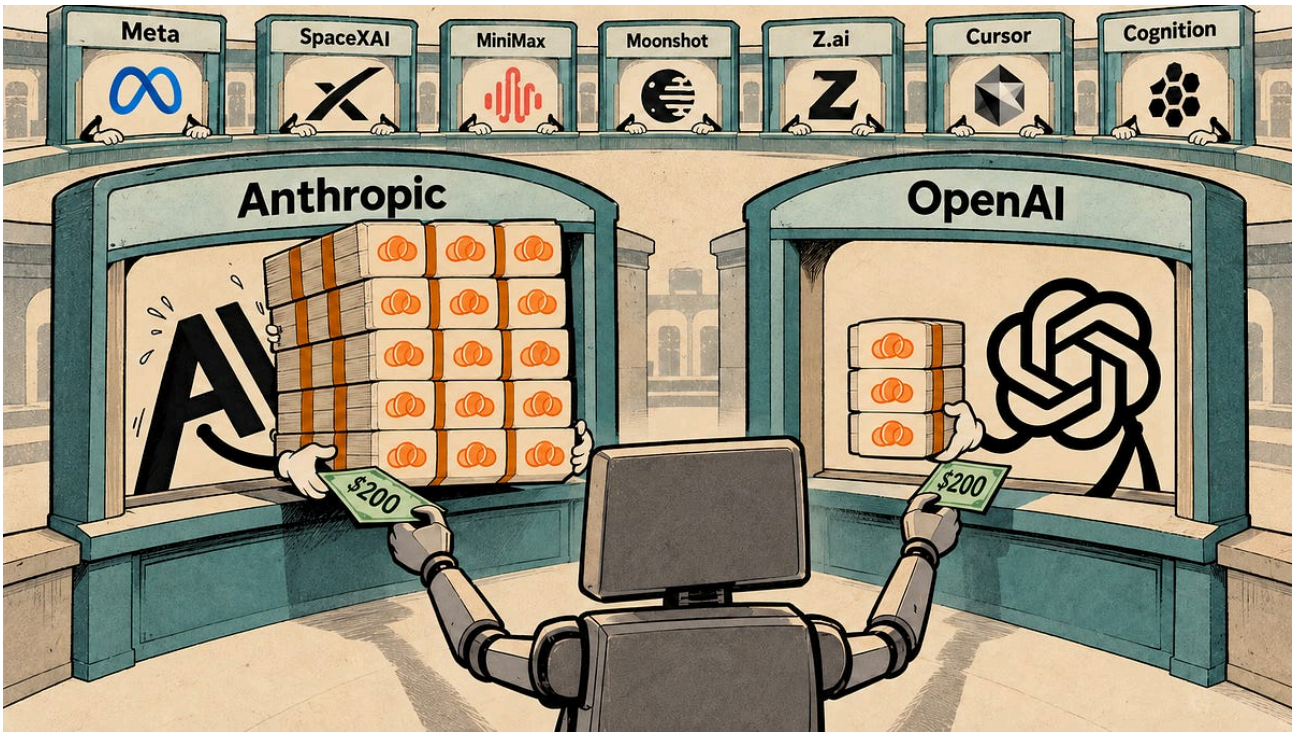
Anthropic 订阅提供的价值是 OpenAI 的 5 倍以上 --- Anthropic Subscriptions Offer 5x+ More Value Than OpenAI

newsletter.semianalysis.com/p/anthropic-subscriptions-offer-5x

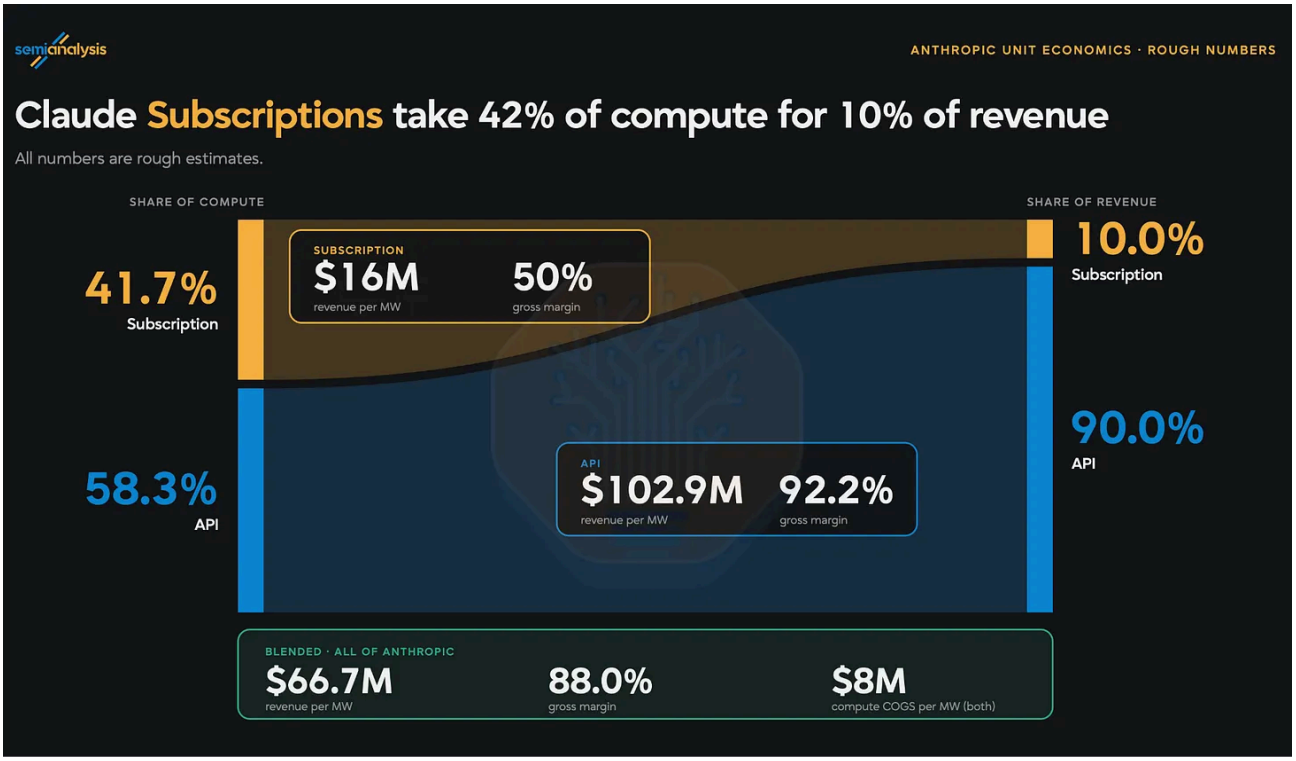
Andrew Megalaa, Max Kan, Dylan Patel

October 5, 2026



Subscription plans are still the primary way consumers and small businesses pay for AI. These plans are highly subsidized—as we previously explained in June—but can still make economic sense as powerful customer acquisition and marketing tools. For example, the goodwill engendered by OpenAI's generous resets is partially responsible for the recent surge in Codex adoption and has forced Anthropic to repeatedly walk back planned subscription nerfs to avoid getting clobbered in the court of public opinion.
订阅计划仍然是消费者和小企业为 AI 付费的主要方式。正如我们在 6 月所解释的，这些计划受到高度补贴，但作为强大的客户获取和营销工具，仍然具有经济意义。例如，OpenAI 慷慨的重置政策所赢得的善意，部分促成了近期 Codex 采用量的激增，并迫使 Anthropic 多次收回计划中的订阅削减措施，以避免在舆论法庭上遭到重创。

Furthermore, because subscription plans are so heavily subsidized, they can also have a large impact on margins and revenue per MW despite only being a small portion of total revenue. Consider the following rough numbers for Anthropic.
此外，由于订阅计划受到如此高额的补贴，尽管它们只占总收入的一小部分，却仍可能对利润率和每兆瓦收入产生重大影响。看看以下 Anthropic 的大致数据。



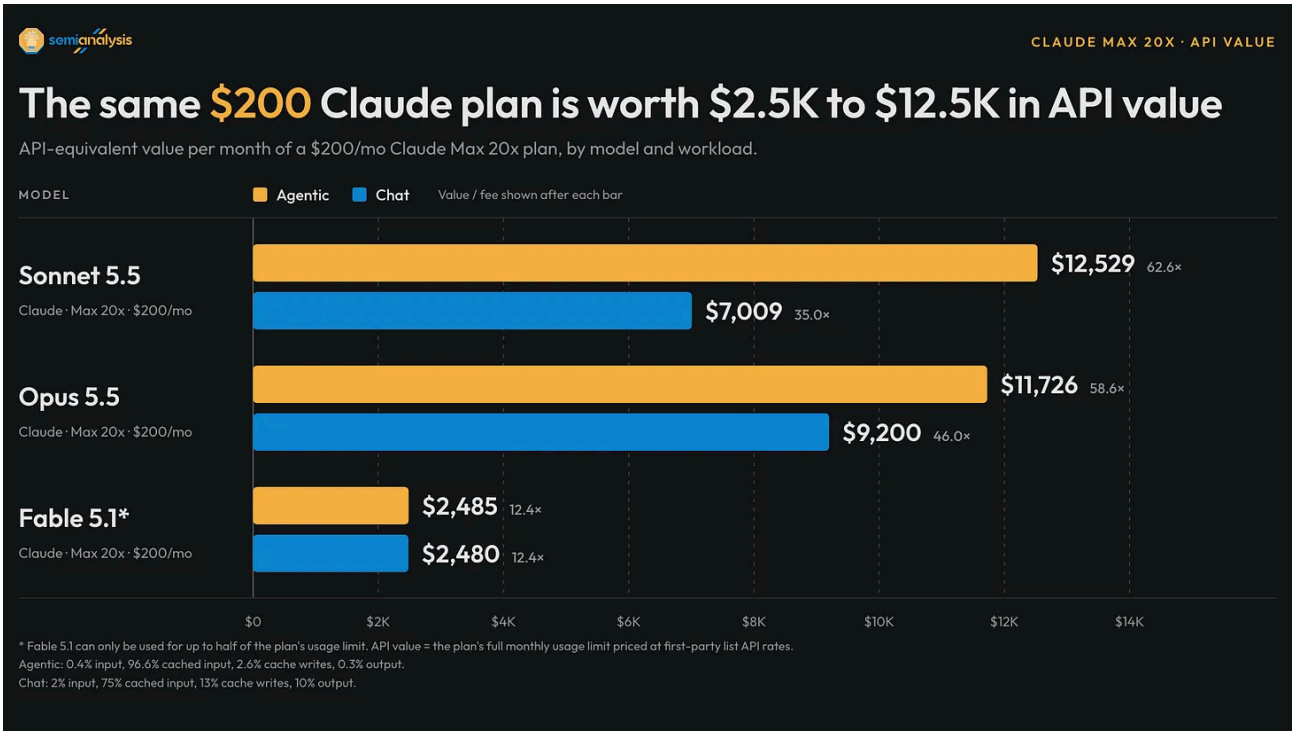
Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

Despite being just 10% of overall revenue, subscriptions can take up over 40% of inference compute and lower blended revenue per MW by ~$36M. Subscriptions are even more important for OpenAI, as they make up a larger portion of their total revenue. For specific numbers, see our Tokenomics Model.
尽管订阅仅占整体收入的 10%，却可能占用超过 40%的推理算力，并将混合每兆瓦收入拉低约~$36M。订阅对 OpenAI 更为重要，因为它们在 OpenAI 总收入中占比更大。具体数据请参见我们的 Tokenomics Model。

In other words, **if you want to accurately model AI lab financials, you need to understand their subscription limits.**
换句话说，如果你想准确建模 AI 实验室的财务状况，就需要了解它们的订阅限制。

So how do limits actually work? The way to think about subscriptions is that your monthly payment grants you some numbers of “credits”. Each (model, token type) combo consumes a different amount of credits. **Because credit cost ratios can differ dramatically from API price ratios, the “value” of the same plan changes depending on what model and workload you’re running.** Put differently, **it doesn’t make sense to say a plan is worth $X in isolation—you need to consider the full (plan, model, workload) tuple**.
那么，限额实际上是如何运作的？理解订阅的方式是：你每月的付款会授予你一定数量的“积分”。每种（模型、token 类型）组合消耗的积分数不同。由于积分成本比可能与 API 价格比存在巨大差异，同一套餐的“价值”会因你运行的模型和工作负载而变化。换句话说，孤立地说某个套餐值 X 美元是没有意义的——你需要考虑完整的（套餐、模型、工作负载）三元组。

Here’s how the API-equivalent value of the same $200/month Claude plan changes depending on the model and workload.
以下是同一个每月 200 美元的 Claude 套餐在不同模型和工作负载下，其 API 等效价值如何变化。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

A single static snapshot isn’t enough either. Labs publicly change their limits all the time with promos and new model releases. They can also silently change limits whenever they want by tweaking credit costs. Ideally, you’d re-check the cost of each (plan, model, token type) daily so you can surface any changes in real-time.
单一的静态快照也不够。实验室会通过促销和新模型发布不断公开调整限额。他们还可以随时通过调整积分成本来悄悄改变限额。理想情况下，你应当每天重新检查每个（套餐、模型、token 类型）的成本，以便实时发现任何变化。

This is exactly what SemiAnalysis has done with our new Subscriptions Dashboard available exclusively to Tokenomics Model subscribers. Besides every OpenAI and Anthropic subscription, we also track Meta, SpaceXAI, Cursor, Cognition, Z.ai, MiniMax, and Moonshot.
这正是 SemiAnalysis 通过我们全新的 Subscriptions Dashboard 所做的，该仪表盘仅向 Tokenomics Model 订阅者开放。除了所有 OpenAI 和 Anthropic 订阅外，我们还追踪 Meta、SpaceXAI、Cursor、Cognition、Z.ai、MiniMax 和 Moonshot。



A screenshot of our dashboard showing a small subset of the available data. Source: SemiAnalysis Tokenomics Model

我们仪表盘的截图，展示了可用数据的一小部分。来源：SemiAnalysis Tokenomics Model

New providers, plans, and models will be added as soon as they're released. The rest of the article will give an overview of everyone's current limits. All token and dollar amounts below assume agentic usage unless otherwise specified.

新的提供商、套餐和模型一经发布就会添加。本文的其余部分将概述各方当前的限额。除非另有说明，下文所有 token 和美元金额均假设为智能体式使用。

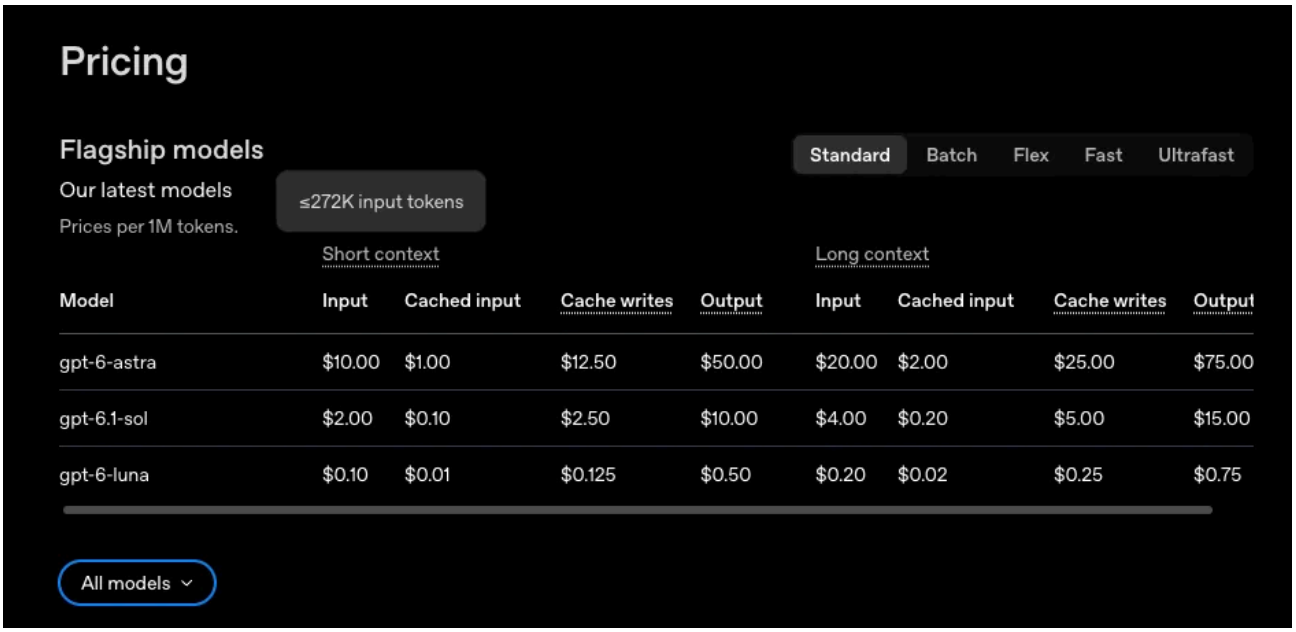
Methodology 方法论

Tokens are generally priced per MTok (million tokens) across the following types:

Token 通常按 MTok（百万 token）计价，涵盖以下类型：

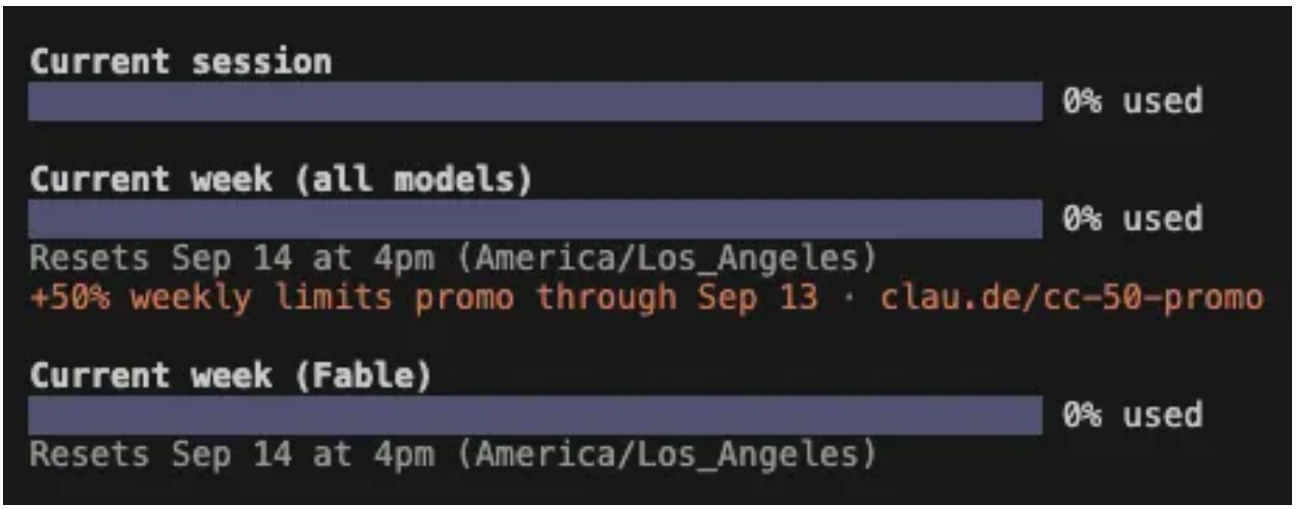
- Input: Fresh tokens added to the LLM's context window that aren't cached
  输入：添加到 LLM 上下文窗口中的全新 token，未经过缓存
- Cache write: Input tokens that are cached for multi-turn conversations. Generally slightly more expensive than regular input tokens.
  缓存写入：为多轮对话缓存的输入 token。通常比普通输入 token 略贵。
- Cache read: Tokens from previous turns that are already cached in the conversation. Generally extremely cheap compared to uncached tokens.
  缓存读取：来自之前轮次、已缓存在对话中的 token。通常比未缓存的 token 便宜得多。
- Output: Tokens generated by the model. The most expensive token type.
  输出：模型生成的 token。最昂贵的 token 类型。

Additionally, some models charge more for tokens above a certain context window. Those models typically compact before that expensive context window is reached, so we also keep our measurements below it.
此外，某些模型对超出特定上下文窗口的 token 收费更高。这些模型通常会在达到该昂贵上下文窗口之前进行压缩，因此我们也将测量值保持在其以下。



Source: OpenAI

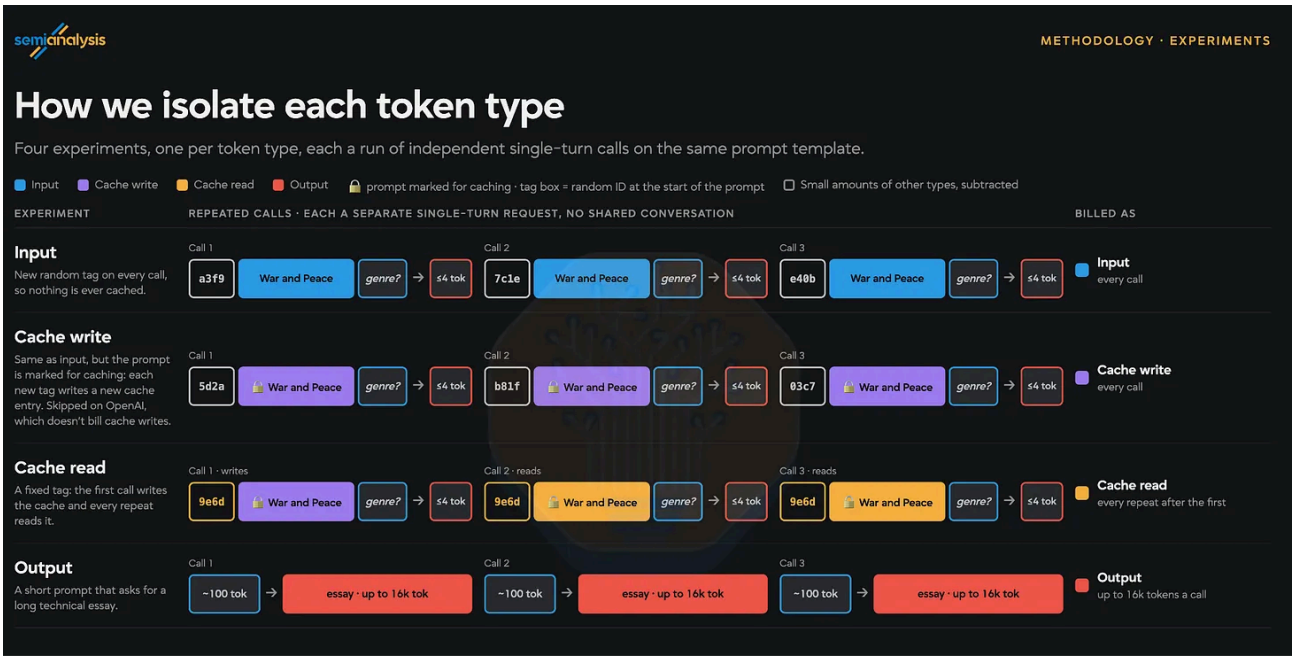
Subscription plans don't expose this fine grained pricing. Instead, they provide a simple 0 to 100% usage meter across 5-hour and 7-day windows, and sometimes a separate meter for a model like Fable. Subscription tiers are then differentiated by usage multipliers relative only to the provider's other plans. OpenAI for example used to advertise "Expanded Codex usage" in their $20 Plus plan, "5x more usage than Plus" in the $100 Pro plan, and "20x more usage" in the $200 Pro plan until they cut their $200 plan usage in half and removed all relative usage from their pricing page.



Screenshot of Claude Code meters. Source: Claude Code CLI

Setup

We compute the subscription-rate of each (plan, model, token type) triple by running experiments that isolate one token type at a time and watching how far the model provider's meter moves. An experiment is some number of repeated calls using a specific prompt that maximizes one token type while minimizing all the others.



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

Input, cache writes, and cache reads share the same prompt template. We use a portion of *War and Peace* since some models will refuse to respond if given large blocks of gibberish. For the input token experiments, we use a random tag on every call to ensure nothing gets cached. The cache write experiments run the same way but with the prompt marked for caching, so each new tag forces a new cache entry. The cache read experiments use a fixed tag, so the first call writes the cache and every repeat reads it.

For the output token experiments, we used a technical essay to force long outputs because models will refuse mechanical prompts like "repeat SemiAnalysis 100,000 times".

Computing the Results

For every call we record two things: how many tokens of each type the provider billed, and what the usage meter read. Turning those into a price takes three steps.

1. Measure the rate of each token type

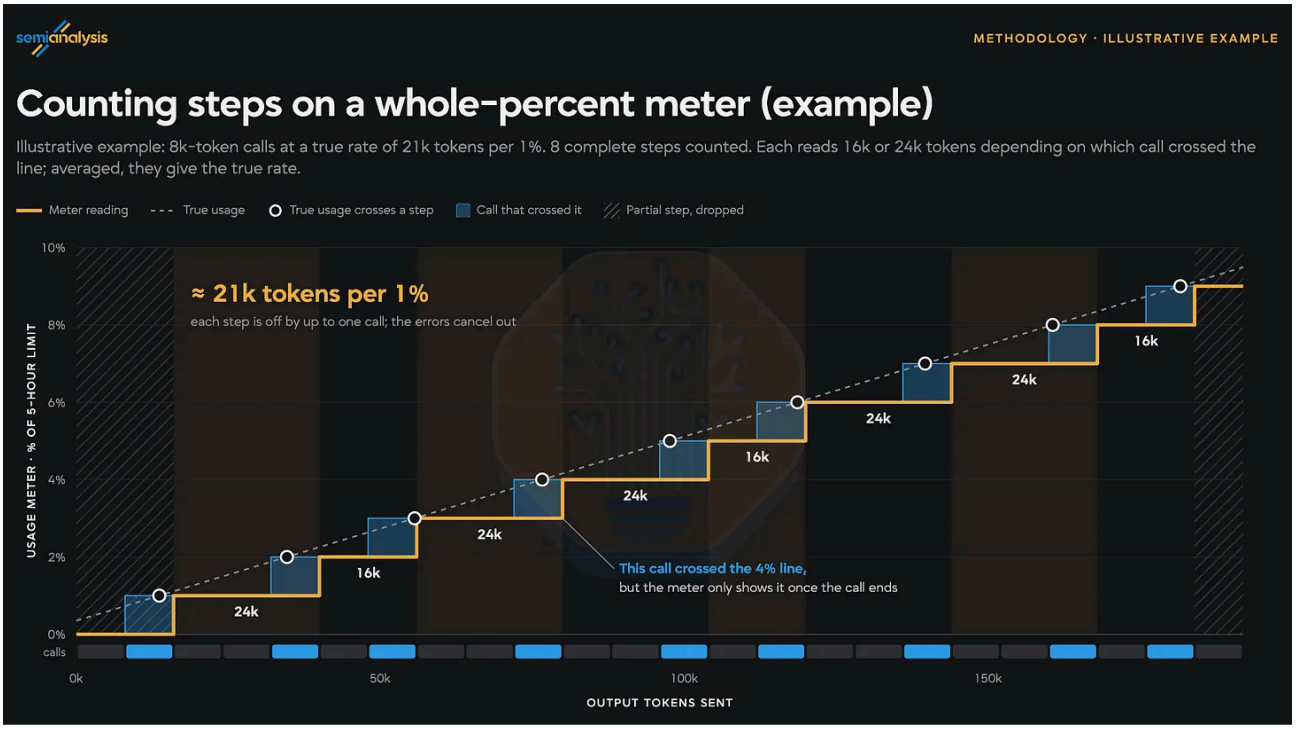
Providers generally report usage on a meter that moves in fixed amounts. This whole percentage point, a credit, a cent etc... A single request often does not r meter at all, so the cost of one request cannot be measured directly.



Instead, we measure in steps. As requests run, we keep a running total of the tokens used. Each time the meter rises, we store that total. The tokens used between two meter moves make up one step, which is the cost of moving the meter by one unit.

We drop two partial steps. When a run starts, the meter is already part of the way to its next move, so the tokens before the first move are less than a full step and would make the rate look higher than it is. The tokens after the last move never finish a step, so we drop those as well.

To get a rate, we add up how far the meter moved over the complete steps and divide by the tokens used in them. Providers charge differently for input, output, and cached tokens, so we calculate a separate rate for each.



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

Caveats: 注意事项：

- Because we see the meter only once per request, a single step can be off by up to one request. These errors don’t add up with more consecutive steps, so we track a range instead of a single number, and the range gets narrower the more steps we count. We keep adding steps until the range is within ±5%, then report the rate across all of them.
  由于我们每次请求只能看到一次计量器读数，单个步骤的误差最多可达一次请求。这些误差不会随着连续步骤增多而累积，因此我们跟踪的是一个范围，而不是单个数值，而且我们统计的步骤越多，这个范围就越窄。我们会持续增加步骤，直到范围在 ±5% 以内，然后报告所有这些步骤的速率。

- No request contains only one type of token. Some plans, for example, require a set of instructions on every request. We subtract that extra amount using the prices we measured for the other types.
  没有任何请求只包含一种类型的 token。例如，有些套餐要求每次请求都附带一组指令。我们会使用为其他类型测得的价格减去这部分额外用量。

- Reading from the cache costs almost nothing on some plans and may not move the meter at all. If the meter does not move after 500M cache read tokens, we assume that cache reads are free.
  在某些套餐中，从缓存读取几乎不产生费用，甚至可能完全不会让计量器移动。如果在 500M 缓存读取 token 后计量器仍未移动，我们就假定缓存读取是免费的。

2. Convert rates into tokens per window and per month
2. 将速率转换为每个窗口和每月的 token 数

We read every meter a plan shows after every request. This could be a 5-hour limit, a weekly limit, and for models like Fable, its own weekly limit. So each token type gets a price on each meter. This is how much of that limit a million tokens uses. From each price we work out how many tokens fit e.g. if a million tokens use 5% of a limit, the whole limit holds 20 million.
我们会读取每次请求后套餐显示的各项额度。这可能是 5 小时限额、每周限额，以及像 Fable 这样的模型还有自己的每周限额。因此，每种 token 类型在每个额度上都有一个价格。这表示 100 万 token 会消耗该限额的多少。根据每个价格，我们算出能容纳多少 token，例如，如果 100 万 token 消耗了某个限额的 5%，那么整个限额可容纳 2000 万 token。



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

The 5-hour limit resets many times within a week, so a month is capped by the weekly limits.

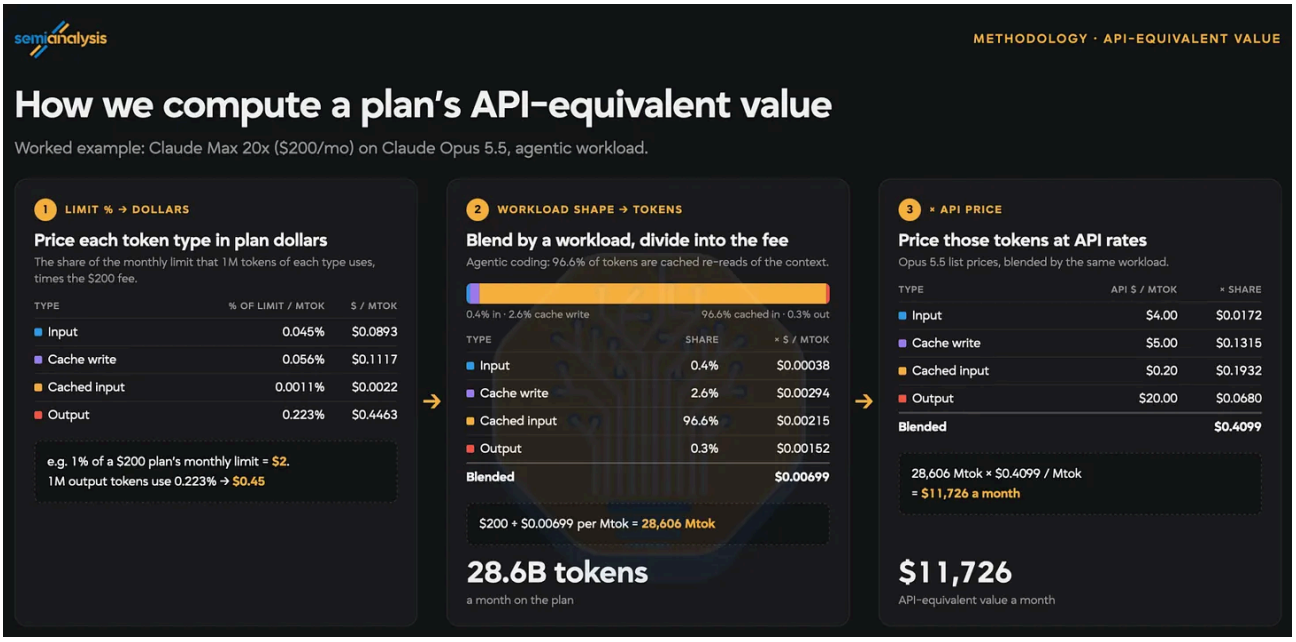
5 小时限额在一周内会重置多次，因此一个月的用量受每周限额限制。

3. Pricing 3. 定价

Lastly, we compute the API-equivalent value by converting the limit % consumed by a million tokens of each type into a dollar amount. For example, if you consume 1% of a $200 plan's monthly limit, that's $2.
最后，我们通过将每种类型 100 万 token 所消耗的限额百分比转换为美元金额，来计算 API 等效价值。例如，如果你消耗了 200 美元套餐月度限额的 1%，那就是 2 美元。

Then, we assume a workload shape, and compute how many total tokens you can get from each plan given this ratio. For the agentic workload, we use our own usage ratios in September as reported in our Tokenomics Model.
然后，我们假设一种工作负载形态，并根据这一比例计算每个方案总共能获得多少 token。对于智能体工作负载，我们使用自家 Tokenomics Model 中报告的 9 月使用比例。

Finally, we multiply by the blended price per MTok at API prices to get the API-equivalent value.
最后，我们乘以 API 价格下每百万 token 的混合价格，得到 API 等效价值。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

Catching a Provider A/B Test
发现服务商的 A/B 测试

While refining our methodology, we ran into a very confusing situation where just 1 of 3 of the same subscription we tested for a particular provider had ~20% lower limits than the other 2. This unlucky account happened to be significantly older as well, and we were worried that the provider had some cursed setup where they changed subscription limits

depending on the age of the account.
在完善我们的方法论时，我们遇到了一个非常令人困惑的情况：我们为某个服务商测试的同一订阅中，3 个里只有 1 个的限额比其他 2 个低 ~20%。这个不走运的账号恰好也明显更旧，我们担心该服务商有什么糟糕的设置，会根据账号年龄更改订阅限额。

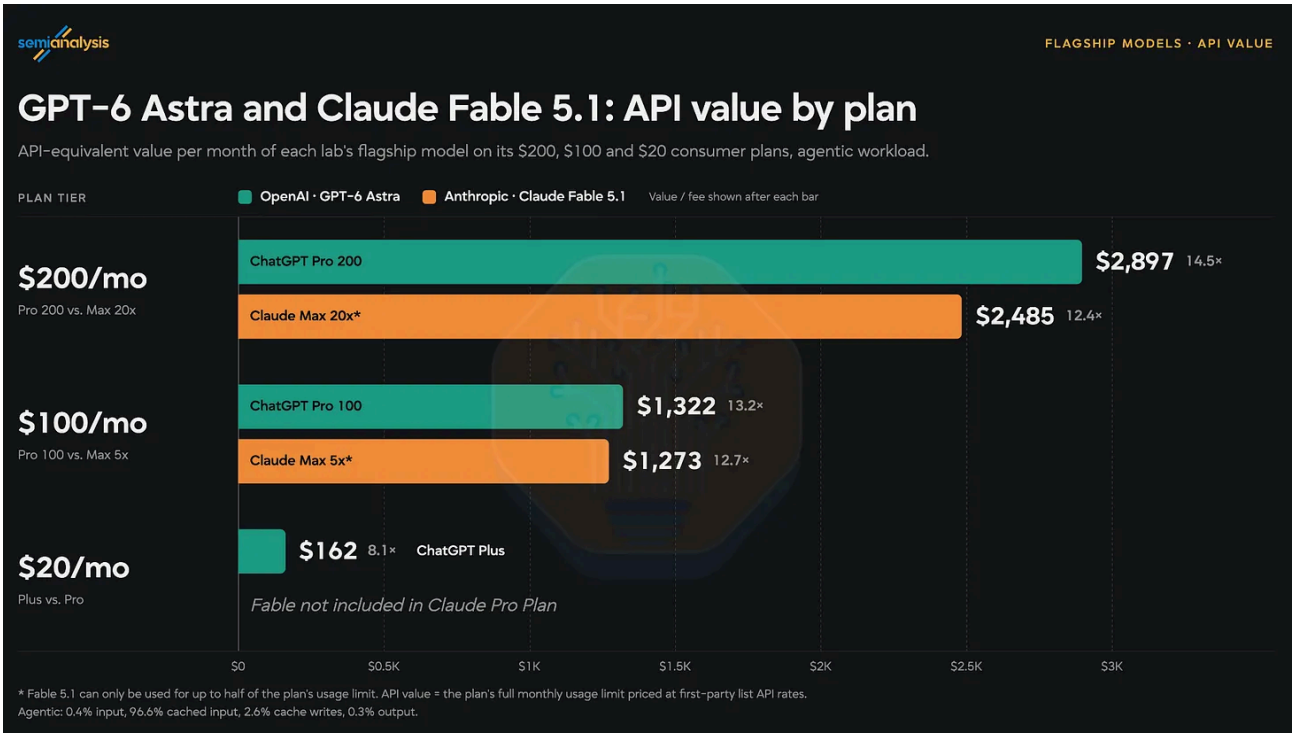
Fortunately, after reaching out to the provider, we were able to confirm this was not the case and that our unlucky account was part of an “extremely tiny” A/B test they were running on limits. Additionally, they wanted to emphasize that they “didn’t just decrease limits wholesale” and were instead testing “how to better balance when people hit limits”.
幸运的是，在联系该服务商后，我们得以确认并非如此，我们那个不走运的账号只是他们正在进行的“极其微小”的限额 A/B 测试的一部分。此外，他们想强调，他们“并不是直接全面降低限额”，而是在测试“如何更好地平衡人们触及限额时的情况”。

This conclusion is interesting for two reasons:
这个结论之所以有趣，有两个原因：

1. It proves providers can silently change subscription limits at any time
   它证明服务商可以随时悄悄更改订阅限额
2. Our method is sensitive enough to detect these subtle changes!
   我们的方法足够灵敏，能够检测到这些细微的变化！

OpenAI vs Anthropic OpenAI 与 Anthropic

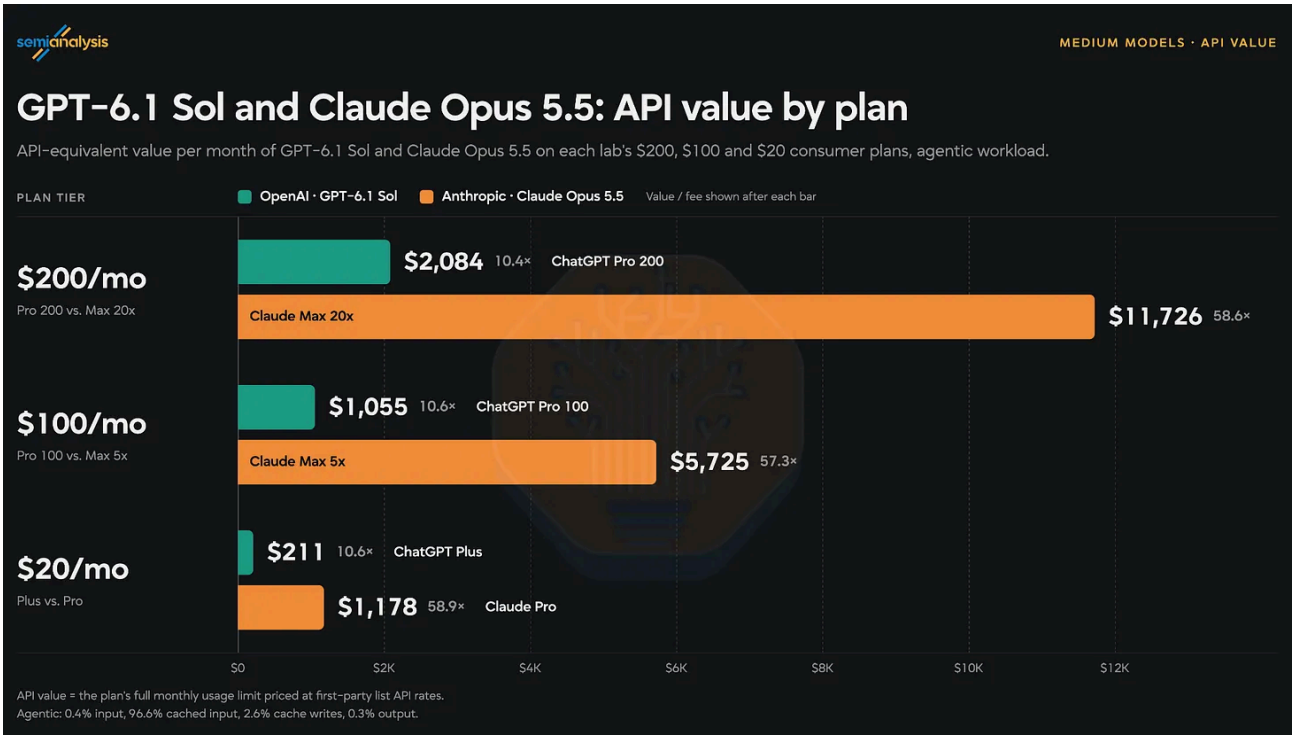
The biggest question in AI subscription land is how Anthropic’s limits compare to OpenAI’s. Note that OpenAI just made some massive changes to their subscriptions last week, halving the API-equivalent value for the $200 plan and introducing a new $500 tier. The following numbers describe the **current state** of OAI vs ANT. Afterwards, we will examine OpenAI’s recent changes in depth.
AI 订阅领域最大的问题是 Anthropic 的限制与 OpenAI 相比如何。请注意，OpenAI 上周刚刚对其订阅做了重大调整，将 200 美元套餐的 API 等效价值减半，并推出了新的 500 美元档位。以下数字描述了 OAI 与 ANT 的当前状况。之后，我们将深入审视 OpenAI 近期的变化。



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

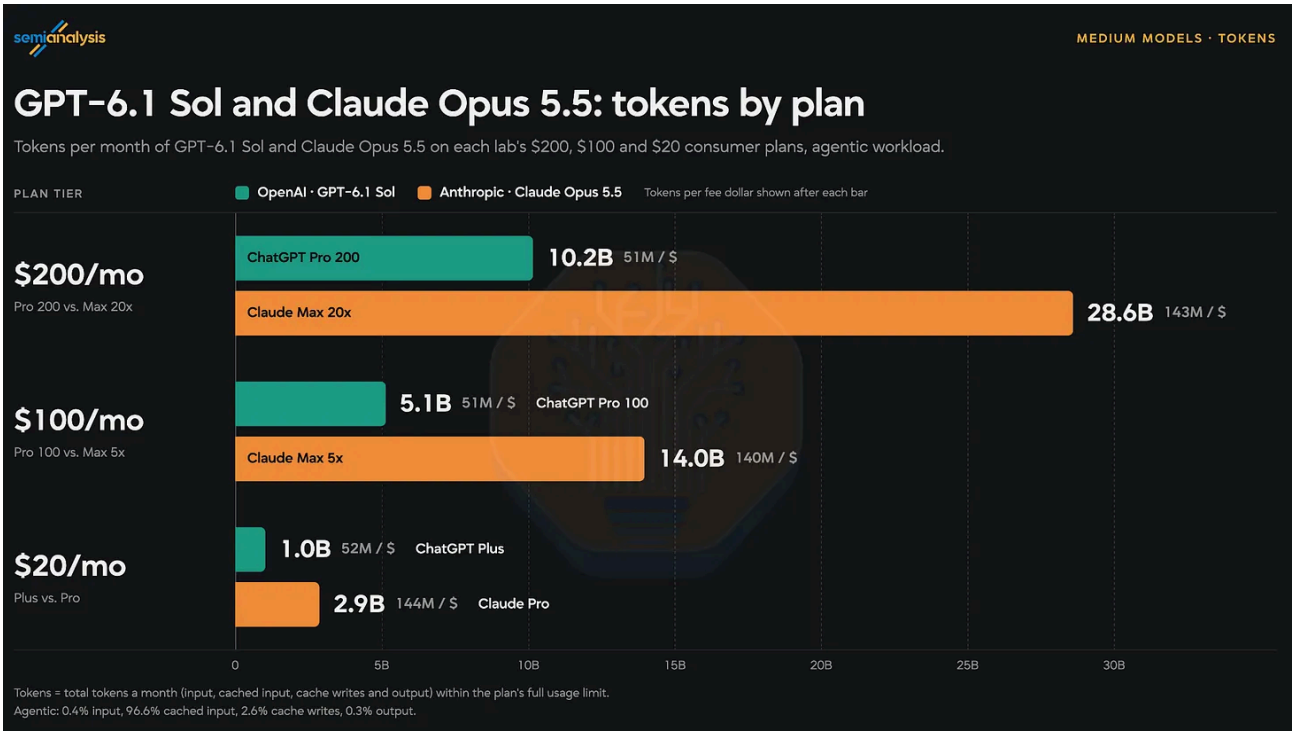
Limits are quite similar across the board for GPT-6 Astra vs Fable 5.1, though remember that Fable can only be used for 50% of your limit. In other words, your $200 ANT plan would still have 50% left after consuming $2,485 worth of Fable 5.1, whereas the equivalent OAI plan would be fully exhausted after $2,897 of Astra.
在 GPT-6 Astra 与 Fable 5.1 之间，各档位的限制相当接近，但请记住 Fable 只能使用你限额的 50%。换句话说，你的 200 美元 ANT 套餐在消耗了价值 2,485 美元的 Fable 5.1 之后仍会剩下 50%，而同等 OAI 套餐在消耗 2,897 美元的 Astra 后就会完全耗尽。

This extra 50% starts becoming very relevant when we compare Opus 5.5 vs GPT 6.1 Sol.
当我们比较 Opus 5.5 与 GPT 6.1 Sol 时，这额外的 50% 就开始变得非常重要。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

At this mid-tier of models—which both companies market as the intended daily driver for most users—Anthropic is an overwhelmingly better deal, offering ~5x the API-equivalent value across the board. You could argue that this is unfair for OAI because 6.1 Sol is much cheaper per token than Opus 5.5, but the gap is still massive even if you switch to comparing the number of tokens.
在这一中端模型层级——两家公司都将其宣传为面向大多数用户的日常主力——Anthropic 的性价比压倒性地更高，在各档位上提供了 ~5 倍的 API 等效价值。你可能会说这对 OAI 不公平，因为 6.1 Sol 每 token 比 Opus 5.5 便宜得多，但即使改用 token 数量来比较，差距仍然巨大。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

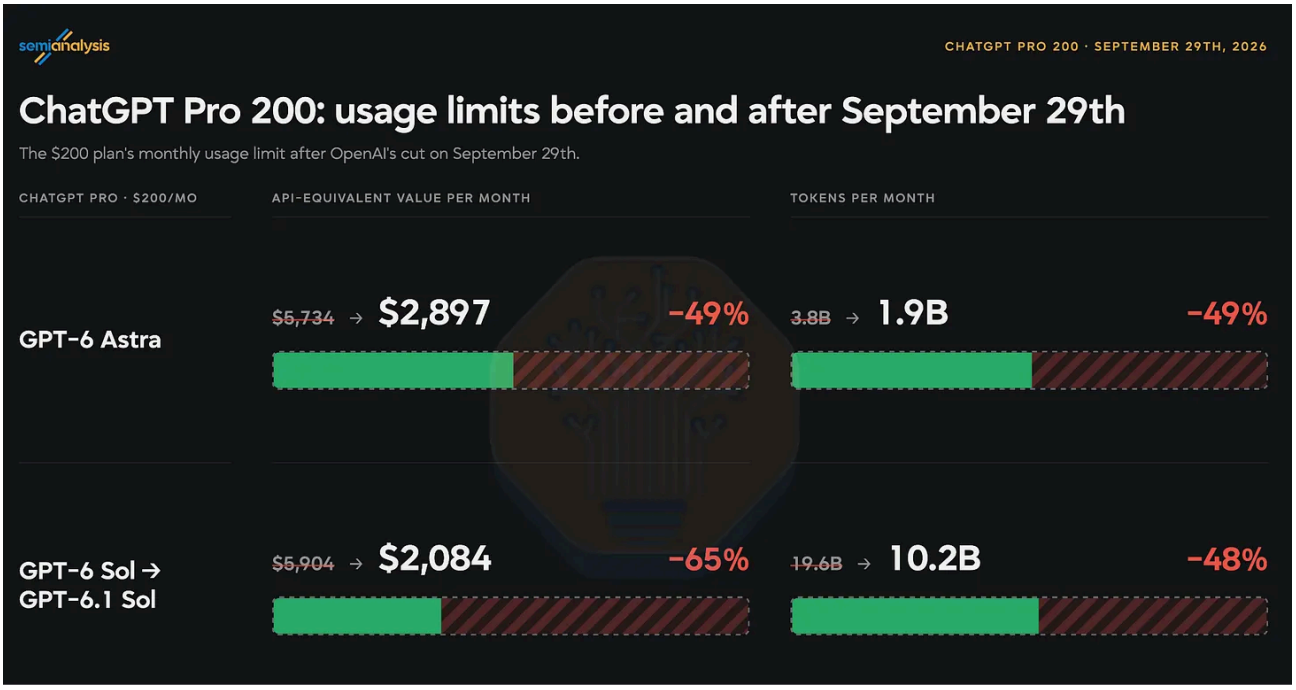
Overall, we believe API-equivalent value is the single best number for capturing the "value" of a subscription plan, but there are times where raw token volumes can be more appropriate. For example, if a model is a particularly bad (or good) deal at API prices, then the API-equivalent value will of course be misleading.
总体而言，我们认为 API 等效价值是衡量订阅套餐"价值"的最佳单一指标，但有时原始 token 量可能更合适。例如，如果某个模型在 API 价格下特别不划算（或特别划算），那么 API 等效价值当然会产生误导。

Token efficiency is also an extremely relevant factor, but the industry unfortunately lacks reliable data here. Many people like to cite this chart from Artificial Analysis, but we do not believe the benchmark tasks in the AA Intelligence Index are at all representative of real work people do with LLMs.
Token 效率也是一个极其相关的因素，但遗憾的是，业界在这方面缺乏可靠数据。许多人喜欢引用 Artificial Analysis 的这张图表，但我们认为 AA 智能指数中的基准任务完全不能代表人们使用 LLM 所做的实际工作。

OpenAI's new $500 plan and massive limit reduction
OpenAI 新的 500 美元套餐与大幅限额削减

Our tracking confirmed the 50% value reduction for OpenAI's $200 plan announced by Tibo. $200 plans purchased before the cut will keep the old higher limits until October 29th, but newly purchased plans immediately start with the lower limits.
我们的追踪证实了 Tibo 所公布的 OpenAI 200 美元套餐价值缩水 50% 的情况。在削减之前

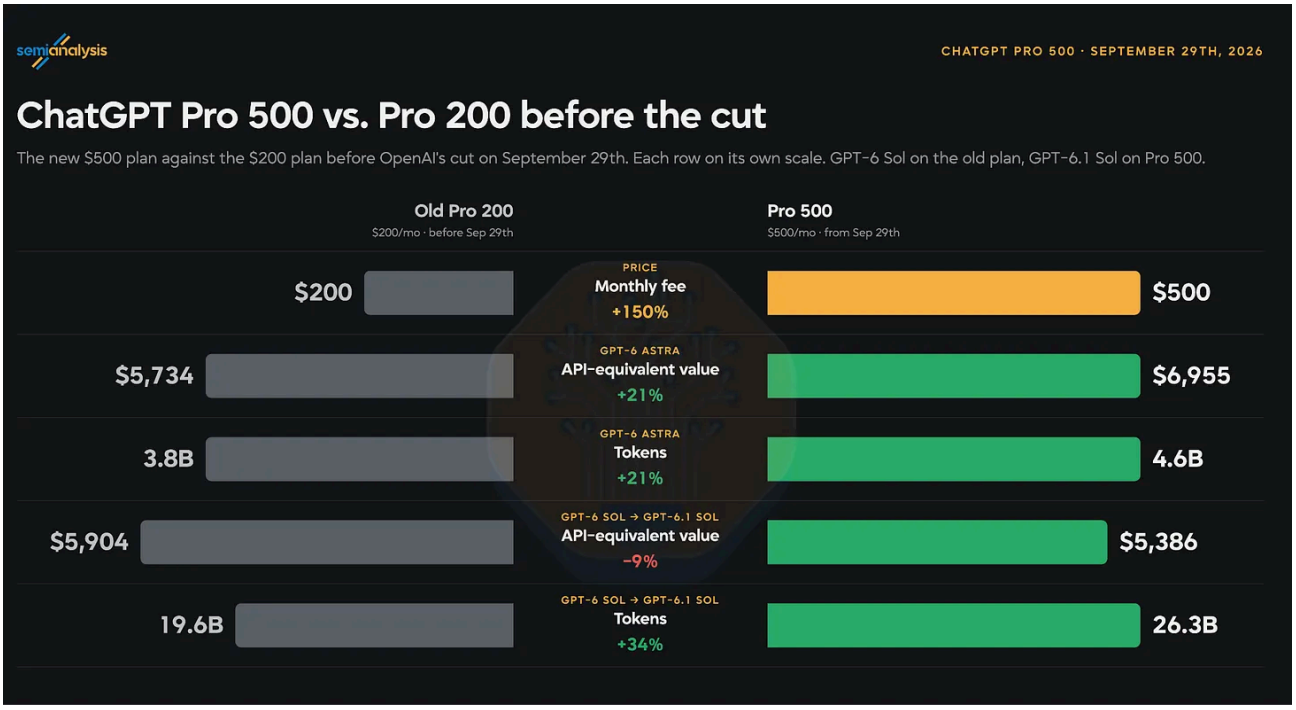
购买的 200 美元套餐将保留原有的较高限额直至 10 月 29 日，但新购买的套餐立即以较低限额开始。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

As you can see, OpenAI specifically halved the number of tokens you get for each model tier. This caused the API-equivalent value for Sol class models to drop by over 50% because OAI also cut cached input token pricing for 6.1 Sol.
如你所见，OpenAI 专门将每个模型层级所获得的 token 数量减半。这导致 Sol 级模型的 API 等值价值下降超过 50%，因为 OAI 还削减了 6.1 Sol 的缓存输入 token 定价。

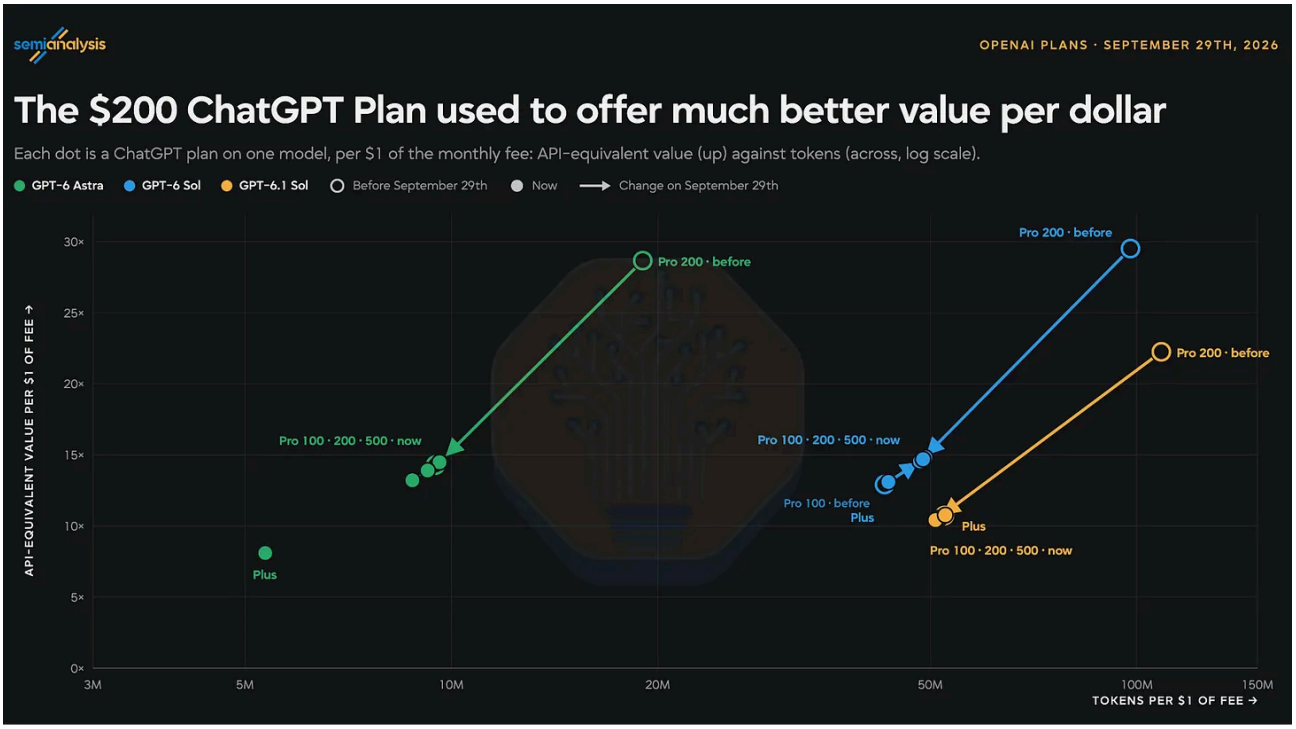
The new $500 plan only offers 21% more Astra than the old $200 plan. Interestingly, the API-equivalent value for Sol class models actually decreased due to the aforementioned 6.1 Sol price cut.
新的 500 美元套餐相比旧的 200 美元套餐仅多提供 21% 的 Astra。有趣的是，由于前述 6.1 Sol 的降价，Sol 级模型的 API 等值价值实际上反而下降了。



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

Of course, 300 TPS Ultrafast was the real headline feature for the $500 plan and we are currently testing its limits. Results will be published to Tokenomics Model subscribers as soon as they're available. We are also actively testing the limits when using your ChatGPT subscription on third party apps like Devin.
当然，300 TPS Ultrafast 才是 500 美元套餐真正的头条功能，我们目前正在测试其极限。结果一经获得将立即发布给 Tokenomics Model 订阅者。我们也在积极测试在 Devin 等第三方应用上使用你的 ChatGPT 订阅时的极限。

The $200/month plan also used to be significantly more subsidized than OpenAI's other subscriptions. If we normalize each (plan, model) combo to the API-equivalent value and tokens per dollar, we can see that Pro 100 offered ~2x more per-dollar value than Plus on Astra and Pro 200 was another ~2x on top of Pro 100 across all models.
每月 200 美元的套餐过去也比 OpenAI 的其他订阅补贴力度大得多。如果我们把每个（套餐、模型）组合都按 API 等价价值和每美元 token 数来标准化，就能看到 Pro 100 在 Astra 上提供的每美元价值是 Plus 的~2 倍，而 Pro 200 在所有模型上又在 Pro 100 的基础上再翻~2 倍。

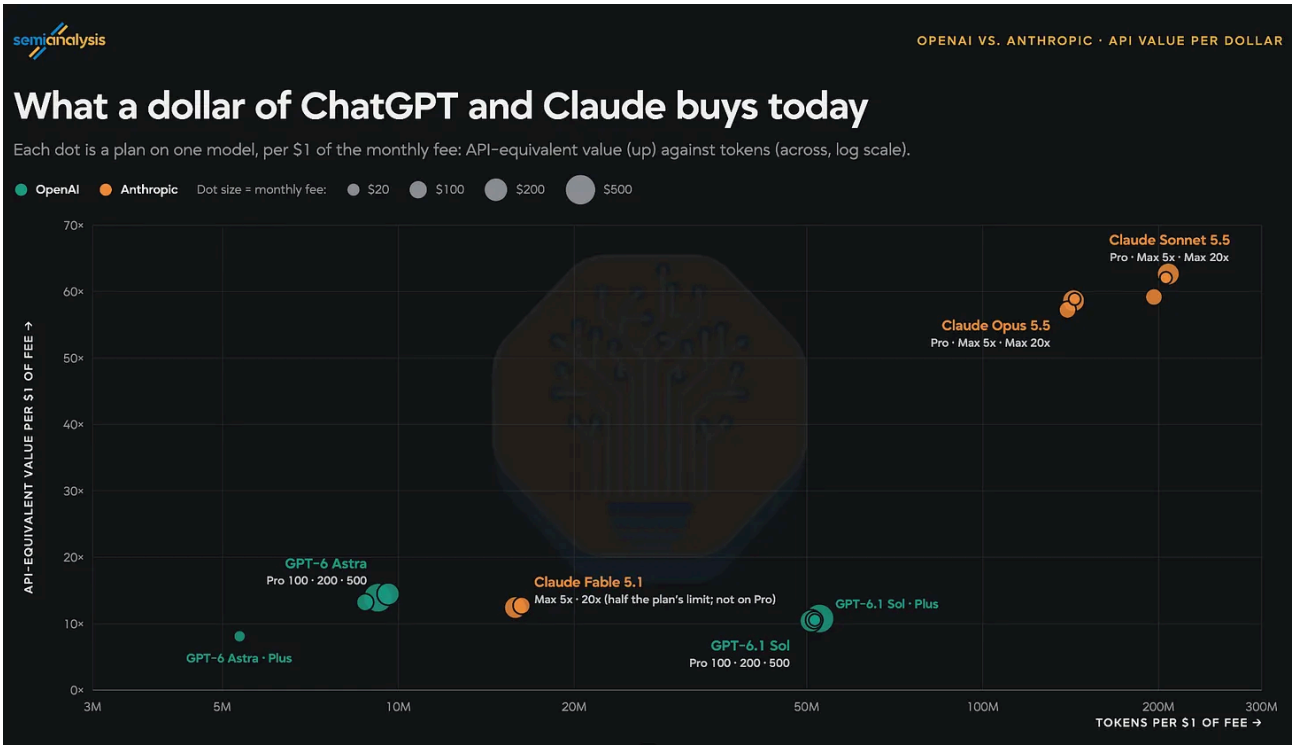


Source: SemiAnalysis [Tokenomics Model]

来源：SemiAnalysis Tokenomics Model

With last week's cut, Pro 100, 200, and 500 all offer the same tokens per dollar for every model. Plus is also comparable for Sol but offers a relatively worse deal on Astra.
随着上周的降价，Pro 100、200 和 500 在所有模型上提供的每美元 token 数都相同。Plus 在 Sol 上也相当，但在 Astra 上的性价比相对更差。

Anthropic, on the other hand, already offered the same per-dollar value for all of their subscription tiers and currently crushes OpenAI.
另一方面，Anthropic 此前就已经为其所有订阅层级提供相同的每美元价值，目前更是碾压 OpenAI。



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

OpenAI used to be lauded by indie developers for their generous subscription limits compared to Anthropic, but this is simply no longer true today. Opus 5.5 on any Claude subscription offers far better value than anything from OpenAI.
OpenAI 过去曾因相比 Anthropic 更慷慨的订阅限额而受到独立开发者的赞誉，但如今这已不再属实。任何 Claude 订阅上的 Opus 5.5 所提供的价值都远超 OpenAI 的任何产品。

The only counter argument for OpenAI is that none of their Pro plans have a 5 hour limit, which makes it easier to consume a higher % of your total monthly limit. However, we don't think this offsets the ~4x higher API-equivalent value offered by Opus 5.5 on Claude plans.
对 OpenAI 唯一有利的反驳是，他们的 Pro 套餐都没有 5 小时限制，因此更容易消耗掉每月总限额中更高的比例。然而，我们认为这无法抵消 Claude 套餐上 Opus 5.5 所提供的~4 倍更高的 API 等价价值。

Do labs adjust subscription limits based on API pricing changes?
实验室会根据 API 定价变化调整订阅限额吗？

We've seen a number of significant price cuts recently to Anthropic and OpenAI models.
我们最近看到 Anthropic 和 OpenAI 的模型出现了一系列大幅降价。

- [Fable 5.1] cut cache reads by 75% vs Fable 5
  Fable 5.1 将缓存读取成本较 Fable 5 降低了 75%

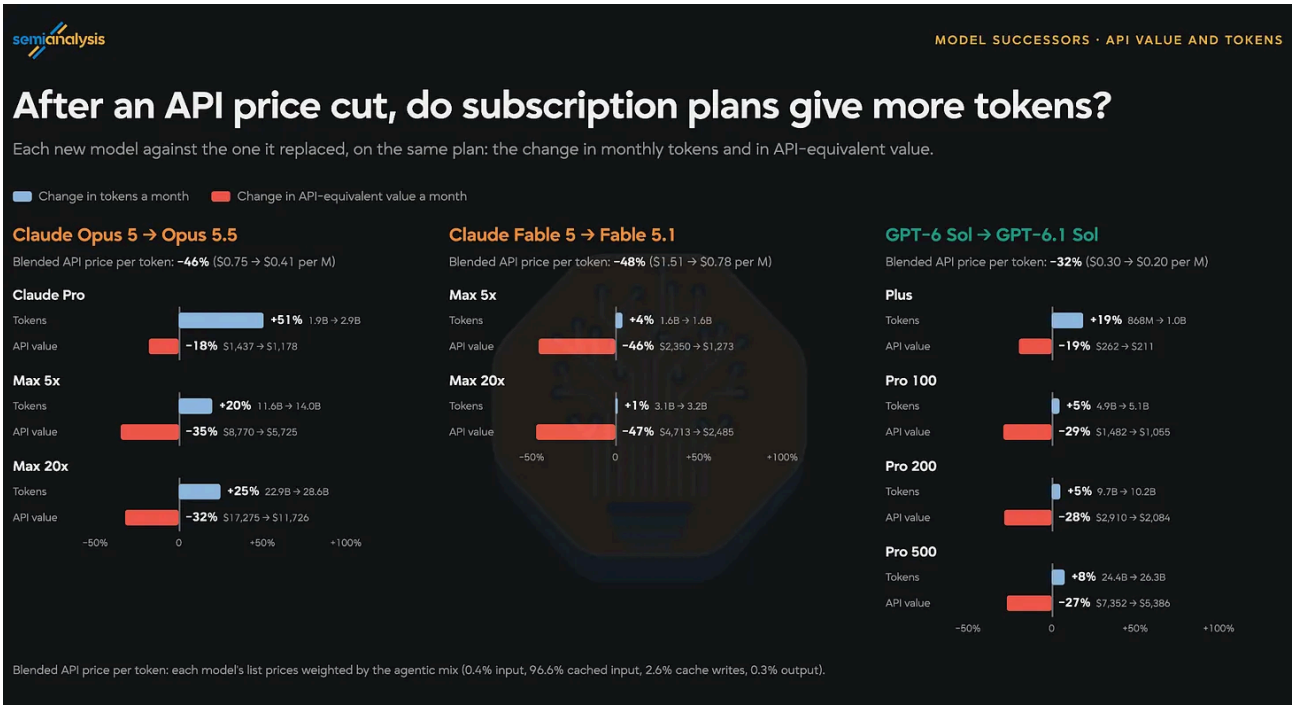
- Opus 5.5 cut input and output token pricing by 20% vs Opus 5 and cache reads by 60%.
  Opus 5.5 将输入和输出 token 价格较 Opus 5 下调了 20%，缓存读取价格下调了 60%。

- GPT 6.1 Sol cut cache reads by 50% vs GPT 6 sol. This was after making GPT 6 Sol 60-67% cheaper vs 5.6 Sol.
  GPT 6.1 Sol 将缓存读取价格较 GPT 6 Sol 下调了 50%。此前，GPT 6 Sol 已较 5.6 Sol 便宜了 60-67%。

One important question is whether the labs adjust token limits after these price cuts, such that the API-equivalent value of your subscription stays constant.
一个重要问题是，这些实验室是否会在降价后调整 token 限额，从而使你订阅的 API 等价价值保持不变。

We already showed earlier that OpenAI didn't change Sol token limits on the $200/month plan when they released 6.1 Sol. This caused the API-equivalent value to drop ~30%.
我们此前已经展示过，OpenAI 在发布 6.1 Sol 时，并未改变每月 200 美元套餐的 Sol token 限额。这导致 API 等价价值下降了 ~30%。

Anthropic also didn't increase Fable token limits at all with the release of 5.1, but they did up Opus token limits with the 5.5 price cut.
Anthropic 在发布 5.1 时也完全没有提高 Fable 的 token 限额，但他们确实在 5.5 降价时提高了 Opus 的 token 限额。



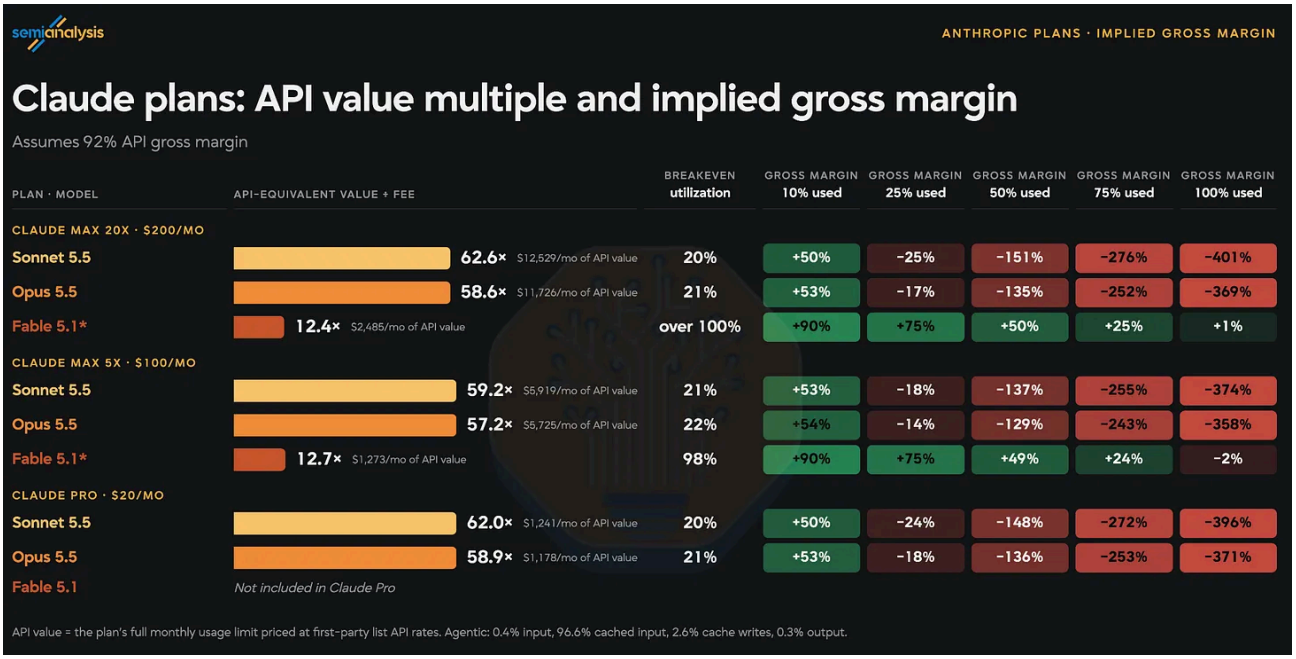
Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

However, the ~20% token increase for Max plans and ~50% increase for the Pro plan weren't enough to fully offset the API price cut.
然而，Max 套餐 token 增加 ~20% 以及 Pro 套餐增加 ~50%，都不足以完全抵消 API 价格的下降。

Two different approaches to increasing subscription margins
提高订阅利润率的两种不同方式

As we explained at the start of this article, subscription gross margins are way lower than API and meaningfully reduce revenue per MW for both OpenAI and Anthropic. Both companies are well aware of this fact but have chosen two different strategies for reducing subscriber subsidies.
正如我们在本文开头所解释的，订阅业务的毛利率远低于 API，并且会显著降低 OpenAI 和 Anthropic 每兆瓦收入。两家公司都非常清楚这一事实，但选择了两种不同的策略来减少订阅用户补贴。

For a given plan, Anthropic decreases the API-equivalent value for more premium models. The drop off is small between Sonnet 5.5 and Opus 5.5, but becomes very meaningful with Fable 5.1.
对于给定套餐，Anthropic 会降低更高端模型的 API 等效价值。Sonnet 5.5 与 Opus 5.5 之间的降幅很小，但到了 Fable 5.1 就变得非常显著。



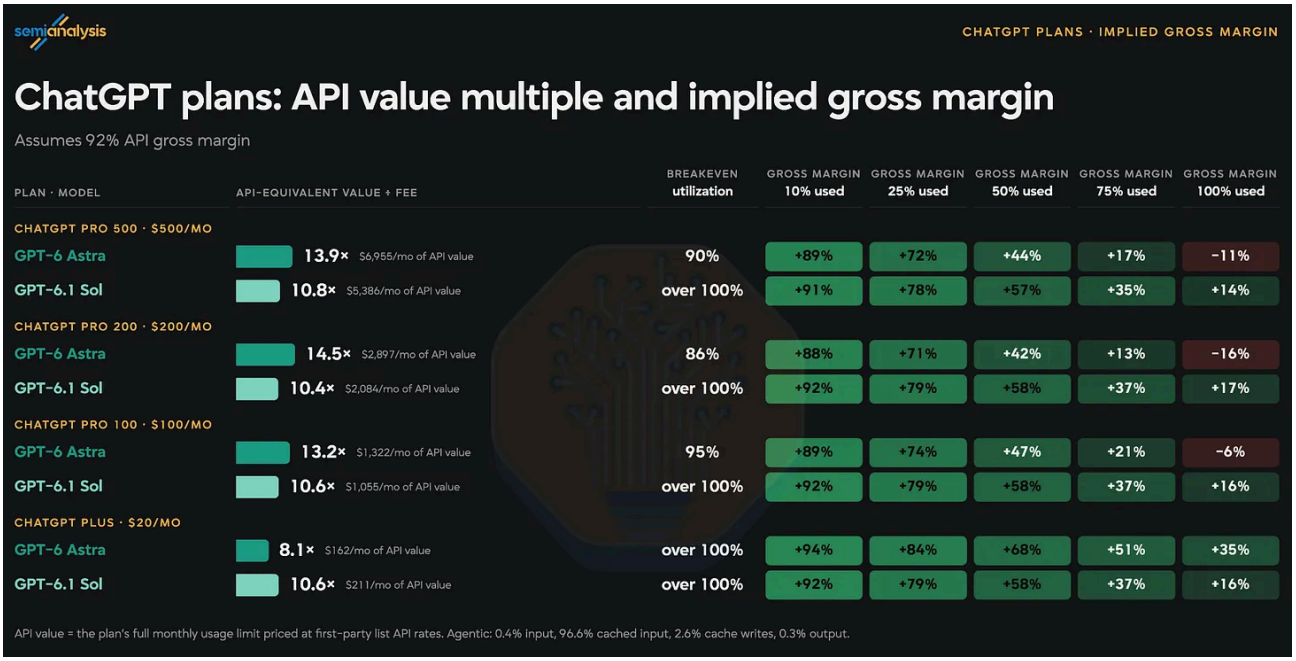
Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

Reducing the API-equivalent value as you release more powerful models at higher list prices is the more subtle way to stop subsidizing subscriptions. Assuming 100% utilization and 92% API gross margins, maxing out Opus 5.5 vs Fable 5.1 usage corresponds to -369% and 1% gross margins respectively. If we were to use a more realistic average

utilization of 20%, then gross margins rise to 6% and 80% respectively.
随着你以更高标价发布更强大的模型，同时降低 API 等效价值，这是一种更隐蔽的停止补贴订阅的方式。假设利用率为 100%、API 毛利率为 92%，那么把 Opus 5.5 与 Fable 5.1 的用量拉满，分别对应 -369% 和 1% 的毛利率。如果我们采用更现实的 20% 平均利用率，那么毛利率分别升至 6% 和 80%。

In other words, **Anthropic would likely already have software like margins on their subscription plans if everyone only used Fable**. Opus class models will continue to become cheaper to serve over time as Anthropic figures out how to make the model smaller, and new model tiers above Fable will have even lower subscription limits (assuming they're included at all).
换句话说，如果所有人都只用 Fable，Anthropic 的订阅套餐很可能已经拥有类似软件业务的利润率。随着 Anthropic 找到让模型更小的方法，Opus 级模型的服务成本将随时间继续下降，而 Fable 之上的新模型层级将拥有更低的订阅限额（假设它们会被纳入的话）。

OpenAI, on the other hand, picked the nuclear option of just immediately cutting to Fable-level limits across the board.
另一方面，OpenAI 选择了核选项，直接一刀切地把所有模型的限额降到 Fable 级别。

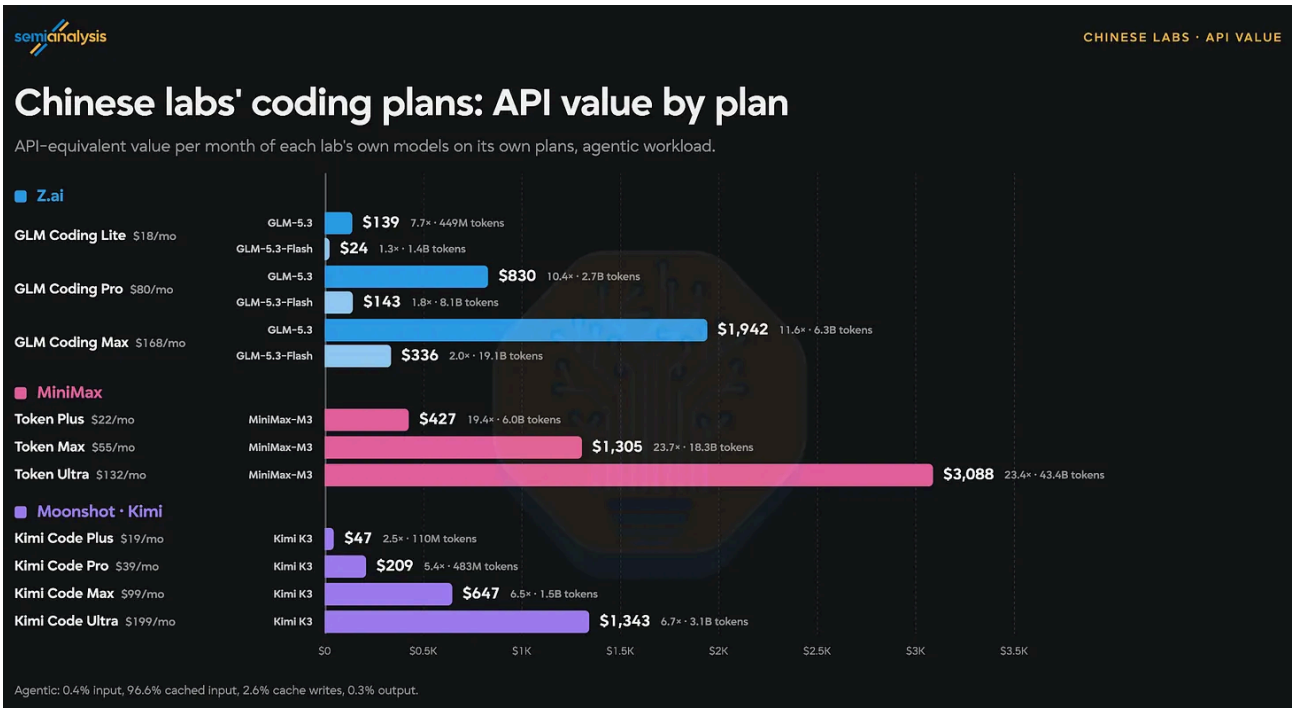


Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

The downside of this approach is supposed to be public backlash, but OpenAI has somehow managed to more or less avoid this entirely. It's likely some combination of all the good DevDay PR drowning out the negative news from the day prior and grandfathering existing plans for an additional month.

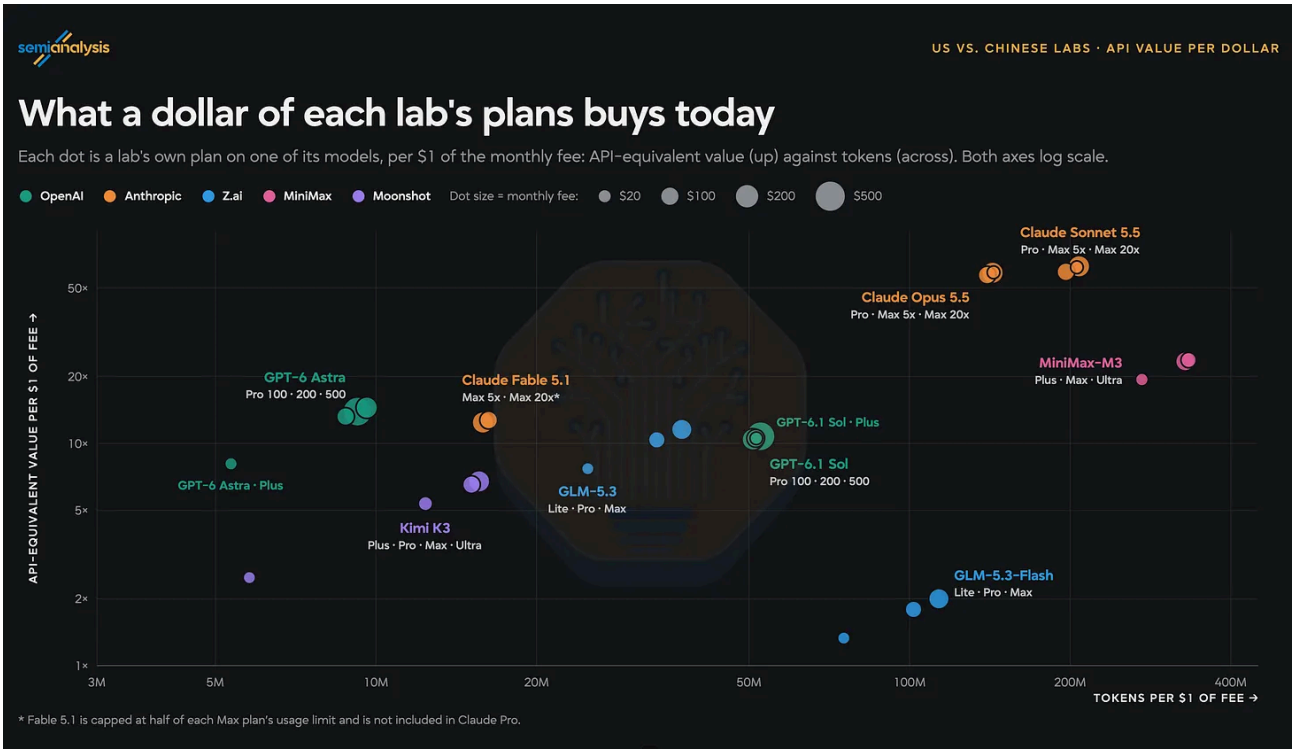
Chinese subscription plans are still a good deal

Despite being compute poor, the Chinese labs still offer subsidized subscription plans, but the level of subsidization varies dramatically between models.



Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

Here’s how they compare to OpenAI and Anthropic after normalizing per dollar.



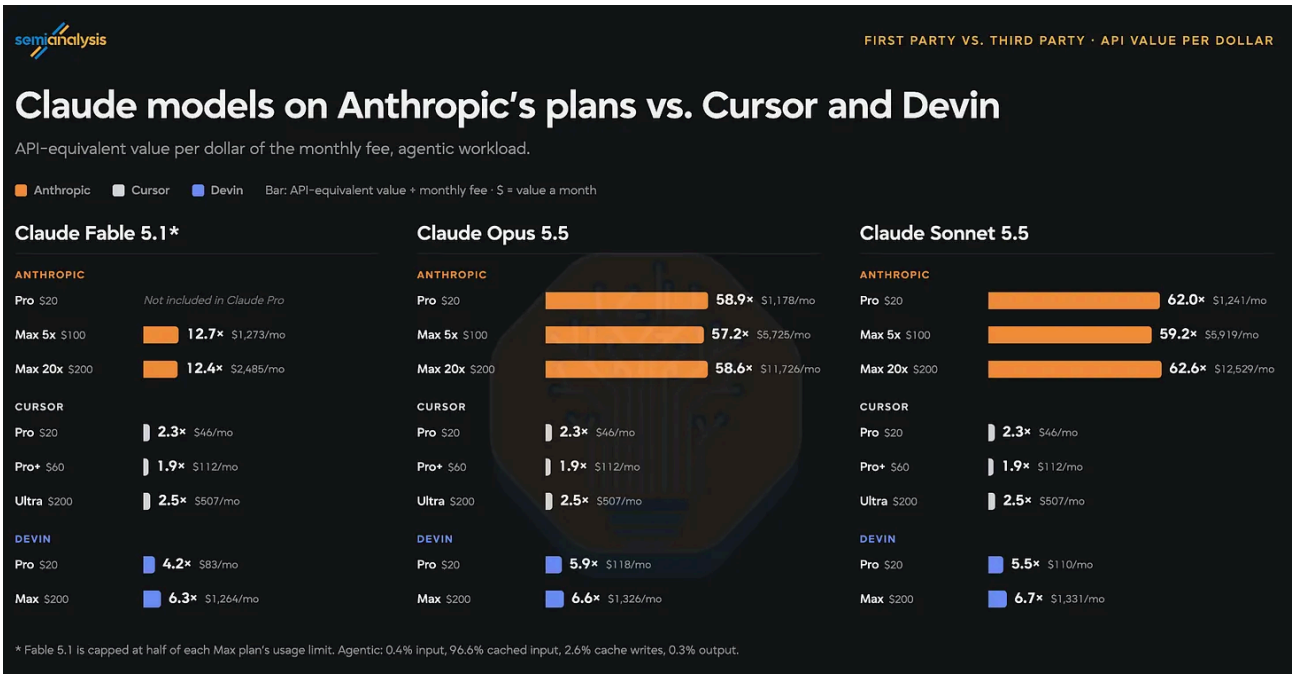
Source: SemiAnalysis [Tokenomics Model]
来源：SemiAnalysis Tokenomics Model

Value per dollar increases as you buy higher tier subscriptions from the Chinese labs. On average, the API-equivalent value per dollar is a little less than the ~12x you get from OpenAI.

Third party plans are worse than first party

The final comparison we'll consider is OpenAI and Anthropic models on their own first party plans vs third party wrappers like Cursor and Cognition (Devin).

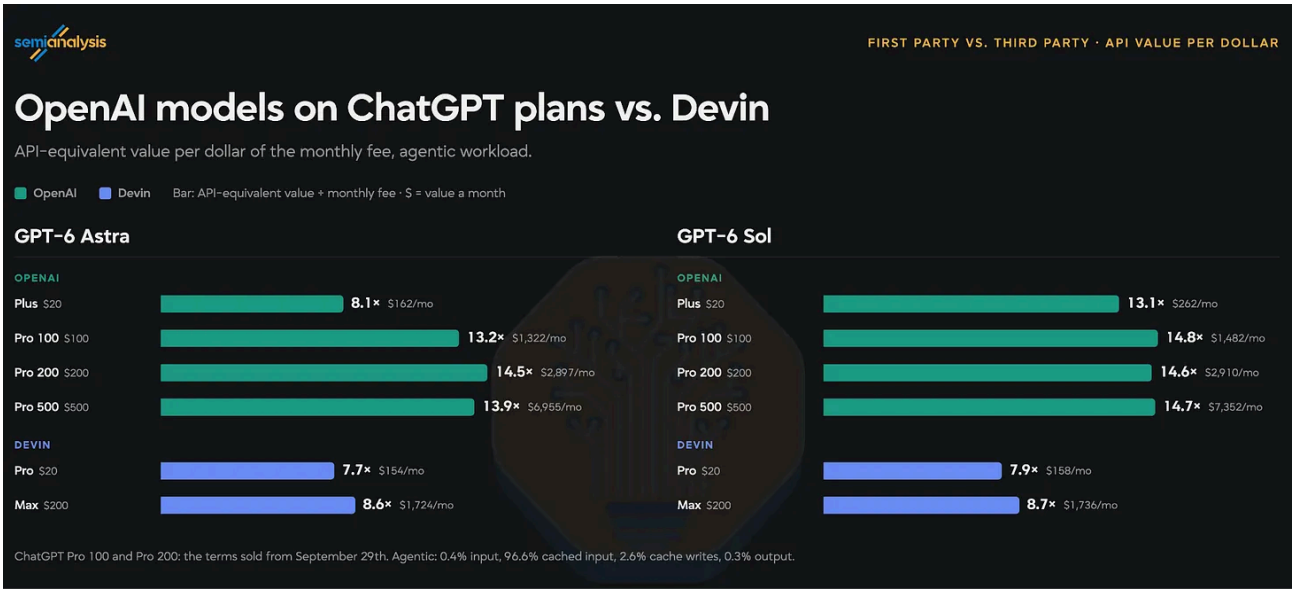
As you'd expect, first party plans are much more generous with their limits.



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

The funny thing is that **Anthropic's subscription margins are likely much higher than both Cursor and Cognition despite offering 9x+ more generous limits**. This is the power of owning the models yourself, as opposed to wrapping someone else's API.

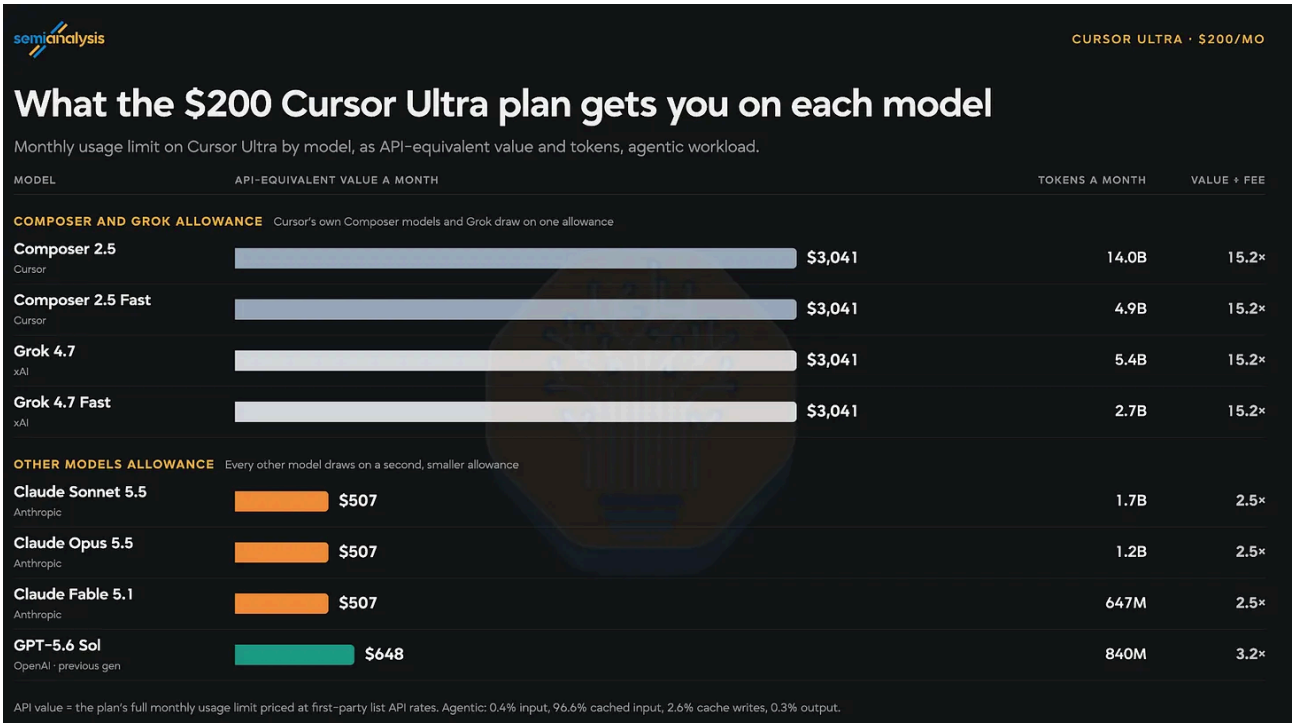
OpenAI's newest models are no longer offered on Cursor because Sam doesn't trust Elon, so here's how OpenAI compares to just Cognition.



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

The gap is interestingly much smaller here, partly because Cognition is more generous with OpenAI models (perhaps they negotiated a higher enterprise discount), but largely because OpenAI subscriptions offer much lower value per dollar than Anthropic.

Lastly, here's how much Composer and Grok 4.7 value you get on the $200/month Cursor plan.



Source: SemiAnalysis Tokenomics Model
来源：SemiAnalysis Tokenomics Model

The 15.2x API-value per dollar is just barely higher than OpenAI's Pro plans and—as you'd expect—far higher than what Cursor offers for third party models.

For more data on Meta and SpaceXAI's subscriptions, as well as live updates as soon as all these providers release new models or inevitably change their limits, see our [Tokenomics Model](#).



Recommend SemiAnalysis to your readers
向你的读者推荐 SemiAnalysis

Bridging the gap between the world's most important industry, semiconductors, and business.
弥合全球最重要的产业——半导体——与商业之间的鸿沟。