency before we hit fundamental CMOS limits, with a 50% chance that improvements cease before a roughly 200× improvement on existing technology. These estimates suggest that CMOS processors are likely sufficiently efficient to power substantially larger AI training runs than today.[7] This implies we have significant headroom to scale using current silicon paradigms through 2030 and beyond, although hardware R&D returns may eventually diminish as we approach physical limits. Beyond these limits, training runs would likely require radical changes to computing paradigms, like a shift to adiabatic computing (Ho et al., 2023 ; Sevilla et al., 2024).

[7]CMOS (Complementary Metal-Oxide-Semiconductor) is the primary paradigm in processor production. The majority of all digital integrated circuits (CPUs, GPUs, RAM, mobile SoCs) produced today are CMOS.

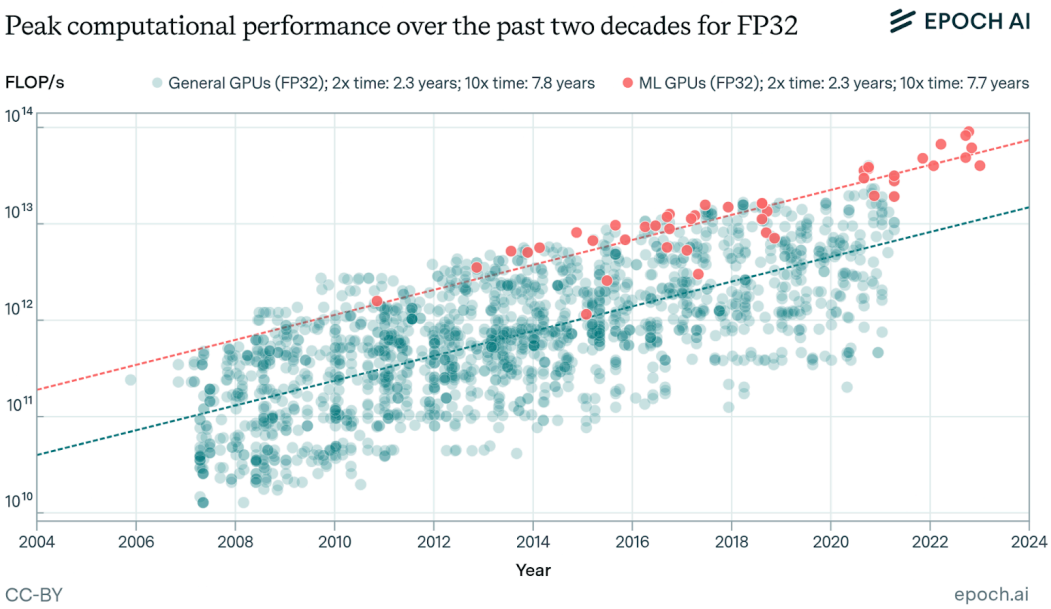


Figure 52: *Currently observed trends in the efficiency of hardware used for ML over the last decades. We can see a clear year by year increase in peak computational performance (Hobbahn et al., 2023).*

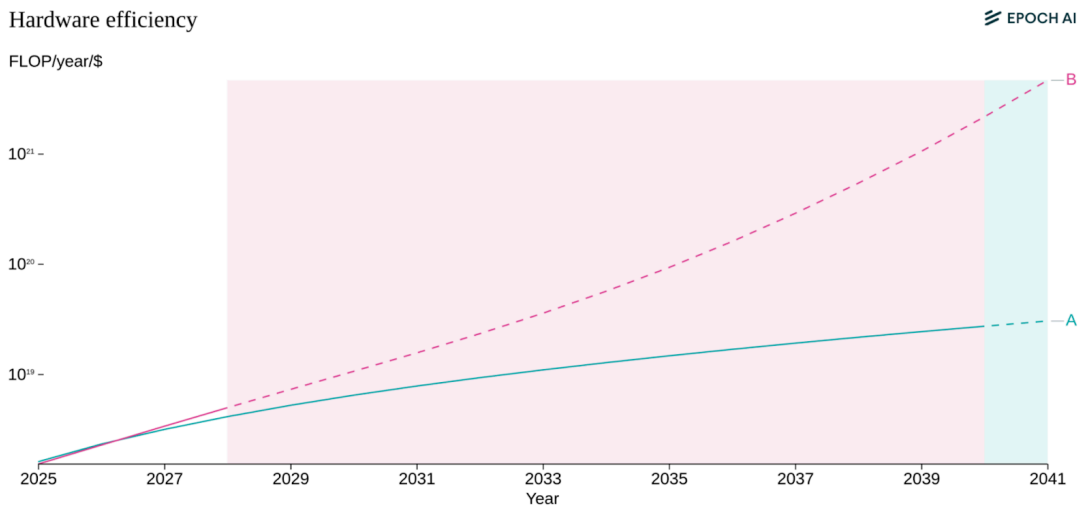


Figure 53: *A forecast of the future using the Growth and AI Transition Endogenous (GATE) playground. This graph shows investments in hardware R&D improving the efficiency of AI chips (measured in units of FLOP/year per $). Eventually, hardware efficiency is capped by physical limits. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

8.1.2 Software and Algorithmic Efficiency

The second factor in increasing effective compute is - how well we can utilize all the existing hardware that we have. This is separate from just making the physical hardware itself more efficient.

Better algorithms have cut the compute needed for a given result by 3x per year. The level

of compute needed to achieve a given level of performance has halved roughly every 8 months[8] . This rapid improvement means that the compute required to achieve specific levels of capability on benchmarks can drop by orders of magnitude over just a few years of algorithmic progress. The improvements to compute efficiency explain roughly 35% of performance improvements in language modeling since 2014, with the other 65% coming from just building more chips and running them longer. Overall, this means we're getting smarter about using available hardware, and not just throwing more compute at problems (Epoch AI, 2023 ; Ho et al., 2024).

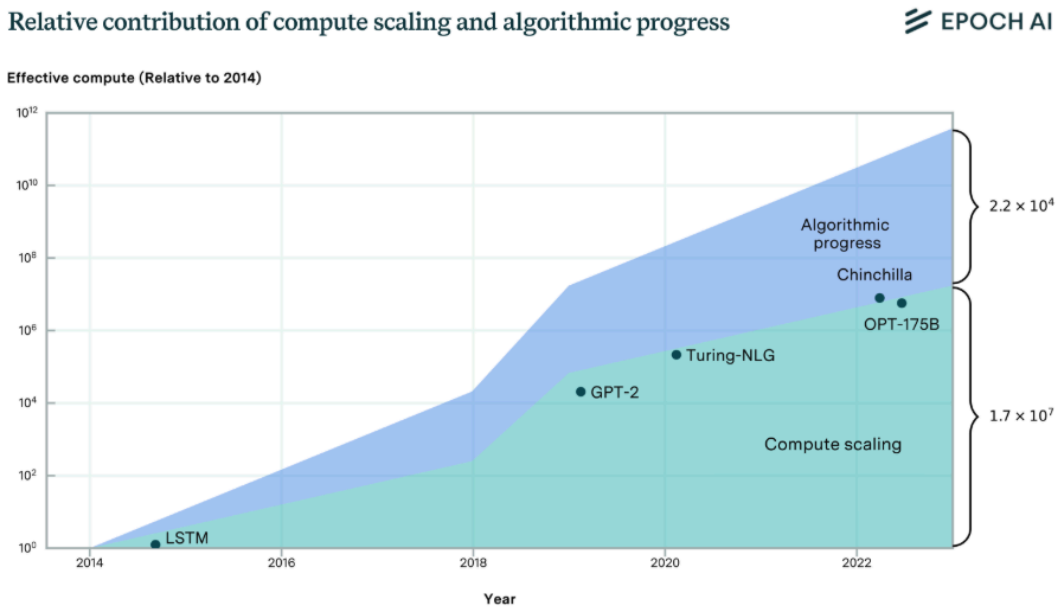


Figure 54: *Estimates of the contributions of scaling and algorithmic innovation in terms of the raw compute that would be naively needed to achieve a state-of-the-art level of performance. The contribution of algorithmic progress is roughly half as much as that of compute scaling (Ho et al., 2024)*

[8] 95% confidence interval of 5 to 14 months

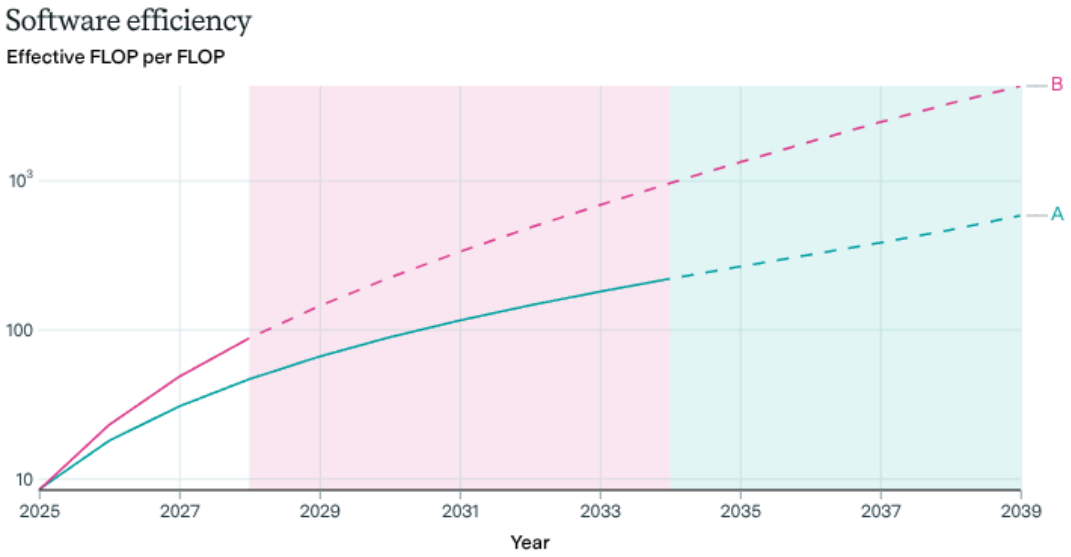


Figure 55: *A forecast of the future using the Growth and AI Transition Endogenous (GATE) playground. This graph shows algorithmic improvements over time. By developing better algorithms and using higher-quality data, it becomes possible to do more with the same hardware (measured in units of "effective FLOP per FLOP"). The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

8.1.3 Semiconductor Production

The third factor in increasing effective compute is - how many chips can you actually build?

Total available computing power from NVIDIA chips has grown by approximately 2.3x per year since 2019, enabling the training of ever-larger models. NVIDIA designs a dominant share of AI training chips, and Taiwan's TSMC serves as the primary chip fab for these manufacturers. The AI chip supply chain is highly concentrated. In 2024 TSMC dedicated roughly 5% of their advanced chip production to AI accelerators—the rest goes to phones, computers, and other electronics (Sevilla et al., 2024). If AI labs want dramatically more chips, TSMC would need to shift priorities, expand capacity, and compete with other customers who also want cutting-edge semiconductors (Epoch AI, 2025).

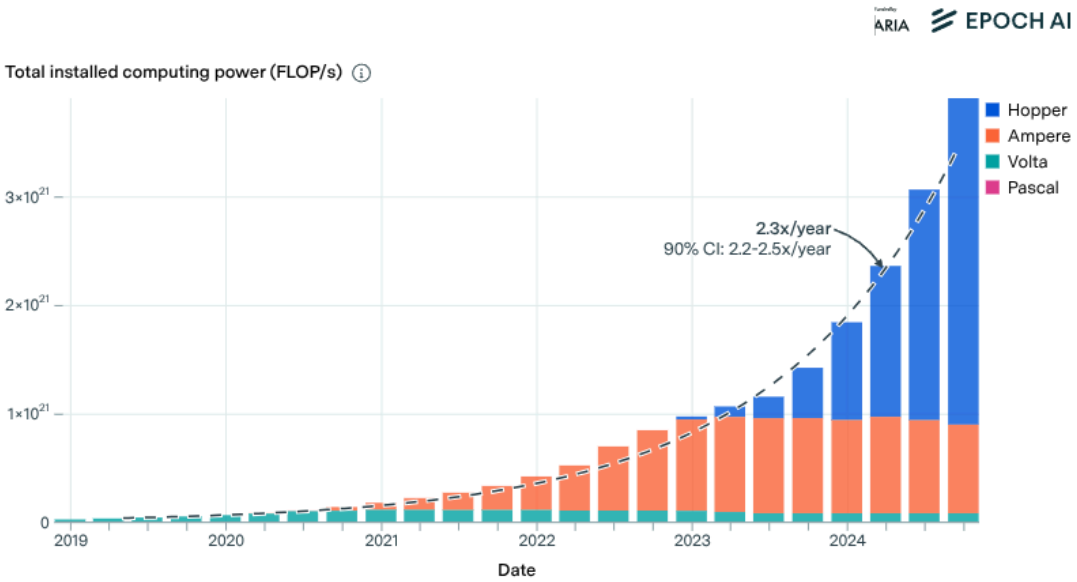


Figure 56: *An estimate the world's installed NVIDIA GPU compute capacity, broken down by GPU model (Epoch AI, 2025).*

Primary constraint	Lab compute investment (real USD)	Estimated lead time (years)
Renting GPUs	<$30 million	~0
Buying GPUs	$30 million to $1 billion	~0.5
Constructing a data center	$3 billion	1-2
Constructing a very large data center/power plant	$10 billion to $30 billion	2-3
Significantly upgrading a fab	$30 billion	2
Building a new cutting-edge fab from scratch	$300 billion	4-5

Figure 57: *Table showcasing estimates costs and times to overcome various constraints in scaling up compute production (Edelman & Ho, 2025).*

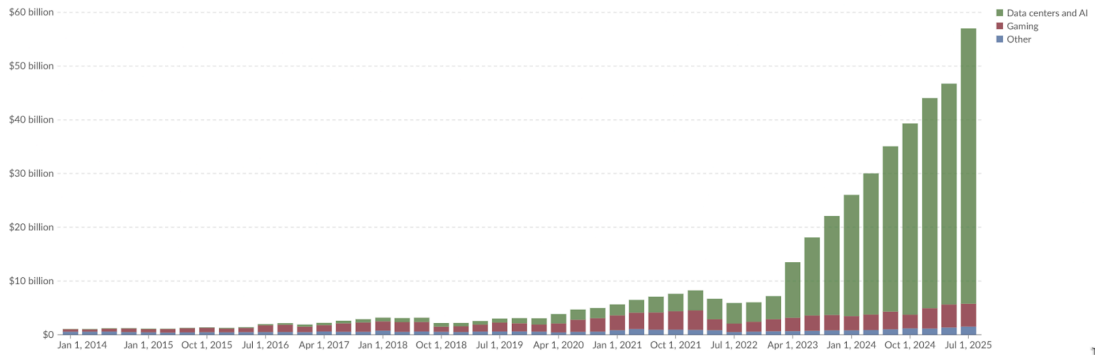


Figure 58: *Quarterly revenue of NVIDIA Corporation across its main market segments, reported in US dollars. NVIDIA manufactures graphics processing units (GPUs), which were originally used for gaming and are now used to train AI models. This data is not adjusted for inflation (Our World in Data, 2023).*
(interactive version on website)

Investment doesn't immediately make chip production go faster. If some country wants to build their own TSMC equivalent, then it can take 4-5 years to build a new cutting-edge fab (accounting for construction and permitting), though upgrading an existing fab takes less time, around 2 years in addition to billions in investment (Edelman & Ho, 2025 ; Epoch AI, 2025). Besides the existing competition for chips from TSMC, another factor is that the machines TSMC needs to manufacture the chips. These are almost exclusively made by a single company: ASML in the Netherlands (Blablová, 2025). These extreme ultraviolet (EUV) lithography machines cost between $150 million to $380 million each. If/When someone (e.g. TSMC) wants to expand production, they can't just order fifty more next month. They join an already long waitlist (Edelman & Ho, 2025 ; Sevilla, AXRP Podcast, 2024).

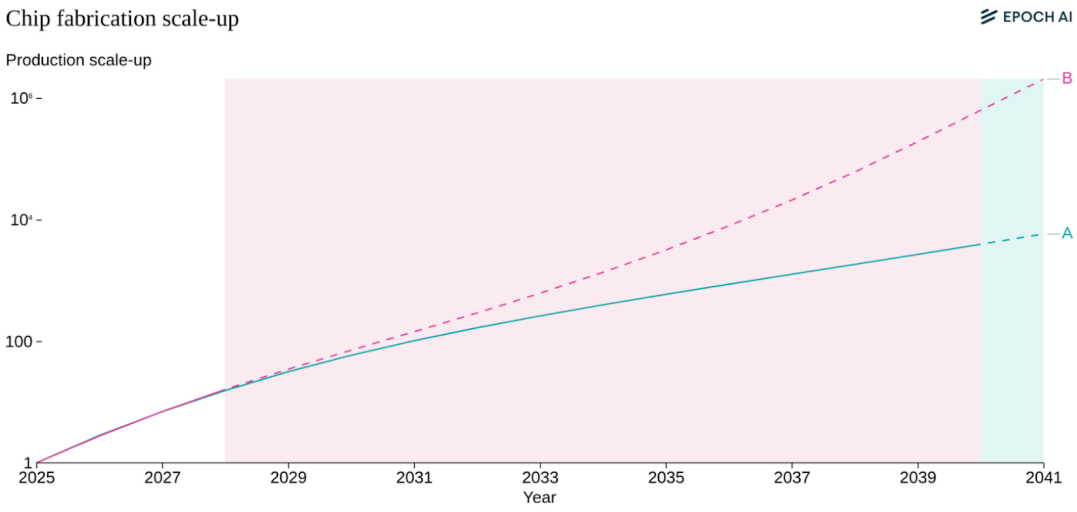


Figure 59: *A forecast using the Growth and AI Transition Endogenous (GATE) playground. This graph shows potential scale up of AI chip fabrication to meet demand for AI workloads. Due to the exponential cost of rapid expansions, growth is reduced by an "adjustment costs" parameter. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

8.2 Investment and Training Costs

The cost in USD of training frontier models has grown by 3.5x per year per year since 2020. If trends continue, we'll see the training cost for a single model approach billion-dollar training runs by 2027. The cost breaks down roughly as: 50-65% for hardware (spread across its useful life), 30-45% for the researchers and engineers, and 2-6% for electricity. This is venture capital and Big Tech money so far, but these numbers are approaching scales where only nations or the largest corporations can compete (Cottier et al., 2025).

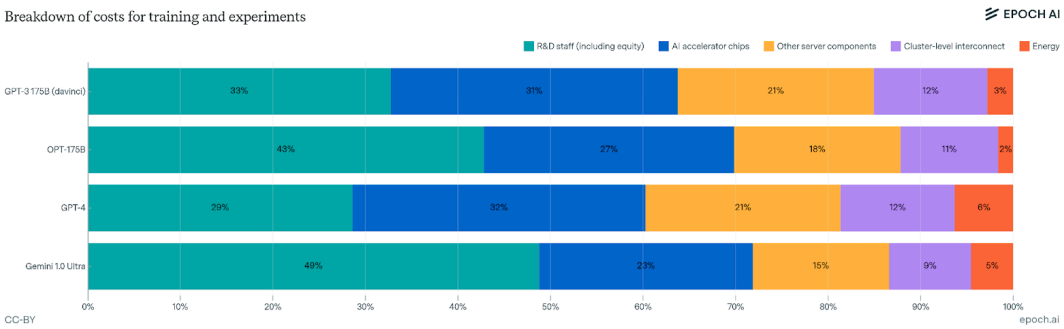


Figure 60: *Breakdown of model development costs for selected models. Hardware costs are amortized to the total number of chip-hours spent on experiments and training. R&D staff costs cover the duration of development, from initial experiments to publication (Cottier et al., 2025).*

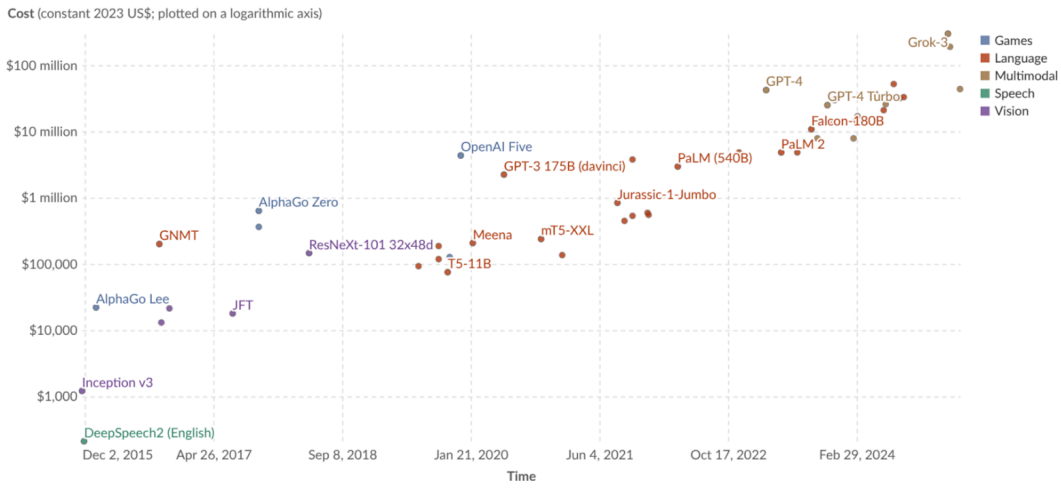


Figure 61: *Hardware and energy cost to train notable AI systems. This data is expressed in US dollars, adjusted for inflation (Our World in Data, 2023).* (interactive version on website)

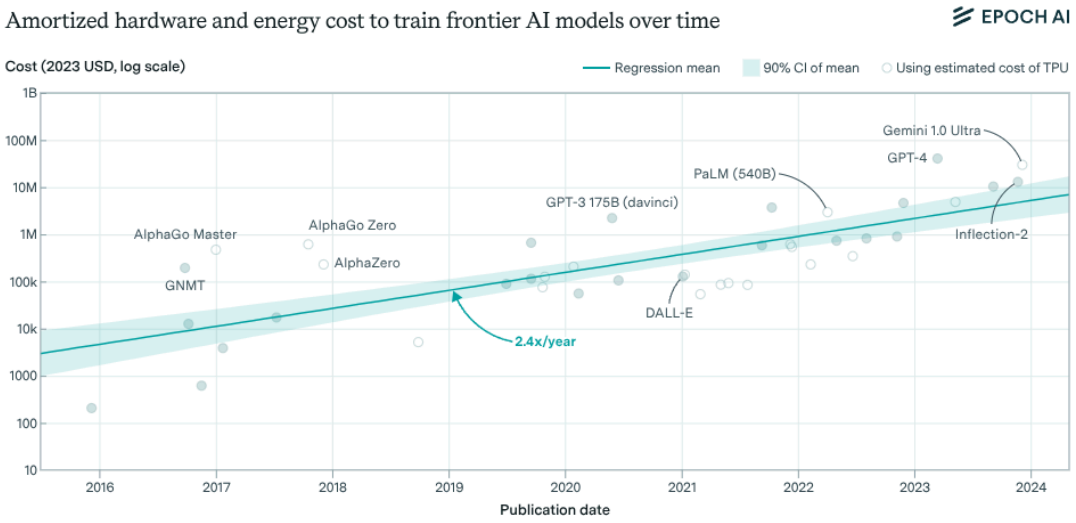


Figure 62: *Amortized hardware cost plus energy cost for the final training run of frontier models. The selected models are among the top 10 most compute-intensive for their time. Amortized hardware costs are the product of training chip-hours and a depreciated hardware cost, with 23% overhead added for cluster-level networking. Open circles indicate costs which used an estimated production cost of Google TPU hardware. These costs are generally more uncertain than the others, which used actual price data rather than estimates (Cottier et al., 2025).*

The acquisition cost in USD of the hardware used to train frontier AI models has grown by 2.5x per year since 2016. To give you a sense of how much training a frontier model costs, the total amortized cost of developing Grok-4 (released in July 2025), including hardware and electricity, is estimated at $480 million USD. The acquisition cost of the hardware to train Grok-3, including GPUs, other server components, and networking, is estimated at $3 billion USD. The hardware used to train Grok 4 may have been even more expensive (Epoch AI, 2025).

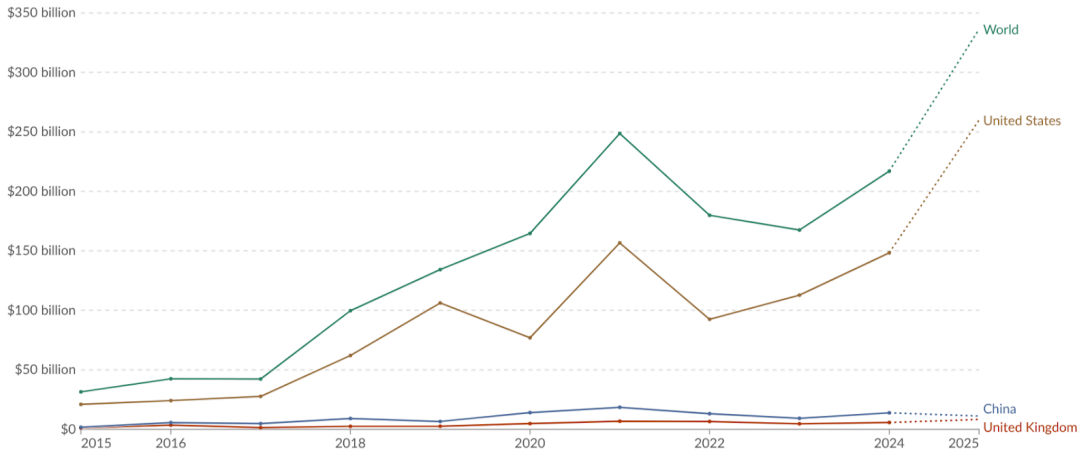


Figure 63: *Money put into privately held AI companies by private investors. This excludes publicly traded companies (e.g., Big Tech companies) and companies' internal spending, such as R&D or infrastructure. Expressed in US dollars, adjusted for inflation (Our World in Data, 2023).* (interactive version on website)

An unknown question is whether the returns to productivity gains will justify continued investment. 2025 saw a lot of news headlines about circular investment and an "AI bubble". If

AI can automate significant portions of cognitive work, capturing even a fraction of the global labor market makes trillion-dollar investments rational. However, massive investments face increasing hurdles due to structural constraints. As compute scales up, the lead time between project initiation and deployment lengthens significantly—roughly one year for every ten-fold increase in scale—creating a feedback loop that naturally slows the pace of scaling (Edelman & Ho, 2025). This uncertainty regarding long-term returns over extended periods may drive investors to prefer incremental scaling, breaking up projects into smaller chunks to gauge returns before committing further, rather than making massive upfront investments (Edelman & Ho, 2025). Several things could break the pattern entirely: models hitting a capability ceiling where more compute doesn't help, regulations capping training runs or data center sizes, energy costs making large runs uneconomical, or economic recession drying up capital. Each represents a distinct way scaling could stop independent of technical feasibility.

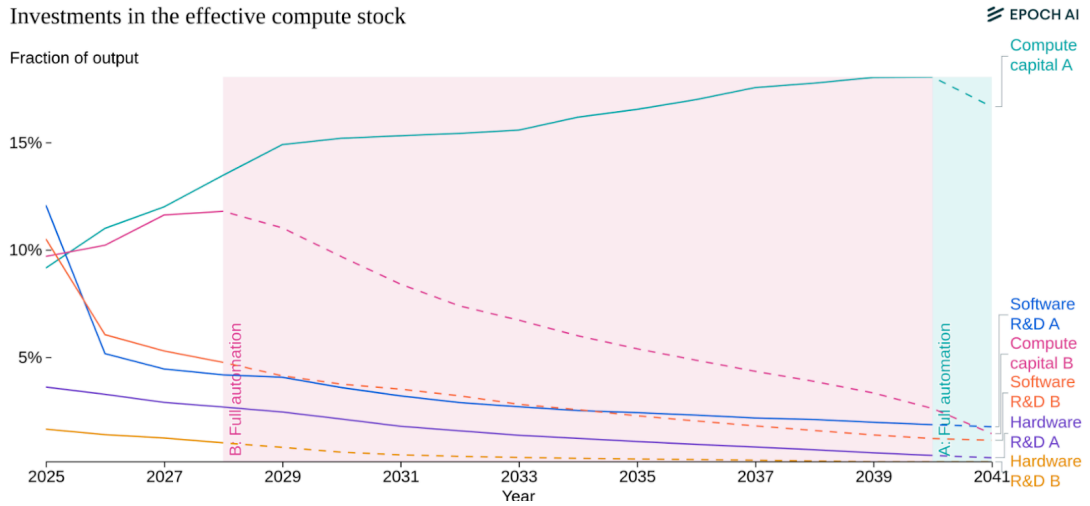


Figure 64: *A forecast using the Growth and AI Transition Endogenous (GATE) playground. GATE predicts high levels of investment in AI-related capital and R&D to support massive expansions of the effective compute stock. This involves major reallocations away from conventional capital investments and consumption, and occurs before AI generates significant economic value, motivated by the large payoffs from AI automation. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

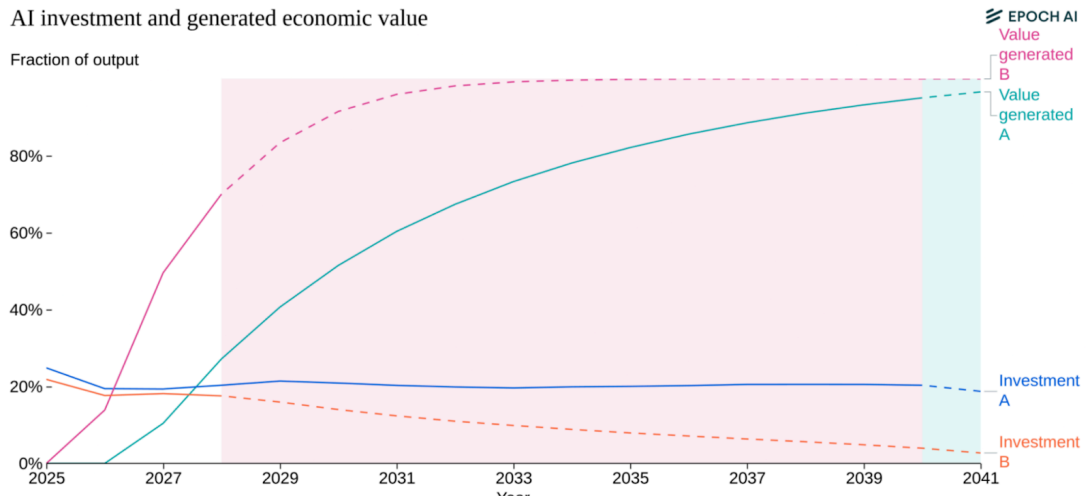


Figure 65: *Large investments in AI precede the economic benefits of automation, driven by expectations of large future payoffs. At the start of simulations, there is a brief period with substantial AI investment but with little generated economic value. However, as automation proceeds the value generated by AI grows rapidly, comprising the majority of output within a few years. The GATE model thus captures the dynamic where it is worth making major upfront investments to enjoy the value generated by AI. The investment is measured as the fraction of yearly economic output that is reinvested in AI, while the benefit is measured as the fraction of yearly output that can be attributed to the deployment of digital workers. The red line showcases aggressive parameter settings, green is conservative parameter settings. The red zone highlights the difference in timelines to full automation between aggressive and conservative models (Epoch AI, 2025).*

8.3 Power Consumption

Training frontier models requires enough electrical power to run a small city. GPT-4's training run likely consumed around 50 megawatts, equivalent to powering roughly 40,000 US homes. If compute scaling continues at 4-5× per year, frontier training runs will require 4-16 gigawatts by 2030—matching the Grand Coulee Dam, America's largest power plant (You et al., 2025).

Energy costs remain small relative to hardware costs but this could change. Currently, electricity costs represent only 2-6% of total training costs, with hardware and labor dominating the budget (Cottier et al., 2024). But if training runs scale to 10+ gigawatts while hardware efficiency improvements slow down, energy could become a much larger fraction of costs. At $0.05 per kilowatt-hour, a 10 GW training run running for 100 days costs roughly $120 million just for electricity. That's still less than the hardware, but it's no longer negligible. And in practice, securing multi-gigawatt power supplies might cost significantly more than wholesale electricity rates suggest.

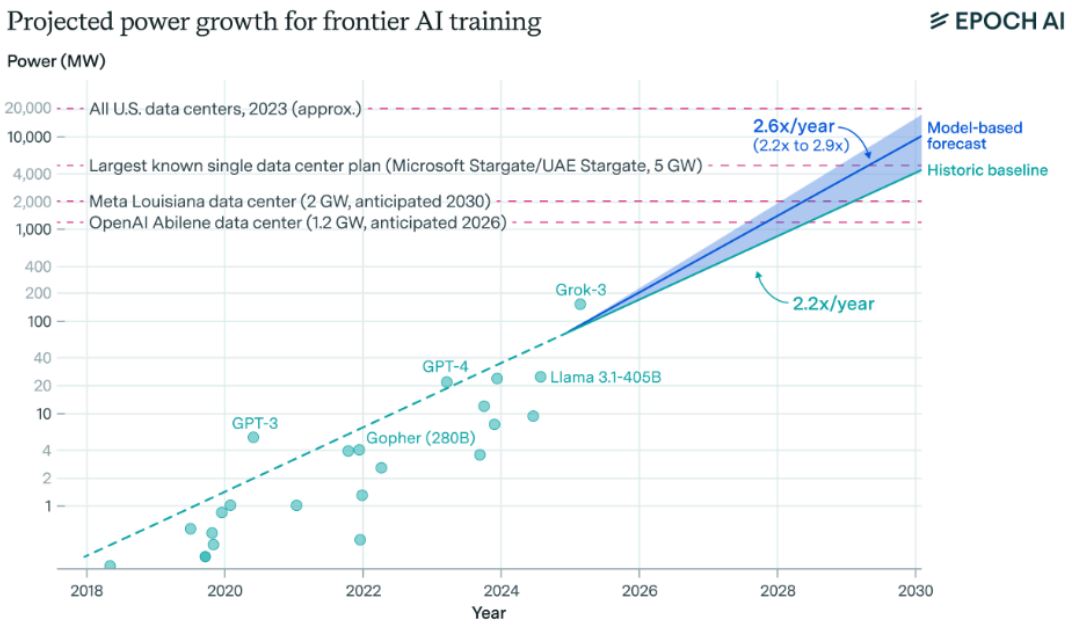


Figure 66: *Historic trend and forecast for the electricity demand of the largest individual frontier training runs (You et al., 2025).*

Figure 67: *The projected power consumption for planned frontier datacenters for selected AI companies (Epoch AI, 2025).*

Scaling AI systems to the forecasted levels of 2030 will require substantial power infrastructure. A training run of 2×10^{29} FLOP would require approximately 6 GW of power, assuming improvements in hardware efficiency. This scale necessitates data center campuses ranging from 1 to 5 GW by the end of the decade (Sevilla et al., 2024). Lead times increase with scale: every additional 10× increase in compute stock adds roughly one year to project timelines. Constructing the necessary large-scale power plants typically takes 2-3 years (Edelman & Ho, 2025). Despite these hurdles, the cost of power remains a small fraction of data center expenses—roughly one-

tenth the cost of the chips - making the capital investment rational given the potential returns (Ho et al., 2025).

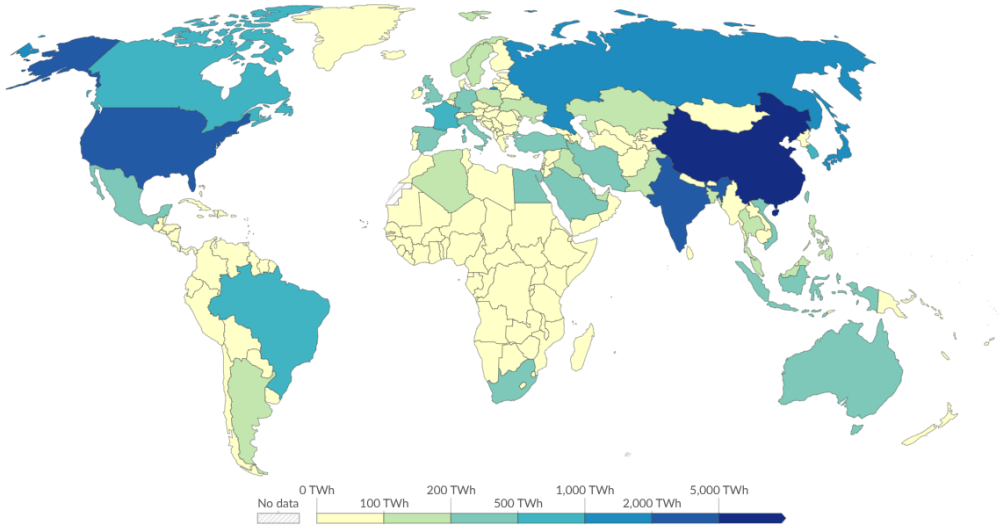


Figure 68: *Total electricity generated in each country or region, measured in terawatt-hours (Our World in Data, 2024).* (interactive version on website)

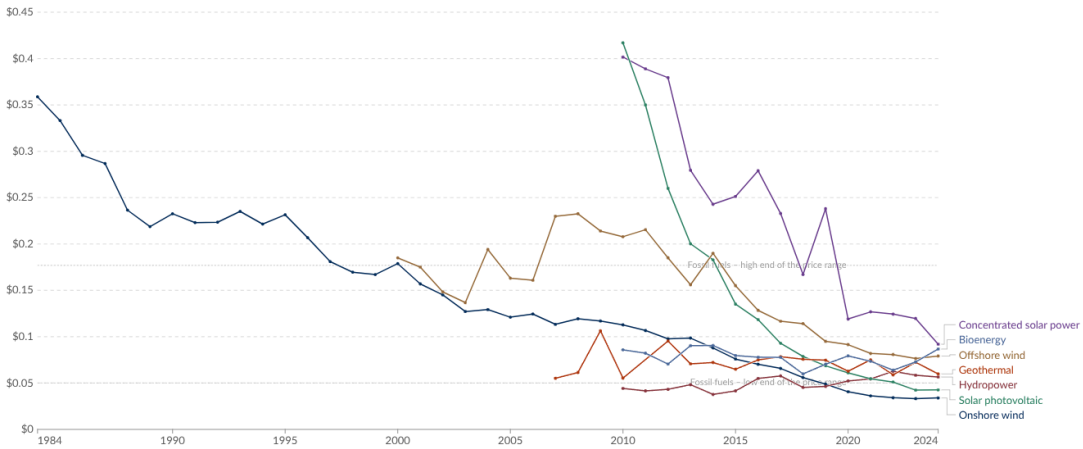


Figure 69: *The levelized cost of energy (LCOE) accounts for everything, the cost of building the power plant, plus the ongoing costs in keeping it operational over the lifetime of the plant. The cost of energy generation per watt has been falling for decades, and even more dramatically for renewable sources (Our World in Data, 2026).* (interactive version on website)

9. Appendix: Takeoff

9.1 Continuity

Continuity describes whether AI capabilities improve smoothly and predictably or through sudden, unexpected jumps. This is different from speed - even a fast takeoff could be continuous if the rapid progress follows predictable patterns, and a slow takeoff could be discontinuous if it involves surprising breakthroughs. Understanding continuity helps us predict whether we can extrapolate from current trends, like the scaling laws we discussed earlier, or if we should expect sudden departures from these patterns. So if you think of speed as a measure of how quickly the AI becomes superintelligent, continuity can be thought of as a measure of "surprise".

In a continuous takeoff, AI capabilities follow smooth, predictable trends. The improvements we've seen in language models provide a good example - each new model tends to be somewhat better than the last at tasks like coding or math, following patterns we can roughly predict from scaling laws and algorithmic improvements. As we saw in the forecasting section, many aspects of AI progress have shown this kind of predictable behavior.

Continuous progress doesn't mean linear or simple progress. It might still involve exponential or even superexponential growth, but the key is that this growth follows patterns we can anticipate. Think of how GPT-4 is better than GPT-3, which was better than GPT-2 - each improvement was significant but not completely surprising given the increase in scale and improved training techniques.

A continuous takeoff suggests that current trends in scaling laws and algorithmic progress might extend even to transformative AI systems. This would give us more warning about upcoming capabilities and more ability to prepare appropriate safety measures. As we'll discuss in the governance chapter, even though progress is fast, this kind of predictability makes it comparatively easier to develop and implement regulation before AI systems become extremely powerful or uncontrollable. Keeping in mind of course that comparatively easier does not mean "easy".

A discontinuous takeoff involves sudden jumps in capability that break from previous patterns. Instead of steady improvements in performance as we add compute or data, we might see the emergence of entirely new capabilities that weren't predicted by existing trends. One hypothetical example would be if an AI system suddenly developed robust general reasoning capabilities after appearing to only handle narrow tasks - this would represent a discontinuity in the pattern of AI development. The terms 'fast takeoff' and 'discontinuous takeoff' are often used interchangeably. However, the images below displaying different takeoff trajectories might help in clarifying the subtle differences between the concepts.

Discontinuities could arise through various mechanisms. We might discover fundamentally new training approaches that are dramatically more efficient than current methods. Or, we might hit tipping points for emergent capabilities - where quantitative improvements in scale lead to qualitative changes in capability. An AI system might even discover such improvements about itself, leading to unexpected jumps in capability.

The historical record provides some precedent for both continuous and discontinuous scientific progress. The development of nuclear weapons represented a discontinuous jump in explosive power, while improvements in computer processing power have followed more

continuous trends. However, as we saw in the forecasting section, technological discontinuities have historically been rare, which some researchers cite as evidence favoring continuous takeoff scenarios.

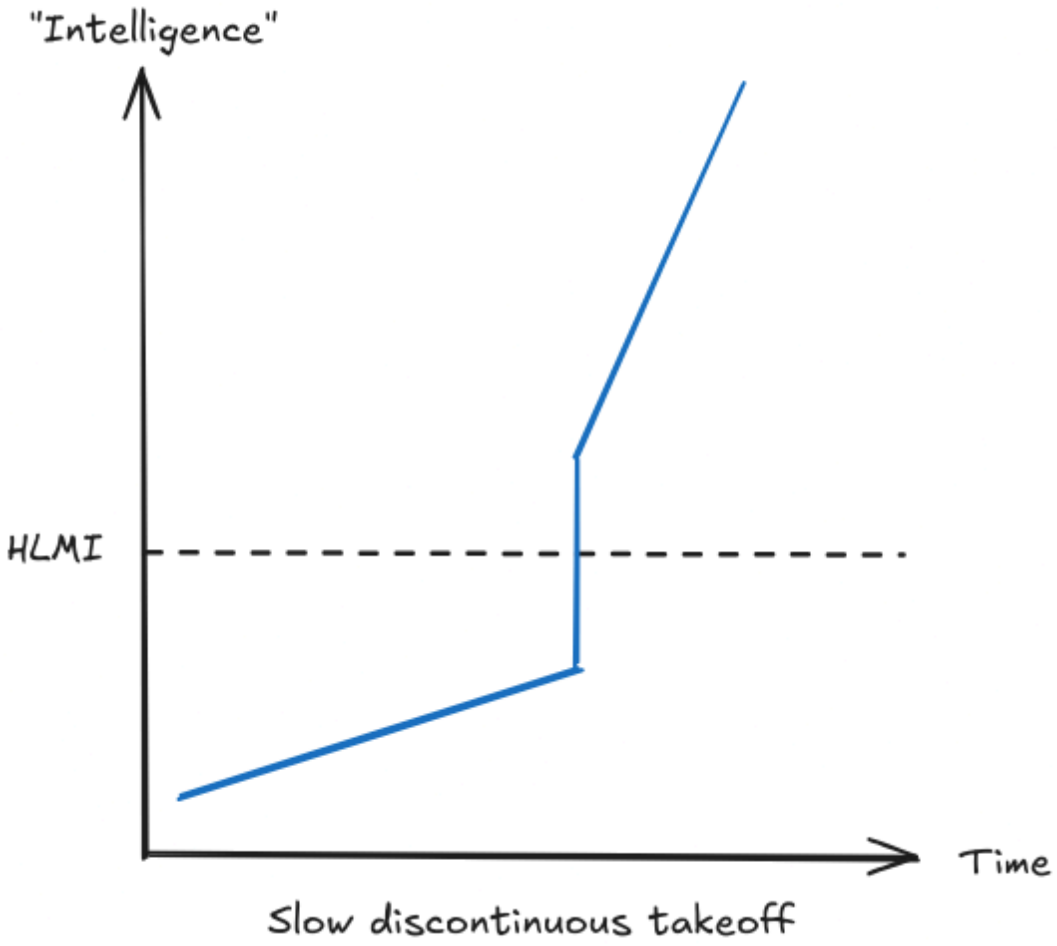


Figure 70: *One example illustration of slow discontinuous takeoff, where even though progress keeps increasing we might see sudden 'jumps' in progress (Martin & Eth, 2021).*

9.2 Homogeneity

Homogeneity describes how similar or different AI systems are to each other during the takeoff period. Will we see many diverse AI systems with different architectures and capabilities, or will most advanced AI systems be variations of the same basic design? This isn't just about technical diversity - it's about whether advanced AI systems will share similar behaviors, limitations, and safety properties (Hubinger, 2020).

In a homogeneous takeoff, most advanced AI systems would be fundamentally similar. We can see hints of this pattern today - many current language models are based on the transformer architecture and trained on similar data, leading to similar capabilities and limitations. In a homogeneous takeoff, this pattern would continue. Perhaps most AI systems would be fine-tuned versions of a few base models, or different implementations of the same core breakthrough in AI design.

A factor that could drive homogeneity is the sheer scale required to train advanced AI systems. If training transformative AI requires massive compute resources, as scaling laws suggest, then only a few organizations might be capable of training base models from scratch. Other organizations would build on these base models rather than developing entirely new architectures, leading to more homogeneous systems.

Homogeneous takeoff could be safer in some ways but riskier in others. If we solve alignment for one AI system, that solution might work for other similar systems. However, if there's a fundamental flaw in the common architecture or training approach, it could affect all systems simultaneously. It's like having a monoculture in agriculture - while easier to manage, it's also more vulnerable to shared weaknesses.

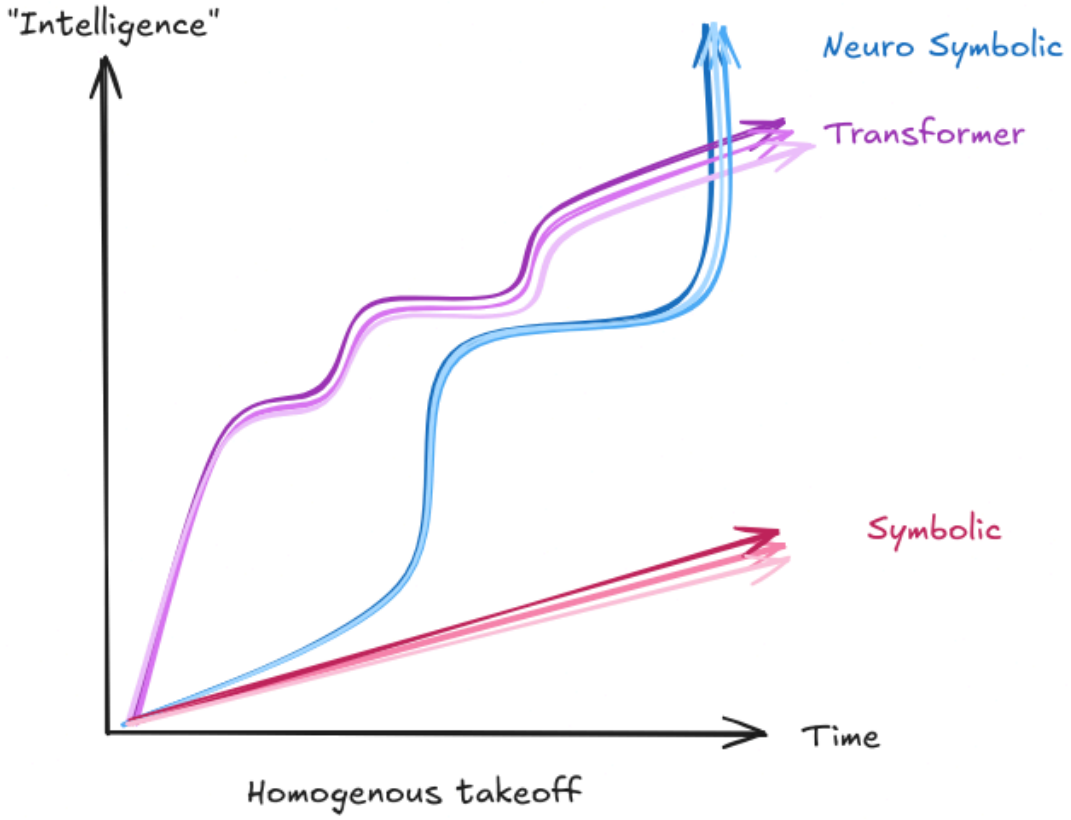


Figure 71: *An illustration of homogeneous takeoff. We can see multiple different overarching model architectures. The figure shows three in different colors. Within each architecture the takeoff is roughly the same due to similarity in design, regulations, and safety mitigations.* ***NOTE****: The curves here with architectures are purely illustrative, and are not meant to indicate predicted growth trajectories and comparisons between different architectures.*

A heterogeneous takeoff is when we see significant architectural diversity among advanced AI systems. Different organizations might develop fundamentally different approaches to AI, leading to systems with distinct strengths, weaknesses, and behaviors. Some might be specialized for specific domains while others are more general, some might be more transparent while others are more opaque, some might be more aligned with human values while others might not be. Competitive dynamics among AI projects could exacerbate diversity, as teams race to achieve breakthroughs without necessarily aligning on methodologies or sharing crucial information. As an example, we might have a future where AI becomes a strategic national asset, and AI development is

closely guarded. In this environment, the pursuit of AI capabilities becomes siloed, each company or country would then employ different development methodologies, potentially leading to a wide range of behaviors, functionalities, and safety levels.

Heterogeneous takeoff creates different challenges for safety. We'd need to develop safety measures that work across diverse systems, and we couldn't necessarily apply lessons learned from one system to others. However, diversity might provide some protection against systemic risks - if one approach proves dangerous, alternatives would still exist.

The degree of homogeneity during takeoff has significant implications for how transformative AI might develop. In a homogeneous scenario, progress might be more predictable but also more prone to winner-take-all dynamics. A heterogeneous scenario might be more robust against single points of failure but harder to monitor and control.

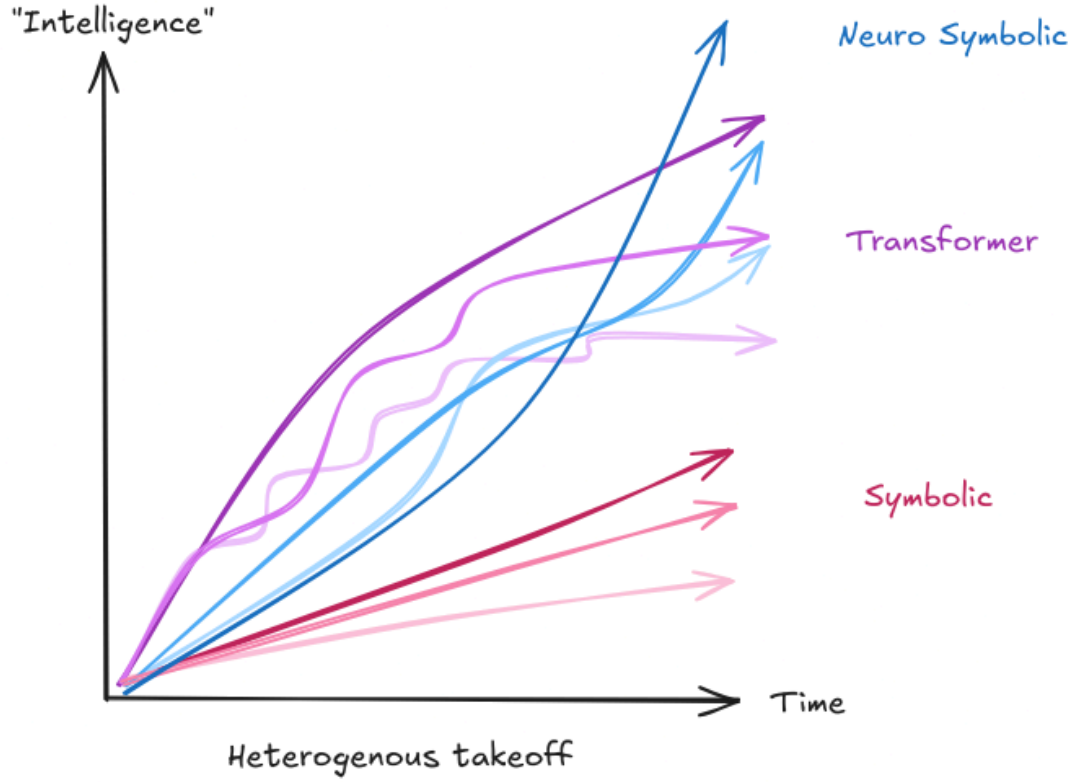


Figure 72: *One example of heterogeneous takeoff. We can see multiple different overarching model architectures. The figure shows three in different colors. Within each architecture the takeoff is different due to differences in design, regulations, and safety mitigations.* ***NOTE****: The curves here with architectures are purely illustrative, and are not meant to indicate predicted growth trajectories and comparisons between different architectures.*

9.3 Polarity

Polarity describes whether power and capability becomes concentrated in a single AI system or organization, or remains distributed among multiple actors. In other words, will one AI system or group pull dramatically ahead of all others, or will multiple AI systems advance in parallel with comparable capabilities?

In a unipolar takeoff, one AI system or organization gains a decisive lead over all others. This could happen through a single breakthrough, exceptional scaling advantages, or recursive self-improvement. For example, if one AI system becomes capable enough to substantially accelerate its own development, it might rapidly outpace all other systems. The mathematics of training compute provide one path to a unipolar outcome. If a doubling of compute leads to reliable improvements in capability, then an organization that gets far enough ahead in acquiring compute could maintain or extend their lead. Their improved systems could then help them develop even better training methods, hardware, and attract investment creating a positive feedback loop that others can't match. But compute isn't the only path to unipolarity. A single organization might discover a fundamentally better training approach, or develop an AI system that's better at improving itself than at helping humans build alternatives. Once any actor gets far enough ahead, it might become practically impossible for others to catch up.

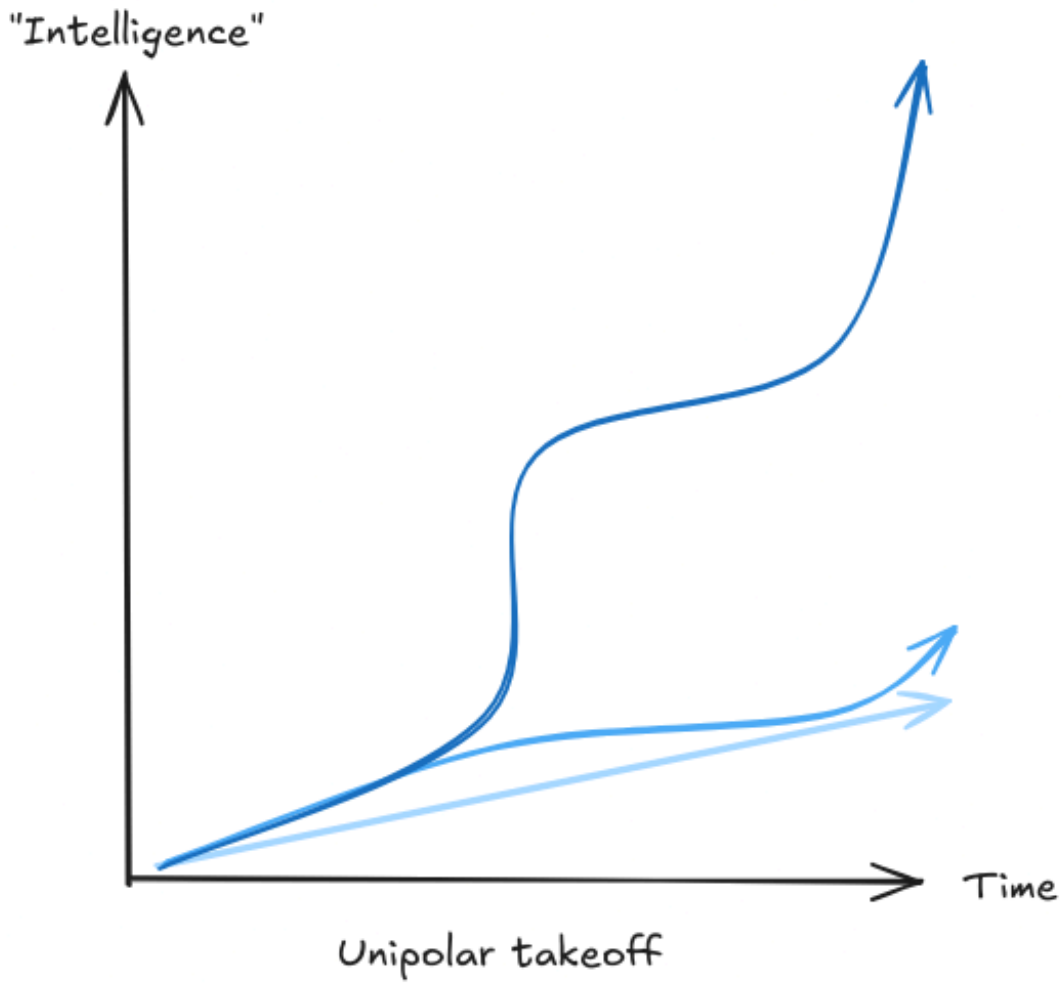


Figure 73: *An illustration of unipolar takeoff. One model (dark blue here) significantly outperforms all others.*

In a multipolar takeoff, multiple AI systems or organizations develop advanced capabilities in parallel. This could look like several large labs developing different but comparably powerful AI systems, or like many actors having access to similar AI capabilities through open source models or AI services. Today's AI landscape shows elements of multipolarity - multiple organizations can train large language models, and techniques developed by one lab are often quickly adopted by

others. A multipolar takeoff might continue this pattern, with multiple groups maintaining similar capabilities even as those capabilities become transformative. A unipolar scenario raises concerns about the concentration of power, while a multipolar world presents challenges in coordination among diverse entities or AI systems. Both unipolar and multipolar worlds have the potential for misuse of advanced AI capabilities by human actors.

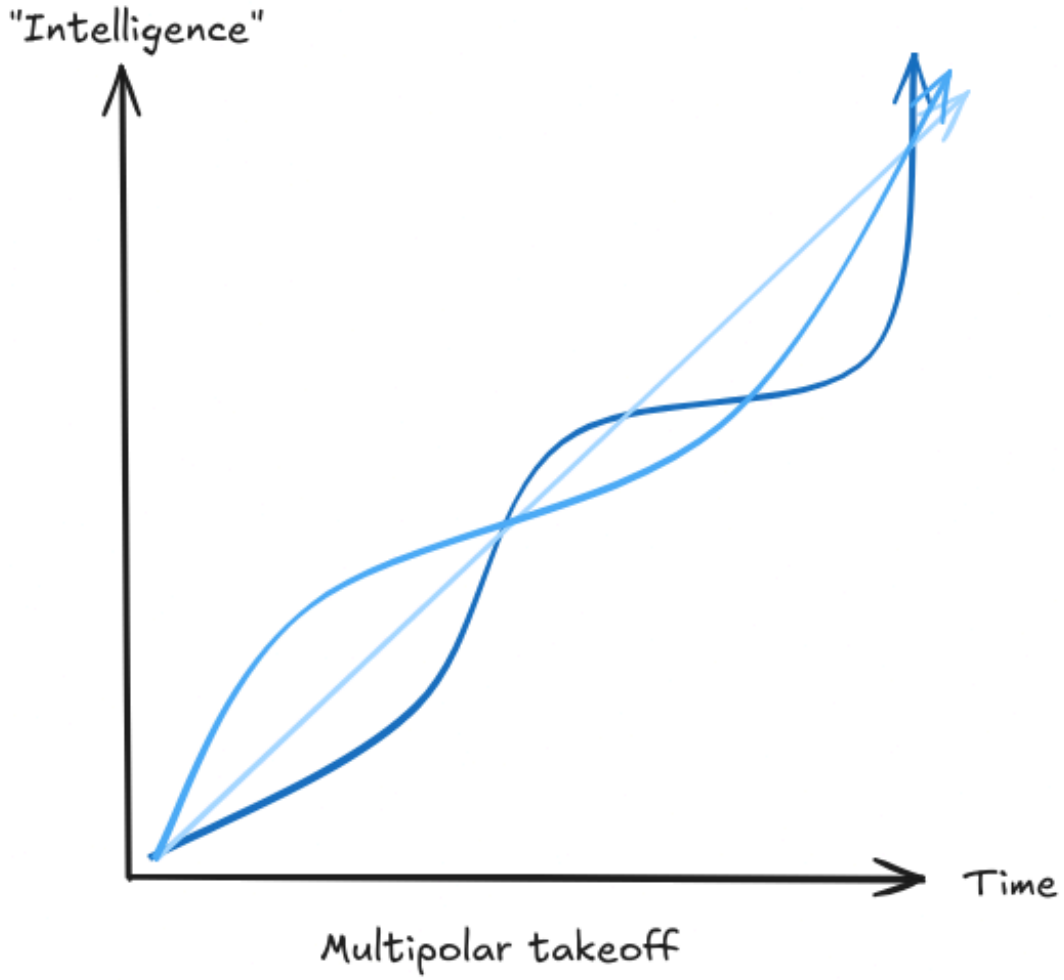


Figure 74: *An illustration of multipolar takeoff. No model significantly outperforms all others, and they all takeoff at a roughly competitive rate relative to each other.*

In a unipolar scenario, the actions and alignment of a single system or organization become crucial - they might gain the ability to shape the long-term future unilaterally. This concentrates risk in a single point of failure, but might also make coordination easier since fewer actors need to agree. A multipolar scenario creates different challenges. Multiple advanced systems might act in conflicting ways or compete for resources. This could create pressure to deploy systems quickly or cut corners on safety. There's also an important interaction between polarity and the other aspects of takeoff we've discussed. A fast takeoff might be more likely to become unipolar, as the first system to make rapid progress could quickly outpace all others. A slow takeoff might tend toward multipolarity, giving more actors time to catch up to any initial leads.

Factors Influencing Polarity . Several key elements influence whether takeoff polarity leans towards a unipolar or multipolar outcome:

- **Speed of AI Development:** A rapid takeoff might favor a unipolar outcome by giving a significant advantage to the fastest developer. In contrast, a slower takeoff could lead to a multipolar world where many entities reach advanced capabilities more or less simultaneously.
- **Collaboration vs. Competition:** The degree of collaboration and openness in the AI research community can significantly affect takeoff polarity. High levels of collaboration and information sharing could support a multipolar outcome, while secretive or highly competitive environments might push towards unipolarity.
- **Regulatory and Economic Dynamics:** Regulatory frameworks and economic incentives also play a crucial role. Policies that encourage diversity in AI development and mitigate against the accumulation of too much power in any single entity's hands could foster a multipolar takeoff.

10. Appendix: Expert Surveys

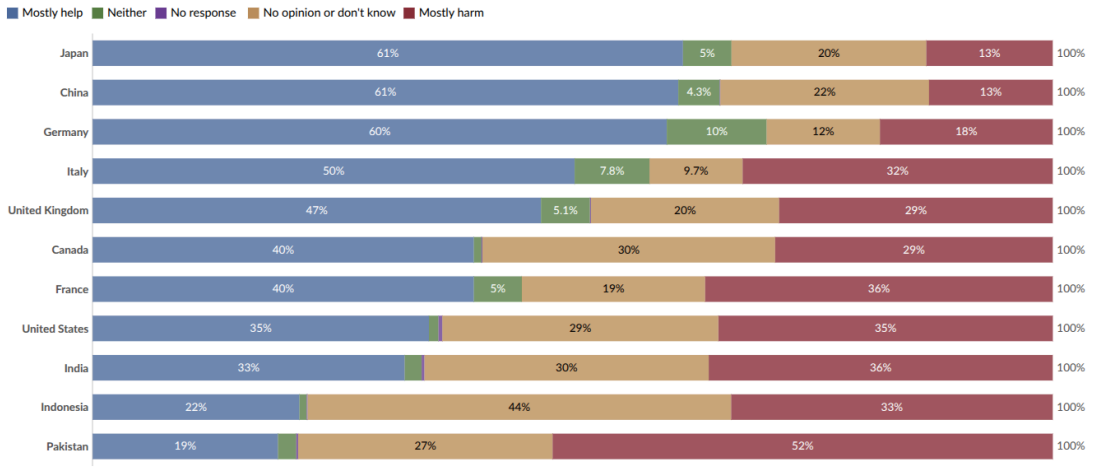


Figure 75: *Views about AI's impact on society in the next 20 years, 2021. Survey respondents were asked, "Will artificial intelligence help or harm people in the next 20 years?" (Giattino et al., 2023).* (interactive version on website)

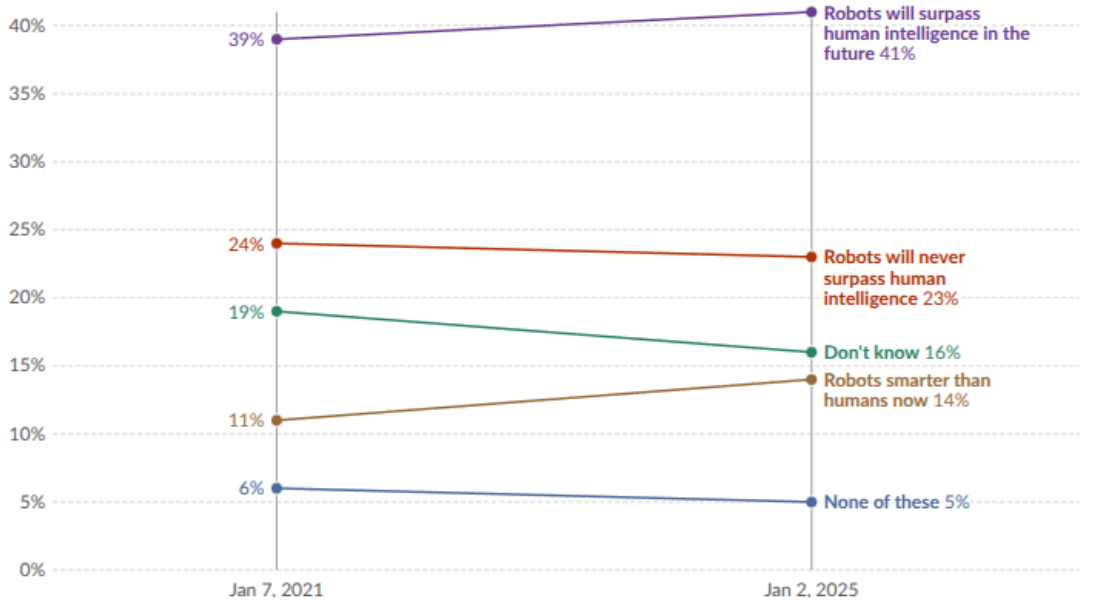


Figure 76: *Views of Americans about robot vs. human intelligence. Survey respondents were asked, "Which ONE, if any, of the following statements do you MOST agree with?" (Giattino et al., 2023).* (interactive version on website)

10.1 Surveys

According to a recent survey conducted by AI Impact (AI Impacts, 2022): " ***Expected time to human-level performance dropped 1–5 decades since the 2022 survey*** . *As always, our questions about 'high-level machine intelligence' (HLMI) and 'full automation of labor' (FAOL) got very different answers, and individuals disagreed a lot (shown as thin lines below), but the aggregate*

forecasts for both sets of questions dropped sharply. For context, between 2016 and 2022 surveys, the forecast for HLMI had only shifted about a year."

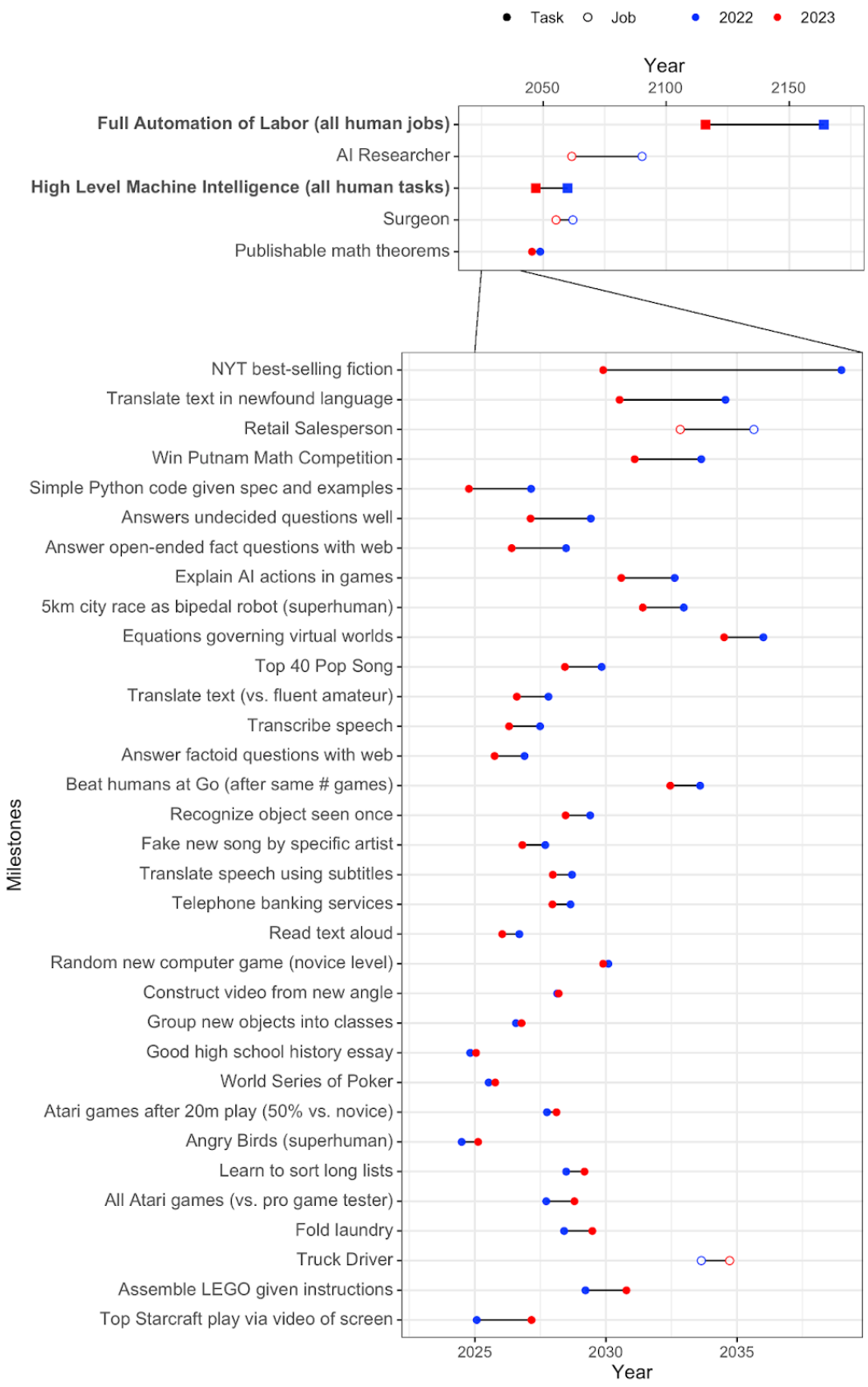


Figure 77: *2024 Survey of AI Experts (AI Impacts, 2022)*

It is also possible to compare the predictions of the same study in 2022 to the current results. It is interesting to note that the community has generally underestimated the speed of progress over the year 2023 and has adjusted its predictions downward. Some predictions are quite surprising. For example, tasks like "Write High School Essay" and "Transcribe Speech" are arguably already automated with ChatGPT and Whisper, respectively. However, it appears that researchers are not aware of these results. Additionally, it is surprising that the forecast for when we are able to build an "AI researcher" has longer timelines than when we are able to build "High-level machine intelligence (all human tasks)". The median of the 2024 expert survey predicts human-level machine intelligence (HLMI) in 2049.

10.2 Quotes

Here are many quotes from people regarding transformative AI.

10.2.1 AI Experts

Note that Hinton, Bengio, and Sutskever are some of the most cited researchers in the field of AI. And that Hinton, Bengio, and LeCun are the recipients of the Turing Award in Deep Learning . Some users on reddit have put together a comprehensive list of publicly stated AI timelines forecasts from famous researchers and industry leaders.

> *The research question is: how do you prevent them from ever wanting to take control? And nobody knows the answer [...] The alarm bell I'm ringing has to do with the existential threat of them taking control [...] If you take the existential risk seriously, as I now do, it might be quite sensible to just stop developing these things any further [...] it's as if aliens had landed and people haven't realized because they speak very good English.*
>
> **Geoffrey Hinton**
> Godfather of modern AI, Turing Award Recipient

> *It's very hard, in terms of your ego and feeling good about what you do, to accept the idea that the thing you've been working on for decades might actually be very dangerous to humanity... I think that I didn't want to think too much about it, and that's probably the case for others [...] Rogue AI may be dangerous for the whole of humanity. Banning powerful AI systems (say beyond the abilities of GPT-4) that are given autonomy and agency would be a good start.*

Yoshua Bengio
One of most cited scientists ever, Godfather of modern AI, Turing Award Recipient

If we pursue [our current approach], then we will eventually lose control over the machines.

Stuart Russell
Co-Author of leading AI textbook, Co-Founder of the Center for Human-Compatible AI

We must take the risks of AI as seriously as other major global challenges, like climate change. It took the international community too long to coordinate an effective global response to this, and we're living with the consequences of that now. We can't afford the same delay with AI [...] then maybe there's some kind of equivalent one day of the IAEA, which actually audits these things.

Demis Hassabis
Co-Founder and CEO of DeepMind

When I think of why am I scared [...] I think the thing that's really hard to argue with is like, there will be powerful models; they will be agentic; we're getting towards them. If such a model wanted to wreak havoc and destroy humanity or whatever, I think we have basically no ability to stop it.

Dario Amodei
Co-Founder and CEO of Anthropic, Former Head of AI Safety at OpenAI

[About a Pause] I don't rule it out. And I think that at some point over the next five years or so, we're going to have to consider that question very seriously.

Mustafa Suleyman
CEO of Microsoft AI, Co-Founder of DeepMind

The future is going to be good for the AIs regardless; it would be nice if it would be good for humans as well [...] It's not that it's going to actively hate humans and want to harm them, but it's just going to be too powerful, and I think a good analogy would be the way humans treat animals [...] And I think by default that's the kind of relationship that's going to be between us and AGIs which are truly autonomous and operating on their own behalf.

Ilya Sutskever
One of the most cited scientists ever, Co-Founder and Former Chief Scientist at OpenAI

Do possible risks from AI outweigh other possible existential risks...? It's my number 1 risk for this century [...] A lack of concrete AGI projects is not what worries me, it's the lack of concrete plans on how to keep these safe that worries me.

Shane Legg

[After resigning at OpenAI, talking about sources of risks] These problems are quite hard to get right, and I am concerned we aren't on a trajectory to get there [...] OpenAI is shouldering an enormous responsibility on behalf of all of humanity. But over the past years, safety culture and processes have taken a backseat to shiny

products. We are long overdue in getting incredibly serious about the implications of AGI.

Jan Leike
Former co-lead of the Superalignment project at OpenAI

[Suggesting about how to ask for a global regulatory body:] "any compute cluster above a certain extremely high-power threshold – and given the cost here, we're talking maybe five in the world, something like that – any cluster like that has to submit to the equivalent of international weapons inspectors" [...] I did a big trip around the world this year, and talked to heads of state in many of the countries that would need to participate in this, and there was almost universal support for it.

Sam Altman
Co-Founder and CEO of OpenAI

The exact way the post-AGI world will look is hard to predict — that world will likely be more different from today's world than today's is from the 1500s [...] We do not yet know how hard it will be to make sure AGIs act according to the values of their operators. Some people believe it will be easy; some people believe it'll be unimaginably difficult; but no one knows for sure.

Greg Brockman
Co-Founder and Former CTO of OpenAI

[Talking about times near the creation of the first AGI] you have the race dynamics where everyone's trying to stay ahead, and that might require compromising on

safety. So I think you would probably need some coordination among the larger entities that are doing this kind of training [...] Pause either further training, or pause deployment, or avoiding certain types of training that we think might be riskier.

John Schulman
Co-Founder of OpenAI

I've not met anyone in AI labs who says the risk [from training a next-gen model] is less than 1% of blowing up the planet. It's important that people know lives are being risked [...] One thing that a pause achieves is that we will not push the Frontier, in terms of risky pre-training experiments.

Jaan Tallinn
Co-Founder of Skype, Future of Life Institute

10.2.2 Academics

An ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

I. J. Good
Cryptologist at Bletchley Park

It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers... They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control.

Alan Turing
Father of Computer Science and AI

The development of full artificial intelligence could spell the end of the human race [...] It would take off on its own, and re-design itself at an ever increasing rate.

Stephen Hawking
Theoretical Physicist

I do not expect something actually smart to attack us with marching robot armies with glowing red eyes where there could be a fun movie about us fighting them. I expect an actually smarter and uncaring entity will figure out strategies and technologies that can kill us quickly and reliably and then kill us.

Eliezer Yudkowsky
AI safety researcher, Co-Founder of Machine Intelligence Research Institute

10.2.3 Tech Entrepreneurs

AI is a rare case where I think we need to be proactive in regulation than be reactive [...] I think that [digital super intelligence] is the single biggest existential crisis that we face and the most pressing one. It needs to be a public body that

has insight and then oversight to confirm that everyone is developing AI safely [...] And mark my words, AI is far more dangerous than nukes. Far. So why do we have no regulatory oversight? This is insane.

Elon Musk
Founder/Co-Founder of OpenAI, Neuralink, SpaceX, xAI, PayPal, CEO of Tesla, CTO of X/Twitter

Superintelligent AIs are in our future. [...] There's the possibility that AIs will run out of control. [Possibly,] a machine could decide that humans are a threat, conclude that its interests are different from ours, or simply stop caring about us.

Bill Gates
Co-Founder of Microsoft

10.2.4 Joint Declarations

Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent. These issues are in part because those capabilities are not fully understood [...] There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models.

The Bletchley Declaration **2023**
Multiple Nations & EU

11. Appendix: Discussion on LLMs

Current LLMs, although trained on abundant data, are still far from perfect.

Will these problems persist in future iterations, or will they disappear? This section examines the main criticisms of those models and tries to determine if they are valid even for future LLMs.

This kind of qualitative assessment is important to know whether LLMs represent the most likely route to AGI or not.

11.1 Empirically insufficiency?

Can LLMs be creative? The creativity of LLMs is often debated, but there are clear indications that AI, in principle, is capable of creative processes in various ways:

- **Autonomous Scientific Research**: Recent advancements have shown that LLMs can indeed make novel discoveries. For example, a study by DeepMind demonstrated that an LLM "*discovered new solutions for the cap set problem, a long-standing open problem in mathematics*" (DeepMind, 2023) which was a favorite open problem of Terence Tao. This indicates that AI can not only understand existing knowledge but also contribute new insights in complex fields like mathematics.
- **Autonomous Discovery**: AI has the capability to rediscover human strategies and openings independently. AlphaGo, for example, rediscovered human Go strategies and openings through self-play (McGrath et al., 2021), without any human data input. This demonstrates an AI's ability to independently learn and innovate within established domains.
- **Creative Optimization**: AI can optimize in surprisingly creative ways. The phenomena of specification gaming, where AI finds unintended solutions to problems, illustrate this. Although this unpredictability poses its challenges, it also shows that AI systems can come up with novel, creative solutions that might not be immediately obvious or intuitive to human problem solvers. DeepMind's blog post on Specification Gaming illustrates this point vividly (Krakovna et al., 2020).

Aren't LLMs just too slow at learning things? Arguments against transformer based language models often state that they are too sample inefficient, and that LLMs are extremely slow to learn new concepts when compared to humans. To increase performance in new tasks or situations, it's often argued that LLMs require training on vast amounts of data — millions of times more than a human would need. However, there's a growing trend towards data efficiency, and an increasing belief that this can be significantly improved in future models.

EfficientZero is a reinforcement learning agent that surpasses median human performance on a set of 26 Atari games after just two hours of real-time experience per game (Ye et al., 2021 ; Wang et al., 2024). This is a considerable improvement over previous algorithms, showcasing the potential leaps in data efficiency. The promise here is not just more efficient learning but also the potential for rapid adaptation and proficiency in new tasks, akin to a child's learning speed. EfficientZero is not an LLM, but it shows that deep learning can sometimes be made efficient.

Scaling laws indicate that larger AIs tend to be more data efficient, requiring less data to reach the same level of performance as their smaller counterparts. Papers such as "Language Models are Few-Shot Learners" (Brown et al., 2020) and the evidence that larger models seem to take less data to reach the same level of performance (Kaplan et al., 2020), suggest that as models scale, they become more proficient with fewer examples. This trend points towards a future where

AI might be able to rapidly adapt and learn from limited data, challenging the notion that AIs are inherently slow learners compared to humans.

Are LLMs robust to distributional shifts? While it is true that AI has not yet achieved maximal robustness, for example being able to perform perfectly after a change in distribution, there has been considerable progress:

- **Robustness correlates with capabilities**: Robustness is closely linked to the capabilities of AI models when AIs are trained on difficult tasks. For instance, there is a significant improvement in robustness and transfer learning from GPT-2 to GPT-4. In computer vision, recent models like Segment Anything (Kirillov et al., 2023) are far more robust and capable of transfer learning than their less capable predecessors. This progression isn't due to any mysterious factors but rather a result of scaling and improving upon existing architectures.
- **Robustness is a continuum, and perfect robustness may be not necessary:** Robustness in AI should not be viewed as a binary concept, but rather as existing on a continuum. This continuum is evident in the way AI models, like those in image classification, often surpass human performance in both capability and robustness (Korzekwa, 2022). However, it's important to recognize that no system is completely immune to challenges such as adversarial attacks. This is exemplified by advanced AIs like Katago in Go, which, despite being vulnerable to such attacks (Wang et al., 2022), still achieves a superhuman level of play. However, the quest for perfect robustness may not be essential to create capable transformative AI, as even systems with certain vulnerabilities can achieve superhuman levels of competence. However, while robustness may not be necessary to create capable AI, the creation of safe, aligned AI will have to solve the problem of misgeneralizing goals.

11.2 Shallow Understanding?

Stochastic Parrots: Do AIs only memorize information without truly compressing it? There are two archetypal ways to represent information in an LLM: either memorize point by point, like a look-up table, or compress the information by only memorizing higher-level features, which we can then call "the world model". This is explained in the very important paper "Superposition, Memorization, and Double Descent" (Anthropic, 2023): it turns out that to store points, initially the model learns the position of all the points (pure memorization), then, if we increase the number of points, the model starts to compress this knowledge, and the model is now capable of generalization (and implements a simple model of the data).

> *Unfortunately, too few people understand the distinction between memorization and understanding. It's not some lofty question like 'does the system have an internal world model?', it's a very pragmatic behavior distinction: 'is the system capable of broad generalization, or is it limited to local generalization?'*
>
> **Francois Chollet** — Chollet, 2023
> Prominent AI Researcher

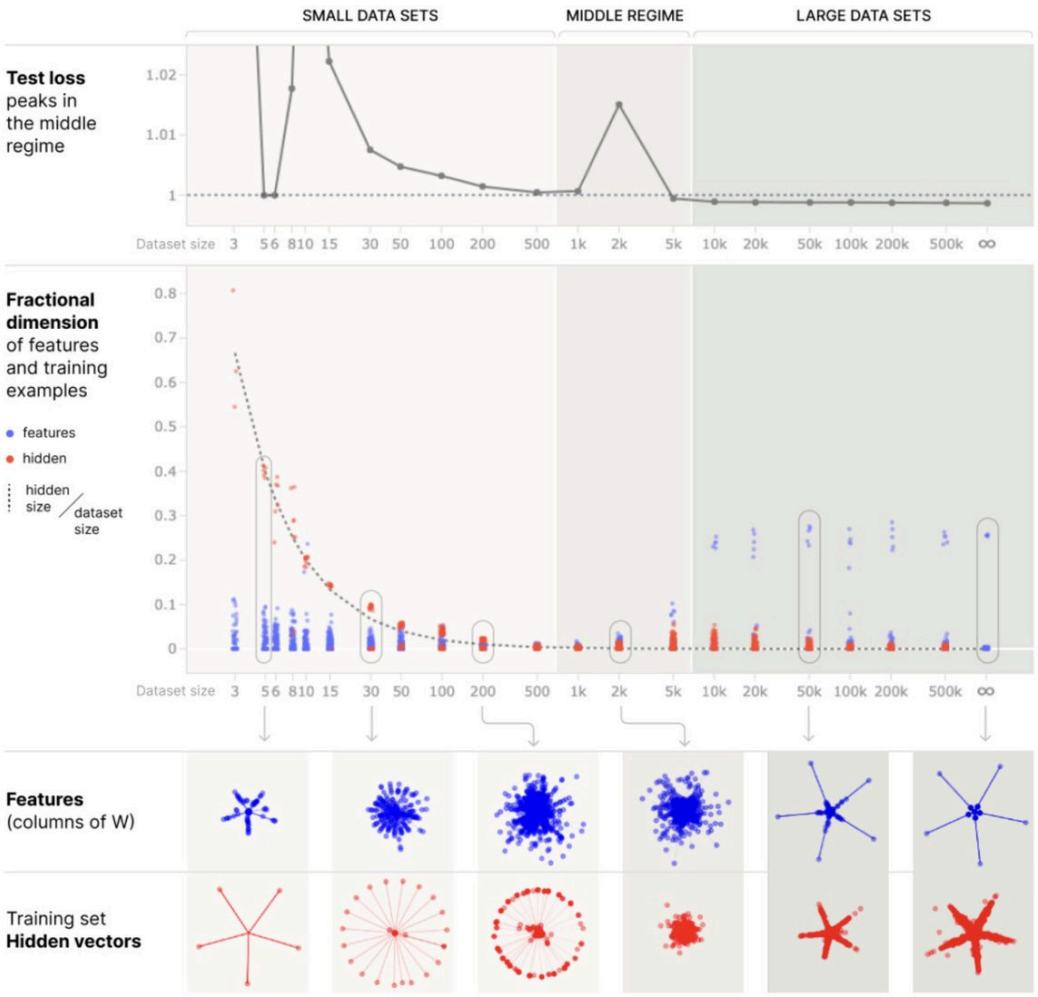


Figure 78: *From Superposition, Memorization, and Double Descent (Anthropic, 2023)*

AI is capable of compressing information, often in a relevant manner. For example, when examining the representations of words representing colors in LLMs like "red" and "blue", the structure formed by all the embeddings of those colors creates the correct color circle (This uses a nonlinear projection such as a T-distributed stochastic neighbor embedding (T-SNE) to project from high-dimensional space to the 2D plane). Other examples of world models are presented in a paper called "Eight Things to Know about Large Language Models" (Bowman, 2023).

Of course, there are other domains where AI resembles more of a look-up table, but it is a spectrum, and each case should be examined individually. For example, for "factual association," the paper "Locating and Editing Factual Associations in GPT" shows that the underlying data structure for GPT-2 is more of a look-up table (Meng et al., 2023), but the paper "Emergent Linear Representations in World Models of Self-Supervised Sequence Models" demonstrates that a small GPT is capable of learning a compressed world model of OthelloGpt. (Nanda et al., 2023) There are more examples in the section dedicated to world models in the paper "Eight Things to Know about Large Language Models" (Bowman, 2023).

It's clear that LLMs are compressing their representations at least a bit. Many examples of impressive capabilities are presented in the work "The Stochastic Parrot Hypothesis is debatable for the last generation of LLMs", which shows that it cannot be purely a memorization. (Feuillade-Montixi & Peigné, 2023)

Will LLMs Inevitably Hallucinate?

LLMs are prone to "hallucinate," a term used to describe the generation of content that is nonsensical or factually incorrect in response to certain prompts. This issue, highlighted in studies such as "On Faithfulness and Factuality in Abstractive Summarization" by Maynez et al. (Maynez et al., 2020) and "TruthfulQA: Measuring How Models Mimic Human Falsehoods" by Lin et al. (Lin et al., 2022), poses a significant challenge. However, it's important to see that these challenges are anticipated due to the training setup and can be mitigated:

- **Inherent Bias in Source Texts**: One of the fundamental reasons LLMs may produce untrue content is training data, which may not always be entirely factual or unbiased. In essence, LLMs are reflecting the diverse and sometimes contradictory nature of their training data. In this context, LLMs are constantly 'hallucinating', but occasionally, these hallucinations align with our perception of reality.
- **Strategies to Enhance Factual Accuracy**: The tendency of LLMs to generate hallucinations can be significantly diminished using various techniques. See the box below for a breakdown of those.
- **Larger models can be more truthful than smaller ones.** This is the case with TruthfulQA. OpenAI reports that GPT-4 is 40% more accurate and factually consistent than its predecessor.

Many techniques can be used to increase the truthfulness of LLMs

OPTIONAL NOTE

Fine-tuning **LLMs for Factuality:** In this paper (Tian et al., 2023), the authors recommend fine-tuning methods using Direct Preference Optimization (DPO) to decrease the rate of hallucinations. By applying such techniques, a 7B Llama 2 model saw a 58% reduction in factual error rate compared to its original model.**Retrieval Augmented Generation (RAG)**. This method works by incorporating a process of looking up real-world information (retrieval, like a Google search) and then using that information to guide the AI's responses (generation, based on the document retrieved). By doing so, the AI is better anchored in factual reality, reducing the chances of producing unrealistic or incorrect content. Essentially, it's like giving the AI a reference library to check facts against while it learns and responds, ensuring its output is more grounded in reality. This approach is particularly useful in the context of in-context learning, where the AI learns from the information and context provided in each interaction.**Prompting techniques** in AI have evolved to include sophisticated methods like- **Consistency checks** (Fluri et al., 2023), that involve comparing the output from multiple instances of the model on the same prompt, identifying and resolving any disagreements in the responses. This method enhances the accuracy and credibility of the information provided. For example, if different iterations of the model produce conflicting answers, this discrepancy can be used to refine and improve the model's understanding.

- **Reflexion.** The Reflexion technique ("Reflexion: Language Agents with Verbal Reinforcement Learning"): It's possible to simply ask the LLM to take a step back, to question whether what it has done is correct or not, and to consider ways to improve the previous answer, and this enhances a lot the capabilities of GPT-4. This technique is emergent and does not work well with previous models. (Shinn et al., 2023).
- **Verification chains**, like **selection inference** (Creswell et al., 2022). Chain-of-Thought has access to the whole context, so each reasoning step is not necessarily causally connected to the last. But selection inference enforces a structure where each reasoning step necessarily follows from the last, and therefore the whole reasoning chain is causal. This process involves the AI model examining its own reasoning or the steps it took to arrive at a conclusion. By doing so, it can verify the logic and consistency of its responses, ensuring they are well-founded and trustworthy.

- **Allowing the AI to express degrees of confidence** in its answers, acknowledging uncertainty when appropriate. For example, instead of a definitive "Yes" or "No," the model might respond with "I am not sure," reflecting a more nuanced understanding akin to human reasoning. This approach is evident in advanced models like Gopher (Rae et al., 2022), which contrasts with earlier models such as WebGPT which may not exhibit the same level of nuanced responses.

Process-based training ensures that the systems are accustomed to detailing their thoughts in much greater detail and not being able to skip too many reasoning steps. For example, see OpenAI's Improving Mathematical Reasoning with process supervision (Lightman et al., 2023).**Training for metacognition**: Models can be trained to give the probability of what they assert, a form of metacognition. For example, the paper "Language Models (Mostly) Know What They Know" (Kadavath et al., 2022) demonstrates that AIs can be Bayesian calibrated about their knowledge. This implies that they can have a rudimentary form of self-awareness, recognizing the likelihood of their own accuracy. Informally, this means it is possible to query a chatbot with "Are you sure about what you are telling me?" and receive a relatively reliable response. This can serve as training against hallucinations.It's worth noting that these techniques enable substantial problem mitigation for the current LLMs, but they don't solve all the problems that we encounter with AI that are potentially deceptive, as we will see in the chapter on goal misgeneralization.

11.3 Structural inadequacy?

Are LLMs missing System 2? System 1 and System 2 are terms popularized by economist Daniel Kahneman in his book "Thinking, Fast and Slow," describing the two different ways our brains form thoughts and make decisions. System 1 is fast, automatic, and intuitive; it's the part of our thinking that handles everyday decisions and judgments without much effort or conscious deliberation. For example, when you recognize a face or understand simple sentences, you're typically using System 1. On the other hand, System 2 is slower, more deliberative, and more logical. It takes over when you're solving a complex problem, making a conscious choice, or focusing on a difficult task. It requires more energy and is more controlled, handling tasks such as planning for the future, checking the validity of a complex argument, or any activity that requires deep focus. Together, these systems interact and influence how we think, make judgments, and decide, highlighting the complexity of human thought and behavior.

A key concern is whether LLMs are able to emulate System 2 processes, which involve slower, more deliberate, and logical thinking. Some theoretical arguments about the depth limit in transformers show that they are provably incapable of internally dividing large integers (Delétang et al., 2023). However, this is not what we observe in practice: GPT-4 is capable of detailing some calculations step-by-step and obtaining the expected result through a chain of thought or via the usage of tools like a code interpreter.

Emerging Metacognition . Emerging functions in LLMs, like the Reflexion technique (Shinn et al., 2023), allow these models to retrospectively analyze and improve their answers. It is possible to ask the LLM to take a step back, question the correctness of its previous actions, and consider ways to improve the previous answer. This greatly enhances the capabilities of GPT-4, enhancing its capabilities and aligning them more closely with human System 2 operations. Note that this technique is emergent and does not work well with previous models.

These results suggest a blurring of the lines between these two systems. System 2 processes may be essentially an assembly of multiple System 1 processes, appearing slower due to involving more

steps and interactions with slower forms of memory. This perspective is paralleled in how language models operate, with each step in a System 1 process akin to a constant time execution step in models like GPT. Although these models struggle with intentionally orchestrating these steps to solve complex problems, breaking down tasks into smaller steps (Least-to-most prompting) or prompting them for incremental reasoning (Chain-of-Thought (CoT) prompting) significantly improves their performance.

Are LLMs missing an internal world model? The notion of a "world model" in AI need not be confined to explicit encoding within an architecture. Contrary to approaches like H-JEPA (LeCun, 2022), which advocate for an explicit world model to enhance AI training, there's growing evidence that a world model can be effectively implicit. This concept is particularly evident in reinforcement learning (RL), where the distinction between model-based and model-free RL can be somewhat misleading. Even in model-free RL, algorithms often implicitly encode a form of a world model that is crucial for optimal performance.

- **Time and geographical coordinates:** Research on Llama-2 models reveals how these models can represent spatial and temporal information (Gurney & Tegmark, 2024). LLMs like Llama-2 models encode approximate real-world coordinates and historical timelines of cities. Key findings include the gradual emergence of geographical representations across model layers, the linearity of these representations, and the models' robustness to different prompts. Significantly, the study shows that the models are not just passively processing this information but actively learning the global geometry of space and time.
- **Board representation:** In the paper "Emergent Linear Representations in World Models of Self-Supervised Sequence Models" (Nanda et al., 2023), the author presents significant findings on the nature of representations in AI models. The paper delves into how the Othello-GPT model, trained to predict legal moves in the game of Othello, develops an emergent world representation of the game board! Contrary to previous beliefs that this representation was non-linear, he demonstrates that it is, in fact, linear. He discovers that the model represents board states not in terms of black or white pieces, but as "my color" or "their color," aligning with the model's perspective of playing both sides. This work sheds light on the potential of AI models to develop complex, yet linear, world representations through simple objectives like next-token prediction.
- **Other examples** are presented in the paper: "Eight Things to know about LLMs". (Bowman, 2023)

Can LLMs learn continuously, and have long term memory? Continual learning and the effective management of long-term memory represent significant challenges in the field of AI in general.

Catastrophic Forgetting . A crucial obstacle in this area is catastrophic forgetting, a phenomenon where a neural network , upon learning new information, tends to entirely forget previously learned information. This issue is an important focus of ongoing research, aiming to develop AI systems that can retain and build upon their knowledge over time. For example, suppose we train an AI on an Atari game. At the end of the second training, the AI has most likely forgotten how to play the first game. This is an example of catastrophic forgetting.

But now suppose we train a large AI on many ATARI games, simultaneously, and even add some Internet text and some robotic tasks. This can just work. For example, the AI GATO illustrates this training process and exemplifies what we call the **blessing of scale** , which is that what is impossible in small regimes can become possible in large regimes.

Other techniques are being developed to solve long-term memory, for example, **Scaffolding-based approaches** have also been employed for achieving long-term memory and continual learning in AI. Scaffolding in AI refers to the use of hard-coded wrappers explicitly programmed structures by humans that involve a for loop to query continuously the model:

- **LangChain** addresses these challenges by creating extensive memory banks. LangChain is a Python library that allows LLM to retrieve and utilize information from large datasets, essentially providing a way for AI to access a vast repository of knowledge and use this information to construct more informed responses. However, this approach may not be the most elegant due to its reliance on external data sources and complex retrieval mechanisms. A potentially more seamless and integrated solution could involve utilizing the neural network's weights as dynamic memory, constantly evolving and updating based on the tasks performed by the network.
- **Voyager:** A remarkable example of a scaffolding-based long-term memory is the AI Voyager, an AI system developed under the "AutoGPT" paradigm. This system is notable for its ability to engage in continuous learning within a 3D game environment like Minecraft. In a single game session, AI Voyager demonstrates the capacity to learn basic controls, achieve initial goals such as resource acquisition, and eventually advance to more complex behaviors, including combat with enemies and crafting tools for gathering sophisticated resources. This demonstrates a significant stride in LLM's ability to learn continually and manage long-term memory within dynamic environments.

It should be noted that scaffold-based long-term memory is not considered an elegant solution, and purists would prefer to use the system's own weights as long-term memory.

Planning

Planning is an area that AIs currently struggle with, but there is significant progress. Some paradigms, such as those based on scaffolding, enable task decomposition and breaking down objectives into smaller, more achievable sub-objectives.

Furthermore, the paper "Voyager: An Open-Ended Embodied Agent with Large Language Models" demonstrates that it is possible to use GPT-4 for planning in Natural language in Minecraft (Wang et al., 2023).

11.4 Differences with the brain

It appears that there are several points of convergence between the LLMs and the linguistic cortex:

- **Behavioral similarities.** LLMs show a close comparison to human linguistic abilities and the linguistic cortex (Canell, 2022). These models have excelled in mastering syntax and a significant portion of semantics in human language. Of course, today, they still lag in aspects such as long-term memory, coherence, and general reasoning - faculties that in humans depend on various brain regions like the hippocampus and prefrontal cortex, but we explained in the last sections that those problems may be solvable.
- **Convergence in internal Representations**: LLMs have a representation that converges with scale toward the brain representation. This is supported by the study, "Brains and algorithms partially converge in natural language processing." (Caucheteux & King, 2022) Additional insights can be found in the works "The Brain as a Universal Learning Machine" (Canell, 2015) and "Brain Efficiency: Much More than You Wanted to Know." (Canell, 2022) At comparable learning stages, LLMs and the linguistic cortex develop similar or equivalent feature representations. In some evaluations, advanced LLMs have been able to predict 100% of the explainable neural variance, as

detailed by Schrimpf, Martin, et al. in "The neural architecture of language: Integrative modeling converges on predictive processing." (Schrimpf et al., 2021)

- **Scale is also important in primates.** The principal architectural difference between human and other primate brains seems to be the number of neurons rather than anything else, as demonstrated in various studies. (Houzel, 2012 Pearson et al., 2023 Charvet, 2021).

11.5 Further reasons to continue scaling LLMs

Following are some reasons to believe that labs will continue to scale LLMs.

Scaling Laws on LLM implies further qualitative improvements. The scaling laws might not initially appear impressive. However, linking these quantitative measures can translate to a qualitative improvement in algorithm quality. An algorithm that achieves near-perfect loss, though, is one that necessarily comprehends all subtleties, and displays enormous adaptability. The fact that the scaling laws are not bending is very significant and means that we can make the model a qualitatively better reasoner.

From simple correlations to understanding. During a training run, GPTs go from basic correlations to deeper and deeper understanding. Initially, the model merely establishes connections between successive words. Gradually, it develops an understanding of grammar and semantics, creating links between sentences and subsequently between paragraphs. Eventually, GPT masters the nuances of writing style.

Exercise: Scaling Laws on LLM implies further qualitative improvements

OPTIONAL NOTE

Let's calculate the difference in loss, measured in bits, between two model outputs: "Janelle ate some ice cream because he likes sweet things like ice cream." and "Janelle ate some ice cream because she likes sweet things like ice cream." The sentence contains approximately twenty tokens. If the model vacillates between "He" or "She," choosing randomly (50/50 odds), it incurs a loss of 2 bits on the pronoun token when incorrect. The loss for other tokens remains the same in both models. However, since the model is only incorrect half the time, a factor of 1/2 should be applied. This results in a difference of (1/2) * (2/20) = 1/20, or 0.05 bits. Thus, a model within 0.05 bits of the minimal theoretical loss should be capable of understanding even more nuanced concepts than the one discussed above.

Text completion is probably an AI-complete test (Wikipedia, 2022).

Current LLMs have only as many parameters as small mammals have synapses, no wonder they are still imperfect. Models like GPT-4, though very big compared to other models, should be noted for their relatively modest scale compared to the size of a human brain. To illustrate, the largest GPT-3 model has a similar number of parameters to the synapses of a hedgehog. We don't really know how many parameters GPT-4 has, but if it is the same size as PALM, which has 512 B parameters, then GPT-4 has only as many parameters as a chinchilla has synapses. In contrast, the human neocortex contains about 140 trillion synapses, which is over 200 times more synapses than a chinchilla. For a more in-depth discussion on this comparison, see the related discussion here .

For a discussion of the number of parameters necessary to emulate a synapse, see the discussion on biological anchors.

GPT-4 is still orders of magnitude cheaper than other big science projects. : Despite the high costs associated with training large models, the significant leaps in AI capabilities provided by scaling justify these costs. For example, GPT-4 is expensive compared to other ML models. It is said to cost 50M in training. But the Manhattan Project cost 25B, which is 500 times more without accounting for inflation, and achieving Human-level intelligence, may be more economically important than achieving the nuclear bomb.

Collectively, these points support the idea that AGI can be achieved by only scaling current algorithms.

Acknowledgements

We would like to express our gratitude to Jeanne Salle, Charles Martinet, Vincent Corruble, Diego Dorn, Josh Thorsteinson, Jonathan Claybrough, Alejandro Acelas, Jamie Raldua Veuthey, Alexandre Variengien, Léo Dana, Angélina Gentaz, Nicolas Guillard, and Leo Karoubi for their valuable feedback, discussions, and contributions to this work.