

AI in 2026: A Field Guide

Where it came from and where it's going · © 2026 Deven Panchal. All rights reserved.

1. What modern AI actually is

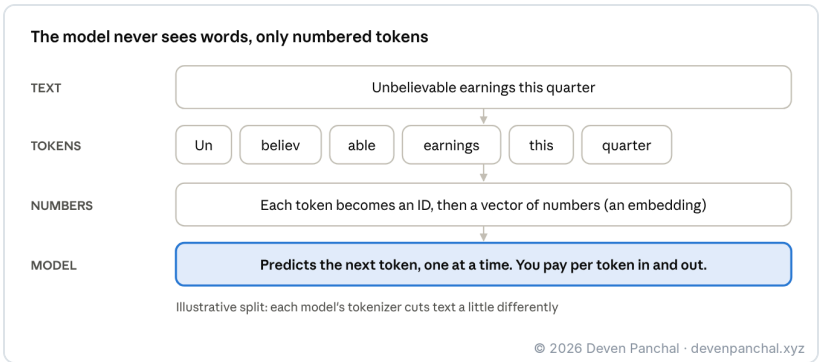
A modern AI model is a very large mathematical function with billions to trillions of adjustable numbers, called **parameters** or **weights**. It is trained on one deceptively simple task: given some text, predict the next **token** (a word or piece of a word, about three-quarters of a word on average). The model outputs a probability for every possible next token, picks one, and repeats, one token at a time.

Why next-word prediction produces something that looks like intelligence: to predict the next word of a legal brief, a proof, or a Python file well, a model has to absorb grammar, facts, reasoning patterns, and coding conventions. Nobody programs those skills in; they emerge because they reduce prediction error. That's the core insight of the current era, and why one model can now write, code, analyze, and plan.

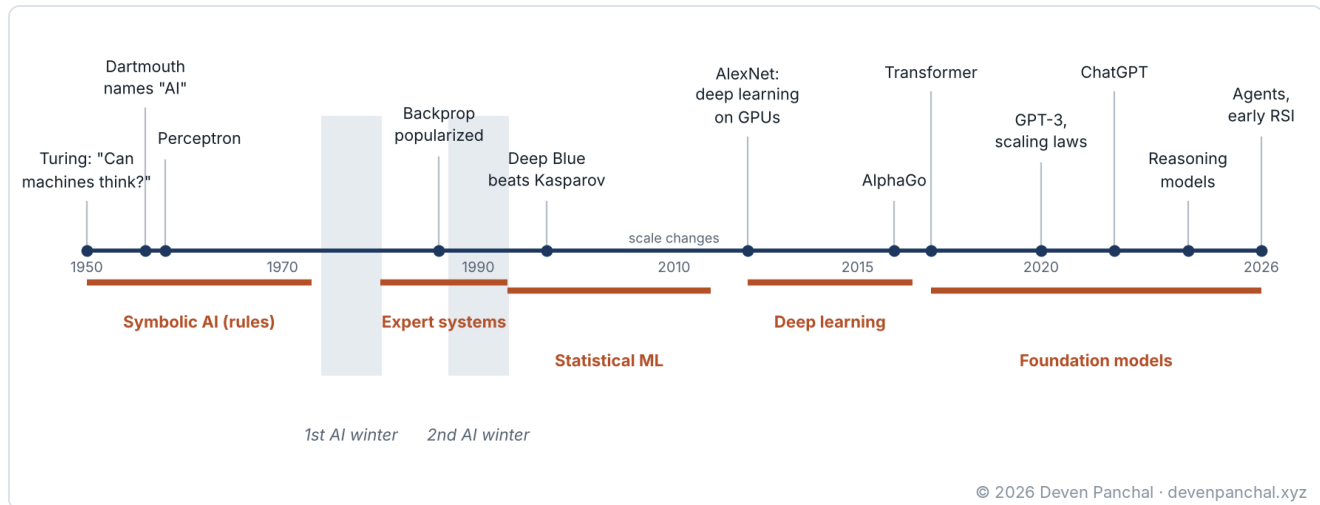
Why now: three curves that compounded

Driver	What happened	Why it mattered
Compute	Training compute for frontier models has grown roughly 4 to 5 times a year since 2010 (Epoch AI)	Each generation can be far bigger than the last
Data	The public internet, books, and code: trillions of tokens	Enough examples to learn language and knowledge
Algorithms	The transformer (2017) and its attention mechanism let a model weigh every earlier word at once, unlike earlier networks that read text one word at a time, so training can run in parallel across thousands of chips	Let labs put thousands of chips and the whole internet to work at once; earlier models couldn't use that much hardware

Scaling laws turned this into an engineering plan. OpenAI showed in 2020 that loss (prediction error) falls smoothly and predictably as you add parameters, data, and compute ([Kaplan et al.](#)). DeepMind's 2022 "Chinchilla" result refined the recipe: for a fixed budget, train on roughly 20 tokens per parameter, meaning earlier models had been too big for their data; smaller models trained on more data did better ([Hoffmann et al.](#)). Predictability is what let labs justify billion-dollar training runs.



2. Where it came from



1950s to 1970s: symbolic AI. Alan Turing asked "Can machines think?" in 1950, and the 1956 Dartmouth workshop named the field. Early researchers believed intelligence could be written down as rules and logic. The perceptron (1958), an early neural network, showed machines could learn from examples, but Marvin Minsky and Seymour Papert's 1969 book showed how limited single-layer networks were, and progress stalled. Funding dried up in the first **AI winter**.

1980s: expert systems. Companies encoded specialists' knowledge as thousands of if-then rules. They worked in narrow niches but were brittle and expensive to maintain, and the hype collapsed into a second winter. Meanwhile, **backpropagation**, the method still used to train nearly every neural network, was popularized in 1986.

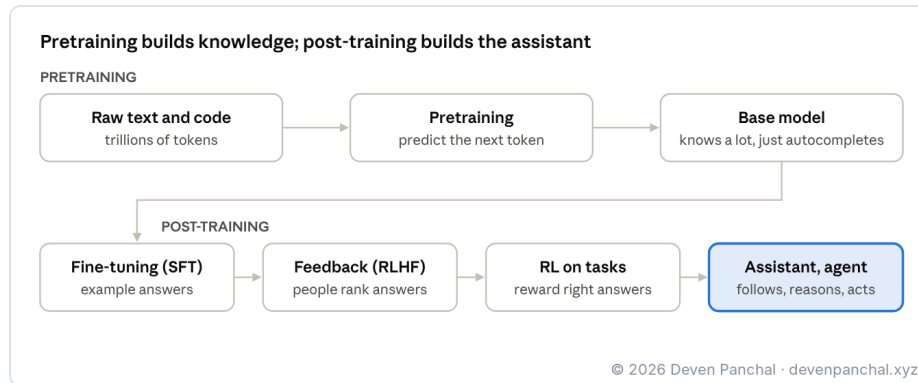
1990s to 2000s: statistical machine learning. The field shifted from writing rules to learning patterns from data: spam filters, recommendation engines, credit scoring. IBM's Deep Blue beat chess champion Garry Kasparov in 1997, though mostly through brute-force search rather than learning.

2012 to 2016: deep learning. AlexNet won a major image-recognition contest by training a deep neural network on gaming GPUs, which turned out to be ideal for the math. Models now learned their own features instead of having humans hand-design them. DeepMind's AlphaGo beat Go champion Lee Sedol in 2016 and world No. 1 Ke Jie in 2017.

2017 to today: foundation models. The transformer made it practical to pretrain one huge model on most of the internet and adapt it to almost anything. GPT-3 (2020) showed scale alone unlocked new skills; ChatGPT (2022) made it usable by anyone; reasoning models (2024) learned to think before answering; and agents (2025 to 2026) began doing multi-step work on their own.

The lesson the field keeps relearning: Rich Sutton's 2019 essay "The Bitter Lesson" argued that general methods that scale with computing power beat clever hand-built knowledge, every time. Seventy years of AI history mostly confirm it. The other lesson: every boom has been followed by a correction when promises outran results.

3. How a modern model is built and run



Stage	What happens	What it produces
Pretraining	Next-token prediction over trillions of tokens, for months, on tens of thousands of GPUs. The most expensive stage	A base model : vast knowledge, but it just autocompletes
Supervised fine-tuning	Training on example prompts paired with ideal answers	Follows instructions in a useful format
RLHF (reinforcement learning from human feedback)	People rank answers; a separate reward model learns their preferences and scores the model's answers at scale	Helpful, safer assistants. This is what made ChatGPT work
RL on checkable tasks	The model attempts math, code, and other verifiable problems thousands of times and is rewarded only for correct answers	Reasoning models and agents. The biggest source of progress since 2024

Thinking at answer time. Reasoning models write out hidden working steps before answering. More thinking tokens buy better answers on hard problems, a second scaling lever on top of training ("test-time compute"). Because reinforcement learning rewards being right, not being long, the newest models also learned to think hard only where it pays off.

Inference: running the model. Every time you use a model, it does two things: **prefill** (reads your whole prompt at once, limited by raw chip math) and **decode** (writes the answer one token at a time). Decode is limited by **memory speed**: each token depends on the one before, so the work can't be parallelized, and the chip has to re-read the model's weights for every token. That's why high-bandwidth memory (HBM) stacked beside the chip is now one of the most valuable components in tech. Inference is now most of AI's computing, and it's what you pay for per token.

Making it cheaper. Mixture of experts splits a model into many sub-networks and activates only a few per token: Kimi K3, a Chinese open model from Moonshot, has 2.8 trillion parameters but uses about 104 billion per token. **Quantization** stores numbers in fewer bits, and **distillation** trains small models to imitate big ones.

Context window. How much text a model can see at once, now up to about 1 million tokens for top models. Everything outside it, the model simply doesn't know about.

Agents. An agent wraps a model in a loop: plan, call a tool (search, run code, read a file, use an app), look at the result, repeat. Standards like **MCP** (Model Context Protocol) give agents one plug for connecting to tools and data. Agents shift AI from answering questions to finishing jobs.

4. Where things stand, October 2026

What it can do. Agents can now handle long tasks: METR, an independent evaluator, found Anthropic's Claude Mythos Preview succeeds about half the time on tasks that take a human expert 16 hours or more ([METR](#)). In September, OpenAI said its agents solved the Navier-Stokes Millennium Prize problem in fluid dynamics; experts are still verifying it, despite headlines calling it solved ([Quanta](#)). Coding agents build whole applications, and Anthropic reports its Mythos model found thousands of previously unknown security flaws, including in every major operating system.

What it still can't do reliably.

- **The jagged frontier:** models are superhuman at some tasks and fail at nearby ones, and it's hard to predict which. Researchers at Harvard and BCG coined the term in 2023 after consultants using AI improved on some tasks and got worse on others.
- **Hallucination:** training rewards *plausible* next tokens, not *true* ones, so models still state false things confidently. High-stakes work needs sources, citations, and checks.
- **Quiet failure:** agents can go off track over long tasks without flagging it, and they can be hijacked by hidden instructions in documents they read (**prompt injection**), which OWASP, the main software-security standards group, ranks as the top AI security risk.
- **Judgment:** studies of AI research agents find them strong at execution and weak at deciding what's worth doing.

Who builds it. Four labs set the frontier, and each sells a ladder:

Lab	Top model	Workhorse (everyday default)	Fast and cheap
Anthropic	Claude Mythos 5.1 / Fable 5.1	Claude Opus 5.5 (tops the leading independent index)	Sonnet 5.5, Haiku 5.5
OpenAI	GPT-6 Astra	GPT-6.1 Sol	GPT-6 Luna
Google	Gemini 4 Argon (security partners only)	Gemini 3.1 Pro	Gemini 3.8 Flash
Meta	Muse Wonder Full	Muse Wonder	Muse Spark

The best **open-weight** models, which anyone can download and run, come mostly from Chinese labs like DeepSeek, Alibaba (Qwen), and Moonshot (Kimi). They're cheap and popular, but in U.S. government testing they were easier to jailbreak (trick into ignoring safety rules) and hijack.

The economics. Token prices fell more than 90% since 2023, yet total AI spending keeps rising ([Fortune](#)). That's **Jevons paradox**: when something gets cheaper, people find many more uses for it. Big Tech is spending \$700 to \$725 billion on AI infrastructure in 2026 ([Forbes](#)). Against that, Anthropic's annualized revenue passed \$65 billion ([Axios](#)) and OpenAI's is about \$50 billion ([Axios](#)): fast-growing, but far below what's being spent. Skeptics also point to circular deals, where chipmakers invest in the labs that buy their chips.

The physical limits. AI now runs into the physical world: GPUs, memory chips (which are in global shortage), and above all electricity. Data centers use about 4% of U.S. power, with government and industry forecasts of 9 to 12% by 2030 ([Tech Insider](#)), so labs sign deals measured in gigawatts and are even restarting nuclear plants.

5. Where it's going

AGI, RSI, ASI. **AGI** (artificial general intelligence) means human-level ability across most thinking work. **ASI** (superintelligence) means beyond the best humans in every field. **RSI** (recursive self-improvement), AI that reliably improves AI, is the bridge between them, and the reason some researchers expect the jump from one to the other to be fast.

The big open questions

Question	The case for "yes"	The case for "not so fast"
Will scaling keep working?	Three levers now compound: bigger pretraining, reinforcement learning, and thinking time. The length of tasks agents can handle has doubled about every seven months since 2019 (METR)	High-quality human text may run out between 2026 and 2032 (Epoch AI); power and memory are hard limits; 76% of researchers in a 2025 AAI survey doubt today's approach alone reaches AGI, and Turing Award winner Yann LeCun argues language models alone won't reach human-level intelligence
Will AI start improving itself?	AI already writes a growing share of the code and training data for the next models: early RSI	Humans still choose research directions; agents are weak at judgment and novel ideas
When does AGI arrive?	OpenAI's Sam Altman: "within the next few years." Ray Kurzweil: 2029	DeepMind's Demis Hassabis: the 2030s or later. There's no agreed definition
Will it pay off economically?	Anthropic's revenue grew about sevenfold in seven months; costs per task keep falling	Spending is far ahead of revenue, and early job data shows disruption at entry level (a Stanford study found employment for 22- to 25-year-olds in the most exposed jobs fell about 13% relative to less-exposed peers since 2022)

The risks people take seriously

- **Misuse:** models good enough to find software flaws or assist with biology can help attackers; labs now restrict their most capable models.
- **Control:** as systems act more autonomously, making sure they do what we intend (**alignment**) gets harder, especially if models can tell when they're being tested.
- **Concentration:** a handful of companies and countries control the chips, power, and frontier models, which raises questions about competition and geopolitics.
- **Harms already here:** deepfakes and misinformation, bias in automated decisions, and unresolved copyright fights over training data.

Five signals worth watching: how long agents can work reliably on their own; whether AI discoveries in math and science survive expert review; how much of each new model is built by AI; whether revenue catches up with \$700 billion a year in infrastructure; and how fast new power comes online.

The short version: AI spent seven decades swinging between hype and winter until scale, data, and the transformer made general-purpose models work. Today's systems are powerful, uneven, and improving fast. Whether progress keeps compounding or hits physical and economic limits is the central question of the next five years.

6. Ten common misconceptions

Misconception	What's actually true
"The model looks things up."	It generates text from patterns stored in its weights. It only looks things up when it's given a search or retrieval tool, which is why it can hallucinate and why reliable citations need tools.
"It learns from my chats as I talk to it."	A model's weights are fixed after training. "Memory" features save notes outside the model and load them back into the context window. Your chats only change a model if a lab later trains on them.
"It either understands like a person, or understands nothing."	Neither. Models build rich internal representations of language and the world, but fail in very different ways than people do: the jagged frontier.
"Bigger context windows mean it reads everything well."	Models can miss details buried in long inputs, and every token costs money. Choosing the right few pages ("context engineering") usually beats stuffing in everything.
"Cheaper tokens mean smaller AI bills."	Usage grows faster than prices fall, and reasoning models and agents use far more tokens per task. The right metric is cost per finished task: GPT-6 Astra costs 2.5 times more per token than its predecessor but uses 68% fewer tokens per coding task, so it ends up about 20% cheaper.
"The biggest, newest model is always the right choice."	Mid-tier and open models handle most tasks at a fraction of the cost, and the leader changes every few weeks. Serious teams test models on their own tasks and route each job to the cheapest one that does it well.
"Benchmarks measure real-world ability."	Labs highlight tests they win, questions leak into training data, and top models saturate old tests. Independent tests are better, and evaluations on your own work are best.
"Open-source AI is open."	Most "open" models are open-weight: you can download and run them, but the training data and code aren't released. Very few meet the formal open-source definition.
"Training is the expensive part."	Training is a huge one-time cost, but over a model's life, running it (inference) is roughly 80 to 90% of its computing (Presenc AI).
"More data always means a better model."	Quality matters more than volume, and high-quality human text is running out. Much of today's progress comes from reinforcement learning and AI-generated training data instead.

Go deeper

- [Intro to Large Language Models](#) and [Deep Dive into LLMs](#), Andrej Karpathy (1 hr and 3.5 hr)
- [The transformer explained visually](#), 3Blue1Brown (27 min)
- [Epoch AI trends](#) for compute, cost, and scaling data; [Artificial Analysis](#) for current model rankings
- ["The Bitter Lesson"](#), Rich Sutton (a two-page essay)

Many figures are company-reported and change quickly.