NO PREP

How AI Got Smart So Fast

The 2022-2026 Boom

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

How AI Got Smart So Fast: The 2022-2026 Boom

Why Did This Happen Now?

If artificial intelligence has been around since the 1950s, why did the tools you use today suddenly appear in just the last few years? This is one of the most important questions in the entire field, and the answer is not simply that someone finally figured out how to make AI work. The technology had been improving steadily for decades. What changed around 2017 to 2022 was the convergence of three separate trends — each individually significant, but together sufficient to trigger a qualitative leap in capability that surprised even many of the researchers who had been building these systems. Understanding those three trends is the key

Name: Period: Date:

How AI Got Smart So Fast: The 2022-2026 Boom

Why Did This Happen Now?

If artificial intelligence has been around since the 1950s, why did the tools you use today suddenly appear in just the last few years? This is one of the most important questions in the entire field, and the answer is not simply that someone finally figured out how to make AI work. The technology had been improving steadily for decades. What changed around 2017 to 2022 was the convergence of three separate trends, each individually significant, but together sufficient to trigger a qualitative leap in capability that surprised even many of the researchers who had been building these systems. Understanding those three trends is the key to understanding why AI seems to have come out of nowhere. The three trends are: the invention of a new type of neural network architecture (the transformer), the availability of vast amounts of data from the internet, and a dramatic improvement in computing hardware, particularly graphics processing units (GPUs) that turned out to be ideal for running AI calculations. Remove any one of these three, and the current generation of AI tools would not exist. Together, they created what researchers sometimes call a 'phase transition': a point where gradual improvement suddenly produced qualitatively different results.

The Transformer Architecture: The Spark

In June 2017, a team of Google researchers published a paper with the understated title 'Attention Is All You Need.' The paper introduced a new type of neural network architecture called the transformer. Before the transformer, the dominant approach to language-related AI tasks used a type of network called a recurrent neural network (RNN), which processed text one word at a time in sequence, making it slow and prone to losing track of earlier context in long texts. The transformer replaced this sequential approach with a mechanism called 'attention,' which allows

the model to consider all parts of a text simultaneously and weigh which parts are most relevant to each other. The practical effect was significant: transformer-based models trained faster, handled longer texts more effectively, and scaled in a way that RNNs did not, meaning that making a transformer larger by adding more parameters consistently produced better performance. This 'scaling law' turned out to be crucial. It meant that the path to more capable AI was, in part, simply to build bigger models and train them on more data. Companies with the resources to do this (Google, OpenAI, Anthropic, Meta) had a clear roadmap for improvement.

The Data Advantage: The Internet as a Training Set

The second ingredient was data. Modern large language models are trained on staggering amounts of text. GPT-3, released in 2020, was trained on a dataset filtered down from roughly 45 terabytes of raw crawled text: a dataset that included large portions of the English-language internet, books, Wikipedia, and other sources. GPT-4 and subsequent models used training sets that were larger still. The key insight is that the internet, which had been growing for 25 years by the time these models were being trained, had essentially created the world's largest annotated corpus of human language and knowledge, available for free. This data advantage was not evenly distributed. Creating a capable large language model requires not just raw text but carefully curated training datasets, human feedback to align the model's behavior, and ongoing safety testing. These requirements mean that the best-performing models are concentrated at a handful of large technology companies with the resources to acquire and process data at scale. Smaller research teams and open-source projects can build capable models, but closing the gap with frontier models requires the kind of infrastructure investment only a few organizations can make. The data story also has a problematic side. Training data scraped from the internet reflects all of the biases, inaccuracies, and harmful content present on the internet. Managing these problems has become one of the central challenges of AI development; companies employ large teams of human reviewers to label data, identify harmful outputs, and provide feedback that steers model behavior away from problematic patterns.

The Hardware Revolution: GPUs Change Everything

The third ingredient was hardware. Training a large neural network requires an enormous number of mathematical operations performed simultaneously — exactly

the kind of parallel computation that graphics processing units (GPUs) were designed for, originally to render video games. Starting around 2012, AI researchers began training neural networks on GPUs instead of traditional central processing units (CPUs), which produced dramatic speedups. An operation that took weeks on CPUs could be completed in days on GPUs. NVIDIA became one of the most valuable companies in the world partly because of this shift. Its H100 GPU chips, released in 2022, became the critical hardware for AI training and inference at scale. The high demand for these chips created supply bottlenecks and intense geopolitical competition — the U.S. government implemented export controls on advanced AI chips to restrict their sale to China, reflecting how central this hardware had become to national AI capabilities. By 2024, AI chip production had become as strategically important as semiconductor manufacturing more broadly.

The Scaling Laws and What Came Next

One of the most surprising findings from the large-model era was how consistent the improvement in model performance was as a function of scale. Researchers at OpenAI published empirical findings — called 'scaling laws' — showing that model performance improved predictably as you increased model size, training data, and compute. This transformed AI development from an experimental search for the right architecture into something closer to an engineering problem: more resources reliably produced better results. The 2022-2026 period saw the consequences of this scaling play out in public. ChatGPT launched in November 2022. GPT-4 launched in March 2023. Within a year, every major technology company had released or announced a competitive frontier model. By 2024 and 2025, models were capable of multimodal inputs — processing images, audio, and video alongside text. By 2026, AI had been integrated into search engines, productivity software, healthcare tools, legal services, and educational platforms worldwide. The pace of change has been fast enough that even researchers who created these systems have expressed surprise at the capabilities that emerged. By 2026, the questions surrounding AI have shifted from whether these systems are capable to who controls them and what values they embed. The rapid concentration of AI capability among a small number of companies — OpenAI, Google, Anthropic, Meta, and Microsoft — reflects the enormous capital requirements for training frontier models. Researchers at universities and smaller organizations can build capable AI, but closing the gap with

the most powerful systems requires the kind of data and compute that only a handful of institutions can assemble. This concentration has prompted serious debates about competition, access, and the governance of technologies that increasingly affect everyone's daily life.

Key Information: The 2022-2026 AI Boom

Factor	Details
Transformer architecture	2017 Google paper; replaced sequential RNNs with parallel 'attention' mechanism
Scaling laws	Larger models + more data + more compute = reliably better performance
Training data source	Primarily internet text; GPT-3 was trained on a dataset filtered from ~45 terabytes of raw crawled text
Key hardware	GPUs (graphics cards) ideal for parallel AI math; NVIDIA H100 chips critical
ChatGPT launch	November 2022; triggered industry-wide response within months
GPT-4 launch	March 2023; major capability jump over prior models
Geopolitical impact	U.S. export controls on advanced AI chips to restrict China's access
What surprised researchers	Emergent capabilities — new abilities appeared that no one specifically programmed

Case Study: The Chip Factory Decision

In early 2023, a semiconductor manufacturing company called Allied Chip Technologies received two large orders simultaneously: one from a major American AI laboratory for 50,000 advanced GPU chips for AI training, and one from an international technology firm for a similar quantity of the same chips. The company's CEO, Sandra Okafor, learned that the U.S. Department of Commerce had added the international firm to an export restriction list, making the second sale potentially illegal. Her legal team confirmed that fulfilling the international order could result in significant federal penalties. 'Five years ago,' Sandra told her board, 'no one thought of us as a national security issue. We made chips for video games and scientific simulations. Now the president is talking about our product in national security briefings.' The board voted to decline the international order and prioritize domestic AI laboratory clients. The company's stock rose 18 percent the following week as investors recognized its strategic position in the AI supply chain.

1. Based on the reading, why did GPU chips go from being primarily a gaming hardware product to a component of national security policy in just a few years? What specific development made this shift happen?

2. The reading describes the convergence of three trends — architecture, data, and hardware — as necessary for the AI boom. How does Sandra's story illustrate the hardware component of this convergence and its real-world consequences?

Multiple Choice

1. According to the reading, what was the key insight behind the transformer architecture that made it better than previous approaches to AI language tasks?

a) It required far less training data than recurrent neural networks

b) It used attention to consider all parts of a text simultaneously instead of sequentially

c) It was specifically designed for use with hardware architecture developed five years later

d) It processed text one word at a time more carefully than older systems

2. What does the term 'scaling laws' refer to in the context of the AI boom?

a) The finding that AI performance improves predictably as model size, data, and compute increase

b) The legal regulations governing how large AI companies can grow

c) The rate at which AI knowledge cutoff dates must be updated to remain current

d) The physical limits on how many GPUs can be fitted inside one standard data-center server rack

3. What geopolitical consequence of the AI hardware boom does the reading describe?

a) The European Union banned the export of all AI software to the United States

b) The U.S. implemented export controls on advanced AI chips to restrict China's access

c) Russia had nationalized all of NVIDIA's manufacturing facilities inside its own borders

d) Japan and South Korea merged their semiconductor industries to compete with NVIDIA

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Transformer	A) Graphics processing unit; ideal for parallel AI calculations originally designed for gaming
2. ___ Recurrent Neural Network	B) 2017 neural network architecture using 'attention' to process all text simultaneously
3. ___ GPU	C) Older AI architecture that processed text sequentially, word by word
4. ___ Scaling Law	D) An unexpected new ability that appears in a large model without being specifically programmed
5. ___ Emergent Capability	E) Finding that more model size, data, and compute reliably produces better AI performance

Fill-in-the-Blank

1. The 2017 Google paper that introduced the transformer architecture was titled 'Attention Is _____ You Need.'

2. GPT-3 was trained on a dataset filtered from roughly 45 _____ of raw crawled text, including large portions of the English-language internet.

3. The company whose GPU hardware became most central to AI training at scale was _____, whose H100 chips became critical AI infrastructure.

Short Answer

Answer each question in 2-4 complete sentences.

1. The reading describes a 'phase transition' caused by the convergence of three trends. In your own words, explain what those three trends are and why all three were necessary — why would two of the three not have been enough?

2. The reading says that the internet essentially created 'the world's largest annotated corpus of human language and knowledge.' What does this mean, and what are both the advantages and the problems that come from using internet text as AI training data?

3. The reading notes that 'emergent capabilities' appeared in large AI models — new abilities that no one specifically programmed. Why should this development be both exciting and concerning to the researchers who build these systems?