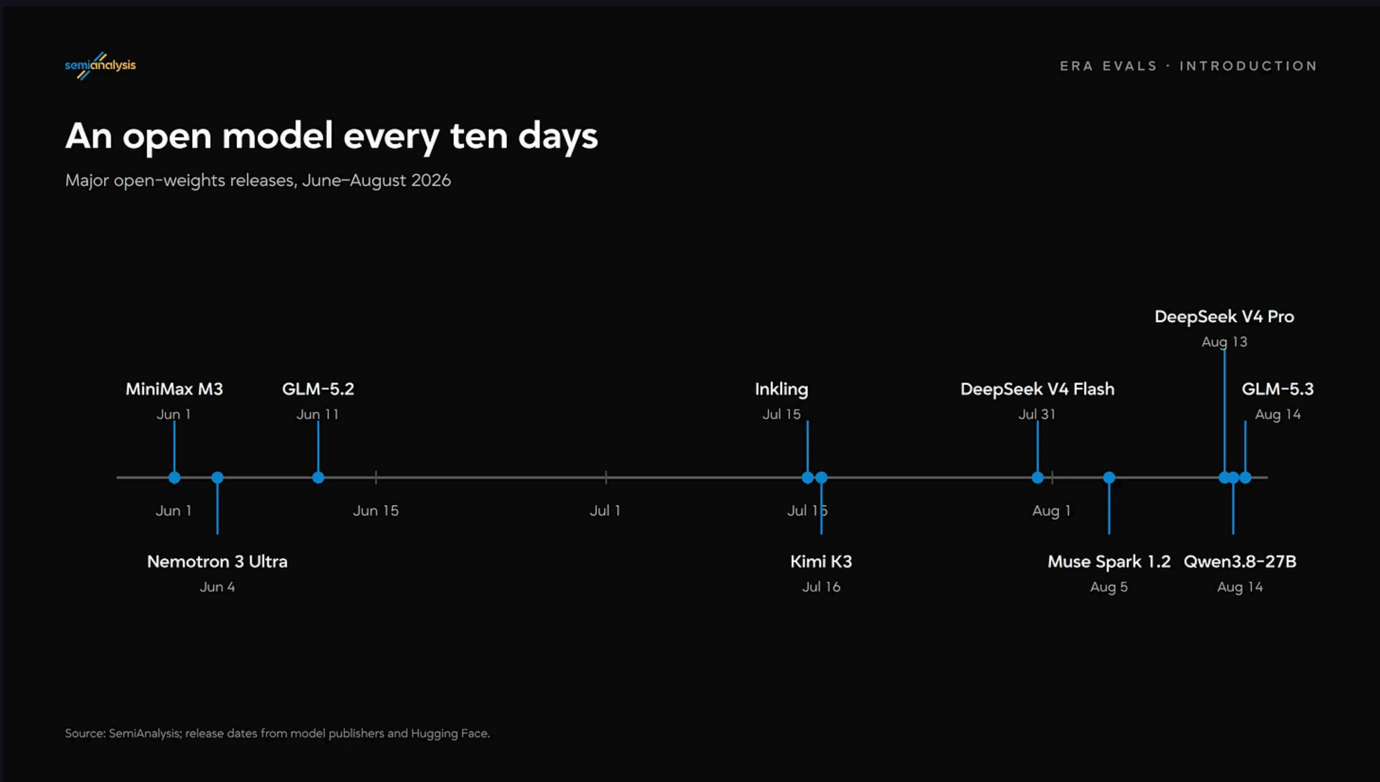
Are Open Models Catching Up?

Comparing open vs. closed models across the eras of frontier models, Is the gap narrowing?

EVAN CLOUTIER, MAX KAN, JORDAN NANOS, AND DYLAN PATEL

AUG 22, 2026 · PAID

The past two months have been a breakout period for open source AI. Yes, there was the "DeepSeek moment" back in January 2025, but no one actually used R1 to do any economically valuable work. In contrast, models like **GLM 5.3 and Kimi K3 are genuinely capable of many of the same coding and agentic tasks that rocketed Anthropic to $65B+ ARR**. Unlike others who inflated ARR, our figures were much closer to reality.



Source: SemiAnalysis

It is an exciting time to be a token consumer. Competition is heating up, usage resets are being doled out, and the battle for your tokens now extends beyond the OpenAI-Anthropic duopoly. Fireworks alone is processing over 40T tokens per day—2x the OpenAI API's volume at the end of March.

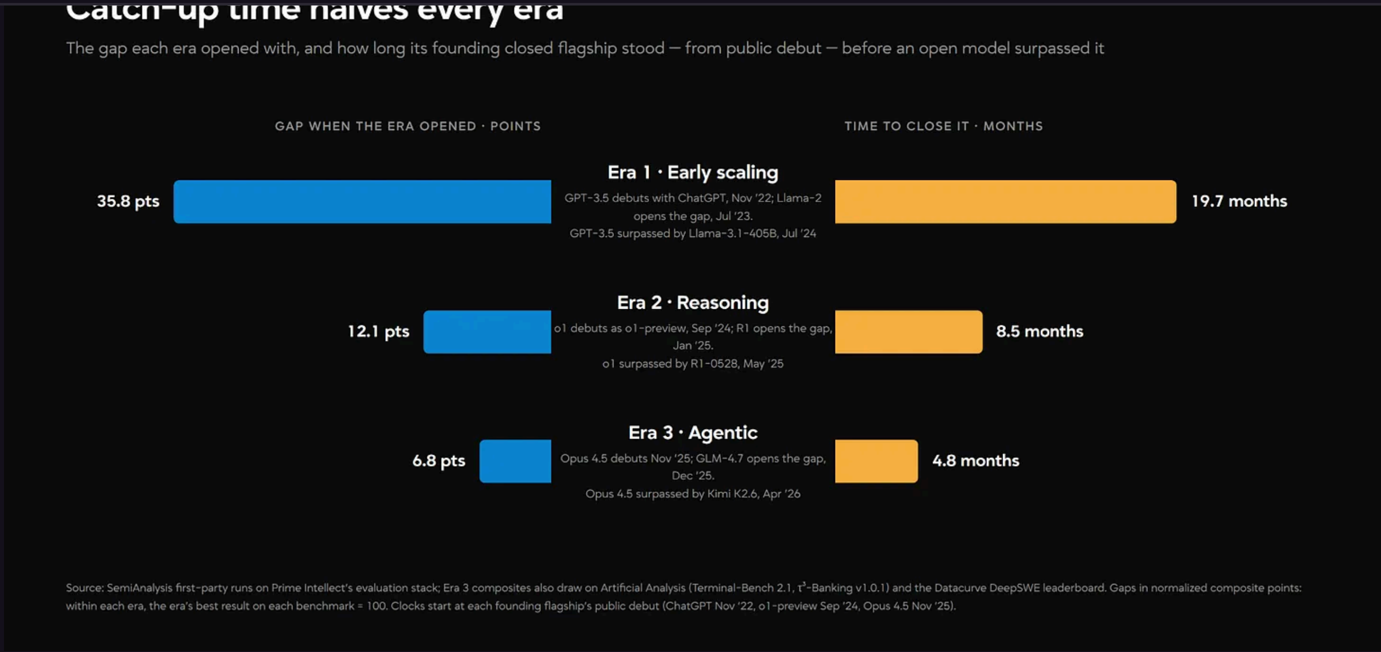
layer become commoditized? This outcome would obviously be disastrous for frontier lab margins. For full details on Anthropic and OpenAI's financials, see our Tokenomics Model.

To project how the open vs closed capability gap will progress in the future, we first need to measure the past. Naively, you might pick a single set of benchmarks to measure all historical models, but this is a mistake. **Every benchmark is a product of a particular era.** When someone creates a new benchmark, their goal is to discern differences in model capabilities at the time. If they're successful, the model makers will climb said benchmark until it becomes saturated. Once that happens, everyone stops caring about the benchmark, and the cycle repeats.

There have been three eras thus far in the history of LLMs: early scaling, reasoning, and agentic. Each era represented a step-function increase in model utility, and rather than trying to plot a single continuous trend, we believe it's better to evaluate the models and benchmarks from each era individually.

When viewed this way, it becomes clear that **the open vs. closed gap moves in cycles**. At the start of each era, a frontier lab completes some promising research, trains an impressive model, deploys it at scale to their users, and jumps ahead. Then, other labs identify the key advances, reverse-engineer what the frontier lab is doing, replicate them in their own models, and close the gap. Nothing stays secret forever—especially when you factor in distillation. It's just a question of how long it takes.

To answer this question, we took all the relevant models from each era and ran a curated set of benchmarks to get a composite capability score. **The result is a clear trend: with each generation, open-source models take half as long to catch up to the first closed-source model of the era.**

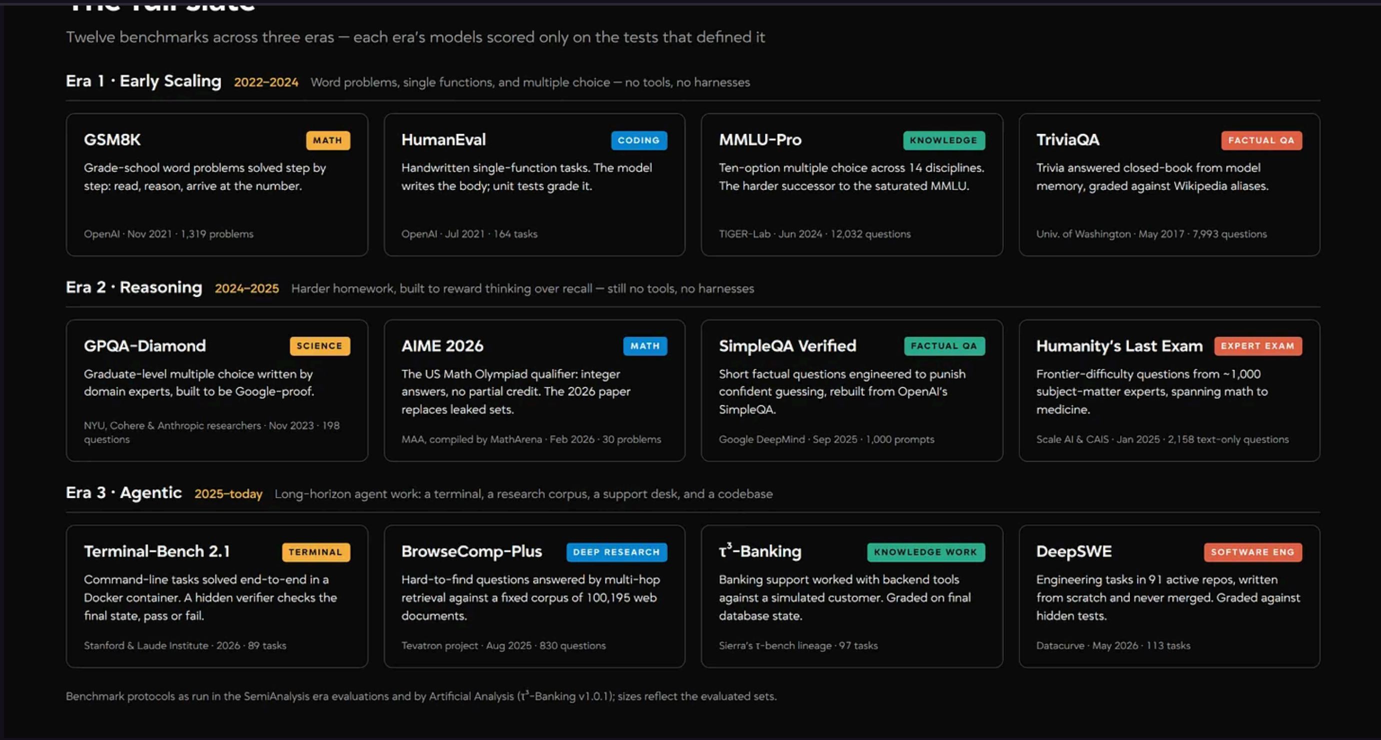


Source: SemiAnalysis

Of course, benchmarks don't tell the full story, and we'll highlight all the relevant caveats below. Finally, we'll extend this analysis into the future, and explain why it's less bearish frontier models than you might initially think.

How we measured

Here's an overview of the models and benchmarks we selected for each era:



Source: SemiAnalysis

Picking a single SOTA closed and open model at a particular time is subjective, but our selections reflect the general consensus among AI experts. In cases where there's debate—e.g. Fable 5 vs GPT 5.6 today—we were conservative and tested both.

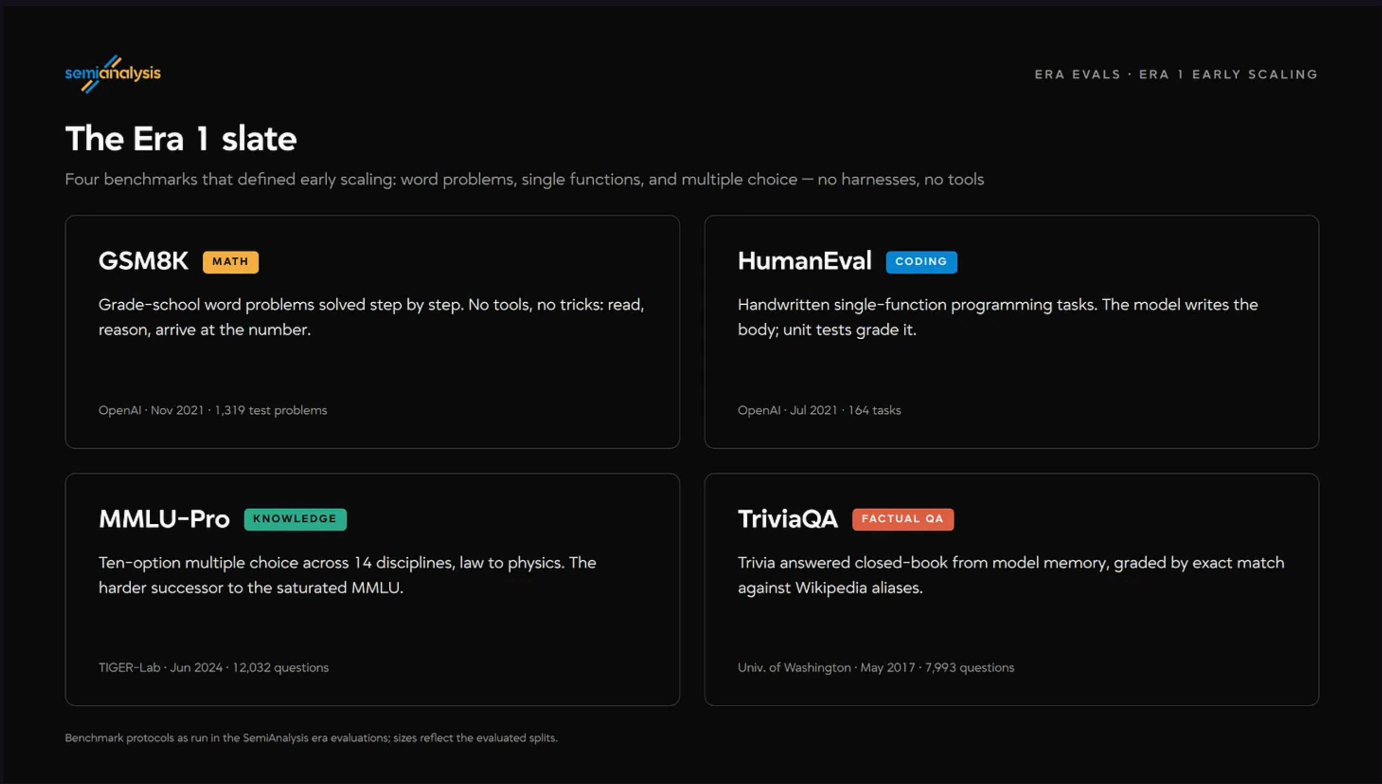
For benchmarks, we relied on a combination of personal taste and popularity. Humanities Last Exam (HLE), for example, is known to have lots of issues, but was also truly one of the defining benchmarks of the reasoning era with no close substitutes. SWE-bench Pro, on the other hand, is similarly popular and problematic, but also closely approximated by DeepSWE.

Most of the benchmark scores here we ran ourselves using Prime Intellect's evaluation stack, specifically their environments hub and the evals harness included in Prime-RL. The rest come from runs by our friends at Artificial Analysis and Datacurve's DeepSWE leaderboard. Open models were served the way they would have been at release: vLLM versions, hardware that was in use at the time, and sampling settings from the model card. For closed models, we ran against their pinned API versions. Where our numbers share a chart with third-party values, we matched their rulesets.

We'd like to give a huge thank you to Florian Brand (@xeophon) from Prime Intellect for helping us pick benchmarks/models, implement evals, and check for correctness.

It's June 2023. The world is reckoning with ChatGPT, and Mark Zuckerberg just agreed to fight Elon Musk at the Colosseum. But while Zuck is training jiu-jitsu and doing Murphs, his company is doing some training of their own. FAIR is about to push past the Mistral exodus and other drama, and successfully ship Llama-2-70B. The first open model that approached the frontier.

How far behind the frontier was it? Four benchmarks helped define SOTA at the time: GSM8K, HumanEval, TriviaQA, and MMLU-Pro:



Source: SemiAnalysis

These benchmarks are representative of what the frontier models were capable of at the time. Simple multiple choice questions, word problems, and programming problems scoped to single functions. How times have changed! Here is how Llama-2 stacked up against GPT-3.5 Turbo in a cage match of their own:

Source: SemiAnalysis

To account for differences in benchmark difficulty, we normalized the scores. Each era's best result is set to 100, and every other model is scored relative to that. The composite score represents the equal-weight average of the four: 75.7 for GPT-3.5 Turbo on the frontier, and 39.9 for Llama-2-70B. A measured, but considerable gap. This initial lag creates the storyline for the rest of the era: Mixtral-8x7B released in December 2023 created momentum towards GPT-4 capability, only to have GPT-4 Turbo and GPT-4o race ahead:

It took until the Llama-3.1-405B release in July 2024 for open models to close the GPT-3.5 Turbo gap, with a composite score of 86. The last frontier model, GPT-4o, was matched in capability by DeepSeek V3 in December 2024, scoring 95.5 and 94.1 respectively. Qwen2.5-72B landed within striking distance of GPT-4o at a sixth of the 405B parameter count, pre-trained on 18T tokens.

This is the first instance of the gap closing. Through this era, we didn't see the frontier rise much beyond the capabilities of GPT-4, but this was mostly due to priorities: Turbo and 4o were built to make GPT-4 cheaper and faster, not smarter.

Meanwhile, OpenAI had been working towards a different kind of model. The process-reward paper and Noam Brown hire both pointed at reasoning, and by mid-2024 every major lab was publishing test-time-compute research. Seven weeks after 405B, on September 12 2024, OpenAI shipped o1-preview: a model that sparked a new era of innovation.

Era 2 | Reasoning (2024-2025)

o1 reset the choice of benchmarks, along with the gap. The elementary evals from Era 1 were no longer difficult enough to test o1's full abilities. Grade school math problems were replaced by the AIME. Scale AI collected some of the most esoteric, PhD-level multiple choice questions in the world and provocatively called it Humanity's Last Exam.

Source: SemiAnalysis

It’s hard to overstate how important the release of o1 was in tech circles. Many consider it the day “we knew for sure we’d get AGI.” However, unlike Llama-2-70B vs GPT-4, the gap between open vs closed source started off much smaller during Era 2. The culprit? A little known model called DeepSeek R1.

Source: SemiAnalysis

open source is good for AI infrastructure. The "we're so back" open model momentum established by R1 was soon squashed by Meta's Llama-4 Maverick and other Chinese models.

Source: SemiAnalysis

Gemini 2.5 Pro and o3 continued to push the reasoning frontier, and the R1-0528 checkpoint closed the initial gap in May 2025 with a score of 78. An 8.5 month window to close a 12.1 point gap:

Source: SemiAnalysis

Notably absent from the charts so far is Anthropic. Their model cards reported these benchmarks like everyone else's, but they never fought for the top of the leaderboard in this era. While OpenAI and Google traded crowns, Anthropic was turning Claude into the default coding agent. This set the terms for the next era: the benchmarks that now matter run in a terminal.

Era 3 | Agentic (2025 - Today)

Prior to Claude Code, agents had their moments (like Cognition's viral demo of Devin in March 2024), but Anthropic was the first to nail a model + harness product—and it paid off. Since the general release of Claude Code in May 2025, Anthropic has added north of $65B in ARR. For in-depth Anthropic and OpenAI ARR projections, see our Tokenomics Model.

With agents came yet another new set of benchmarks. Fancy math problems were no longer the best test of model capabilities. Instead, people wanted to know how well models could write code, do web search, and generally use a computer like a human.

Source: SemiAnalysis

Terminal-Bench 2.1, BrowseComp-Plus, τ^3-banking, and DeepSWE cover the long-horizon work agents are used for today: software engineering, deep research, and knowledge work. We also picked benchmarks that skew newer by design to limit memorization.

Source: SemiAnalysis

benchmark suite, but this didn't correspond to a better user experience. The full agentic product (model + harness) was now what mattered, and Anthropic had been laser-focused on iterating towards a harness that excelled at general agentic work. Codex, in contrast, was comparatively crude, and OpenAI was simultaneously pursuing side quests like web browsers.

Model releases also compressed between the two frontier labs. OpenAI and Anthropic created their duopoly by releasing a model every 51 days on average throughout this era. Compared to the 213 and 120 day release averages throughout Era 1 and Era 2, respectively, this is a massive speed up.

Source: SemiAnalysis

Yet despite the massive explosion in economic value created by frontier models, the gap closed faster in Era 3 than either era before it. Kimi K2.6 surpassed Opus 4.5 with a score of 56.3 in 4.8 months, and GLM-5.2 cleared GPT-5.2 with a score of 72.4 in 6 months. The trend of the closing time halving with each subsequent era is remarkably consistent.

Source: SemiAnalysis

Looking forward

So, what does this all mean for the future of closed vs open source models?

First, we want to acknowledge that **benchmarks are not the end all be all**. Kimi K3 may score higher than Fable 5 on our curated composite, but we still prefer using Fable at SemiAnalysis for our day to day work. This is partly because Anthropic has done a better job productizing their model via things like Claude Code and Claude Tag, but also largely because benchmarks aren't a perfect proxy for real work. **This is especially true for public benchmarks, which model makers can easily hill climb by simply creating a bunch of RL environments that closely mimic the benchmark tasks.**

Second, you may argue that the closing time for Era 3 is artificially deflated due to Anthropic and OpenAI spending more time on safety testing than Moonshot and Zhipu, but this is not a new phenomenon. GPT-4, for example, finished training 218 days before release. Even if we assume Mythos finished training in mid February, that's still only a 114 day delay before the Fable release.

The Upcoming Era

Previous