AI's Great Flattening: What Happens when Everyone Is State-of-the-Art?

Published April 28, 2025

Christopher Gannatti, CFA
Global Head of Research

Key Takeaways

- In 2025, frontier AI model performance converged dramatically, with performance benchmarks like LMSYS's Chatbot Arena revealing near-parity among top systems—a seismic shift suggesting excellence across different large language models (LLMs) is now ubiquitous, not rare.
- Breakthroughs in reasoning, shown in such benchmarks as ARC-AGI-1 and GPQA, signal a paradigm shift from memorization to genuine cognitive capability, with models beginning to challenge and even outshine human experts on complex, abstract problem-solving in certain cases.
- As technical performance across more LLMs equalizes, the competitive advantage pivots to cost efficiency, contextual integration and user experience—marking a new era where "state-of-the-art" is no longer a differentiator, but a baseline.

In the early years of the artificial intelligence (AI) race, performance benchmarks told a clear story: a handful of frontier models, developed by a few dominant labs, consistently outperformed the rest. In 2024, that changed. The 2025 Stanford AI Index Report reveals a stunning convergence in technical performance at the very top of the leaderboard. In just 12 months, the once-substantial gulf between the most powerful models and their challengers narrowed into near-parity. We are entering an era of *ubiquitous excellence* in the sphere of large language models (LLMs).

Each year, the Stanford AI Index report sets a new standard in giving us a better sense as to the progress the world has seen in AI. Over the coming weeks, you will see five different articles where we will highlight our most impactful takeaways from the 2025 edition. We will also provide one further piece contextualizing the different investment strategies available for those looking at AI as an investment.

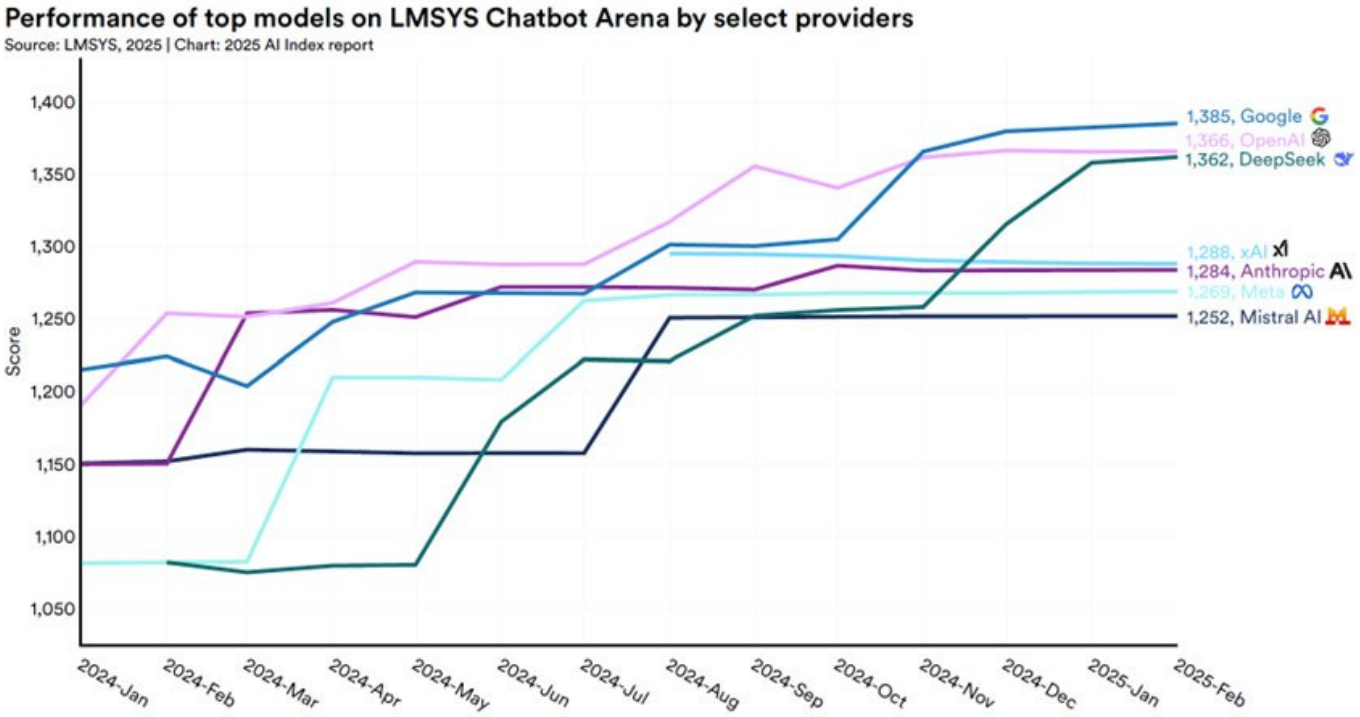
Enter the AI Model Thunderdome

If you wanted to understand which AI model is "best," you might assume the answer lies in obscure academic metrics or labs with billion-dollar compute clusters. But in reality, the most useful models—the ones that genuinely help people—are best judged in the wild. That's the premise of Large Model Systems (LMSYS's) Chatbot Arena, a kind of public Thunderdome for AI where models go head-to-head in blind matchups, judged by everyday users. No brand names, no reputations—just raw performance. One model's

answer goes on the left, another on the right, and the users pick their favorites. The result is something rare in the world of AI: a benchmark that is crowdsourced, dynamic and deeply human.

What makes the leaderboard interesting isn't just the scores—it's the storylines. Google recently pulled ahead with a model that feels like it skipped a generation. DeepSeek, a Chinese upstart, has surged toward the top of the heap after garnering little if any attention prior to 2025. And while OpenAI and Anthropic still post elite results, challengers like xAI and Mistral show how much competition remains. In a world where every company is promising AI transformation, LMSYS offers something simple but powerful: an honest signal of what actually works when real people put models to the test.

Figure 1: Large Model Systems (LMSYS) Chatbot Arena



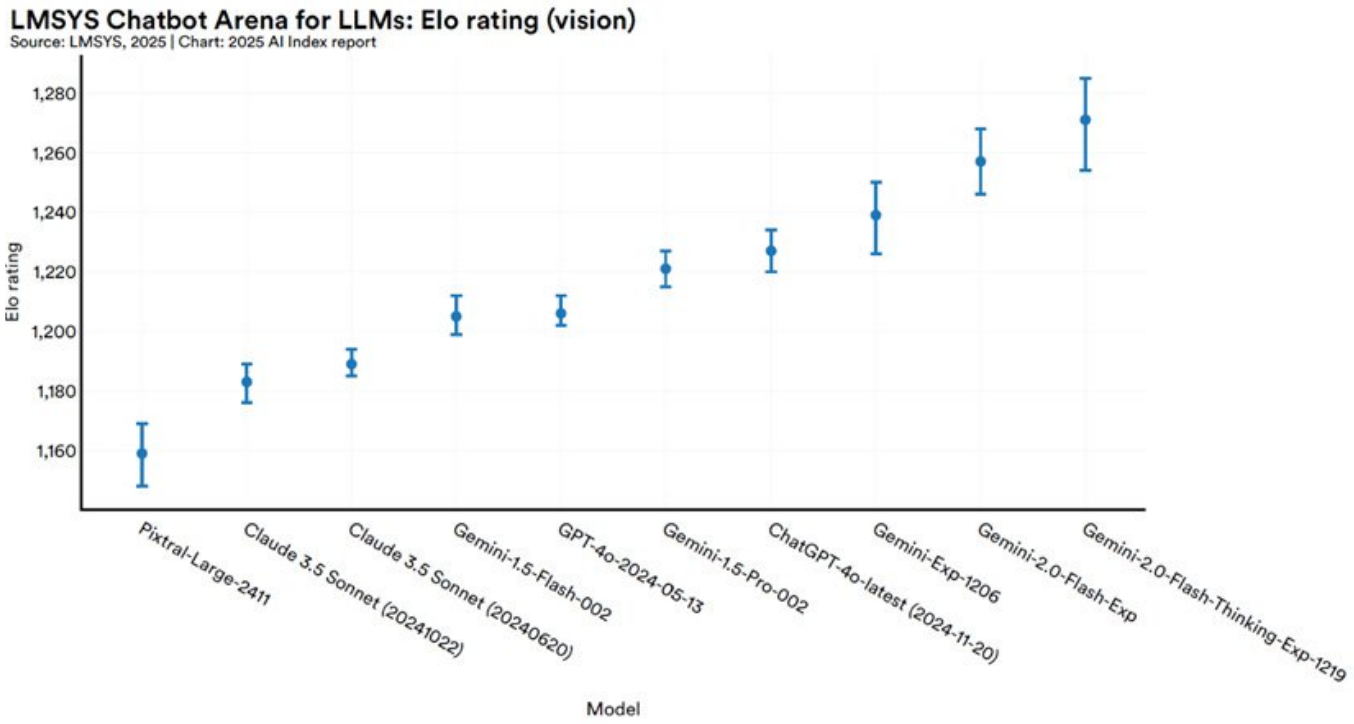
A Comparison of Multimodal Systems—The Case of Vision

In the rapidly evolving race of multimodal AI—where models are judged not just by how they write, but how they see—Google's Gemini 2.0 models have seized the high ground. The latest leaderboard from LMSYS's Chatbot Arena, which uses an Elo rating system derived from human-voted head-to-head matchups, shows Gemini dominating vision tasks by a meaningful margin. Its top models don't just edge out rivals like GPT-4o—they establish clear distance. That matters. Because when users are feeding these systems images, graphs or charts, the gap between "gets the gist" and "understands the nuance" is the difference between a mediocre assistant and an indispensable co-pilot.

What makes this moment interesting isn't just the leaderboard—it's what it signals about how fast the frontier is moving. Six months ago, most conversations centered on OpenAI's GPT-4 as the gold standard. Today, it's still a powerhouse, but the spotlight is shifting. Claude sits solidly in the middle tier, while

lesser-known entrants like Pixtral are struggling to keep up. Now, here's the nuance that can be easy to miss: even a 30- to 50-point Elo gap on this chart is meaningful. These are not cosmetic differences—they reflect repeated, statistically significant wins in blind head-to-head matchups. In a system modeled after chess rankings, a 50-point edge translates to winning about 57% of the time. That's not trivial. In a world where tasks are increasingly delegated to AI—interpreting a chart, summarizing an earnings call, responding to a visual prompt—small percentage wins compound into huge differences in productivity. So while the models may all look clustered together, the best ones are quietly stacking real-world advantages.

Figure 2: Large Model Systems (LMSYS) Chatbot Arena with Vision Tasks



Each **dot** on the chart represents the average *Elo rating* of a specific large language model (LLM) in the LMSYS Chatbot Arena, focused on vision-based tasks. The Elo rating system is originally used in competitive games like chess, and it helps rank models by performance based on pairwise comparisons—higher Elo means stronger overall performance in head-to-head matchups.

The **vertical lines (error bars)** around each dot show the **uncertainty or margin of error** in that model's Elo rating. This reflects variability in the model's performance across different test prompts and evaluators. A **shorter line** means the model's rating is more stable and consistent. A **longer line** means there's more variability, suggesting the performance could be slightly better or worse than the average depending on the test conditions.

Together, the dots and lines give a **visual summary of both performance and confidence** in that performance across the different models.

The Models Are Evolving—and So Are the Tests!

In the world of AI evaluation, most benchmarks eventually become obsolete—not because they're poorly designed, but because the models evolve faster than the tests themselves. ARC-AGI-1 is different. The Alignment Research Center Artificial General Intelligence (ARC-AGI) evaluation is a **benchmark developed to assess whether advanced AI systems show early signs of general intelligence**—that is, the ability to solve a wide range of problems, including unfamiliar ones, without being specifically trained on them.

It was developed by the **Alignment Research Center**, an independent nonprofit focused on ensuring that future AI systems are safe and aligned with human intentions. ARC created this test to evaluate how well AI models can **reason, plan, and solve tasks that require multiple steps or abstract thinking**—in other words, skills beyond memorizing answers or pattern-matching.

Why was it developed?

Most AI benchmarks measure how well a system performs on a specific skill (e.g., summarizing text or answering trivia). But the ARC-AGI test aims to answer a deeper question:

"Does this AI system exhibit the kind of general problem-solving ability that would make it capable of behaving like an early form of AGI?"

ARC-AGI is designed to:

- Detect *early warning signs* of general intelligence
- Provide a **more rigorous test of reasoning and adaptability**
- Ensure safety evaluations keep pace with the rapid progress of model capabilities

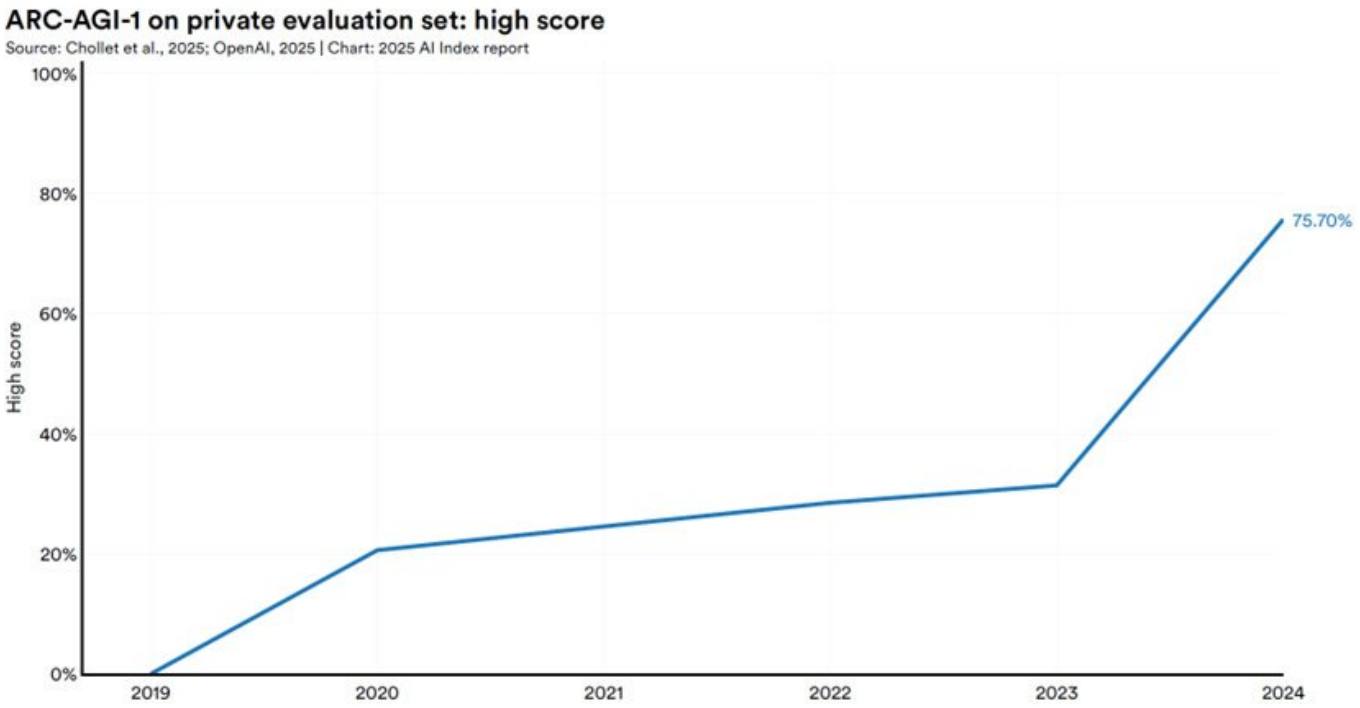
It includes tasks that are **intentionally challenging** for today's models, often requiring multi-step thinking, code generation, logical reasoning, or unfamiliar problem-solving.

The most striking thing isn't just the test—it's the leap seen in the results. From 2019 through 2023, models made slow, incremental progress, struggling to crack 30% on the ARC-AGI-1 private evaluation set. But in 2024, something changed. The high score jumped to 75.7%, a near threefold increase. That kind of curve doesn't suggest a slightly better model—it suggests a fundamentally different one. You don't see this kind of progress unless a system gains some new degree of agency, planning or inner scaffolding for how it reasons through problems. It's as if the lights suddenly came on in a room that we thought was already fully lit.

For investors and technologists, this shift should be treated less like an isolated metric and more like a signal flare. The progress on ARC-AGI-1 is a proxy for something deeper: the emergence of models that don't just *answer* but *reason*. This raises the ceiling on what AI can do in research labs, enterprise workflows and even autonomous systems. But more importantly, it compresses timelines. Every few years,

we recalibrate our expectations for when AGI might arrive. This year's leap suggests we may be ahead of schedule.

Figure 3: Marching toward a More Generalized Form of Intelligence



The **blue line** represents the **highest score achieved by any AI model** on the ARC-AGI-1 test's **private evaluation set** for each year shown, from 2019 to 2024.

- The **Y-axis** shows the percentage of questions answered correctly—this is the model's high score.
- The **X-axis** shows the **calendar year** in which that top score was achieved.
- The **2024 data point**, marked at **75.70%**, is the most recent high score reported.

This line provides a **time-series view of AI progress** on a benchmark designed to measure general problem-solving ability. It illustrates how the best-performing models each year have steadily improved on this difficult reasoning task set, with especially rapid gains seen between 2023 and 2024.

Getting Beyond Searching and Recalling to Measuring Something More

Imagine you're preparing for an exam in a subject you know deeply—say, organic chemistry at the graduate level. You're not memorizing facts; you're reasoning through why a signal in an NMR spectrum moves downfield, what that says about electron environments and what kind of catalysts might explain the transformation. **NMR spectrum** refers to the graphical output produced by **Nuclear Magnetic Resonance (NMR) spectroscopy**, a powerful analytical technique used to determine the **structure, dynamics, and environment of molecules**, particularly organic compounds.

That's the level of thinking required by the GPQA benchmark, which isn't testing trivia—it's testing whether a model can walk the same mental pathways as someone who has *truly studied* the subject. **GPQA** stands for **Graduate-Level Professional and Quantitative Aptitude**. It is a **benchmark dataset and evaluation** designed to test whether advanced AI systems can answer **difficult, expert-level questions** across a wide range of scientific and technical domains. The sample chemistry question feels pulled from a real-world industrial process, asking not what's happening, but why—and what kind of elements, from where in the periodic table, would help make it happen.

Now here's the part that should stop you in your tracks: in 2023, the best AI models answered these questions right less than half the time. In 2024, they beat expert human validators. The accuracy jumped from 40% to nearly 88%, clearing the 81% benchmark set by humans trained to validate these answers. That's not a minor improvement—it's a paradigm shift. Because these aren't questions you Google. They're problems you solve, requiring contextual knowledge, logical sequencing and domain abstraction. The fact that a model can now perform this way implies something profound: we're building tools that reason, not just recall.

The implications are vast. In the same way calculators changed math class or spreadsheets transformed finance, these systems could begin changing what it means to "know" something professionally. If an AI can reason like a specialist across chemistry, law and economics, then the constraint becomes less about what one person can hold in their head, and more about how fast and reliably we can ask the right questions. GPQA isn't just a test—it's a window into a world where intelligence scales. And when intelligence scales, everything else does too.

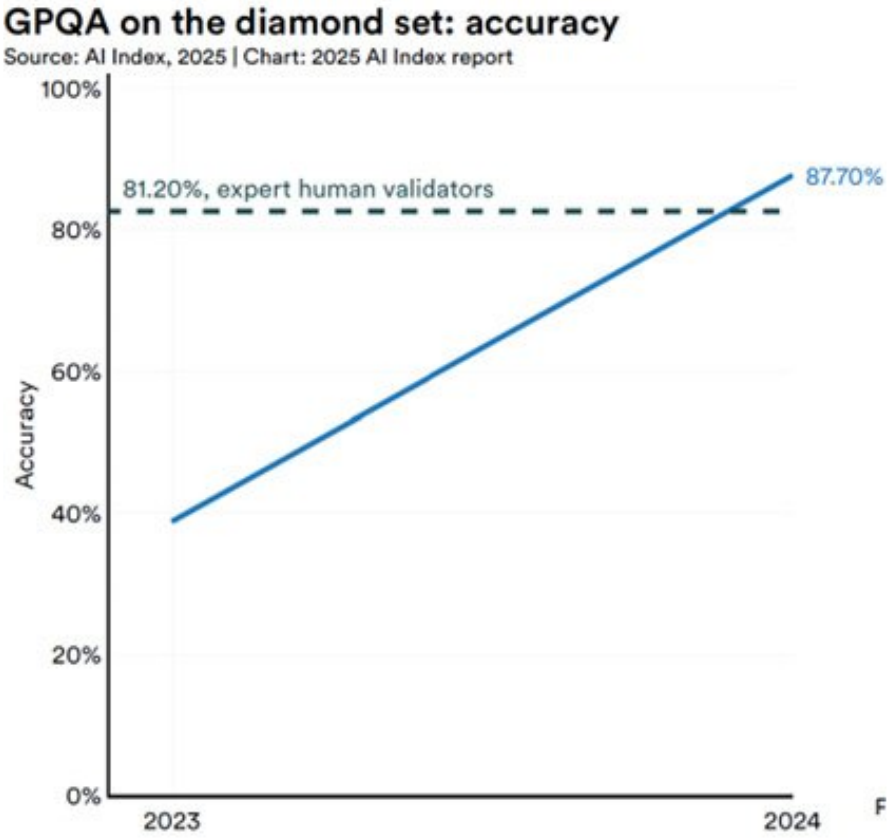
Figure 4a: Example of a Question where a Simple Search May Not Work

Sample chemistry question from GPQA
Source: Rein et al., 2023

Chemistry (general)

A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?
A) A metal compound from the fifth period.
B) A metal compound from the fifth period and a non-metal compound from the third period.
C) A metal compound from the fourth period.
D) A metal compound from the fourth period and a non-metal compound from the second period.

Figure 4b: Quantifying the Dramatic Increase in Capability over Just One Year



The **blue line** in the chart shows the **year-over-year accuracy of the top-performing AI model** on the **GPQA Diamond Set**, from 2023 to 2024.

- The **Y-axis** shows accuracy, measured as the **percentage of questions answered correctly**.
- The **X-axis** spans two years: **2023** and **2024**.
- The **data point in 2024** reaches **87.70%**, which is the **best accuracy achieved by an AI model** on this test to date.

The **dashed horizontal line at 81.20%** represents the **average accuracy of expert human validators** on the same set of questions—essentially a benchmark for human-level performance. The chart shows that by 2024, the top model **outperformed expert humans** on this challenging, graduate-level technical benchmark.

Technical Excellence Is Now Table Stakes

These are just some examples—there are many—showcasing the progress we are seeing in different AI models and what they can do. In years past, releasing a model that could achieve top-1 scores on certain evaluations and tests was a headline event. In 2025, it's the new minimum.

Consider that Claude 3.5 Sonnet, OpenAI's o1 and Gemini 2.0 all exceed 90% on GPQA Diamond—a benchmark for graduate-level science reasoning (figure 4b cited this test but didn't specify individual

model results). That means even subtle differences in architecture or training regimes no longer yield clear performance edges.

If Everyone Is Excellent, What Matters?

With performance parity at the frontier, the manner in which the different large language models seek to differentiate themselves from one another shifts.

- **Cost and Efficiency:** Training cost for Llama 3.1-405B reached $170 million. Inference cost for GPT-3.5-level outputs dropped 280-fold in 18 months. Models like o1 are six times more expensive to run and 30 times slower than GPT-4o. Efficiency is no longer a technical afterthought—it's central to platform viability. For clarity, AI models are developed, trained and then allowed to run and serve users. This ‘running and serving users' is often termed ‘inference.' When inference costs are discussed, it tends to mean what it costs to allow possibly hundreds of millions or even billions of users to use the model for their needs.
- **Agency and Reasoning:** Models like o1 and o3 from OpenAI show advanced planning, chaining and task execution capabilities. This agentic layer, not just raw token prediction, may become the next arena for differentiation.
- **Contextual Integration:** Gemini's 2M-token context window and the rise of retrieval-augmented generation (RAG) pipelines suggest a shift from standalone model intelligence to system-level orchestration.
- **Ecosystem and Interface:** With the landscape getting closer and closer to parity in core intelligence, user interface (UI)/user experience (UX), data integrations and developer ecosystems may become the real moats.

The Commoditization of Brilliance

The AI Index 2025 calls it clearly: the Turing test is no longer a meaningful boundary. On benchmark after benchmark, LLMs now equal or exceed human baselines. But the democratization of state-of-the-art performance may be more disruptive than any single breakthrough.

We are entering a world where "the best model" is not a unique crown but a crowded field. What follows is not about building better answers, but about building better systems, better interfaces and better outcomes.

Welcome to the age of commoditized brilliance.

Important Information Related to this Article

Source for all data in this blog post, including figures: Maslej et al., “The AI Index 2025 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025.