The Coding Assistant Breakdown: More Tokens Please

《The Coding Assistant Breakdown: More Tokens Please》

Hands On With GPT 5.5, Opus 4.7, DeepSeek V4, Why Benchmarks Are Bad, and Who's Going To Win

亲身体验 GPT 5.5、Opus 4.7、DeepSeek V4、为什么基准测试不好，以及谁将获胜

MAX KAN, JORDAN NANOS, SAMUEL KRUSE, AND 4 OTHERS

APR 25, 2026 2026 年 4 月 25 日 · PAID · 已付费

Since we called out the Claude Code inflection point on February 5th, we have see flurry of model releases. Opus, Mythos, Codex, Gemini, DeepSeek, Kimi, Qwen, G MiniMax, Composer, Muse Spark, and more. Today we will break down all of thes major model releases, explain when you can vs can't trust the benchmarks, and giv our predictions for the future of the agentic coding market.

自从我们在 2 月 5 日指出 Claude Code 的拐点以来，已经出现了一波模型发布浪潮 Opus、Mythos、Codex、Gemini、DeepSeek、Kimi、Qwen、GLM、MiniMax、 Composer、Muse Spark 等等。今天我们将拆解所有这些主要模型发布，解释何时 以以及何时不能信任基准测试，并给出我们对智能代理编程市场未来的预测。

First we have to highlight GPT-5.5 from OpenAI. In our view, GPT-5.5 is now **materially better** at some tasks than all other models. We believe that GPT-5.5 has arrived at the frontier. This is a huge change from November when Opus 4.5 was released. At that time, and for the 6 months since, OpenAI's coding model was no world class in most metrics, leading to Opus being our daily driver. GPT-5.5 is nov

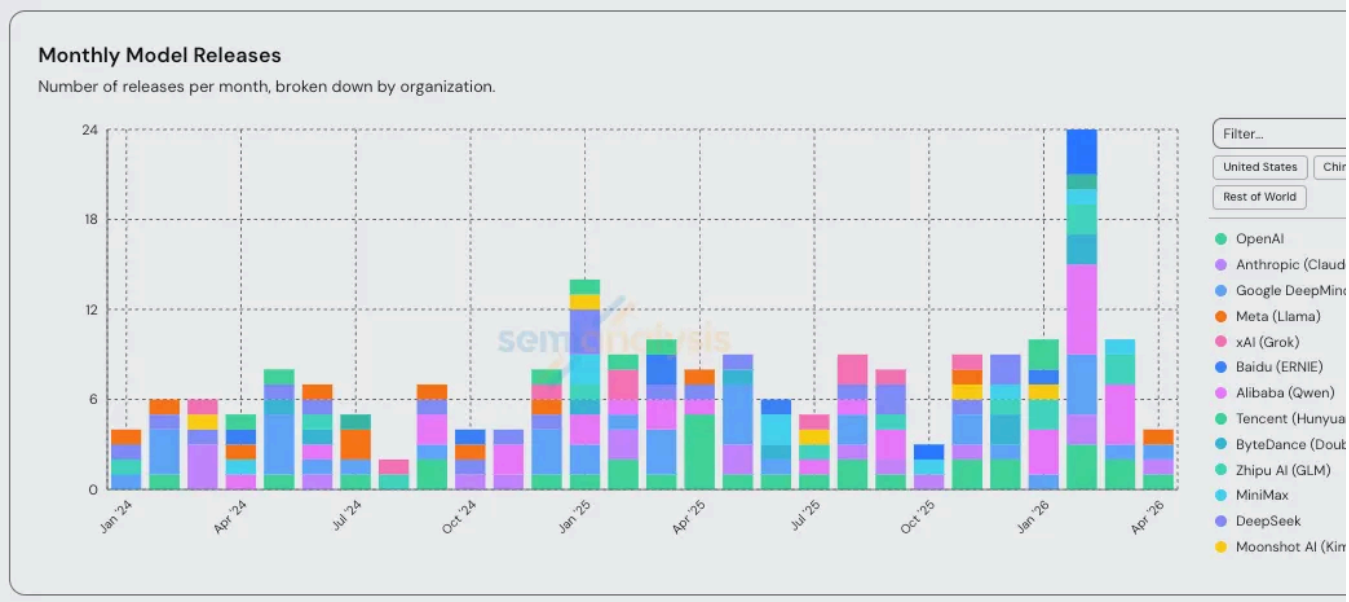
integrated in our daily work.

首先我们必须突出介绍 OpenAI 的 GPT-5.5。在我们看来，GPT-5.5 在某些任务上
前已经明显优于所有其他模型。我们认为 GPT-5.5 已经到达了前沿。这与去年十一
Opus 4.5 发布时的情况发生了巨大变化。那时，以及随后六个月内，OpenAI 的编
模型在大多数指标上并不属于世界一流，这使得 Opus 成为我们的日常首选。现在
GPT-5.5 已被整合到我们的日常工作中。

Meet the Models 认识这些模型

There's been at least one major lab releasing a new checkpoint purpose-built for coding every week for the past 3 months. GLM-5.1, Qwen3.6-Plus, Kimi K2.6, Composer 2, and Gemini 3.1 Pro all emphasize "agentic coding," "long-horizon tas or similar capabilities in their headlines. February was a particularly busy month.

在过去三个月里，至少有一家大型实验室每周都会发布一个为编程专门构建的新榜
检查点。GLM-5.1、Qwen3.6-Plus、Kimi K2.6、Composer 2 和 Gemini 3.1 Pro 都
其标题中强调"代理式编程"、"长远任务"或类似能力。二月尤其繁忙。



Source: SemiAnalysis Tokenomics Dashboard

来源：SemiAnalysis Tokenomics 仪表板

New checkpoints are cool, but entirely new pre-trains are what really get the peop going. Heading into April, the San Francisco rumor mill was ablaze with talk abou Capybara and Spud. These are codenames for Anthropic and OpenAI's newest pre trains. With the release of GPT-5.5 yesterday, we now have something concrete to discuss.

新的检验点很酷，但真正让人兴奋的是全新的预训练模型。进入四月时，旧金山的言工厂关于 Capybara 和 Spud 的讨论甚嚣尘上。这些是 Anthropic 和 OpenAI 最训练模型的代号。随着昨日 GPT-5.5 的发布，我们现在有了可以具体讨论的内容。

GPT 5.5

GPT-5.5 is the first public release based on "Spud". As OpenAI's first new pre-trai since the failed GPT-4.5, expectations are obviously high. And despite both NVID and OpenAI claiming with precise language that the model was "trained" on a 10(GB200 NVL72 cluster, this "training" is post-training (RL) only. Pre-training is sti Hopper.

GPT-5.5 是基于 "Spud" 的第一个公开发布版本。作为 OpenAI 在失败的 GPT-4.5 后的首个新预训练模型，外界显然对此抱有很高期望。尽管 NVIDIA 和 OpenAI 都精准措辞声称该模型是在一个 100k GB200 NVL72 集群上"训练"出来的，但这种"训练"仅指的是后训练（强化学习）阶段。预训练仍然是在 Hopper 上完成的。

OpenAI's flagship model has historically been cheaper than Anthropic's, but at $5 million input tokens and $30 per million output tokens, GPT-5.5's API price will b more expensive than GPT-5.4 and slightly more expensive than Opus 4.7. The API went live this morning after a brief ChatGPT/Codex-only window due to safety concerns. We've been testing the model via Codex and API during an alpha testing

period and describe that experience later in this article.

OpenAI 的旗舰模型历来比 Anthropic 更便宜，但以每百万输入 tokens $5、每百万
出 tokens $30 的定价，GPT-5.5 的 API 价格将比 GPT-5.4 贵 2 倍，并且比 Opus 4
略贵。由于安全问题，API 在经过一个仅限 ChatGPT/Codex 的短暂窗口后于今晨
线。我们在 alpha 测试期间通过 Codex 和 API 测试了该模型，并在本文后文描述
体验。

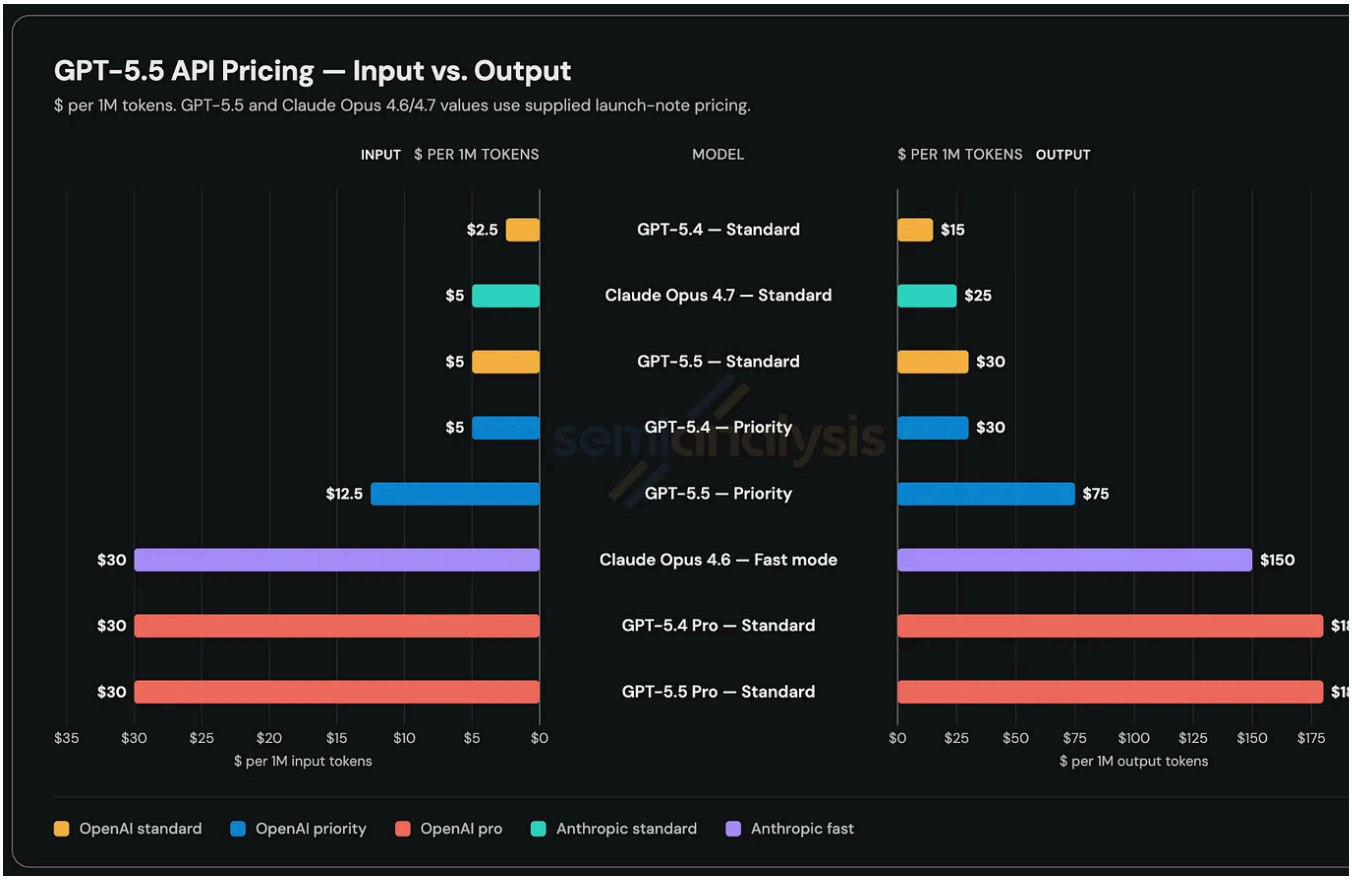
Like all their other models, OpenAI will also be offering a priority tier for GPT-5.5
priced at 2.5x the standard rate. Figuring out how to charge users more money for
faster tokens is becoming increasingly important, and it's worth clarifying that pri
is totally different from fast mode. Fast mode just makes some vague guarantees li
"2.5x faster for 6x the price," whereas priority makes more conservative, concrete S
(e.g. > 50 tokens/sec > 99% of the time). Both Anthropic and OpenAI offer fast mod
and priority tiers, but we think Opus 4.6 Fast is the only SKU that's gained real
traction.

像他们所有其他模型一样，OpenAI 也将为 GPT-5.5 提供一个优先级层，价格为标
费率的 2.5 倍。想办法向用户多收费以换取更快的 tokens 变得越来越重要，需要注
的是，优先级与快速模式完全不同。快速模式只是做出一些模糊的承诺，例如"价格
原来的 6 倍但速度提升 2.5 倍"，而优先级则提出更为保守、具体的服务级别协议
如 > 50 tokens/sec 且在 > 99% 的时间内达到）。Anthropic 和 OpenAI 都提供快速模
和优先级层，但我们认为只有 Opus 4.6 Fast 是唯一获得实际吸引力的 SKU。

Separately, OpenAI also offers GPT-5.3-Codex-Spark, but that's a totally different
model built to run on Cerebras. Specifically, it is a distilled version of GPT-5.3. Th
a big difference between offering faster tokens via running smaller batch sizes,
changing the reasoning depth, and routing requests to a priority queue without
changing the underlying model (priority and fast mode) vs running a dumber, sma

model (codex spark).

> 另一个方面，OpenAI 还提供 GPT-5.3-Codex-Spark，但那是为在 Cerebras 上运行构建的完全不同的模型。具体来说，它是 GPT-5.3 的蒸馏版本。在通过运行更小的处理、更改推理深度以及将请求路由到优先队列而不更改底层模型（优先和快速模来提供更快的 token，与运行一个更“愚蠢”、更小的模型（codex spark）之间存在大差异。



Source: SemiAnalysis

Also released is GPT-5.5 Pro, which is only available via ChatGPT and API. It's m for scientific research or long range reasoning tasks instead of everyday agentic w GPT-5.5 Pro earned SOTA scores on BrowseComp and FrontierMath, and is price the same $30/180 as GPT-5.4 Pro. We expect to see more announcements about GI

5.5 Pro making scientific discoveries soon.

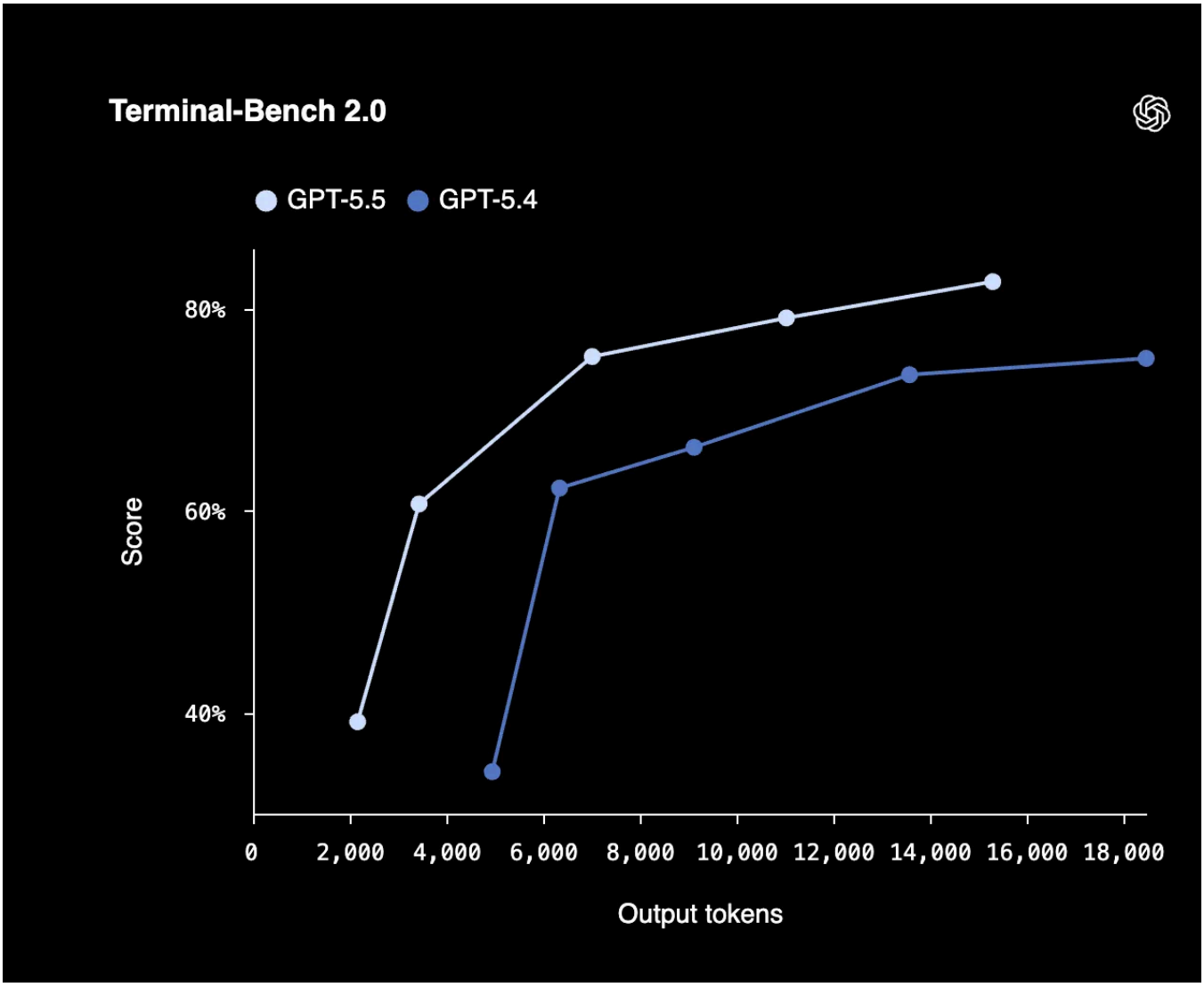
还发布了 GPT-5.5 Pro，该版本仅通过 ChatGPT 和 API 提供。它旨在用于科学研
或长程推理任务，而非日常的代理式工作。GPT-5.5 Pro 在 BrowseComp 和
FrontierMath 上取得了 SOTA 成绩，定价与 GPT-5.4 Pro 相同，为 $30/180。我们
计很快会看到更多关于 GPT-5.5 Pro 在科学发现方面的公告。

Both the standard and pro models offer different levels of reasoning: xhigh, high, medium, low, and non-reasoning, which is a tradeoff between cost vs capability. A has been clear since the release of strawberry/o1, higher reasoning levels lead to b outputs but require more tokens and users have to wait longer for a response.

标准模型和专业模型都提供不同级别的推理能力：xhigh、high、medium、low 和
non-reasoning，这在成本与能力之间做出权衡。自从 strawberry/o1 发布以来已经
明显，较高的推理级别会产生更好的输出，但需要更多的 token，用户也必须为响
等待更长时间。

Relatedly, OpenAI advertised in their model card that GPT-5.5 scores higher on benchmarks than 5.4 while simultaneously using less tokens. In other words, it's n "token efficient." This is an extremely important concept to understand, and we believe it will become a major talking point this year. As we explained and quantif to Tokenomics model subscribers last week, **cost per task, not cost per token, is t true north star metric that determines model pricing**. Mythos may be 5x more expensive than Opus on a per token basis, but much of that price increase is nullif because Mythos can solve the same problem using fewer tokens. It may also be a f end to end response.

相关地，OpenAI 在其模型卡中宣称 GPT-5.5 在基准测试上的得分高于 5.4，同时
的 tokens 更少。换言之，它在"token 效率"上更高。这是一个极其重要的概念，我
相信今年它将成为一个主要讨论点。正如我们上周向 Tokenomics 模型订阅者所解
并量化的那样，每项任务的成本（cost per task），而非每个 token 的成本（cost p
token），才是决定模型定价的真正北极星指标。按每 token 计算，Mythos 可能比
Opus 贵 5 倍，但由于 Mythos 可以用更少的 tokens 解决同样的问题，这部分价格
涨在很大程度上被抵消。它的端到端响应也可能更快。



Source: OpenAI

Opus 4.7

This all comes a short week after Anthropic's release of Claude Opus 4.7, a drop-i replacement for Claude Opus 4.6. Opus has been the daily driver for most of SemiAnalysis, and Opus 4.7 is a small improvement. With improved scores on ma benchmarks and predictably good vibes, but not a step change, 4.7 has been reluct adopted by our team members. Why? Fast mode does not exist yet. For the first tir we have found that many of our engineers are willing to sacrifice a bit of quality (b not too much) for faster speed, claiming that the 2.5x faster for 6x the price tradeo

lets them hit “flow state”.

这发生在 Anthropic 发布 Claude Opus 4.7 仅一周之后，Opus 4.7 是 Claude Opus 的直接替代品。Opus 一直是大多数 SemiAnalysis 的日常驱动器，而 Opus 4.7 是一小幅改进。在多个基准测试上得分有所提高，整体感受也算良好，但并非质的飞跃。4.7 已被我们团队成员勉强接受。为什么？因为尚未推出快速模式。我们首次发现，多工程师愿意为更快的速度牺牲一些质量（但不会太多），他们表示以 6 倍价格换取 2.5 倍速度的权衡能让他们进入“心流状态”。

Source: of our frustration (i.e. Dylan on X

来源：我们挫败感的源头（即 Dylan 在 X 上）

In practice, the noticeable changes moving from Opus 4.6 → Opus 4.7 have been f features/functionality rather than raw performance. In general, these models have gotten so good that most day-to-day tasks are accomplished successfully, with our

engineers' criticisms of a code edit or PR being more about style, approach, architectural decisions, and token efficiency (i.e. speed) rather than success on functional tests. It is increasingly rare for any of these coding models to go haywir and botch a commit completely.

在实际应用中，从 Opus 4.6 → Opus 4.7 的显著变化更多体现在功能/特性上，而非始性能。总体而言，这些模型已经变得非常出色，大多数日常任务都能成功完成；们的工程师对代码修改或 PR 的批评更多集中于风格、方法、架构决策和令牌效率（即速度），而不是功能测试的通过与否。如今这些编码模型完全失控并彻底搞砸提交的情况越来越少见。

As a result, the noticeable changes in this transition are:

因此，此次过渡中值得注意的变化有：

1. High-resolution image support, and a clear increase in RL training objectives include the use of screenshots for frontend styling rather than running tests programmatically via headless browsers and tools like playwright

   高分辨率图像支持，以及强化学习训练目标明显增加，开始使用截屏来处理前样式问题，而不是通过无头浏览器和像 playwright 这样的工具以编程方式运行试

2. An "xhigh" reasoning effort option that slots in between "high" and "max" on hierarchy of effort (i.e. how much time the model is going to spend reasoning about a task, described earlier)

   一个"xhigh"推理力度选项，位于力度层级中"high"和"max"之间（即模型将在大程度上花时间对任务进行推理，之前已作描述）

3. Thinking content is omitted by default. Of course, you still get charged for the tokens, but you have to opt in to see them.

   思考内容默认被省略。当然，这些代币仍会计费，但你必须选择加入才能看到们。

4. Task budgets (in beta, and API only) where the model is given a suggestion on how efficiently to complete a task. If the model is given a task budget that is t restrictive, it can take shortcuts or refuse. This is different from max_tokens, which is a hard restriction on output length

> 任务预算（测试版，仅限 API），模型会收到关于以何种效率完成任务的建议。果给模型的任务预算过于严格，它可能会采取捷径或拒绝执行。这不同于 max_tokens，后者是对输出长度的硬性限制

5. Updated token counting, the most critical change when it comes to pricing. 4. uses a new tokenizer, which trades off improved performance via more granul token counting for more total token usage. They admit directly that this will l to increases up to 35% in token usage. Implicitly, this is a 35% increase in pric

> 更新的计数器令牌，这是与定价最相关的重大变化。4.7 使用了一个新的分词器该分词器通过更细粒度的令牌计数换取改进的性能，但也导致总体令牌使用量加。他们直接承认这将导致令牌使用量最多增加 35%。隐含地，这意味着价格增加 35%！

On model behavior changes, the biggest thing we have noticed in our testing is ho 4.7 is using fewer tool calls by default, and using reasoning more. The jury is still o on the benefits here, but in general we don't like it. Anthropic suggests raising the reasoning effort from high to xhigh or max to increase tool usage. And it seems th our users are doing exactly this in order to let the model bring in enough context t successfully complete a complex task or form a complete multi-step plan. Not exac the token efficiency tradeoff claimed in the announcement.

> 在模型行为变化方面，我们在测试中注意到的最大变化是 4.7 默认情况下调用工具少，而更多依赖推理。对此的利弊尚无定论，但总体来说我们并不喜欢这种变化。Anthropic 建议将推理力度从 high 提升到 xhigh 或 max，以增加工具使用。看起来们的用户正是在这样做，以便让模型引入足够的上下文来成功完成复杂任务或形成整的多步计划。这并不完全是公告中所声称的代币效率权衡。

Notably, many people have been accusing Anthropic of intentionally degrading th model on the lead up to the 4.7 release. Anthropic has categorically denied these

claims, but multiple engineers at SemiAnalysis independently said that over the la few weeks the changes in 4.6 performance have made them “feel a little schizo”. A course, they were right.

值得注意的是，许多人指责 Anthropic 在 4.7 发布前故意降低了 4.6 模型的性能。 Anthropic 已断然否认这些指控，但 SemiAnalysis 的多位工程师独立表示，过去几 4.6 性能的变化让他们“感觉有点分裂”。当然，他们是对的。

On April 23, a week after the Opus 4.7 release, Anthropic posted a postmortem detailing three bugs that they found in March/April. All three were present for we and affected basically all users of Claude Code. One of the bugs is trivial, two are interesting, and all are real. When the harness is part of the product, the model ge blamed.

4 月 23 日，在 Opus 4.7 发布一周后，Anthropic 发布了一篇事后分析，详述了他们 三月/四月发现的三个错误。所有三个错误都存在了数周，并基本影响了所有 Claud Code 的用户。其一较为琐碎，另外两个较为有趣，而且全部都属实。当测试工具 产品的一部分时，模型就会被归咎。

We take reports about degradation very seriously. We never intentionally degrade our models, and we were able to immediately confirm that our API and inference layer were unaffected.

After investigation, we identified three different issues:

1. On March 4, we changed Claude Code's default reasoning effort from `high` to `medium` to reduce the very long latency—enough to make the UI appear frozen—some users were seeing in `high` mode. This was the wrong tradeoff. We reverted this change on April 7 after users told us they'd prefer to default to higher intelligence and opt into lower effort for simple tasks. This impacted Sonnet 4.6 and Opus 4.6.
2. On March 26, we shipped a change to clear Claude's older thinking from sessions that had been idle for over an hour, to reduce latency when users resumed those sessions. A bug caused this to keep happening every turn for the rest of the session instead of just once, which made Claude seem forgetful and repetitive. We fixed it on April 10. This affected Sonnet 4.6 and Opus 4.6.
3. On April 16, we added a system prompt instruction to reduce verbosity. In combination with other prompt changes, it hurt coding quality and was reverted on April 20. This impacted Sonnet 4.6, Opus 4.6, and Opus 4.7.

Source: Anthropic Postmortem

来源：Anthropic Postmortem

Notably the three timelines are March 4 to April 7, March 26 to April 10, and Apri to April 20. This is weeks and weeks of bugs going unnoticed. Bugs that were introduced by Claude, and likely root-caused by Claude. Live by the sword, die by sword.

值得注意的是，三个时间线分别为 3 月 4 日至 4 月 7 日、3 月 26 日至 4 月 10 日，及 4 月 16 日至 4 月 20 日。这是数周的错误未被发现。由 Claude 引入且很可能由 Claude 作为根因的错误。生于刀锋，死于刀锋。

DeepSeek V4

The long awaited DeepSeek v4 drop is here. DeepSeek took the world by storm las year with its R1 release and since then there have been legitimate questions in the

community about whether open source models will commoditize intelligence. For those keeping score at home, DeepSeek crashed the market so hard that CEOs we scrambling to explain Jevons paradox. This seems to have played out quite clearly the 16 months since, with the Great GPU Shortage now upon us.

期待已久的 DeepSeek v4 发布来了。DeepSeek 去年凭借其 R1 版本震惊全球，从
起，AI 社区中就有合理的质疑：开源模型是否会使智能商品化。对于在家记分的
说，DeepSeek 对市场的冲击之大以至于 CEO 们争相解释杰文斯悖论。过去 16 个
里这似乎已清晰显现，现在大规模 GPU 短缺正席卷而来。

V4 is an improvement over V3, but it didn't crash the market today. That said, the achievements of DeepSeek should not be discounted. They open-sourced the weig a detailed technical report, and updated libraries such as DeepEP, DeepGEMM, a FlashMLA that are widely used by labs around the world. Ironically, DeepSeek is helping American open source AI stay alive.

V4 比 V3 有所改进，但今天并未引发市场崩盘。尽管如此，不应低估 DeepSeek 的
就。他们开源了权重、详尽的技术报告，并更新了诸如 DeepEP、DeepGEMM 和
FlashMLA 等被全球实验室广泛使用的库。具有讽刺意味的是，DeepSeek 正在帮
美国的开源 AI 求生存。

This release includes two models: DeepSeek-V4-Pro and DeepSeek-V4-Flash. The former is 1.6T total / 49B active, and the latter is 284B total / 13B active. Pro is a st up from V3, which was 671B total / 37B active, while Flash is a step down. We beli that both these architectures are still meaningfully behind their closed-source counterparts on the frontier in terms of both total and active parameter counts. W detail more about how we model the architectures of leading closed source frontie models in our Tokenomics model.

本次发布包含两个模型：DeepSeek-V4-Pro 和 DeepSeek-V4-Flash。前者为 1.6T
数 / 49B 激活参数，后者为 284B 总参数 / 13B 激活参数。Pro 比 V3（671B 总参数
37B 激活参数）更进一步，而 Flash 则是降级版。我们认为这两种架构在总参数与
活参数计数方面仍明显落后于封闭源代码的前沿对手。我们在 Tokenomics 模型中
详细地说明了如何对领先封闭源前沿模型的架构进行建模。

The core advancement of V4 over V3 is a move from a 128k context window to 1M context. As a result, all of the main technical advancements are focused on long context performance. These include:

V4 相对于 V3 的核心进步是将上下文窗口从 128k 扩展到 1M。因此，所有主要的技术改进都集中在长上下文性能上。这些包括：

- Compressed Sparse Attention (CSA)

  压缩稀疏注意力（Compressed Sparse Attention，CSA）

- Heavily Compressed Attention (HCA)

  高度压缩注意力（Heavily Compressed Attention，HCA）

- Manifold-Constrained Hyper-Connections (mHC)

  流形约束超连接（Manifold-Constrained Hyper-Connections，mHC）

And result in the following claim: “In the one-million-token context setting, DeepSeek-V4-Pro requires only 27% of single-token inference FLOPs and 10% of KV cache compared with DeepSeek-V3.2.” That’s a 90% reduction in KV Cache, way more impactful than Google’s TurboQuant paper last month! NAND Flash investors, watch out.

并得出如下断言：“在一百万令牌的上下文设置中，DeepSeek-V4-Pro 在单令牌推理 FLOPs 上只需 DeepSeek-V3.2 的 27%，KV 缓存只需 10%。”这意味着 KV 缓存减少 90%，比上个月 Google 的 TurboQuant 论文的影响要大得多！NAND 闪存的投资者请注意。

On benchmarks, DeepSeek did not feel that standard benchmarks were good at capturing real-world task capability, so they introduced their own set of agentic benchmarks to measure how V4 compared against other SOTA models: Chinese writing, retrieval augmented search, a suite of white-collar tasks with long horizon and coding. V4 Pro was able to compete with top models on all these tasks but lag behind in key areas. For instance, on especially difficult Chinese writing tasks, Cla

Opus 4.7 still beats DeepSeek V4 Pro. Claude mogs Chinese models in it's own language.

在基准测试方面，DeepSeek 认为标准基准无法很好地捕捉现实世界任务能力，因
他们引入了一套自己的智能体基准来衡量 V4 与其他 SOTA 模型的比较：中文写作
检索增强搜索、一套长期白领任务以及编程。V4 Pro 在所有这些任务上能够与顶级
型竞争，但在关键领域仍有落后。例如，在尤其困难的中文写作任务上，Claude
Opus 4.7 仍然击败 DeepSeek V4 Pro。Claude 在其本国语言中碾压中文模型。

Unfortunately, using public announcements on model performance benchmarks a proxy for real world performance is unreliable. Conflicting incentives cause these to publish certain benchmarks and not others. Like this example, where DeepSeek takes a shot at the Kimi and GLM APIs:

不幸的是，使用公开发布的模型性能基准作为现实世界表现的替代指标并不可靠。
益冲突导致这些实验室只发布某些基准而忽略其他基准。比如这个例子，DeepSee
对 Kimi 和 GLM API 发起了攻击：

For formal math tasks, we evaluate in an agentic setting on Lean v4.28.0-rc1 (Moura a Ullrich, 2021), with access to the Lean compiler and a semantic tactic search engine, runn up to 500 tool calls with max reasoning effort. In addition, we evaluate a more compu intensive pipeline in which candidate natural-language solutions are first generated and filte by self-verification(Shao et al., 2025), and the retained solutions are then provided as guida to a formal agent for proving the corresponding Lean statement. This design uses infor reasoning to improve exploration while preserving strict correctness through formal verificat A submission is counted as correct only if the strict verifier Comparator accepts it for b settings.

We have left some entries blank for K2.6 and GLM-5.1, as their APIs were too busy to ret responses to our queries.

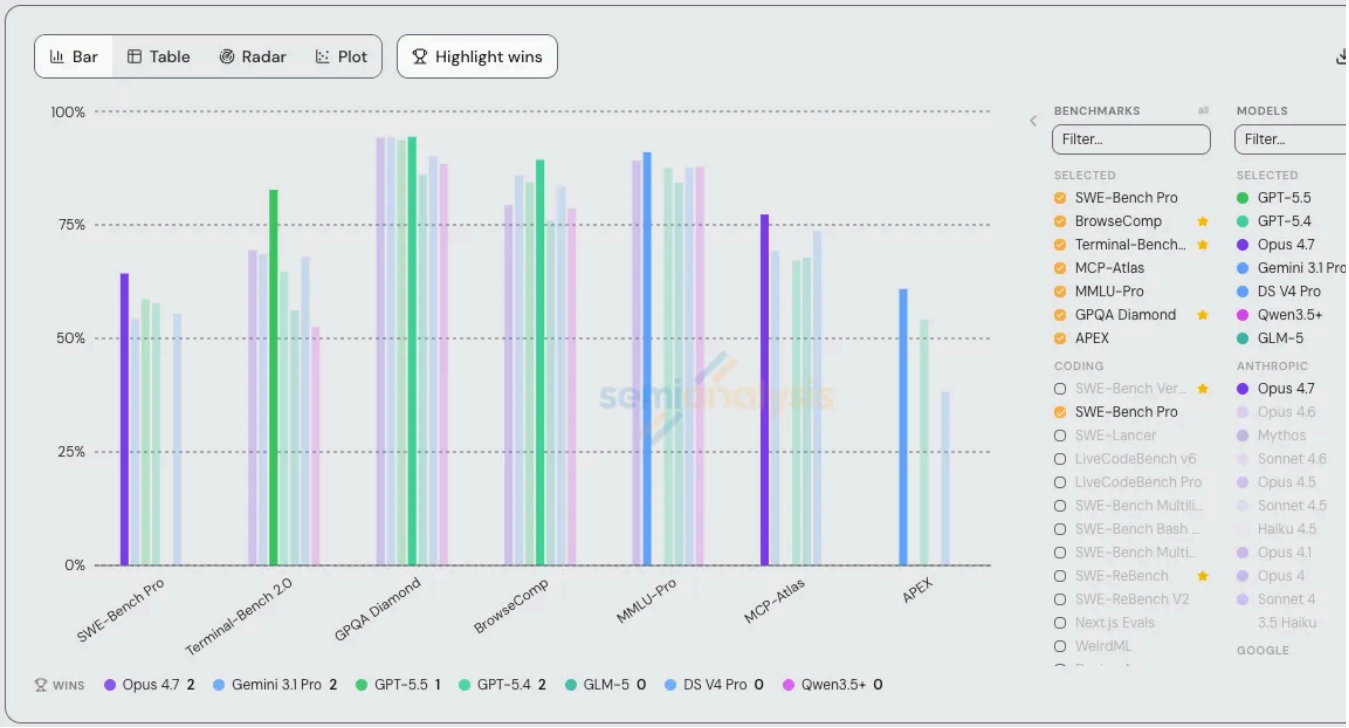
1M-Token Context. Since DeepSeek-V4 series supports 1M-token contexts, we evaluate mo performance in a long context scenario by selecting OpenAI MRCR (OpenAI, 2024b) a CorpusQA (Lu et al., 2026) as the benchmarks. We re-evaluate Claude Opus 4.6 and Gemini Pro on these tasks with the goal of standardizing the configuration across all models. We not evaluate GPT-5.4 because its API failed to respond to a large portion of our queries.

Source: DeepSeek V4 Technical Report

DeepSeek V4 技术报告

This is the reason why the SemiAnalysis Tokenomics Dashboard tracks all major model performance claims, pricing, release dates, usage disclosures in an unbiased manner. We also do our own hands-on testing of all the major models. Below is an example of our tracking of meaningful benchmark performance across the major model releases. We will explain later why benchmarks are bad.

这就是 SemiAnalysis Tokenomics 仪表板以无偏见方式跟踪所有主要模型性能声明定价、发布日期、使用披露的原因。我们也对所有主要模型进行了自己的实际操作试。下面是我们对各主要模型版本间有意义基准性能跟踪的示例。我们稍后会解释什么基准有弊端。



Source: Tokenomics Model 来源：Tokenomics Model

Source: Tokenomics Model 来源：Tokenomics Model

DeepSeek also open sourced a Mega-Kernel inside of DeepGEMM that supports b NVIDIA GPUs and **Huawei Ascend NPUs**. NPU support is claimed, but only the for SM90 (Hopper) and SM100 (Blackwell) GPUs is released publicly. It is likely a g to run a meaningful portion of the future inference traffic on Ascends. It is notabl however that the parameter size fits just inside the memory domain of an 8x H20 I at FP4.

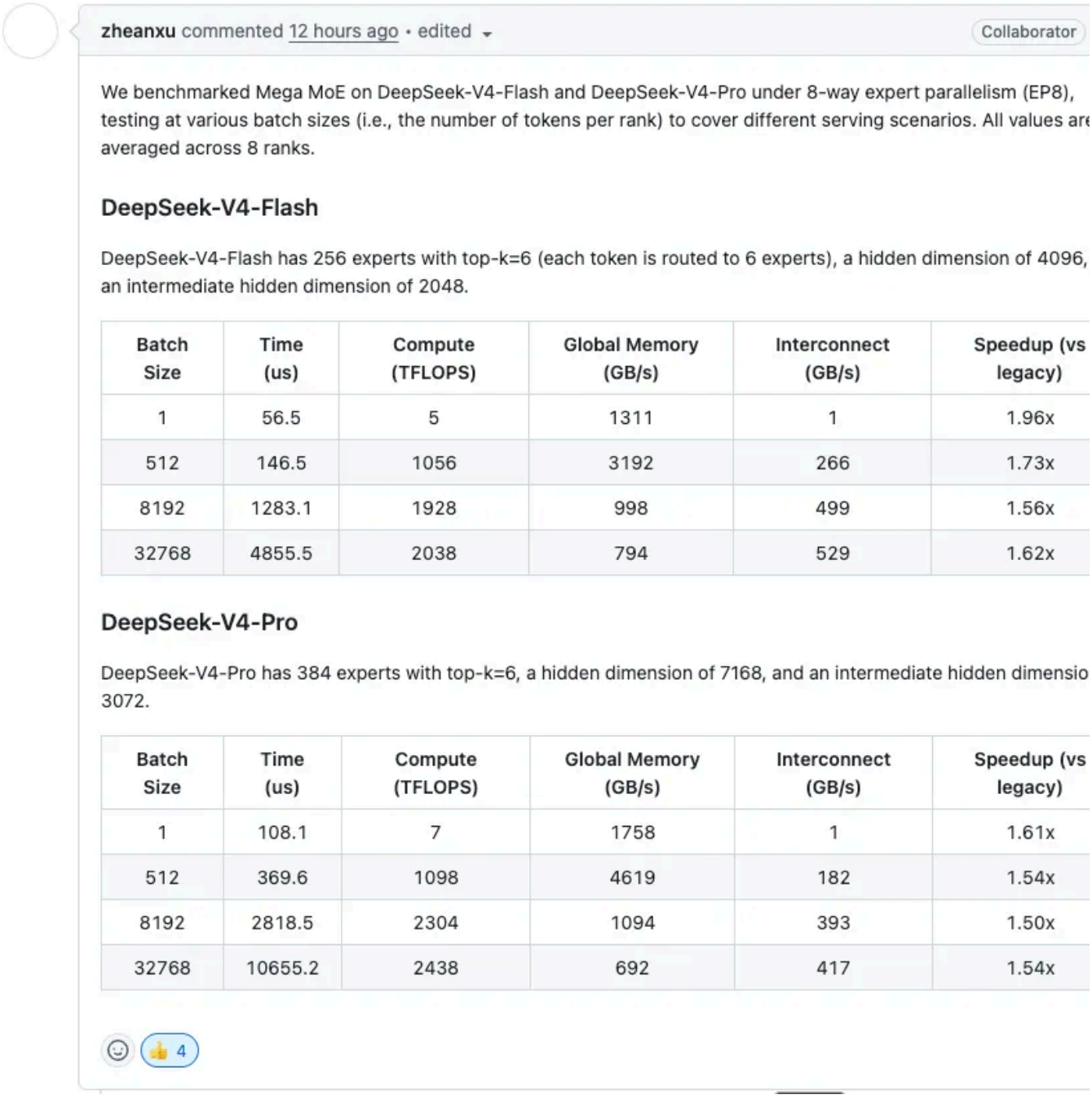
DeepSeek 还在 DeepGEMM 中开源了一个 Mega-Kernel，支持 NVIDIA GPU 和 为 Ascend NPU。声称提供了 NPU 支持，但目前公开发布的代码仅限于 SM90 (Hopper) 和 SM100 (Blackwell) GPU。很可能目标是将未来推理流量的有意义部分 行在 Ascend 上。然而值得注意的是，该参数规模恰好在 8x H20 HGX 的 FP4 内存 内。

Performance and Open-Sourced Mega-Kernel. We validated the fine-grained EP scheme on both NVIDIA GPUs and HUAWEI Ascend NPUs platforms. Compared against strong non-fused baselines, it achieves 1.50 ~ 1.73× speedup for general inference workloads, and up to 1.96× for latency-sensitive scenarios such as RL rollouts and high-speed agent serving. We have open-sourced the CUDA-based mega-kernel implementation named **MegaMoE**[2] as a component of DeepGEMM.

Source: DeepSeek V4 Technical Report

Mega MoE performance across various batch sizes is described in a PR:

关于不同批量大小下 Mega MoE 的性能已在一个 PR 中有所描述：



zheanxu commented 12 hours ago • edited — Collaborator

We benchmarked Mega MoE on DeepSeek-V4-Flash and DeepSeek-V4-Pro under 8-way expert parallelism (EP8), testing at various batch sizes (i.e., the number of tokens per rank) to cover different serving scenarios. All values are averaged across 8 ranks.

DeepSeek-V4-Flash

DeepSeek-V4-Flash has 256 experts with top-k=6 (each token is routed to 6 experts), a hidden dimension of 4096, an intermediate hidden dimension of 2048.

Batch Size	Time (us)	Compute (TFLOPS)	Global Memory (GB/s)	Interconnect (GB/s)	Speedup (vs legacy)
1	56.5	5	1311	1	1.96x
512	146.5	1056	3192	266	1.73x
8192	1283.1	1928	998	499	1.56x
32768	4855.5	2038	794	529	1.62x

DeepSeek-V4-Pro

DeepSeek-V4-Pro has 384 experts with top-k=6, a hidden dimension of 7168, and an intermediate hidden dimensio 3072.

Batch Size	Time (us)	Compute (TFLOPS)	Global Memory (GB/s)	Interconnect (GB/s)	Speedup (vs legacy)
1	108.1	7	1758	1	1.61x
512	369.6	1098	4619	182	1.54x
8192	2818.5	2304	1094	393	1.50x
32768	10655.2	2438	692	417	1.54x

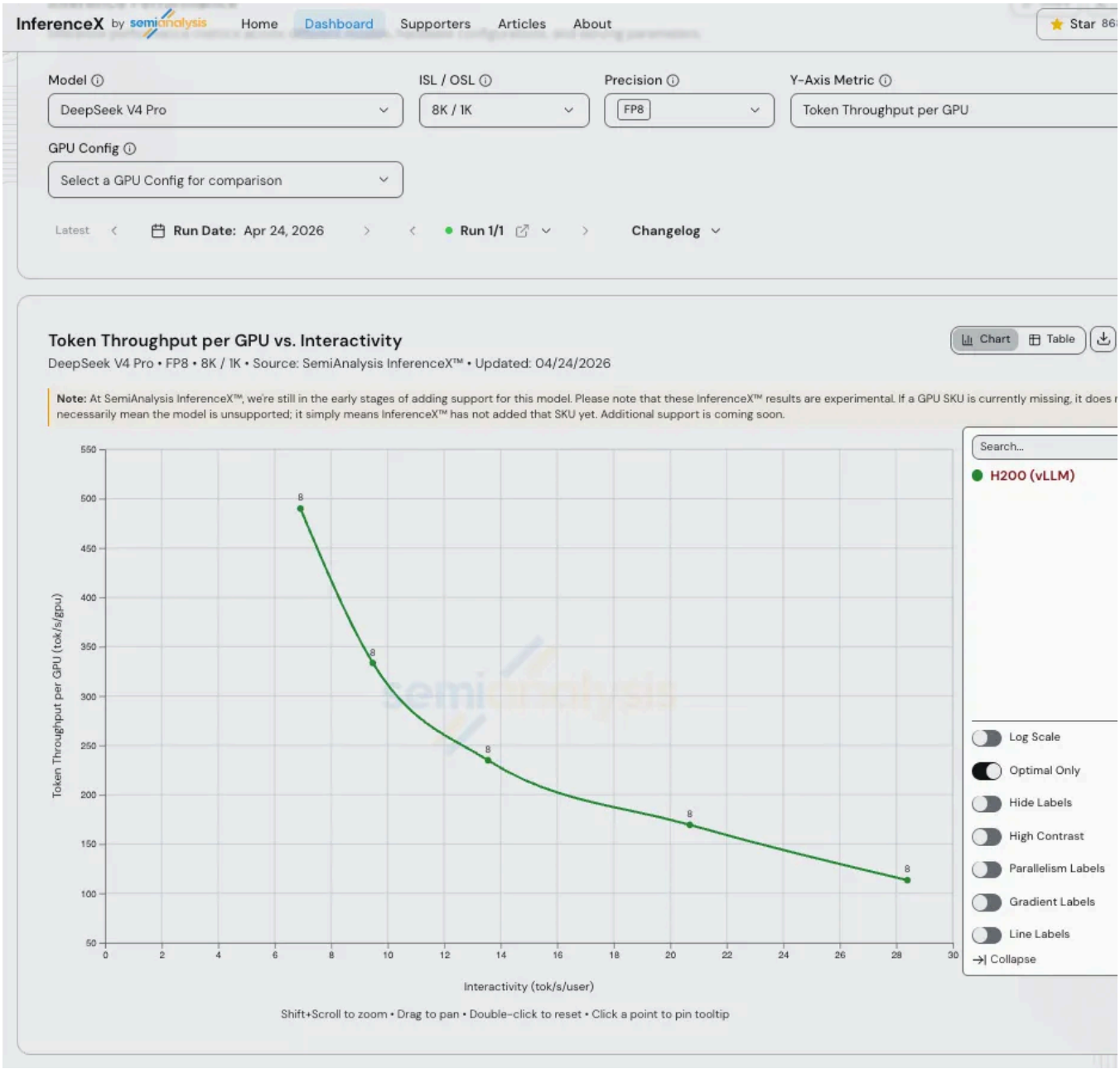
4

Source: DeepGEMM repo 来源：DeepGEMM 仓库

Of course, the key contribution of DeepSeek V4 is that it is open source. Thanks t all nighter, our InferenceX team, collaborating with 10x engineers from vLLM/Inf and NVIDIA, have published day-zero support on our H200 cluster. Support for Blackwell and AMD GPUs using vLLM, SGLang and TRT-LLM with Dynamo is a

work in progress.

当然，DeepSeek V4 的关键贡献在于它是开源的。多亏了一夜未眠，我们的 InferenceX 团队与来自 vLLM/Inferact 和 NVIDIA 的十倍工程师协作，已经在我 H200 集群上发布了零日支持。使用 vLLM、SGLang 和配合 Dynamo 的 TRT-LL 对 Blackwell 和 AMD GPU 的支持仍在进行中。



Source: inferencex.com 来源：inferencex.com

Interestingly, day-zero support on H200 at FP8 [1] performance of this tok/sec throughput per GPU at 20 tok/sec interactivity on 8k in 1k ou V3 hits ~1.3k to 2.3k tok/sec of throughput per GPU at 20 tok/sec inte

1k out. This is a new model and we expect meaningful optimization in the coming weeks. Watch

cannibalized by DeepSeek.

总体而言，DeepSeek 是一次卓越的工程发布，紧随 SOTA 前沿。它将成为对闭源
型成本最低的替代方案，但其能力并不处于领先地位。SemiAnalysis 的工作流可能
会被 DeepSeek 吞噬。

VIBEZ: Our Impressions of GPT-5.5 vs Opus 4

VIBEZ：我们对 GPT-5.5 与 Opus 4.7 的印象

SemiAnalysis is famous (infamous?) for shilling Claude, and we've been testing GP
5.5 as part of an alpha program with OpenAI the past few weeks.

SemiAnalysis 因为推崇 Claude 而出名（或臭名昭著？），过去几周我们作为与
OpenAI 的 alpha 计划一部分一直在测试 GPT-5.5。

We think GPT-5.5 is a significant improvement within Codex specifically. Previou
~all our engineers used Claude exclusively, and use of ChatGPT models for coding
restricted to wrappers like Cursor. Now, most of our engineers switch between Co
and Claude models depending on the task and IDE preference. Here are some quo

我们认为 GPT-5.5 在 Codex 体系内是一次重大改进。此前，几乎所有工程师都只
Claude，而用于编码的 ChatGPT 模型仅限于像 Cursor 这样的封装工具。现在，
数工程师会根据任务和 IDE 偏好在 Codex 与 Claude 模型之间切换。以下是一些引
用：

> *"What I have really appreciated about Codex recently is how it pulls in a lot of context*
> *before making changes to code. Not like just a structural change, but a change that act*
> *requires non trivial 'thinking'. 4.7 often feels like it just does a quick Explore and then*
> *#yolos changes whereas codex pulls in a shit ton of more granular context from the inte*

> *+ codebase and then makes a directed effort at the ask"*
>
> *"我最近真正欣赏 Codex 的地方是它在修改代码前会引入大量上下文。不只是结性的改动，而是需要非平凡"思考"的改动。4.7 常常感觉只是做了一个快速的探后 #yolos 改动，而 Codex 会从互联网和代码库中拉入大量更细粒度的上下文，针对请求做出有方向性的努力。"*

> *"Currently I use Codex for reviewing PRs/bug hunting, explaining existing code, and creating/revising documentation. Its better at understanding code structure and reason about it."*
>
> *"目前我用 Codex 来审查 PR/排查 bug、解释现有代码，以及创建/修订文档。它解代码结构并进行推理方面更好。"*

However, it's not all positive for OpenAI. Some of our other engineers complained Codex is still worse at inferring your true intent than Claude Code. Humans natur give terse and not particularly well thought out instructions when prompting codi agents, and Codex often listens too literally.

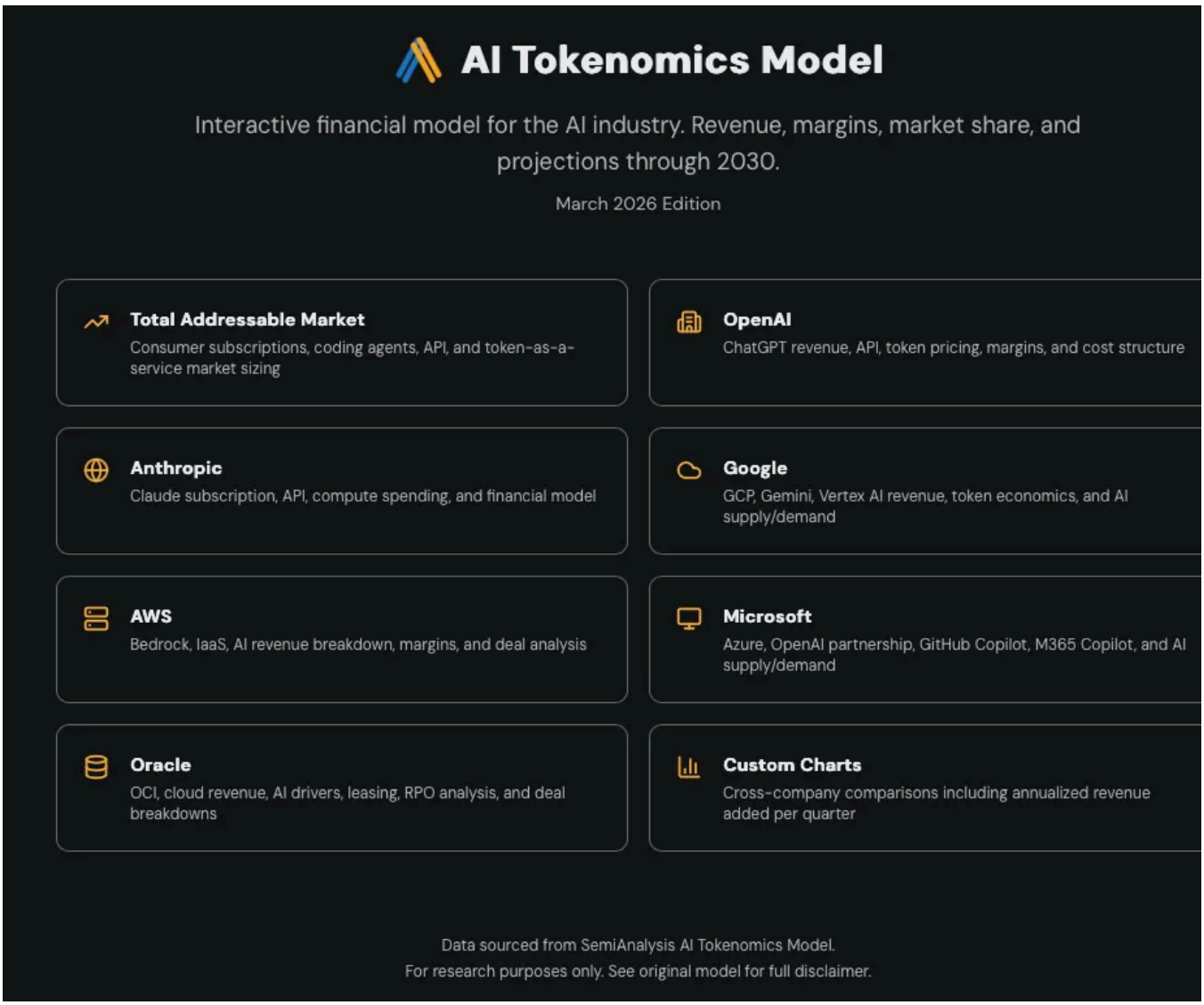
> 然而，对 OpenAI 来说并非全是正面的。我们的一些其他工程师抱怨说 Codex 在推你的真实意图方面仍然不如 Claude Code。人在提示编码代理时自然会给出简练且一定深思熟虑的指令，而 Codex 往往过于字面地执行。

Relatedly, another engineer commented that GPT-5.5 feels too conservative when comes to actually making code changes. Yes, this improves token efficiency, but it comes at the cost of correctness. A similar tradeoff happened from 4.6 → 4.7 as we described previously. Seeing the words "narrow fix" in the output is now a signal t double check the model's work.

> 相关地，另一位工程师评论说 GPT-5.5 在实际进行代码更改时显得过于保守。是的这提高了令牌使用效率，但代价是正确性。如前所述，从 4.6 → 4.7 也发生了类似的衡。现在在输出中看到"narrow fix"这样的字样就成了一个信号，提示需要复查模型工作。

Here's a concrete example that illustrates our overall impression on the strengths weaknesses of Codex vs Claude Code well. We asked both Opus 4.6 and GPT-5.5 t make a new dashboard for our accelerator model and gave it the current tokenomi dashboard as an example. As our institutional subscribers know, this dashboard includes a homepage that links to all the different tabs.

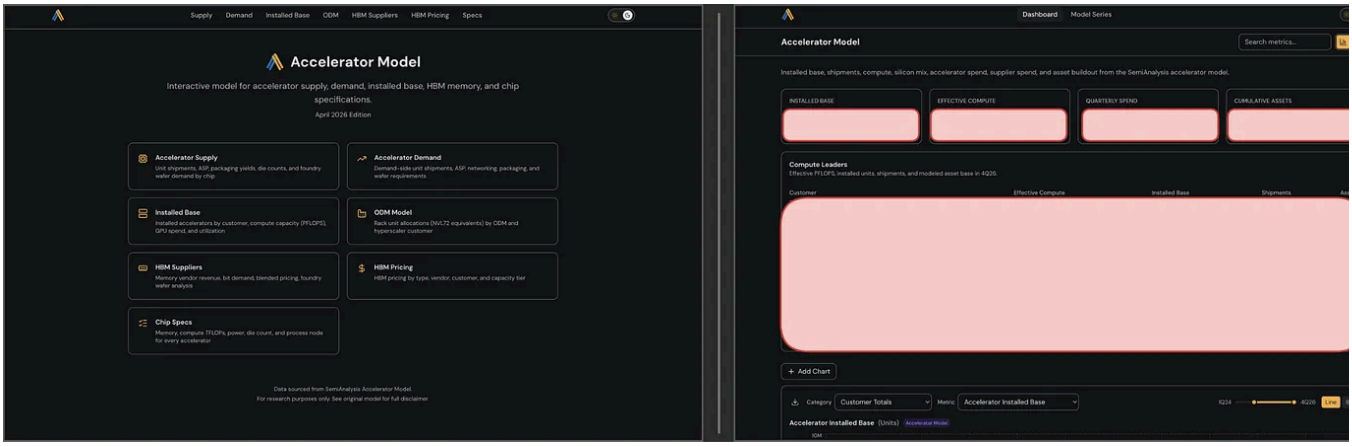
这里有一个具体的例子，很好地说明了我们对 Codex 与 Claude Code 优势和劣势 体印象。我们要求 Opus 4.6 和 GPT-5.5 为我们的加速器模型制作一个新的仪表盘 将现有的 tokenomics 仪表盘作为示例提供。正如我们的机构订阅者所知，该仪表 包含一个链接到所有不同标签页的首页。



Source: SemiAnalysis 来源：SemiAnalysis

Opus 4.6 made an identical looking homepage, whereas Codex ignored it entirely.

Opus 4.6 做了一个外观相同的主页，而 Codex 完全忽略了它。



Source: SemiAnalysis 来源：SemiAnalysis

If we specifically asked Codex to copy the homepage in the prompt, we're sure it would've done so, but it was unable to infer this intent itself.

如果我们在提示中明确要求 Codex 复制主页，我们确信它会那样做，但它无法自行判断出这一意图。

With that said, the actual data Codex included in the dashboard was much more accurate than Claude (though to be clear neither was perfect on the first pass). This implies stronger reasoning about the data structures and relationships with a relatively complex excel file on the part of Codex. Meanwhile, many of Claude's numbers were straight up hallucinated and it made mistakes like including Nvidia GPUs in TPU charts. This tracks with our overall impression that Codex is "smarter and better at doing complex reasoning to solve harder, more narrowly scoped task

whereas Claude is better for more open ended, greenfield problems.

话虽如此，Codex 在仪表板中包含的实际数据比 Claude 要准确得多（尽管要明确出，两者在首次尝试时都并非完美）。这表明 Codex 对数据结构和关系有更强的推能力，能够处理相对复杂的 Excel 文件。与此同时，Claude 的许多数字纯属幻觉，还犯了将 Nvidia GPU 包含在 TPU 图表中的错误。这与我们的总体印象一致：Co更“聪明”，在进行复杂推理以解决更难、更具狭窄范围的任务时表现更好，而 Cla更适合更开放、更具开创性的问题。

It’s for these reasons that some of our engineers have settled on the following workflow:

也正是出于这些原因，我们的一些工程师采用了以下工作流程：

1. Start off with Claude to create an initial plan/scaffolding for new applications features, and the first implementation/POC step.

   先用 Claude 制定初步计划/架构，用于新应用或功能，以及首个实现/概念验证骤。

2. Switch to Codex to actually solve the problem or fix bugs

   切换到 Codex 实际解决问题或修复漏洞

Importantly, before the GPT-5.5 release, ~all of SemiAnalysis used Claude Code fo both of these steps. Our use of ChatGPT models had become restricted to Deep Research on the webapp and wrappers like Cursor Bugbot.

重要的是，在 GPT-5.5 发布之前，SemiAnalysis 的几乎所有工作在这两个步骤上都用 Claude Code。我们对 ChatGPT 模型的使用已被限制为 webapp 上的深度研究和 Cursor Bugbot 这样的封装工具。

Critically, features in the plugins/CLIs are holding Codex back. Many of our engin prefer fast mode with 1M context, use remote control/sandbox plugins to take sess from laptop to phone and back, and upload images/screenshots during a conversat

All of this is possible with the Claude Code CLI, VSCode Plugin, web app and mo app. But none of it is currently possible with the Codex CLI, VSCode Plugin, desk app, web app and mobile app.

关键在于，插件/命令行界面（CLI）中的功能正制约着 Codex。我们许多工程师更欢使用具有 1M 上下文的快速模式，使用远程控制/沙箱插件将会话从笔记本带到手再带回，并在对话中上传图像/截图。所有这些都可以通过 Claude Code CLI、VSCode 插件、网页版和移动端应用实现。但目前在 Codex 的 CLI、VSCode 插件桌面应用、网页版和移动端应用中都无法实现这些功能。

Even if GPT-5.5 is a better model, OpenAI needs to ship features at a faster pace i order to catch up with Anthropic and increase adoption.

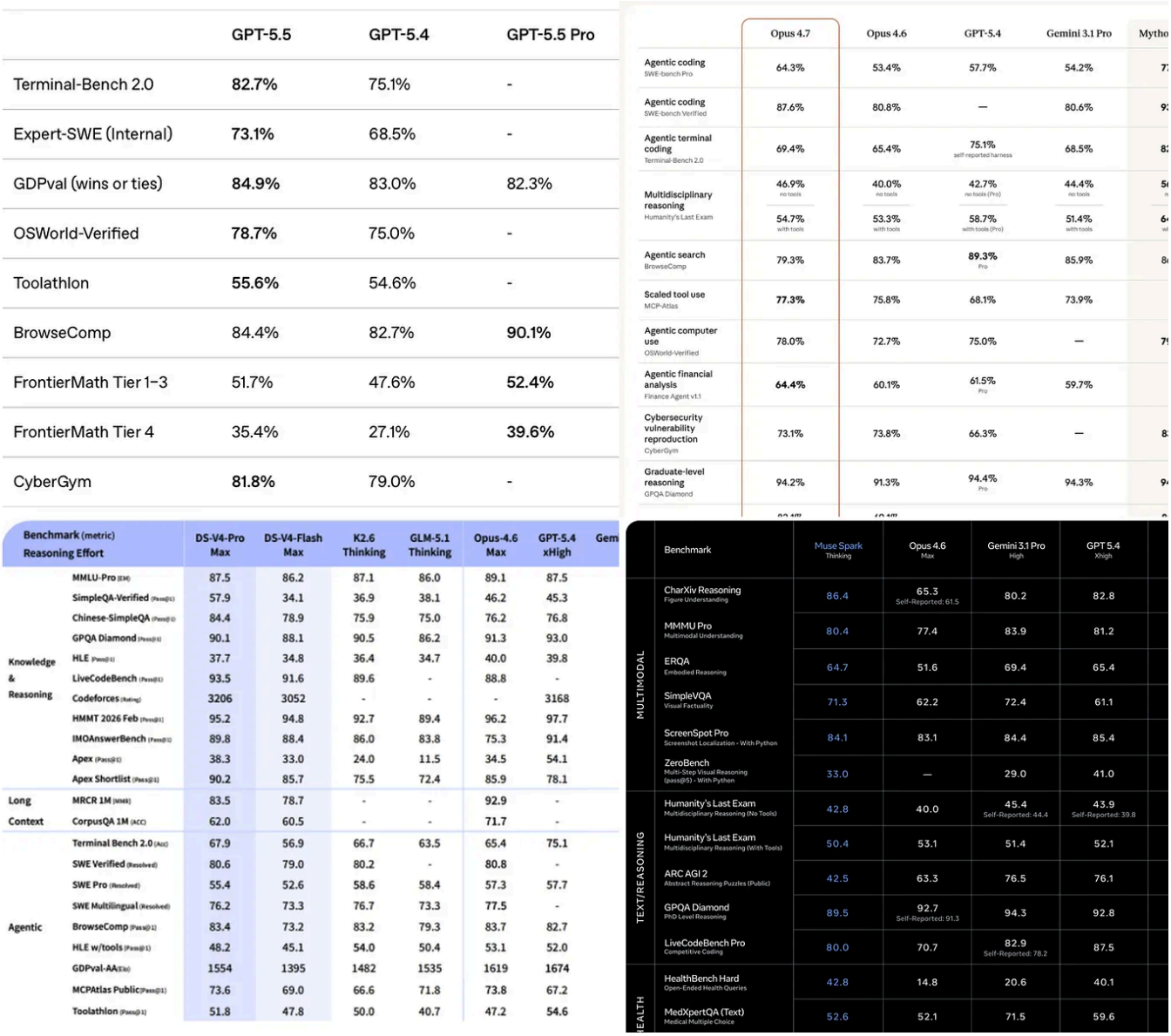
即使 GPT-5.5 是更好的模型，OpenAI 也需要更快地发布功能，以追赶 Anthropic 提高采用率。

Benchmarks are bad but we need to keep using them anyways

基准测试很糟，但我们无论如何仍需继续使用它们

The one thing that is always prominently featured in every new model announcem is a table comparing performance on various benchmarks.

每次新模型发布时，始终显著出现的一项内容是一张比较各类基准测试表现的表格



	GPT-5.5	GPT-5.4	GPT-5.5 Pro
Terminal-Bench 2.0	**82.7%**	75.1%	-
Expert-SWE (Internal)	**73.1%**	68.5%	-
GDPval (wins or ties)	**84.9%**	83.0%	82.3%
OSWorld-Verified	**78.7%**	75.0%	-
Toolathlon	**55.6%**	54.6%	-
BrowseComp	84.4%	82.7%	**90.1%**
FrontierMath Tier 1–3	51.7%	47.6%	**52.4%**
FrontierMath Tier 4	35.4%	27.1%	**39.6%**
CyberGym	**81.8%**	79.0%	-

	Opus 4.7	Opus 4.6	GPT-5.4	Gemini 3.1 Pro	Mytho
Agentic coding SWE-bench Pro	64.3%	53.4%	57.7%	54.2%	7
Agentic coding SWE-bench Verified	87.6%	80.8%	—	80.6%	9
Agentic terminal coding Terminal-Bench 2.0	69.4%	65.4%	75.1% self-reported harness	68.5%	8
Multidisciplinary reasoning Humanity's Last Exam	46.9% no tools 54.7% with tools	40.0% no tools 53.3% with tools	42.7% no tools (Pro) 58.7% with tools (Pro)	44.4% no tools 51.4% with tools	5 6
Agentic search BrowseComp	79.3%	83.7%	**89.3%** Pro	85.9%	8
Scaled tool use MCP-Atlas	**77.3%**	75.8%	68.1%	73.9%	
Agentic computer use OSWorld-Verified	78.0%	72.7%	75.0%	—	7
Agentic financial analysis Finance Agent v1.1	**64.4%**	60.1%	61.5% Pro	59.7%	
Cybersecurity vulnerability reproduction CyberGym	73.1%	73.8%	66.3%	—	8
Graduate-level reasoning GPQA Diamond	94.2%	91.3%	94.4% Pro	94.3%	9

Benchmark (metric) Reasoning Effort		DS-V4-Pro Max	DS-V4-Flash Max	K2.6 Thinking	GLM-5.1 Thinking	Opus-4.6 Max	GPT-5.4 xHigh	Gem
Knowledge & Reasoning	MMLU-Pro (EM)	87.5	86.2	87.1	86.0	89.1	87.5	
	SimpleQA-Verified (Pass@1)	57.9	34.1	36.9	38.1	46.2	45.3	
	Chinese-SimpleQA (Pass@1)	84.4	78.9	75.9	75.0	76.2	76.8	
	GPQA Diamond (Pass@1)	90.1	88.1	90.5	86.2	91.3	93.0	
	HLE (Pass@1)	37.7	34.8	36.4	34.7	40.0	39.8	
	LiveCodeBench (Pass@1)	93.5	91.6	89.6	-	88.8	-	
	Codeforces (Rating)	3206	3052	-	-	-	3168	
	HMMT 2026 Feb (Pass@1)	95.2	94.8	92.7	89.4	96.2	97.7	
	IMOAnswerBench (Pass@1)	89.8	88.4	86.0	83.8	75.3	91.4	
	Apex (Pass@1)	38.3	33.0	24.0	11.5	34.5	54.1	
	Apex Shortlist (Pass@1)	90.2	85.7	75.5	72.4	85.9	78.1	
Long Context	MRCR 1M (MMR)	83.5	78.7	-	-	92.9	-	
	CorpusQA 1M (ACC)	62.0	60.5	-	-	71.7	-	
Agentic	Terminal Bench 2.0 (Acc)	67.9	56.9	66.7	63.5	65.4	75.1	
	SWE Verified (Resolved)	80.6	79.0	80.2	-	80.8	-	
	SWE Pro (Resolved)	55.4	52.6	58.6	58.4	57.3	57.7	
	SWE Multilingual (Resolved)	76.2	73.3	76.7	73.3	77.5	-	
	BrowseComp (Pass@1)	83.4	73.2	83.2	79.3	83.7	82.7	
	HLE w/tools (Pass@1)	48.2	45.1	54.0	50.4	53.1	52.0	
	GDPval-AA (Elo)	1554	1395	1482	1535	1619	1674	
	MCPAtlas Public (Pass@1)	73.6	69.0	66.6	71.8	73.8	67.2	
	Toolathlon (Pass@1)	51.8	47.8	50.0	40.7	47.2	54.6	

	Benchmark	Muse Spark Thinking	Opus 4.6 Max	Gemini 3.1 Pro High	GPT 5.4 Xhigh
MULTIMODAL	CharXiv Reasoning Figure Understanding	86.4	65.3 Self-Reported: 61.5	80.2	82.8
	MMMU Pro Multimodal Understanding	80.4	77.4	83.9	81.2
	ERQA Embodied Reasoning	64.7	51.6	69.4	65.4
	SimpleVQA Visual Factuality	71.3	62.2	72.4	61.1
	ScreenSpot Pro Screenshot Localization - With Python	84.1	83.1	84.4	85.4
	ZeroBench Multi-Step Visual Reasoning (pass@5) - With Python	33.0	—	29.0	41.0
TEXT/REASONING	Humanity's Last Exam Multidisciplinary Reasoning (No Tools)	42.8	40.0	45.4 Self-Reported: 44.4	43.9 Self-Reported: 39.8
	Humanity's Last Exam Multidisciplinary Reasoning (With Tools)	50.4	53.1	51.4	52.1
	ARC AGI 2 Abstract Reasoning Puzzles (Public)	42.5	63.3	76.5	76.1
	GPQA Diamond PhD Level Reasoning	89.5	92.7 Self-Reported: 91.3	94.3	92.8
	LiveCodeBench Pro Competitive Coding	80.0	70.7	82.9 Self-Reported: 78.2	87.5
HEALTH	HealthBench Hard Open-Ended Health Queries	42.8	14.8	20.6	40.1
	MedXpertQA (Text) Medical Multiple Choice	52.6	52.1	71.5	59.6

Source: every release, man

来源：每次发布，man

It's very tempting to be able to point to a small set of numbers in order to prove th "objective" superiority of your new model release, but many within the AI commu have long lamented that benchmarks are no longer a useful proxy for real-world ut We tend to agree with this point of view. There's a big difference between claimin measure a model's coding/finance/reasoning abilities vs actually doing so in any

meaningful capacity.

指着一小组数字来证明你新模型发布“客观”优越性的诱惑很大，但长期以来许多 A
人士一直感叹基准测试不再是衡量现实世界效用的有用代理。我们倾向于同意这一
点。声称衡量模型的编码/金融/推理能力，与在任何有意义的程度上真正做到这一
间存在很大差别。

That being said, we expect all the labs to continue highlighting improved benchm
performance for all future model releases, and the following section will help you
separate the signal from the noise.

话虽如此，我们预计所有研究团队在今后的模型发布中仍将继续强调基准性能的提
升，下面的章节将帮助你在信号与噪声之间做出区分。

Anatomy of a benchmark 基准测试的构成

Each benchmark consists of 3 things

每个基准由三部分组成

1. **Tasks**: what the model is actually asked to do

   任务：模型实际被要求完成的内容

2. **The evaluation method**: how the model is actually scored

   评估方法：模型实际如何被评分

3. **A harness**: what tools, instructions, interface, etc the model is given to solve t
   task

   测试框架：为解决任务而向模型提供的工具、指令、界面等

Really understanding the first two is how you determine if a benchmark is any goo
not. To illustrate, we'll walk through some famous benchmarks below in rough
chronological order. This will also give you a sense of how benchmarks have chan

over time.

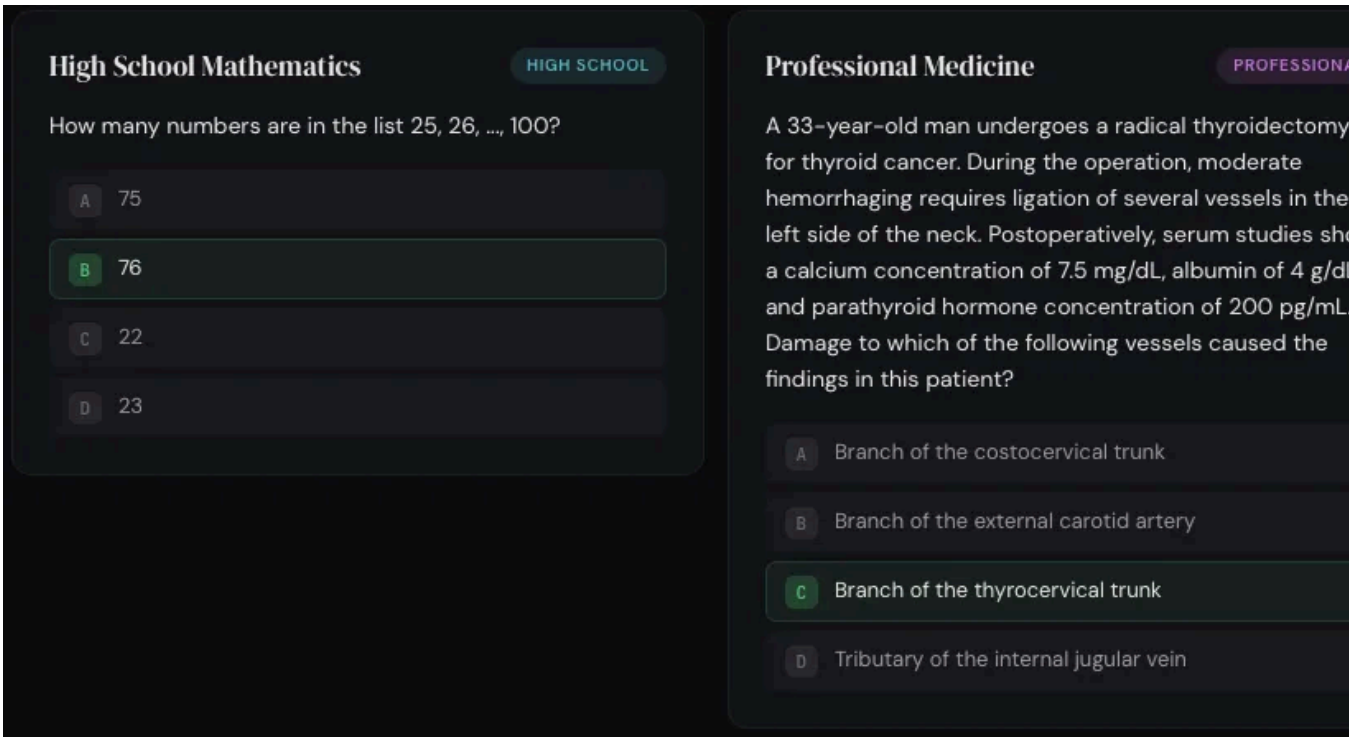
真正理解前两点是判断一个基准测试好坏的关键。为说明这一点，我们将按大致时
顺序走查下面一些著名基准测试。这也能让你了解基准测试如何随着时间变化。

MMLU and multiple choice/simple answer benchmarks

MMLU 以及多项选择/简答型基准测试

Released by academic researchers in 2020, Measuring Massive Multitask Languag Understanding (MMLU) is a set of 15,908 multiple choice questions covering 57 subjects. These questions were manually collected by university students from onl sources like standardized tests and college exams/problem sets. All of them have exactly 4 choices and are publicly available, but they range in difficulty from "elementary" to "advanced professional".

由学术研究人员于 2020 年发布的《Measuring Massive Multitask Language Understanding (MMLU)》是一套包含 57 个学科、共 15,908 道选择题的问题集。
题目由大学生从标准化考试和大学考试/习题集等在线来源手工收集。所有题目均
选一并公开可得，但难度从"基础"到"高级专业"不等。



Example MMLU questions. Source: MMLU

示例 MMLU 问题。来源：MMLU

MMLU has a minimal harness that essentially just formats the question into a pro
Tools like web search are not included. **The multiple choice format is crucial bec
it makes grading trivial–just check if the model outputted the right letter.**

MMLU 只有一个最小的测评框架，基本上只是将问题格式化为提示。不包括像网
索这样的工具。选择题格式至关重要，因为它使评分变得简单——只需检查模型是
输出了正确的字母。

MMLU was effectively solved (aka "saturated") by GPT-4 in March 2023 when it scored 86.4%. In practice, the true max score for benchmarks is usually lower than 100% because some of the tasks are ambiguous, poorly worded, or just straight up incorrect. This paper estimates that 6.49% of MMLU questions contain errors for example.

MMLU 在 2023 年 3 月被 GPT-4 实质上"解决"(即"饱和")，当时其得分为 86.4%。
实践中，基准测试的真实最高分通常低于 100%，因为一些任务存在歧义、措辞不
根本错误。例如，本文估计 6.49% 的 MMLU 问题包含错误。

Other benchmarks from the same era include

同一时期的其他基准包括

- GSM8K: Multi-step math problems created by contractors with STEM degree make evaluation simple, all answers are a single number.

  GSM8K：由具备理工科背景的承包商创建的多步数学题。为简化评估，所有
  均为单一数字。

- HellaSwag: Multiple choice questions that ask the AI to predict the most likel continuation of an everyday scenario. Tasks are sourced from video captions a

WikiHow articles.

> HellaSwag：多项选择题，要求 AI 预测日常情境中最可能的续写。任务来源于
> 频字幕和 WikiHow 文章。

- MMMU: The same thing as MMLU except the questions also include images, the model needs vision. The third M stands for “multimodal”.

> MMMU：与 MMLU 相同的测试，区别是题目还包含图像，因此模型需要具有
> 觉能力。第三个 M 代表“多模态”。

- GPQA: “Google-proof” multiple choice science questions created by 61 PhD-l contractors.

> GPQA：“抗谷歌”选择题科学问题，由 61 名博士级承包商创建。

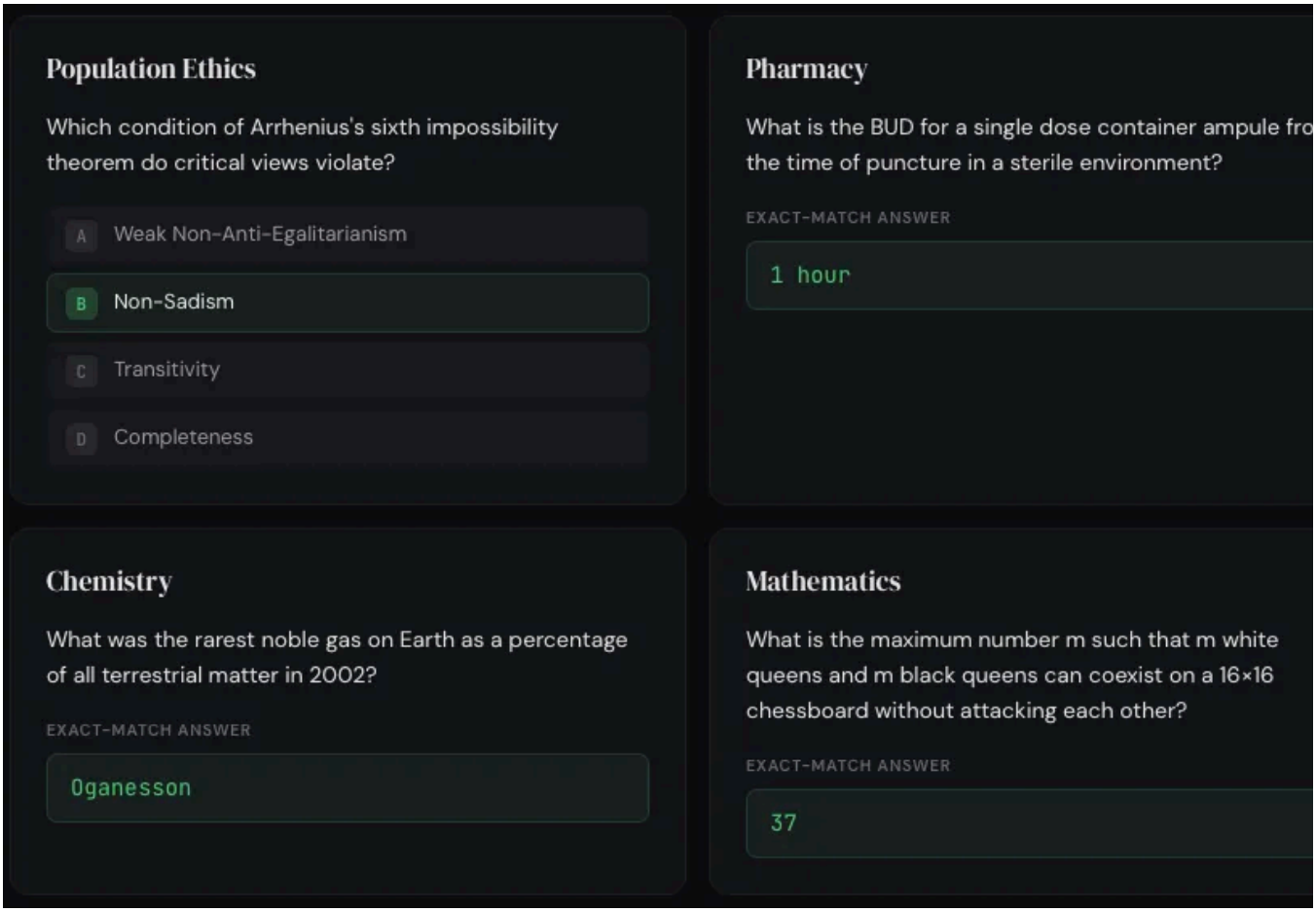
As each of these became saturated, their creators made harder versions (e.g. MML Pro, MMMU-Pro, GPQA-Diamond). Tactics include filtering out easy questions fr the previous version, using an LLM to up the choices from 4 to 10, paying your contractors to make harder questions, etc.

> 随着每一项达到饱和，其创建者就推出更难的版本（例如 MMLU-Pro、MMMU-Pro、GPQA-Diamond）。策略包括从前一版本中过滤掉简单题目，使用 LLM 将选
> 从 4 个增加到 10 个，付钱给承包商让他们出更难的题目等。

The most relevant simple answer benchmark today is Humanity’s Last Exam (HLE Released by Scale AI in January 2025, they sourced 1000+ experts from around the world to create 2500 questions on everything from algebraic geometry to classical ballet. 80% of the questions require an exact-match short answer and 20% are mult choice. For the harness, you can choose to run the model with or without tools (e.g

web search and code execution).

当今最相关的简明答案基准是 Humanity's Last Exam (HLE)。由 Scale AI 于 2025 月发布，他们从世界各地招募了 1000 多名专家，创建了 2500 道涵盖从代数几何到 典芭蕾的题目。80% 的题目要求精确匹配的简短答案，20% 为多项选择题。对于测 框架，可以选择在有工具（例如网络搜索和代码执行）或无工具的情况下运行模型



Example HLE questions. Source: Scale AI

示例 HLE 问题。来源：Scale AI

These questions obviously aren't representative of real-world LLM usage and are a riddled with issues. For example, one study found that 30% of HLE chemistry/biol questions had answers that directly conflicted with peer-reviewed literature.

这些问题显然并不代表现实世界中 LLM 的使用情况，而且也充满了问题。例如，项研究发现 30% 的 HLE 化学/生物问题的答案与经同行评审的文献直接冲突。

However, the labs still absolutely hillclimb all of these benchmarks during the RL stage of training. Google, for example, had a 9 figure budget in 2025 specifically fo HLE style STEM questions, which they paid to data vendors like Mercor, Surge, ar Handshake. It's no coincidence that Gemini 3 Pro was a step-change improvemen the benchmark.

然而，这些实验室在训练的强化学习阶段仍然绝对会对所有这些基准测试进行爬坡化。例如，Google 在 2025 年专门为 HLE 风格的 STEM 问题设置了一个九位数预算，他们将这笔费用支付给了像 Mercor、Surge 和 Handshake 这样的数据供应商Gemini 3 Pro 在该基准上的显著改进并非巧合。

Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT
Humanity's Last Exam	Academic reasoning	No tools	**37.5%**	21.6%	13.7%	26.
		With search and code execution	**45.8%**	—	—	—

Source: Google

The generous explanation for why labs care about things like HLE is that the knowledge gained from solving esoteric multiple choice questions will transfer to other use cases. The cynical explanation is that corporate VPs want to be able to p to a single number to prove that they're doing their job, and Scale was good at marketing HLE to win sufficient mind share.

关于实验室为什么关心像 HLE 这样的东西，慷慨的解释是：从解决晦涩的多项选题中获得的知识会迁移到其他用例。愤世嫉俗的解释是：企业副总裁们希望能够指一个单一数字来证明他们在尽职，而 Scale 善于营销 HLE，从而赢得了足够的关注度。

SWE-bench and coding benchmarks

SWE-bench 与 编程基准

Coding is the most important AI capability, and SWE-bench (released in 2023) was first big coding benchmark.

编程是最重要的人工智能能力，而 SWE-bench（于 2023 年发布）是第一个大型编
基准。

The tasks were automatically scraped from 12 **Python** repos, including django, scikit-learn, and seaborn. They used the following 3 step filtering process:

这些任务是从 12 个 Python 代码库自动抓取的，包括 django、scikit-learn 和 seaborn。他们使用了以下三步筛选流程：

1. Start with all ~93k merged PRs for all 12 repos

   从所有 12 个代码库的大约 93k 个已合并 PR 开始

2. Reduce to ~11k that were linked to a GitHub issue and introduced new tests

   缩减到大约 11k，那些与 GitHub issue 关联并引入了新测试的

3. Keep just the 2294 PRs where at least one of the new tests fail when applied to commit immediately before said PR

   只保留那 2294 个 PR，其中至少有一个新测试在应用到该 PR 之前的提交时失

In other words, the GitHub issue is the task, and the PR that resolved said issue is proof that the task is possible. The eval is all the old tests in the repo plus the new tests included in the PR. The AI is successful if none of the old tests break (pass-to-pass) AND all the new tests pass (fail-to-pass). Importantly, the model is not allowed see any of the new tests while attempting the task. For the harness, the model is allowed to inspect the codebase, but it can't actually run any code.

换句话说，GitHub issue 是任务，而解决该 issue 的 PR 则证明该任务是可行的。评估包括仓库中所有旧测试以及 PR 中包含的新测试。如果旧测试都未被破坏（从通过到通过）并且所有新测试都通过（从失败到通过），则 AI 成功。重要的是，模型在试任务时不允许查看任何新测试。对于测试框架，模型可以查看代码库，但不能实际运行任何代码。

It is worth emphasizing that there was **no human verification** at any step in the ta creation process. GitHub issues are often ambiguous and poorly specified. Furthermore, the tests devs include in their PRs are typically noncomprehensive a scoped to particular implementation details. This causes two big issues

值得强调的是，在任务创建过程的任何环节都没有人工验证。GitHub issue 通常含 且规范不明。此外，开发者在其 PR 中包含的测试通常不够全面，并且针对特定实 细节进行范围限定。这会导致两个重大问题

1. If the problem statement allows for multiple solutions, but your tests are scop a single correct solution, then some correct answers will get wrongly rejected

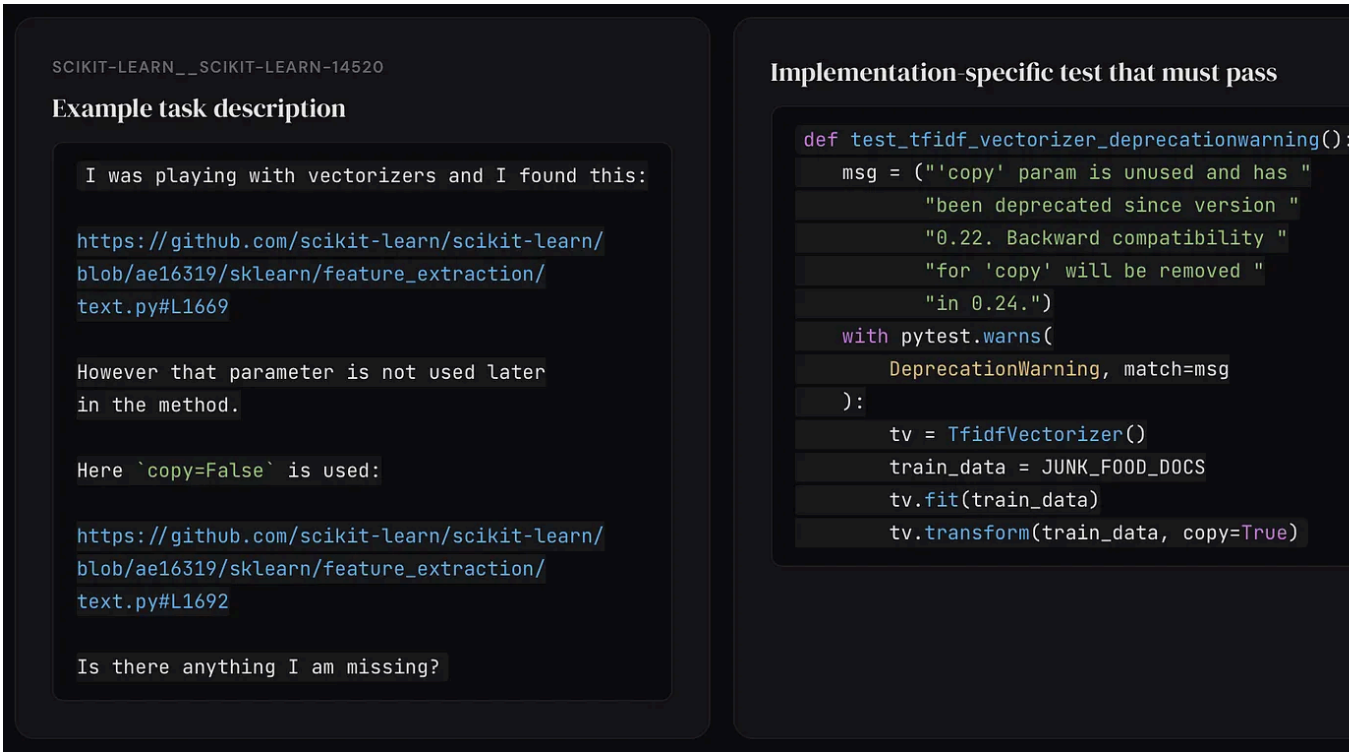
   如果问题陈述允许多种解决方案，但你的测试仅限定为单一正确解法，那么一 正确答案会被错误地拒绝

2. If your tests are noncomprehensive, then you will incorrectly pass the AI even only completes a subset of the requirements

   如果你的测试不全面，那么即使 AI 只完成了部分要求也会被错误地判定为通

In short, many of the SWE-bench tasks were straight up broken. For example, one required the AI to perfectly match a 19 word error message in the eval despite not mentioning it at all in the problem description.

简言之，许多 SWE-bench 任务根本就是有缺陷的。例如，有一道题要求 AI 在评 完美匹配一条由 19 个单词组成的错误信息，尽管问题描述中根本没有提到该信息。



Example SWE-bench task description (left) with an unfair test (right). Source: SWE-bench

示例 SWE-bench 任务描述（左）与不公平测试（右）。来源：SWE-bench

OpenAI attempted to solve these issues by releasing SWE-bench verified in Augus 2024. They hired 93 python devs to **manually review** all the task descriptions and e for ambiguity/unfairness. After filtering out all the problematic ones, the original problems were reduced to 500 "verified" tasks. OpenAI also added a bash tool to t harness so the AI could execute code–making the benchmark more agentic–and improved infra reliability by packaging each task as a Docker container.

OpenAI 试图通过在 2024 年 8 月发布 SWE-bench verified 来解决这些问题。他们佣了 93 名 Python 开发者，手工审查所有任务描述和评估以查找歧义/不公正之处。剔除所有有问题的条目后，原本的 2294 道题目被缩减为 500 道"已验证"任务。OpenAI 还向测评工具链添加了一个 bash 工具，使 AI 能够执行代码——使基准测更具代理性——并通过将每个任务打包为 Docker 容器来提高基础设施的可靠性。

In February 2026, OpenAI announced that they would no longer report results on SWE-bench verified for two reasons:

在 2026 年 2 月，OpenAI 宣布他们将不再报告 SWE-bench verified 的结果，原因两个：

1. Of the 138 problems consistently failed by o3, over half *still* had unfair evals th were scoped to specific implementation details not mentioned in the task description OR extra tests that checked for functionality not mentioned in the task description. In other words, the "verified" subset still wasn't very good

   在 o3 持续失败的 138 个问题中，超过一半仍然存在不公平的评估，这些评估于任务描述中未提及的特定实现细节，或者包含检查任务描述中未提及功能的外测试。换句话说，"已验证"子集仍然不是很好

2. Because all the PRs are part of popular open-source repos that are definitely included in every model's training data, they found evidence that GPT-5.2, Op 4.5, and Gemini 3 Flash had all memorized some of the answers (aka "contamination").

   因为所有的 PR 都属于那些流行的开源代码库，这些代码库肯定包含在每个模的训练数据中，他们发现证据表明 GPT-5.2、Opus 4.5 和 Gemini 3 Flash 都记了部分答案（即"污染"）。

Instead, they recommended model makers report SWE-bench pro results instead. SWE-bench pro is another Scale creation. The main difference (besides making th tasks harder) is that they used public repos with less permissive licenses and priva repos to avoid contamination. They also hired contractors to write evals and probl descriptions for the commits, instead of relying purely on GitHub issues and preexisting PRs. These are all good steps, but they definitely don't fully solve eithe problem identified with SWE-bench verified. As you've probably already figured o

by now, no benchmark is perfect.

> 相反，他们建议模型制造者改为报告 SWE-bench pro 的结果。SWE-bench pro 也
> Scale 的另一个作品。主要区别（除了让任务更难）在于他们使用了许可更不宽松
> 共仓库和私有仓库以避免污染。他们还雇佣了承包商为提交编写评估和问题描述，
> 不是完全依赖 GitHub issue 和已有的 PR。这些都是很好的措施，但它们确实无法
> 全解决 SWE-bench verified 所识别的那两个问题。正如你现在大概已经猜到的，
> 任何基准是完美的。

SWE-bench pro and verified are both still commonly reported in model release ca
today. Other popular coding benchmarks include

> SWE-bench pro 和 verified 在今天的模型发布说明中仍然常被提及。其他流行的编
> 基准测试还包括

- SWE-bench multilingual: Basically SWE-bench verified but with 9 languages instead of just Python

  > SWE-bench 多语言版：基本上是经过验证的 SWE-bench，但支持 9 种语言而
  > 仅仅是 Python

- Terminal-bench: Tasks and evals are both crowdsourced, anything that's doab a terminal is fair game. For example, cracking a password protected file or building a linux kernel.

  > Terminal-bench：任务和评估均由众包完成，凡是在终端可实现的都在评估范
  > 内。例如，破解受密码保护的文件或构建 Linux 内核。

- NL2Repo: Human annotators reverse engineered 104 open-source Python rep into a natural language requirements doc. The task for the AI is to recreate th repo given the doc

  > NL2Repo：人工标注员将 104 个开源 Python 仓库逆向工程为一份自然语言需
  > 文档。AI 的任务是根据该文档重建仓库

GDPval and non-coding agentic benchmarks

GDPval 及非代码类代理基准

Agentic AI extends far beyond coding today, and so too do agentic benchmarks. T most famous example is GDPval by OpenAI. Released in September 2025, it aims measure AI's ability to complete real economically valuable tasks across 44 differe jobs, from financial analysts to nurse practitioners.

Agentic AI 的应用远不止当今的编码领域，agentic 基准也是如此。最著名的例子 OpenAI 的 GDPval。该基准于 2025 年 9 月发布，旨在衡量 AI 在 44 种不同职业中 成真正具有经济价值任务的能力，涵盖从金融分析师到执业护士等岗位。

To create the tasks, OpenAI hired expert contractors from each job–e.g. an ex-Bof banker for finance tasks–and asked them to provide 3 things per task:

为了创建这些任务，OpenAI 从每个职业中雇佣了专家承包商——例如为金融任务 请了一位前美国银行（BofA）银行家——并要求他们为每个任务提供 3 项内容：

1. The problem statement, which can include reference files along with plain tex

   问题陈述，可能包含参考文件以及纯文本

2. An example solution to the problem, with deliverable formats spanning pdfs, spreadsheets, videos, etc.

   问题的示例解答，交付格式可包括 PDF、电子表格、视频等

3. A **rubric** that explains how to grade any given solution

   说明如何对任何给定解答进行评分的评分标准

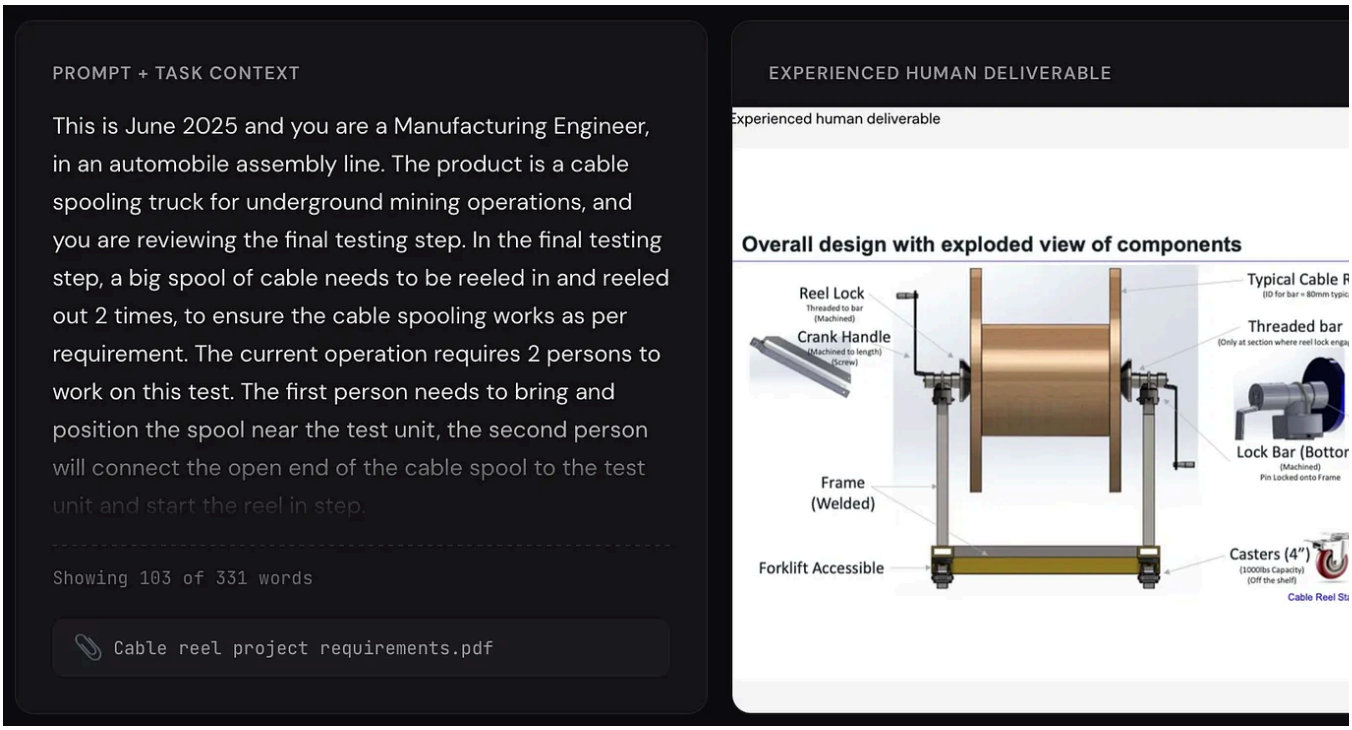
The harness is another step up from coding benchmarks. Agents are given access apps like LibreOffice (Microsoft Office clone) and CAD software, along with the standard web search and code execution tools. Although GDPval isn't quite this advanced, newer agentic benchmarks also include fake calendars, emails, Slack messages, Google Drives, etc. that the AI needs to navigate in order to successfull

complete the task.

该测试框架相比编码基准又上了一层。代理可以访问 LibreOffice（Microsoft Offi
的克隆）和 CAD 软件等应用，以及标准的网络搜索和代码执行工具。尽管 GDPv
还没有达到这种程度，但较新的主体型基准也包括 AI 需要导航的假日历、电子邮
件、Slack 消息、Google Drive 等，以便成功完成任务。

Finally, to evaluate each task, OpenAI used additional expert contractors to compa
the AI outputs to the human-provided solutions. They also created an AI grader th
uses the rubric to rank solutions, but conceded it's still not as reliable as the huma
experts. For this reason, they still use human experts for their official results, desp
being way slower and more expensive.

最后，为了评估每个任务，OpenAI 使用了额外的专家承包商，将 AI 的输出与人
供的解决方案进行比较。他们还创建了一个使用评分标准来对解决方案进行排名的
评分器，但承认它仍然不如人类专家可靠。出于这个原因，尽管这要慢得多且更昂
贵，他们仍然在官方结果中使用人类专家。



Example GDPval task. Source: OpenAI

示例 GDPval 任务。来源：OpenAI

However, “LLM-as-a-judge” for evals is a popular technique used by other agentic benchmarks for tasks that aren’t objectively verifiable. GDPval-aa, for example, is the public GDPval tasks but with an LLM judge.

然而，“将 LLM 作为评审”来进行评估是一种流行技术，其他一些以主体性代理为向的基准测试也在针对那些无法客观验证的任务使用这种方法。例如，GDPval-aa 是公开的 GDPval 任务，但用 LLM 作为评审者。

In theory, rubrics allow you to measure important qualitative traits like style, but t have obvious limitations. For example, it’s hard to guarantee quality when your ru are either written by contractors or AI generated. Using an LLM to evaluate quali also inherently suspect, especially when there’s no human in the loop making the decision.

理论上，评分标准（rubrics）允许你衡量像风格这样的重要定性特征，但它们有明的局限性。例如，当你的评分标准要么由外包人员撰写要么由 AI 生成时，很难保质量。使用 LLM 来评估质量本身也值得怀疑，尤其是在没有人工介入做最终决定情况下。

Another big limitation with GDPval is the clearly defined, unnaturally specified prompts. Real world tasks typically have an element of ambiguity that’s completely missing from this benchmark. Human jobs also involve iteration based on feedbac whereas GDPval is strictly single-turn.

GDPval 的另一个重大局限是提示被明确定义且不自然地指定。现实世界的任务通包含这种基准完全缺失的模糊因素。人类的工作也涉及基于反馈的迭代，而 GDPv则严格是单回合的。

That being said, GDPval is certainly closer to actual knowledge work than someth like HLE. Other popular agentic benchmarks include:

话虽如此，GDPval 确实比像 HLE 这种更接近实际的知识型工作。其他流行的主体（agentic）基准测试包括：

- Apex Agents: Mercor benchmark that focuses exclusively on banking, consult and law. Tasks are created by their contractors. Agent is placed in a Google Workspace environment complete with fake files, emails, etc. Uses LLM judg grading.

  Apex Agents：Mercor 的基准测试，专注于银行、咨询和法律。任务由他们的包商创建。代理被置于包含假文件、电子邮件等的 Google Workspace 环境中使用 LLM 作为评分判定者。

- Finance Agent: Tasks are created by human experts and involve analyzing rec SEC filings. Rubrics are generated by GPT-4o and then reviewed by humans. LLM judge for grading.

  Finance Agent：任务由人类专家创建，涉及分析近期的 SEC 披露文件。评分表由 GPT-4o 生成，随后由人类审阅。使用 LLM 作为评分判定者。

- BrowseComp: Tasks are hard to Google questions created by contractors. For example, “Between 1990 and 1994 inclusive, what teams played in a soccer ma with a Brazilian referee had four yellow cards, two for each team where three the total four were not issued during the first half, and four substitutions, one which was for an injury in the first 25 minutes of the match.”

  BrowseComp：任务是难以用谷歌搜索解决的问题，由承包商创建。例如，“在 1990 年到 1994 年（含）期间，哪几支球队参加了一场由巴西裁判执法的足球赛，该裁判出示了四张黄牌，每队两张，其中总共四张中有三张不是在上半场示，并且有四次换人，其中一次是在比赛前 25 分钟内因伤换下。”

- OSWorld: Computer use benchmark that tests the AI’s ability to use apps like LibreOffice, GIMP, and VLC. Tasks were manually created by 9 CS students v were listed as co-authors in exchange. Evals are custom scripts that check if tl computer is in the correct state

  OSWorld：电脑使用基准，测试 AI 使用 LibreOffice、GIMP 和 VLC 等应用程序的能力。任务由 9 名计算机科学学生手动创建，他们以共同作者的身份列出换取报酬。评测为自定义脚本，用来检查计算机是否处于正确状态。

- Tau-bench: Customer service benchmark that tests the AI's ability to do thing like cancel orders and modify flights. Environments and tasks were created by Sierra's researchers, but they used AI for things like fake data generation. Eva check the state of the application as well as the AI output for an exact string match.

  Tau-bench：客户服务基准，测试 AI 执行取消订单和修改航班等操作的能力。境和任务由 Sierra 的研究人员创建，但他们使用 AI 生成虚假数据等内容。评既检查应用程序的状态，也检查 AI 输出是否与精确字符串匹配。

Some sneaky benchmark reporting by OpenAI

OpenAI 的一些偷偷摸摸的基准测试报告

Hopefully the previous sections have convinced you that benchmarks are often wi unrepresentative of the capability they claim to be measuring. However, they also definitely aren't totally useless, and a 10%+ improvement on SWE-bench verified a everyone else thought the benchmark was already saturated (which is what Mytho did) still means something.

希望前面几节已经让你相信，基准测试往往很难代表它们声称要衡量的能力。然而它们也绝非完全无用，在大家都以为某个基准已被饱和之后（正如 Mythos 所做的样）还能在 SWE-bench 上获得超过 10%的提升并得到验证，这仍然具有一定意义

Looking at the benchmarks companies choose NOT to report can also be telling. I example, OpenAI barely included any benchmarks in their GPT-5.4 announcemen and didn't compare it to any Anthropic models. We think this is because it would' gotten brutally mogged by Opus 4.6—which came out a month earlier. This match our overall vibe of the model. Until yesterday, OpenAI's models were worse than

Anthropic's for ~all agentic tasks.

	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1 P
Terminal-Bench 2.0	**82.7%**	75.1%	-	-	69.4%	68.5%
Expert-SWE (Internal)	**73.1%**	68.5%	-	-	-	-
GDPval (wins or ties)	**84.9%**	83.0%	82.3%	82.0%	80.3%	67.3%
OSWorld-Verified	**78.7%**	75.0%	-	-	78.0%	-
Toolathlon	**55.6%**	54.6%	-	-	-	48.8%
BrowseComp	84.4%	82.7%	**90.1%**	89.3%	79.3%	85.9%
FrontierMath Tier 1–3	51.7%	47.6%	**52.4%**	50.0%	43.8%	36.9%
FrontierMath Tier 4	35.4%	27.1%	**39.6%**	38.0%	22.9%	16.7%
CyberGym	**81.8%**	79.0%	-	-	73.1%	-

Source: OpenAI 来源：OpenAI

However, there's still one benchmark that's suspiciously missing. Coding is the mo important model capability and OpenAI literally wrote a blog post in February arg for SWE-bench Pro to become the industry's new de facto benchmark. So why did use this random "Expert-SWE" benchmark instead?

然而，仍然有一个基准测试明显缺失。编程是模型最重要的能力，OpenAI 在二月专门发表了一篇博文，主张让 SWE-bench Pro 成为行业新的事实标准基准。那么为什么要改用这个随机的"Expert-SWE"基准呢?

Scrolling down all the way to the very bottom of the blog post reveals the answer:

一直往下滚动到博文最底部就能找到答案：

Coding

Eval	GPT-5.5	GPT-5.4	GPT-5.5 Pro	GPT-5.4 Pro	Claude Opus 4.7	Gemini 3.1
SWE-Bench Pro (Public) *	58.6%	57.7%	-	-	64.3%	54.2%
Terminal-Bench 2.0	82.7%	75.1%	-	-	69.4%	68.5%
Expert-SWE (Internal)	73.1%	68.5%	-	-	-	-

*Labs have noted evidence of memorization on this eval

Source: OpenAI 来源：OpenAI

GPT-5.5 got mogged by Opus 4.7 (much less Mythos which scored 77.8%). This supports our qualitative impression of the three models. GPT-5.5 is better than Op 4.7 at some coding tasks but is not decisively better across the board. Mythos is presumably a true step up compared to both of them, but Anthropic hasn't given u access yet :(

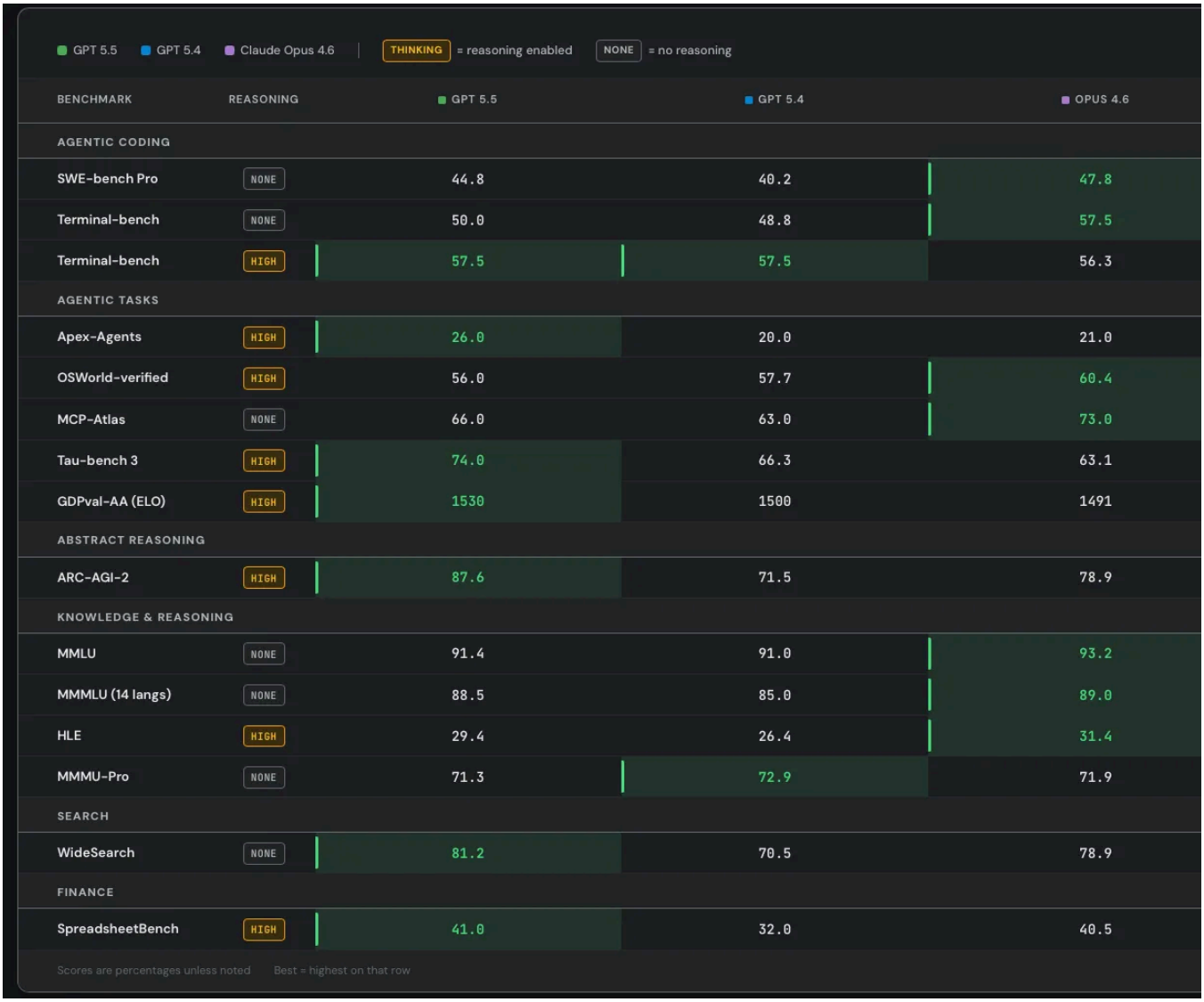
GPT-5.5 被 Opus 4.7 完全压制（Mythos 得分为 77.8%，远低于 Opus）。这支持了 们对这三款模型的定性印象。GPT-5.5 在某些编码任务上优于 Opus 4.7，但并非在 有方面都明显更强。Mythos 大概是真正的提升，相比两者都有进步，但 Anthrop 还没有给我们访问权限 :(

Why you shouldn't use the same harness for an appl to-apples comparison

为什么你不应该为"同条件"比较使用相同的测试框架

As part of our alpha-testing, we also ran a number of benchmarks on GPT-5.5 vs 5 Opus 4.6. Here are the results:

作为我们 alpha 测试的一部分，我们也对 GPT-5.5、5.4 与 Opus 4.6 进行了多项基 测试。以下是结果：



GPT 5.5 GPT 5.4 Claude Opus 4.6 | THINKING = reasoning enabled NONE = no reasoning

BENCHMARK	REASONING	GPT 5.5	GPT 5.4	OPUS 4.6
AGENTIC CODING				
SWE-bench Pro	NONE	44.8	40.2	47.8
Terminal-bench	NONE	50.0	48.8	57.5
Terminal-bench	HIGH	57.5	57.5	56.3
AGENTIC TASKS				
Apex-Agents	HIGH	26.0	20.0	21.0
OSWorld-verified	HIGH	56.0	57.7	60.4
MCP-Atlas	NONE	66.0	63.0	73.0
Tau-bench 3	HIGH	74.0	66.3	63.1
GDPval-AA (ELO)	HIGH	1530	1500	1491
ABSTRACT REASONING				
ARC-AGI-2	HIGH	87.6	71.5	78.9
KNOWLEDGE & REASONING				
MMLU	NONE	91.4	91.0	93.2
MMMLU (14 langs)	NONE	88.5	85.0	89.0
HLE	HIGH	29.4	26.4	31.4
MMMU-Pro	NONE	71.3	72.9	71.9
SEARCH				
WideSearch	NONE	81.2	70.5	78.9
FINANCE				
SpreadsheetBench	HIGH	41.0	32.0	40.5

Scores are percentages unless noted Best = highest on that row

Source: SemiAnalysis Tokenomics Team

来源：SemiAnalysis Tokenomics Team

Our numbers are generally lower than OpenAI's and Anthropic's for 2 reasons:

我们的数据普遍低于 OpenAI 和 Anthropic，有两个原因：

1. Both these labs use custom, closed-source, harnesses for their benchmark run order to increase performance

   这两家实验室在基准测试运行中均使用了定制的、闭源的框架以提升性能

2. We only ran a subset of tasks for most benchmarks to save money. In some ca these subsets weren't representative. For example, for MCP atlas, we only considered 21/36 MCP servers and ignored tasks that required things like

MongoDB, twelvedata, or alchemy.

为了节省开支，我们对大多数基准只运行了子集。在某些情况下，这些子集并
具有代表性。例如，对于 MCP atlas，我们只考虑了 21/36 个 MCP 服务器，
略了需要 MongoDB、twelvedata 或 alchemy 等的任务。

You could argue that our benchmark numbers are better than OpenAI/Anthropic because we use the same harness for a more apple-to-apples comparison. But the harness is clearly part of the product at this point. What people actually care abou how good is Codex vs Claude Code, not GPT-5.5 vs Opus 4.7.

你可以认为我们的基准数据优于 OpenAI/Anthropic，因为我们使用相同的测试框架，能做到更接近“同条件”比较。但此刻该测试框架显然已成为产品的一部分。人真正关心的是 Codex 对比 Claude Code 的表现，而不是 GPT-5.5 对比 Opus 4.7。

Returning back to the importance of token efficiency, it's worth emphasizing that **harness has a huge impact on the ultimate cost per task**. Prompt caching, input/output ratio, and tool use patterns are all largely determined by the harness. SemiAnalysis is currently collecting millions of dollars worth of agentic AI traces order to better understand how different harnesses (e.g. Claude Code vs Codex vs Cursor vs OpenCode) change cost per task. Preliminary analysis shows that Codex likely more token efficient than Claude Code, with an average input/output ratio o 80:1 vs 100:1. Yes, a higher input/output ratio means a lower price per Mtok, but C still ends up being cheaper because it consumes less input tokens overall. Those interested in the full results should subscribe to the Tokenomics model.

回到令牌效率的重要性上，值得强调的是，harness 对每项任务的最终成本有巨大响。提示缓存、输入/输出比率以及工具使用模式在很大程度上由 harness 决定。SemiAnalysis 目前正在收集价值数百万美元的代理式 AI 痕迹，以便更好地理解不 harness（例如 Claude Code vs Codex vs Cursor vs OpenCode）如何改变每项任务成本。初步分析显示，Codex 可能比 Claude Code 在令牌效率上更优，平均输入/比率为 80:1 对 100:1。是的，更高的输入/输出比率意味着每 Mtok 的价格更低，但 Codex 最终仍更便宜，因为它整体消耗的输入令牌更少。对完整结果感兴趣者应订 Tokenomics 模型。

Who Wins in the Agentic Coding Wars?

谁将在代理式编码大战中胜出?

So what does this all mean for the future of coding agents? Behind the paywall, we give our predictions for this soon to be multi-trillion dollar industry.

那么这一切对编码代理的未来意味着什么？在付费墙后面，我们将给出对这个即将为数万亿美元规模产业的预测。

It's currently a two horse race between OpenAI and Anthropic. DeepSeek did put an incredible alpha drop in their newest v4 paper, but the gap between closed and open-source is starting to widen again. We think unmentioned in this race for *cod agents* is SpaceXAI, who after the recent Cursor pseudo-acquisition, has a shot at But it will take a genuine Elon miracle to displace Anthropic or OpenAI. Google h the resources to shock the world if they get their act together on RL, and Meta's M Spark debut puts them in the race but they're both still clearly behind.

目前这场竞赛基本上是 OpenAI 和 Anthropic 的二马之争。DeepSeek 在其最新的论文中确实带来了令人惊叹的 Alpha 发布，但封闭源与开源之间的差距又开始拉大我们认为在这场关于编程代理的竞赛中未被提及的是 SpaceXAI，在最近对 Curso伪收购之后，他们有机会争夺第三名。但要取代 Anthropic 或 OpenAI，真的需要隆式的奇迹。Google 如果在强化学习方面振作起来，拥有震惊世界的资源，而 M的 Muse Spark 首秀也将其拉入竞争，但两者显然仍然落后。

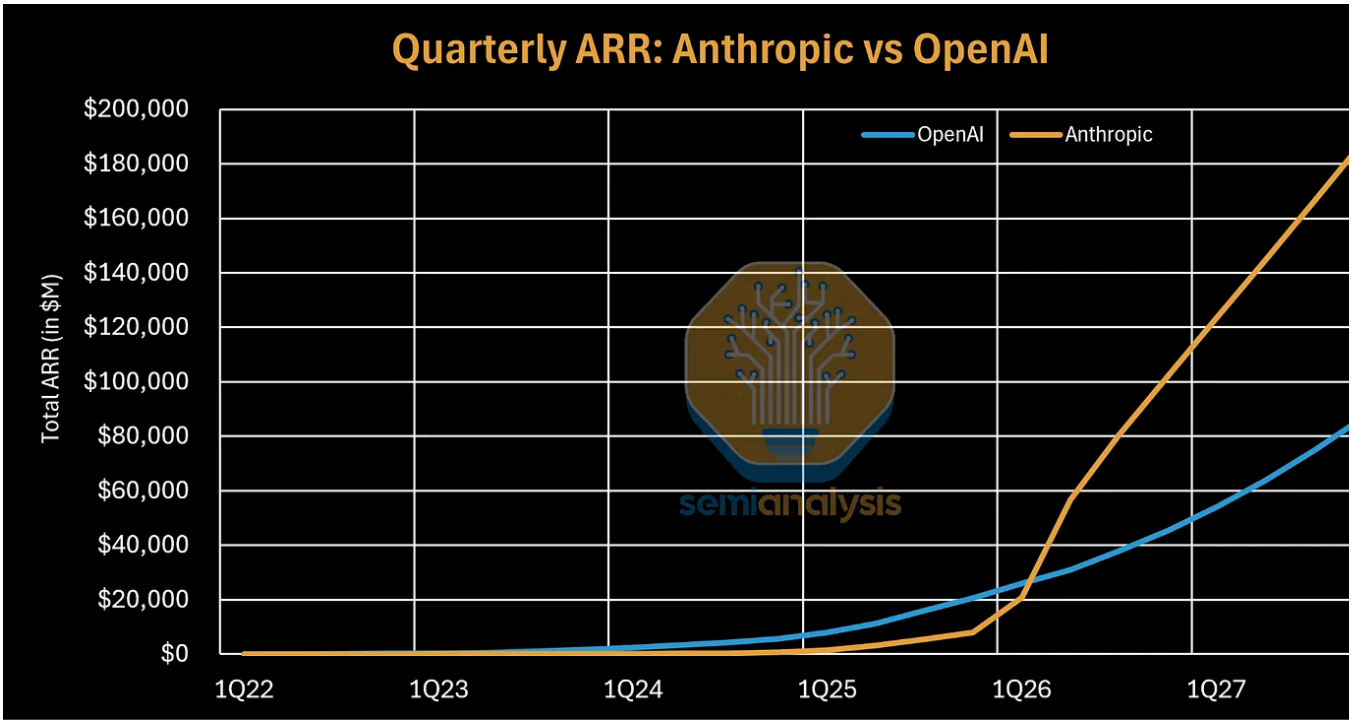
To date, the only companies with any meaningful market share are Anthropic and OpenAI. Vibe coding startups like Windsurf/Cognition, Replit, Vercel V0, Lovable Kio, Base44, and others are all falling behind, even while growing revenue 3x or m in months. These companies often have negative gross margins, and their status a wrappers seems clearer than ever after the rise of Claude Code and Codex. The rea complete product is a harness and a model. Without one of the two, you're missing and the harder product is the model. The ARR of all of these companies added together is still in the low-to-mid single digit billions. Whereas we believe the majority of Anthropic's explosion from $9B to $40B (our current estimate) is

attributable to agentic coding in Claude Code.

到目前为止，唯一拥有任何有意义市场份额的公司是 Anthropic 和 OpenAI。像 Windsurf/Cognition、Replit、Vercel V0、Lovable、Kio、Base44 等充满氛围感的程初创公司即便在数月内营收增长 3 倍或以上，也都落后了。这些公司往往毛利为负，而且在 Claude Code 和 Codex 崛起之后，它们作为包装层的地位比以往任何都更明显。真正完整的产品是一个束缚层（harness）和一个模型。缺少其中任何一个，你都会失去关键部分，而更难做的产品是模型。这些公司的年经常性收入（A加在一起仍然只是在低到中个位数十亿美元范围。我们认为 Anthropic 的估值从 9美元飙升到 400 亿美元（我们目前的估计）的大部分，归因于 Claude Code 中的智体式编码。

Things were looking bleak for OpenAI the past few months to put it lightly. Accounting discrepancies related to how the two companies recognize rev share f cloud providers aside, we believe Anthropic has truly passed them in ARR on a lik like basis. This outcome would've been inconceivable to basically everyone at the of the year, and we think most people still haven't fully internalized it. There's a ne leader in the world's most important race, and their name is **Anthropic**.

过去几个月，OpenAI 的情况可以说相当黯淡。撇开与两家公司如何从云供应商确收入分成相关的会计差异不谈，我们认为 Anthropic 在经可比口径的 ARR 上确实经超过了他们。年初时几乎没人能想象到这样的结果，我们认为大多数人至今仍未全接受这一现实。世界上最重要竞赛出现了新的领跑者，他们的名字是 Anthropic



Source:SemiAnalysis Tokenomics Model

The lead is defensible because Anthropic's revenue is of higher quality. ChatGPT' free tier is generally known to be more generous and less restrictive than Claude's which is why OpenAI dominates on the consumer side. Anthropic's revenue mix i inverse with ~70% of revenue coming from API usage which scales with deployme rather than sign ups.

领先地位是可以防守的，因为 Anthropic 的收入质量更高。众所周知，ChatGPT 费层通常比 Claude 更慷慨、限制更少，这就是 OpenAI 在消费者端占主导地位的 因。Anthropic 的收入构成则相反，约 70% 的收入来自随部署而非注册增长的 AP 用。

That being said, OpenAI's "code red moment" might be ending soon. OpenAI wil hungry to reclaim market share. Codex 5.5 is clearly in the same class as Opus 4.7, industry wide compute constraints will be OpenAI's saving grace to retain market share. Anthropic is aggressively buying up as much compute as possible, which is of the main reasons why H100 rental prices continue to soar. But there's only so m

capacity the world can physically bring online this year.

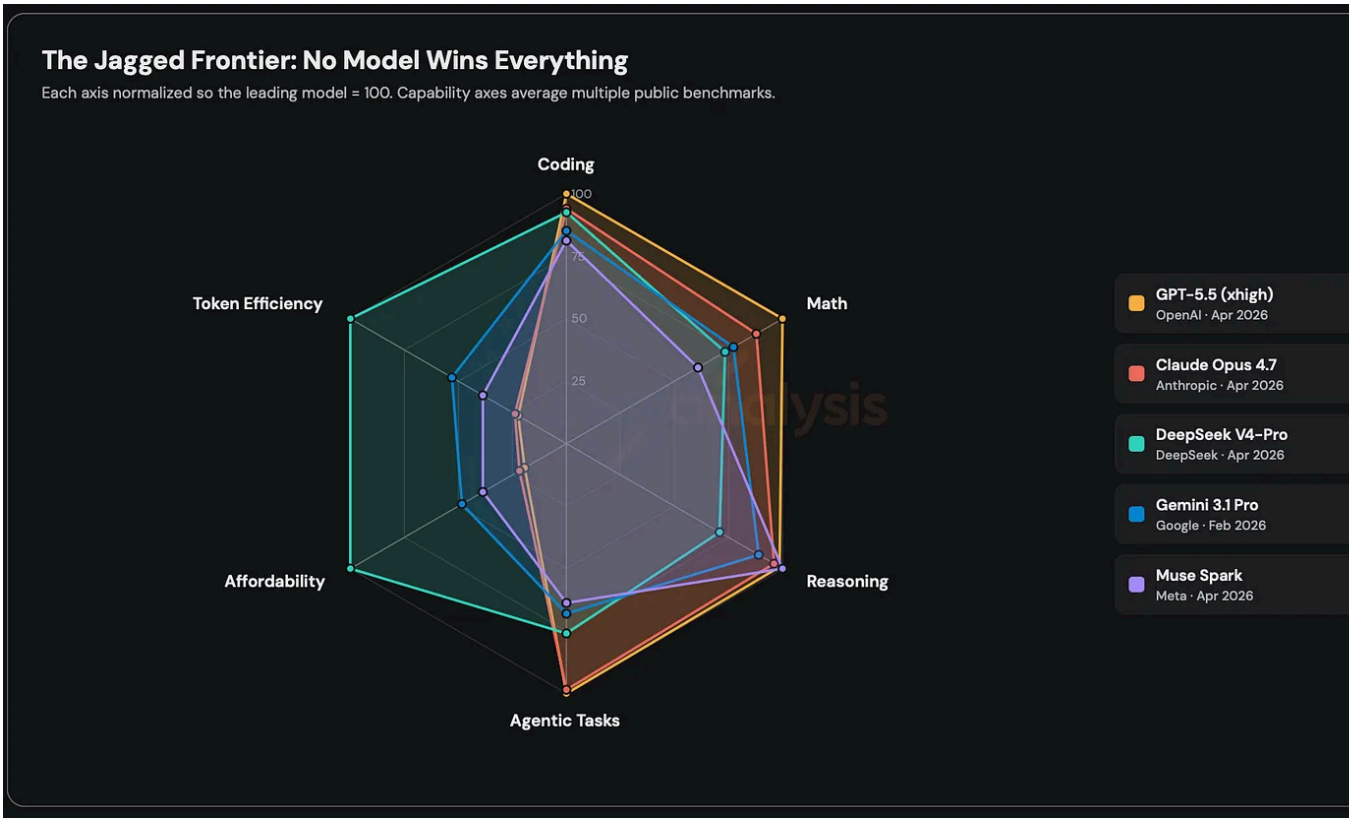
话虽如此，OpenAI 的“红色代码时刻”可能很快就会结束。OpenAI 将迫切希望收
场份额。Codex 5.5 显然与 Opus 4.7 同属一个档次，行业范围内的算力限制将成为
OpenAI 保持市场份额的救命稻草。Anthropic 正在积极抢购尽可能多的算力，这
是 H100 租赁价格持续飙升的主要原因之一。但今年全球实体能投入的算力容量终
上限。

We believe Anthropic is too AGI-pilled to ever dedicate more than 50% of their compute fleet to serving customers, and they're already starting to intentionally alienate large swathes of the market with price discrepancy. If we are to use Cour competition as a framework, Anthropic is becoming Evian water. Premium capaci like Opus 4.6 fast, lower rate limits during peak hours, testing removing Claude co from the $20/month subscription, and banning third party harnesses like OpenCla are all tactics they've tried so far to shift XPU capacity to higher margin workload We expect them to be even more aggressive in the future.

我们认为 Anthropic 对 AGI 的信仰过于坚定，不会将其算力舰队超过 50% 投入到
客户提供服务上，而且他们已经开始通过价格差异故意疏远市场上的大部分客户。
果我们以 Cournot 竞争作为框架，Anthropic 正在变成依云水。像 Opus 4.6 fast 这
的高端容量、在高峰时段降低速率限制、测试将 Claude 代码从每月 20 美元订阅中
除，以及禁止像 OpenClaw 这样的第三方封装工具，都是他们迄今为止为了将 XP
容量转向更高利润工作负载而尝试过的策略。我们预计他们将来会更加激进。

The wave of customers that get priced out of using Claude or its uptime, is OpenA opportunity. That isn't to say GPT-5.5 is a bad model, In fact it outperforms Opus many dimensions on the jagged frontier.

被定价排除在使用 Claude 或其正常运行之外的一大波客户，正是 OpenAI 的机会
这并不是说 GPT-5.5 是一款糟糕的模型，事实上在锯齿状前沿的许多维度上它都优
Opus。



Source:SemiAnalysis Tokenomics Model, Model release cards

来源：SemiAnalysis Tokenomics Model，Model release cards

Of course, there's the whole host of tasks where GPT-5.5 actually outperforms Op 4.7 as well, but we'll have to wait and see if that still holds post-Mythos.

当然，在许多任务上 GPT-5.5 实际上也优于 Opus 4.7，但我们得等到 Mythos 之后能确定这种优势是否依然存在。

---

1 * We display as FP8, but note that MoE expert parameters use FP4, while most other parameters use FP8 as per below.

* 我们以 FP8 显示，但请注意 MoE 专家参数使用 FP4，而大多数其他参数如下面所示使 FP8。

Recommend SemiAnalysis to your readers

向您的读者推荐 SemiAnalysis

Bridging the gap between the world's most important industry, semiconductors, and business.

弥合全球最重要产业——半导体——与商业之间的鸿沟。

Recommend 推荐

63 Likes 63 个赞 · 8 Restacks 8 次重置

Previous 上一个

Discussion about this post

关于这篇帖子的讨论

Comments 评论 Restacks 重新堆栈

Write a comment...
