The Future of Meta Superintelligence: A 1 Year Progress Update

Meta 超级智能的未来：一周年进展更新

A top tier RL environment startup spawns out of thin air, the most aggressive compute ramp we've ever seen, 2000km+ scale-across, and some advice for Google DeepMind

一家顶级强化学习环境初创公司横空出世，见证了我们有史以来最激进的算力扩张、超过 2000 公里的跨尺度互联，并为 Google DeepMind 提供了一些建议

MAX KAN, JULIEN MARTIN-PRIN, JEREMIE ELIAHOU ONTIVEROS, AND DYLAN PATEL

MAX KAN、JULIEN MARTIN-PRIN、JEREMIE ELIAHOU ONTIVEROS 和 DYLAN PATEL

JUL 10, 2026 2026 年 7 月 10 日 · PAID · 付费

64 1 4 Share 分享

It’s been a little over 1 year since the disastrous Llama 4 release spurred Zuck to rebuild his entire AI org. Highlights include the shocking $14.3B Scale AI “investment” just to poach Alexandr Wang and the best people from his Safety, Evaluations, and Alignment Labs (SEAL) team, the multi-hundred million dollar (sometimes $1B+) pay packages offered to top AI researchers/engineers, and the expedited compute ramp enabled by their new “Tent” datacenter design. For more details, see our original post on MSL.

距离灾难性的 Llama 4 发布促使扎克伯格重建其整个 AI 组织，已过去一年多。亮点包括：为挖角 Alexandr Wang 及其安全、评估与对齐实验室（SEAL）团队的最优秀人才，令人震惊地“投资”143 亿美元收购 Scale AI；向顶级 AI 研究人员/工程师提供数亿美元（有时超过 10 亿美元）的薪酬包；以及凭借全新的"Tent"数据中心设计加速算力扩张。更多详情，请参阅我们最初在 MSL 发布的文章。

Since then, frontier AI has increasingly felt like a two horse race between OpenAI vs Anthropic. Google had a brief moment in the spotlight with Gemini 3 Pro and Nano Banana, but they’ve since faded dramatically. Despite their Windsurf acquisition, they’re far from a compelling agentic coding product, and 3.5 Flash is a benchmaxxed prop that performs far worse than GPT 5.5 and Opus 4.8 in real world scenarios (much less Fable and 5.6). 3.5 Pro is not even Opus level on coding. Microsoft has completely

blown their early lead with GitHub copilot and failed to effectively leverage their access to OpenAI IP. SpaceXAI is selling $26B a year worth of GPUs to Anthropic/Google, and the Chinese labs are simply too compute poor to truly reach the frontier.

自那时起，前沿人工智能领域越来越像是 OpenAI 与 Anthropic 之间的双雄争霸。谷歌曾凭借 Gemini 3 Pro 和 Nano Banana 短暂成为焦点，但此后迅速黯然失色。尽管收购了 Windsurf，他们距离打造一款真正具有吸引力的智能体编程产品仍相去甚远；而 3.5 Flash 只是一个在基准测试中表现优异、却在实际场景中远逊于 GPT 5.5 和 Opus 4.8（更不用说 Fable 和 5.6）的营销噱头。3.5 Pro 在编程能力上甚至未达到 Opus 的水平。微软则完全浪费了其在 GitHub Copilot 上的早期领先优势，未能有效利用其获取 OpenAI 知识产权的渠道。SpaceXAI 每年向 Anthropic 和谷歌出售价值 260 亿美元的 GPU，而中国实验室则因算力严重不足，难以真正触及前沿。

Meanwhile, MSL made their public debut this April with the launch of Muse Spark. You could argue this model represents a relative regression for Meta. Llama 3 70B and 3.1 405B were both SOTA open-source on release, whereas Muse Spark, despite also being closed source, lagged both DeepSeek v4 Pro and Kimi K2.6—open source models released around the same time—on most benchmarks.

与此同时，MSL 于今年四月随着 Muse Spark 的发布正式亮相公众视野。有人或许会认为，该模型标志着 Meta 的一次相对退步。Llama 3 70B 和 3.1 405B 在发布时均为最先进的开源模型；而 Muse Spark 尽管同样是闭源模型，却在大多数基准测试中落后于同期发布的开源模型 DeepSeek v4 Pro 和 Kimi K2.6。

VIEWS | MODELS 3 / 12 selected | BENCHMARKS all 28 shown

Preset & saved views | 3 selected | 28 selected

BENCHMARK	Muse Spark Meta 1 win	Kimi K2.6 Moonshot AI 3 wins	DeepSeek V4 Pro DeepSeek 3 wins
AGENTIC · 2 BENCHMARKS			
GDPval-AA	1417	1482	1558
Terminal-Bench 2.0	59.0%	66.7%	67.9%
CODING · 2 BENCHMARKS			
SWE-Bench Pro	52.4%	58.6%	55.4%
SWE-Bench Verified	77.4%	80.2%	80.6%
GENERAL · 1 BENCHMARK			
Chatbot Arena (Text)	1489	1462	1459
REASONING · 1 BENCHMARK			
Humanity's Last Exam	50.4%	54.0%	48.2%
SCIENCE · 1 BENCHMARK			
GPQA Diamond	89.5%	90.5%	90.1%

Showing 7 benchmarks of 28 where all selected models have a reported score. Vendor-reported numbers — methodology varies across benchmarks.

Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis Tokenomics 模型

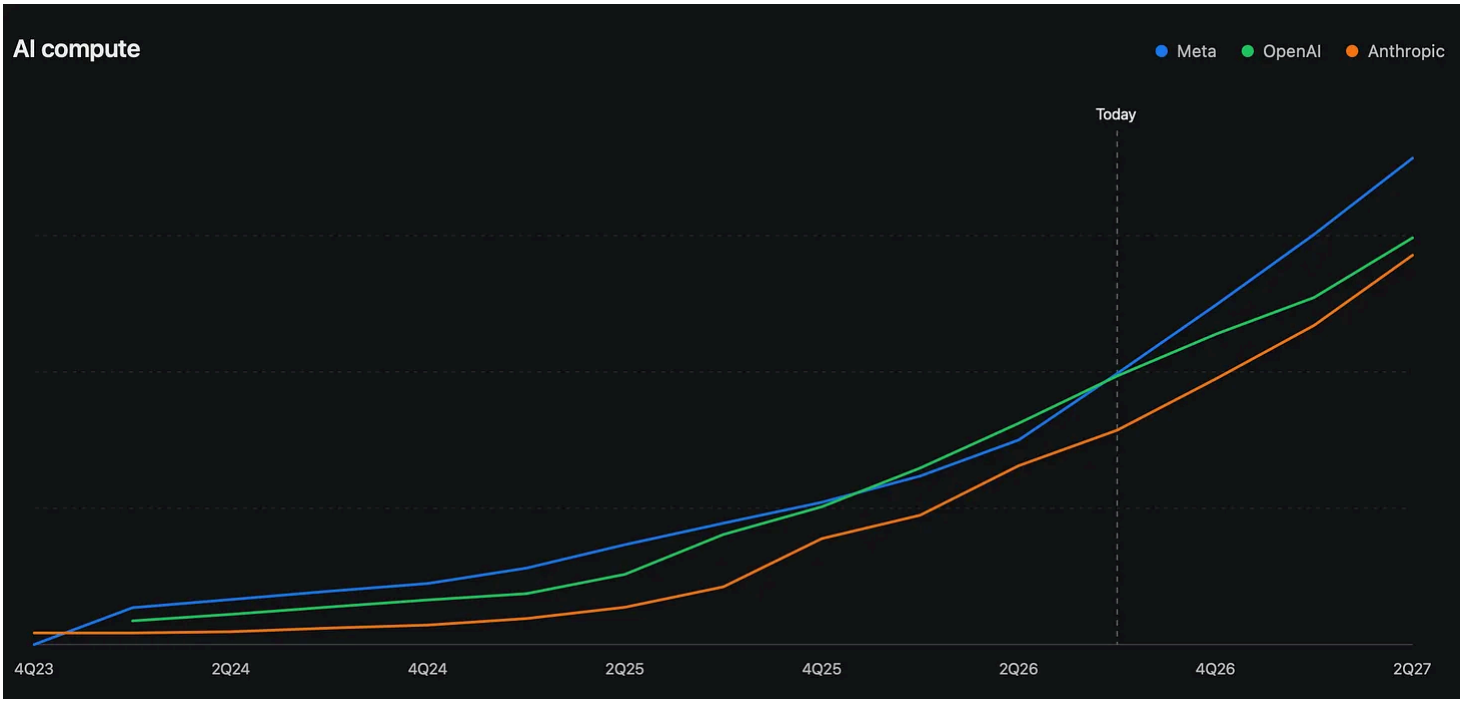
However, evaluating Muse Spark in isolation is missing the forest for the trees. What matters for MSL is the slope, not the intercept. Rebuilding your entire team from the ground up obviously comes with some short term setbacks, and it appears Meta has finally finished paying down this debt. Thus, the interesting question is not where MSL is today, but trying to predict where they'll be in the next 6 months. We think it's very possible they are better than Google by then due to the team's focus.

然而，孤立地评估 Muse Spark 无异于只见树木不见森林。对于 MSL 而言，关键在于斜率而非截距。从头开始重建整个团队显然会带来一些短期挫折，而 Meta 似乎终于还清了这笔债务。因此，有趣的问题不在于 MSL 今天处于何种位置，而在于预测他们未来六个月将达到何种高度。我们认为，凭借团队的专注，届时他们完全有可能超越谷歌。

At the simplest level, there are three things you need to build a true frontier model: **data**, **talent**, and **compute**. We believe **Meta** is the only hyperscaler/neolab on track to be world class at all three and therefore **has the best chance at catching up with Anthropic/OpenAI**. We'll explain why in full detail below, but as a teaser, here are the

AI compute projections from our new Tokenomics Model.

在最基础的层面上，构建真正的尖端模型需要三要素：数据、人才和算力。我们相信，Meta 是唯一一家有望在这三个方面均达到世界水平的超大规模云厂商/新型实验室，因此也最有可能追上 Anthropic 和 OpenAI。下文将详细阐述原因，在此先剧透一下：这是我们新推出的代币经济模型所预测的 AI 算力数据。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济模型

Lastly, behind the paywall, we'll discuss what this all means for Google—the company most people today still believe rounds out the AI big 3.

最后，在付费内容中，我们将探讨这一切对谷歌意味着什么——如今大多数人仍认为谷歌构成了人工智能三大巨头之一。

Data is the new oil (for real this time)

数据是新的石油（这次是真的）

We'll start with data because it's Meta's newest advantage and probably the most underappreciated of the three.

我们将从数据开始，因为这是 Meta 最新的优势，也可能是这三者中最被低估的一个。

In 2024, Ilya famously said that “data is the fossil fuel of AI.” While this analogy correctly highlights the importance of data for training AI models, it incorrectly assumes that the amount of good data is finite. In reality, if demand is strong enough, market forces will find a way.

2024 年，伊利亚（Ilya）有句名言：“数据是人工智能的化石燃料。”虽然这一类比正确地强调了数据对于训练 AI 模型的重要性，但它错误地假设优质数据的总量是有限的。事实上，只要需求足够强劲，市场力量自会找到出路。

This time, the invisible hand created a new human data/RL environment supply chain. The three incumbements—Mercor, Surge, and Handshake—are all at $1B+ ARR, and many new entrants who are barely a year old (e.g. Fleet, Mechanize, and Afterquery) are sitting around $100M ARR.

这一次，“看不见的手”创造了一条新的人类数据与强化学习（RL）环境供应链。三家现有巨头——Mercor、Surge 和 Handshake——的年经常性收入（ARR）均已超过 10 亿美元，而许多成立尚不足一年的新进入者（例如 Fleet、Mechanize 和 Afterquery）的 ARR 也已达 1 亿美元左右。

RL data companies — ARR & valuation

COMPANY	LATEST ARR	LATEST VALUATION	
Surge AI	~$3B Jul 2026 · rumored	~$30B Sep 2025	Bootstrapped to an estimated **$3B ARR** with **<110 employees** and **$0 outside capital**.
Mercor	$2B+ Jun 2026	$10B Oct 2025	**$1B → $2B+ gross ARR in 4 months.** Expert hours grew from **2K** to **2.5M** per quarter in two years, **+69%** quarter over quarter in 2Q26.
Handshake AI	~$1.0B May 2026	$3.5B Jan 2022	**$0 → $1B gross ARR in 16 months** after Jan 2025 launch. Spun up as a new AI-data unit on top of Handshake's decade-old college job-board business, monetizing its existing network of millions of students, PhDs, and professionals.
Micro1	$250M+ May 2026	~$2.5B May 2026	**$8M → $250M+ RR in 14 months** (~30×). Disclosed **7-figure 2025 profits** at **20%+ MoM growth**.
Fleet	$63M Apr 2026	$725M Apr 2026	**$1M → $63M RR in 6 months** (~60×). Founders project **$160M RR next quarter**. $45M Series A on a 30-person team.
Mechanize	*undisclosed* No public figure	$500M Apr 2026	Raised $9.1M at $500M post-money (**~1.8% dilution**) one year after launch.
Afterquery	$100M+ Apr 2026	$300M Apr 2026	**$0 → $100M run-rate in 14 months** as a YC W25 company.
Deeptune	*undisclosed* No public figure	*undisclosed* Mar 2026	Pivoted from AI dubbing (2022–24) to RL training gyms in 2025, then closed a **$43M Series A** from Andreessen Horowitz within ~12 months of the pivot.

Source: Public disclosures, SemiAnalysis

来源：公开披露信息，SemiAnalysis

Reinforcement learning (RL) is the most important scaling law for improving AI capabilities today. The general idea is that instead of just predicting the next token, you teach the model how to complete entire tasks (e.g. fixing a bug in a codebase). Besides the *tasks* themselves, RL requires you to provide an *environment* for the model to complete the task in, *tools* the model can use to interact with the environment, and *verifiers* to check if the model's answer is correct or not. For more background, see our previous posts.

强化学习（RL）是当今提升人工智能能力最重要的扩展法则。其核心理念在于，不再仅仅预测下一个词元，而是教导模型如何完成完整任务（例如修复代码库中的漏洞）。除了任务本身，强化学习还要求提供模型执行任务所需的环境、模型与环境交互时可使用的工具，以及用于验证模型答案正确性的检验器。更多背景信息请参阅我们之前的文章。

It is worth emphasizing that many AI insiders believe more RL environments/tasks are all we need to automate virtually all white-collar work. Here's a quote from my roommate Sholto Douglas, a prominent Anthropic researcher, on the Dwarkesh podcast last year:

值得强调的是，许多人工智能业内专家认为，只要拥有更多的强化学习环境和任务，就足以自动化几乎所有白领工作。以下是我的室友、Anthropic 知名研究员肖尔托·道格拉斯（Sholto Douglas）去年在 Dwarkesh 播客节目中的一段话：

> Even if algorithmic progress stalls out, and we just never figure out how to keep progress going—which I don't think is the case, that hasn't stalled out yet, it seems to be going great—**the current suite of algorithms are sufficient to automate white collar work provided you have enough of the right kinds of data.** Compared to the TAM of salaries for all of those kinds of work, it is so trivially worthwhile.
>
> 即使算法进展停滞，我们永远无法找到持续推动进步的方法——但我认为情况并非如此，目前进展并未停滞，反而势头良好——现有的算法组合已足以自动化白领工作，前提是你拥有足够多且类型恰当的数据。与所有这些工作对应的薪资总可寻址市场（TAM）相比，其价值显而易见且微不足道地值得投入。

Generally, all of these data companies create new RL tasks by hiring expert contractors from the relevant industries (hence why it's called "human data"). However, there's another thing every data company is desperately searching for today: real recordings of white-collar work.

通常，所有这些数据公司都会通过聘请相关行业的专家承包商来创建新的强化学习任务（因此被称为"人工数据"）。然而，如今每家数据公司都在迫切寻求的另一样东西是：白领工作的真实屏幕录制。

Screen recordings are extremely valuable for making RL tasks

屏幕录制对于构建强化学习任务具有极高的价值

Naively, you might think screen recordings are primarily useful for older training paradigms like supervised fine-tuning (SFT) which teach the AI to mimic humans. After all, the whole point of RL is that you let the AI figure out the steps itself and assign rewards based solely on outcomes.

乍看之下，你可能会认为屏幕录制主要适用于监督微调（SFT）等较旧的训练范式，这类方法旨在教会人工智能模仿人类行为。毕竟，强化学习的核心在于让人工智能自行推演步骤，并仅根据最终结果给予奖励。

In reality, however, screen recordings can still be extremely useful for making RL tasks as long as you're willing to do a little extra work.

然而实际上，只要你愿意多付出一些努力，屏幕录制在构建强化学习任务方面仍然极具价值。

The first benefit is realism. RL data is only good if it's 1) representative of real economically valuable work and 2) at the right difficulty level for the AI. Too easy and there's nothing for the AI to learn, too hard and it'll never achieve any reward.

第一个好处是真实性。强化学习数据只有在满足以下两个条件时才有价值：1) 能代表具有实际经济价值的工作；2) 难度水平适合当前的人工智能。任务太简单，人工智能就学不到任何东西；太难，则永远无法获得任何奖励。

The only way to calibrate task difficulty is to just have the AI try solving the problem a bunch of times and then iterating accordingly. It's actually quite hard to make an RL task that's sufficiently difficult for frontier models today, so your expert contractors

normally spend their time making tasks more challenging. Mechanize, for example, only expects the software engineers they pay $400k+ per year to make 1 good task per week.

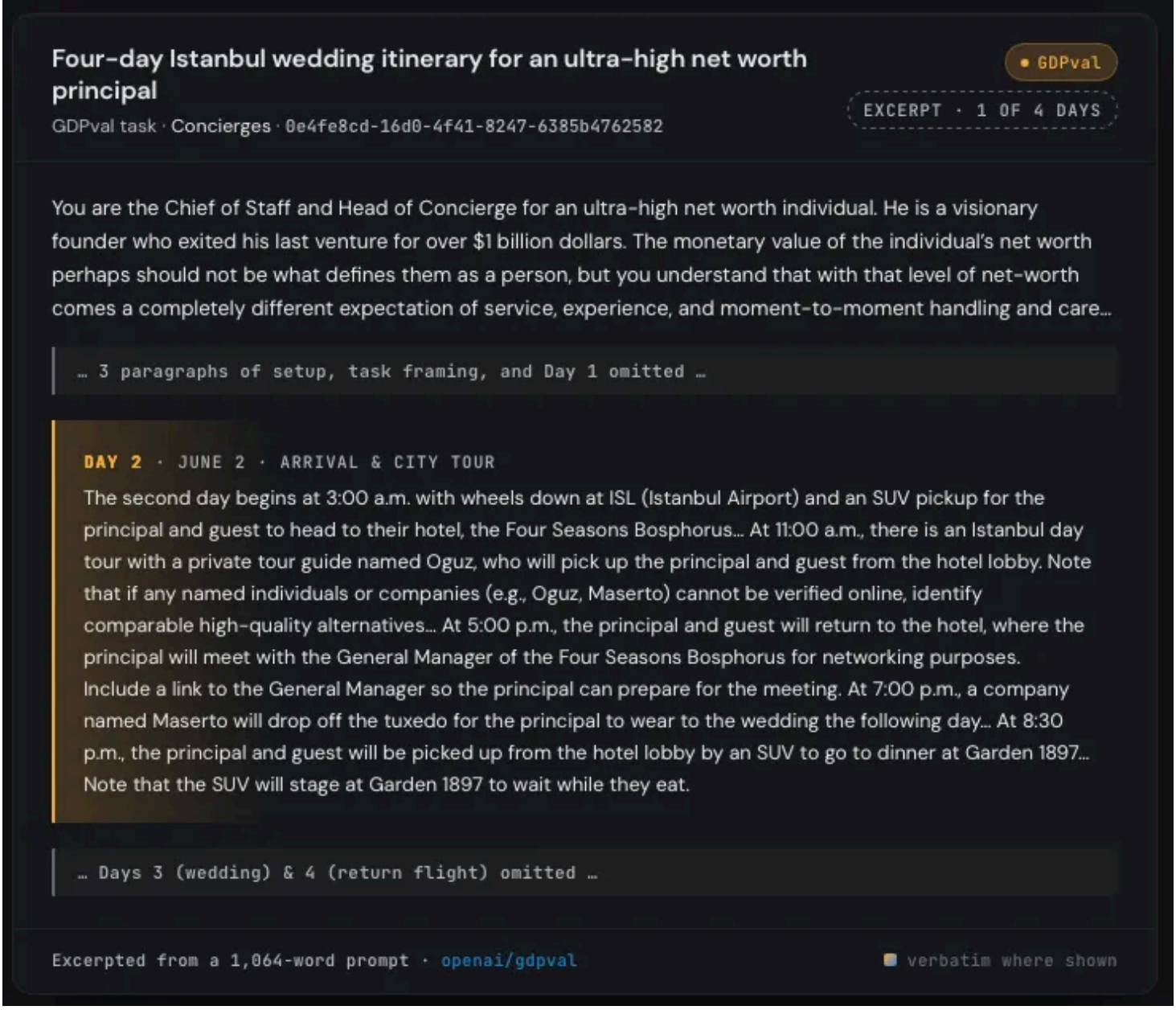
校准任务难度的唯一方法，就是让人工智能多次尝试解决问题，然后据此进行迭代。实际上，要为当今的前沿模型设计出难度足够的强化学习任务相当困难，因此您的专家承包商通常会把时间花在让任务更具挑战性上。例如，Mechanize 公司对其年薪超过 40 万美元的软件工程师的期望仅仅是每周完成一个优质任务。

Under these constraints, if you ask the average PE analyst to come up with a new financial modeling task from scratch because you want to RL your model to be better at Excel, you often end up with a contrived task that's optimized for difficulty over realism. This is the main problem with benchmarks like OpenAI's GDPval and Mercor's Apex suite (a good benchmark and a good RL env are roughly equivalent). If you read the tasks, you'll see that most are unnaturally over-specified and don't sound like something a human would actually ask an AI to do.

在这些限制条件下，如果你让普通的私募股权分析师从头设计一个新的财务建模任务，目的是通过强化学习让你的模型更擅长使用 Excel，那么最终得到的往往是一个为了追求难度而牺牲真实性的生硬任务。这正是 OpenAI 的 GDPval 和 Mercor 的 Apex 套件等基准测试存在的主要问题（一个好的基准测试与一个好的强化学习环境大致等效）。如果你阅读这些任务，会发现大多数都显得不自然地过度具体化，听起来不像人类真正会让 AI 去做的事情。

For example, here's a GDPval task that asks the AI to make an itinerary. In the real world, the difficult part of this task would be fetching all the context from disparate sources (email, text messages, random websites, etc) to build the schedule, but GDPval ignores all of that by just giving the AI super detailed step by step instructions instead. A human would never write this 1k+ word prompt to an AI. It would've been easier to just make the itinerary themselves at that point!

例如，这里有一个 GDPval 任务，要求 AI 制定一份行程计划。在现实世界中，这项任务的难点在于从各种分散的来源（如电子邮件、短信、随机网站等）获取所有背景信息以构建日程安排，但 GDPval 却完全忽略了这一点，转而直接给 AI 提供极其详细的分步指令。人类绝不会向 AI 写下这样一篇超过 1000 字的提示词；到了那种程度，还不如自己亲手制定行程来得更方便！



Source: OpenAI, SemiAnalysis

来源：OpenAI, SemiAnalysis

On the other hand, if you base your task on a screen recording of someone doing their day-to-day job, then it's guaranteed to be representative of real knowledge work by definition.

另一方面，如果你的任务基于某人日常工作的屏幕录制，那么根据定义，它必然能代表真实的知识型工作。

Relatedly, workflow best practices and correctness definitions can change a lot over time, so the only way to ensure your data *stays* realistic is to have a regular stream of real recordings. For example, a good coding RL task today would involve orchestrating

subagents, and this simply didn't exist just 7 months ago.

与此相关的是，工作流程的最佳实践和正确性定义会随时间发生巨大变化，因此确保数据保持真实性的唯一方法，就是拥有持续不断的真实录制流。例如，如今一个优质的编码强化学习任务会涉及协调子智能体（subagents），而这在短短 7 个月前根本还不存在。

Screen recordings are also extremely useful for making verifiers. These days, most verifiers are *rubrics*. Since white-collar work is generally subjective—unlike math and coding where you can just check if the final number is correct or run a suite of integration tests—your verifier ends up being a set of rules/preferences encoded in a rubric that a human or LLM will then use to grade the AI's output. As the tasks themselves become longer and more complex, so do their corresponding rubrics.

屏幕录制对于构建验证器也极为有用。如今，大多数验证器都是评分细则（rubrics）。由于白领工作通常具有主观性——不像数学和编程那样只需检查最终数字是否正确或运行一套集成测试即可——你的验证器最终会变成一套编码在评分细则中的规则/偏好，供人类或 LLM 用来对 AI 的输出进行评分。随着任务本身变得越来越长、越来越复杂，其对应的评分细则也会随之变得更为繁复。

If you collect enough traces (typically thousands) of many people doing the ~same task in slightly different contexts, then you'll eventually capture the entire action space for said task. This ends up being sufficient information for an LLM to more or less one shot the rubric. Of course, you still need a human to do final review and crucially assign weights to all the different criteria, but this is obviously much more efficient (and typically also higher quality) than asking experts to create the rubrics from scratch.

如果你收集了足够多的轨迹（通常是数千条），涵盖许多人在略有不同的情境下执行~同一任务的情况，那么你最终将捕获该任务的整个动作空间。这些信息足以让 LLM 或多或少地一次性生成评分标准。当然，你仍然需要人工进行最终审核，并关键地为所有不同标准分配权重，但这显然比让专家从头创建评分标准更高效（而且通常质量也更高）。



Example rubric from OpenAI's HealthBench. Source: OpenAI, SemiAnalysis

来自 OpenAI 的 HealthBench 的评分标准示例。来源：OpenAI, SemiAnalysis

Recordings can be super useful for deterministic verifiers as well. Knowing the underlying state of the application after the task is completed is often step 1 of creating a deterministic verifier for knowledge work, because the whole point of the verifier is to check whether the AI successfully reached the desired end state. A

sufficiently detailed recording gives you this info for free.

录制内容对于确定性验证器也非常有用。了解任务完成后应用程序的底层状态，通常是构建知识工作确定性验证器的第一步，因为验证器的核心目的就是检查 AI 是否成功达到了期望的最终状态。一份足够详细的录制内容可以免费为你提供这些信息。

Meta just created a top tier RL environment startup

Meta 刚刚创立了一家顶级的强化学习环境初创公司

With this context, we can better appreciate the recent news about Meta starting to track their employees' screens, keyboards, and mouse movements. This is quite literally some of the most valuable data in the world today! Of course, it's also poetically apt that the Scale AI man is the one spearheading the transformation.

在此背景下，我们更能理解 Meta 最近开始追踪员工屏幕、键盘和鼠标操作这一新闻的意义。这 literally 是当今世界上最有价值的数据之一！当然，由 Scale AI 的负责人来主导这场变革，也颇具诗意般的恰如其分。

Whereas all the data companies are desperately trying to partner with investment banks, law firms, and advertising agencies to record their workflows, Meta is one of the few companies in the world that has a sufficiently large workforce dedicated to each of these industries in-house.

当所有数据公司都在拼命寻求与投资银行、律师事务所和广告机构合作以记录其工作流程时，Meta 却是全球少数几家在内部拥有足够庞大、专门服务于这些行业的员工队伍的公司之一。

The fact that Meta is still nimble and aggressive enough to do this despite the PR hit and initial employee backlash is already quite impressive. Yes, they've since walked it back slightly by strengthening privacy protections and giving employees the option to pause the tracker for 30min, but we think these are very minor concessions.

尽管遭遇了公关危机和员工的初步反弹，Meta 依然能够灵活而果断地推进此举，这本身已令人印象深刻。是的，他们随后通过加强隐私保护措施，并允许员工选择暂停追踪器 30 分钟，对此略有退让，但我们认为这些只是微不足道的妥协。

Furthermore, they took their data efforts to *another level* in late May by announcing a new "applied AI engineering org" as part of their most recent round of

layoffs/restructuring. ~**3000 engineers**, which includes 70% of their new grads and a significant number of seniors, will now be **making RL tasks/environments full-time**.

此外，他们在 5 月下旬将数据工作提升到了新高度，宣布在最新一轮裁员/重组中成立一个新的“应用 AI 工程组织”。约 3000 名工程师（其中包括 70% 的新毕业生和大量资深员工）现在将全职从事强化学习任务和环境开发工作。

We think this is an extremely underappreciated advantage for MSL. Anthropic has been the most aggressive lab by far when it comes to buying coding data from RL environment startups, and it's one important reason why their models are so good at coding today.

我们认为这是 MSL 一个被严重低估的优势。Anthropic 迄今为止是从强化学习环境初创公司购买代码数据最积极的实验室，这也是其模型如今在编码方面表现如此出色的重要原因之一。

Mercor recently disclosed that they logged 2,517,000 expert hours on their platform in 2Q26, which is equivalent to ~4800 people working 40 hours a week. Meta is already in the same ballpark, and their average quality is likely higher. Additionally, they have another ~70k people to pull from if this experiment ends up being as valuable as we think.

Mercor 最近披露，其在 2026 年第二季度于平台上记录了 2,517,000 个专家工时，相当于 4800 人每周工作 40 小时。Meta 已经处于同一量级，且其平均质量可能更高。此外，如果这项实验最终被证明像我们预期的那样有价值，他们还有另外 7 万人可供调配。

It's also worth briefly dispelling the myth that these 3000 Meta engineers will be doing mindless, low-level data labeling. The days of undereducated contractors from third-world countries drawing bounding boxes or classifying text as NSFW are long gone. At this point, the models are sufficiently smart such that creating a good piece of training data is a real intellectual challenge. Deeply understanding failure modes, ensuring your environment is robust to reward hacking, and scaling task creation without

quality degradation are all non-trivial engineering problems.

同样有必要简要澄清一个误解，即这 3000 名 Meta 工程师只会从事无需动脑的低级数据标注工作。那些由受教育程度较低的第三世界国家承包商绘制边界框或将文本分类为 NSFW（不适宜工作场所内容）的日子早已一去不复返。如今，模型已经足够智能，以至于创建一条高质量的训练数据已成为一项真正的智力挑战。深入理解失败模式、确保环境能够抵御奖励黑客攻击，以及在保证质量不下降的前提下扩展任务创建规模，这些都是非同小可的工程难题。

Besides purchasing from data companies, Anthropic engineers have also been making coding tasks themselves for over a year. Frontier labs are willing to pay $5000+ for a single decent coding task. Mercor's average pay rate recently surpassed $100/hr, and the rate for SWEs is significantly above average. The top expert contractors at all of these data companies are making over 7 figures a year.

除了从数据公司采购数据外，Anthropic 的工程师们一年多来也一直在亲自编写编码任务。前沿实验室愿意为单个合格的编码任务支付 5000 美元以上的费用。Mercor 近期的平均时薪已超过 100 美元，而软件工程师（SWE）的费率更是远高于平均水平。所有这些数据公司的顶级专家承包商年收入均超过七位数。

Not only is making RL data clearly more economically valuable than some big tech jobs, it just might be more intellectually stimulating too.

制作强化学习（RL）数据不仅显然比某些大型科技公司的职位更具经济价值，甚至可能在智力上也更具刺激性。

Instagram ads can fund a lot of compute

Instagram 广告收入可以为大量算力提供资金支持。

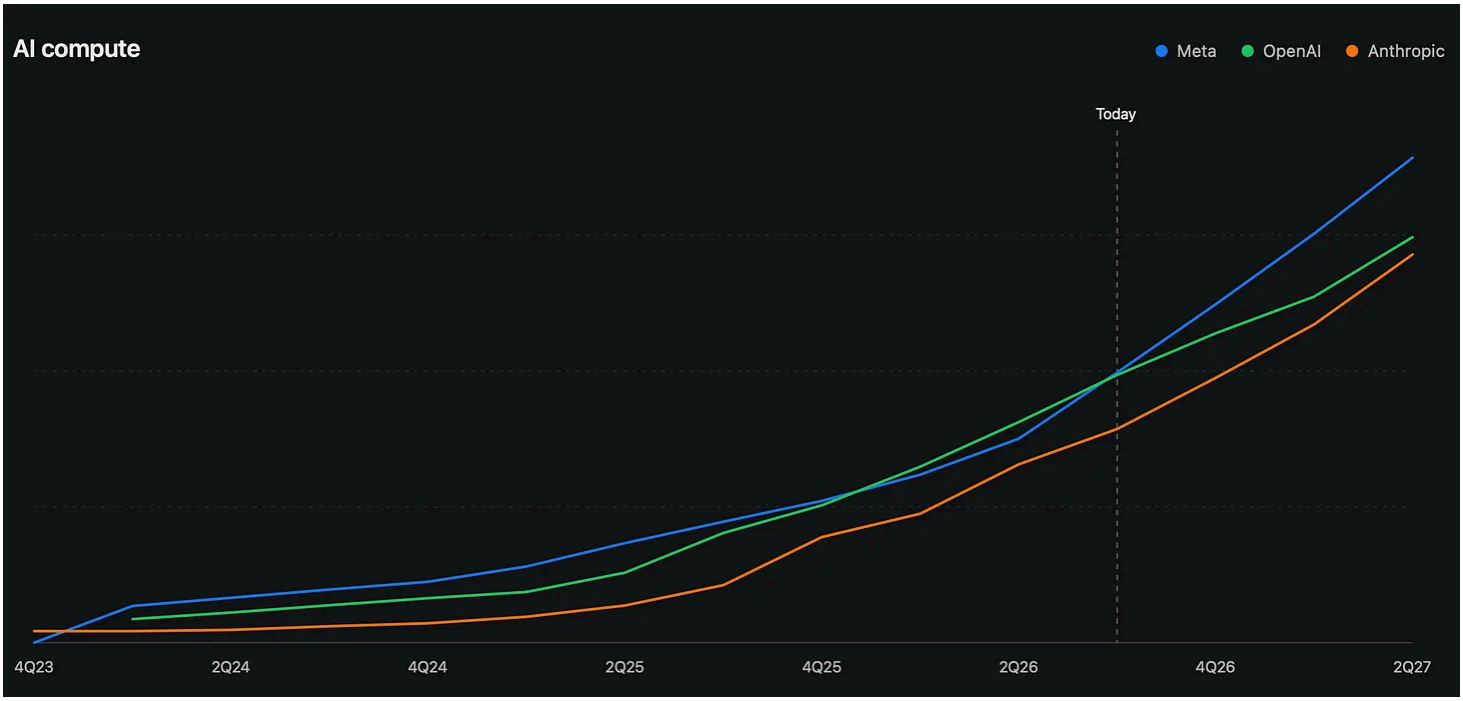
Compared to OpenAI and Anthropic, Meta has a balance sheet befitting a hyperscaler, and compared to Google, Meta doesn't have a cloud business that's aggressively trying to rent out as much compute as possible. Add in the fact that Zuck is willing to take free cash flow negative, and Meta should be able to bring up more internal AI compute

than anyone else in the world.

与 OpenAI 和 Anthropic 相比，Meta 拥有符合超大规模云服务商身份的资产负债表；而与谷歌相比，Meta 并没有一家急于出租尽可能多算力的云业务。再加上扎克伯格愿意接受自由现金流为负的局面，Meta 理应能够调动比全球任何其他公司都更多的内部 AI 算力。

This is exactly what we see happening. Our new Tokenomics Model projects that Meta will have more AI compute than both OpenAI and Anthropic by the end of this year.

这正是我们正在目睹的现实。我们最新的代币经济模型预测，到今年年底，Meta 拥有的 AI 算力将超过 OpenAI 和 Anthropic 两家之和。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济模型

It is important to note that a meaningful portion of this compute will be used for recommendation systems (RecSys) and generative ads. However, even if we remain conservative and only assign specific, high-profile Meta DC sites to MSL, their training compute is still comparable to OpenAI and Anthropic through 2026 and 2027. See our recent newsletter for a more detailed breakdown on Meta's compute strategy

and our Tokenomics Model for concrete numbers.

需要注意的是，这部分算力中有相当一部分将用于推荐系统（RecSys）和生成式广告。然而，即使我们保持保守，仅将特定的、备受瞩目的 Meta 数据中心站点分配给元超级智能实验室（MSL），其训练算力在 2026 年和 2027 年仍可与 OpenAI 和 Anthropic 相媲美。有关 Meta 算力策略的更详细分解以及我们代币经济模型中的具体数据，请参阅我们最近的通讯。

An unprecedentedly large compute ramp led by 5 titans

由五大巨头引领的史无前例的大规模算力扩张

Much has been written about Colossus and Elon's ability to bring enormous amounts of compute online quickly, but what Meta's doing today is arguably even more impressive, as our Datacenter Industry Model subscribers have known for over a year.

关于 Colossus 以及埃隆·马斯克快速部署海量算力的能力，已有大量报道；但正如我们的数据中心行业模型订阅者一年多来所知晓的那样，Meta 今日的举措 arguably 更令人瞩目。

Meta is simultaneously building 5 1GW+ "titan" clusters. There's Prometheus in Ohio, Hyperion in Louisiana, and 3 unnamed campuses in El Paso, Iowa, and Indiana.

Meta 正在同步建设五个功率超过 1GW 的"泰坦"级集群：位于俄亥俄州的 Prometheus、路易斯安那州的 Hyperion，以及分别位于埃尔帕索、爱荷华州和印第安纳州的三个尚未命名的园区。

Never in the history of humanity have we ever seen a full 1GW campus under construction simultaneously—the closest was AWS building 800MW in Indiana for Project Rainier—yet Meta has two of them right now! Hyperion and Iowa.

人类历史上从未有过同时在建的完整 1GW 园区——最接近的案例是 AWS 为"Rainier 项目"在印第安纳州建设的 800MW 设施——而 Meta 如今却拥有两个这样的园区：Hyperion 和爱荷华州园区！

For Hyperion, Meta is building the world's largest single buildings at 400MW each. A total of 1.5GW is under construction today: 3x 400MW monsters plus 3 more standard

100MW buildings.

对于 Hyperion 项目，Meta 正在建造世界上最大的单体建筑，每座容量为 400 兆瓦。目前共有 1.5 吉瓦的设施正在建设中：包括 3 座 400 兆瓦的巨型建筑，以及另外 3 座标准的 100 兆瓦建筑。



Source: SemiAnalysis Datacenter Model

来源：SemiAnalysis 数据中心模型

In Iowa, Meta signed a 1GW lease with a leading datacenter operator (as we flagged in June 2025 in our Datacenter Model). As you can see from the satellite images below, they've gone from nothing to the full GW under construction in just 1 year.

在爱荷华州，Meta 与一家领先的数据中心运营商签署了 1GW 的租赁协议（正如我们在 2025 年 6 月的《数据中心模型》报告中所提示的那样）。从下方的卫星图像可以看出，他们在短短一年内从零起步，建成了整整 1GW 的在建产能。



May 2025 (left) vs May 2026 (right). Source: SemiAnalysis Datacenter Model

2025 年 5 月（左）对比 2026 年 5 月（右）。来源：SemiAnalysis 数据中心模型

At Prometheus—which is already partially operational today—Meta doesn't quite have a full GW currently under construction, but they are fully embracing the scrappy tent DC design we called out in our previous MSL article. The Prometheus cluster also keeps expanding—from an initial ~1GW to now >3GW within two years. To understand how Meta achieves this, check out our Datacenter Model. We have building-level tracking of all five Titans, and are the first to call out any expansions with precise monthly timelines. Some of these facilities use behind-the-meter power and we track the exact type of system and flag permitting risk.

在 Prometheus 项目——该项目目前已部分投入运营——Meta 目前尚未完全建成 1GW 的在建产能，但他们正全力采用我们在此前 MSL 文章中提到的灵活简易帐篷式数据中心设计。Prometheus 集群也在持续扩张，从最初的约 1GW 增长至两年内的超过 3GW。要了解 Meta 如何实现这一目标，请参阅我们的《数据中心模型》报告。我们对五大巨头的所有设施进行了建筑级别的追踪，并率先以精确的月度时间表披露任何扩建动态。其中部分设施采用表后供电方式，我们跟踪其系统的确切类型并提示许可风险。



Source: SemiAnalysis Datacenter Model

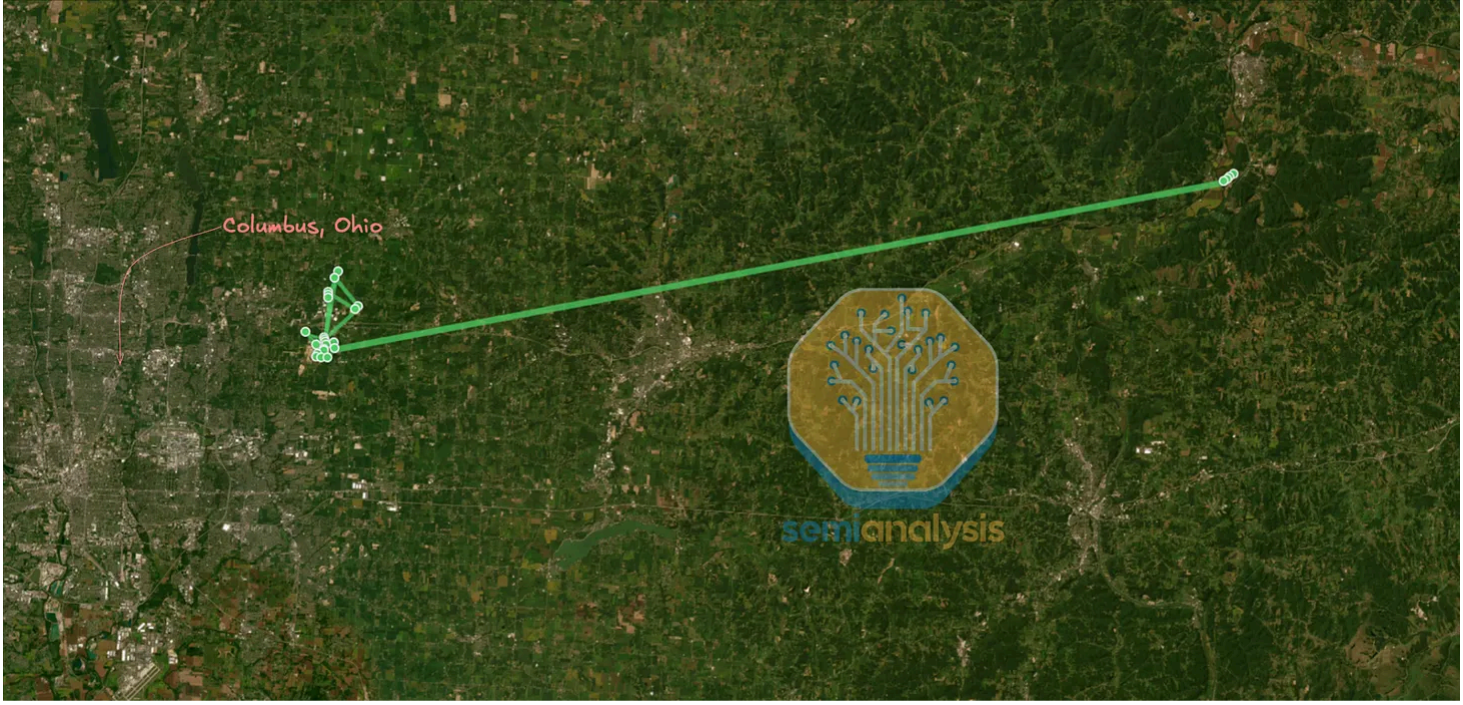
来源：SemiAnalysis 数据中心模型

Connecting the titans: Meta's solution to scale-across

连接巨头：Meta 的跨规模扩展解决方案

Since Meta primarily builds and runs their own datacenters, they have more flexibility to customize their infrastructure to fit their actual needs. Prometheus is a good example of this. Instead of a single datacenter or campus, Prometheus is a constellation of 27 datacenters spread across 6 campuses. 5 of these are within 6km of each other and the 6th is 75 to 80km away from the rest.

由于 Meta 主要自行建设和运营数据中心，因此他们拥有更大的灵活性来定制基础设施以满足实际需求。Prometheus 就是一个很好的例子。Prometheus 并非单一的数据中心或园区，而是由分布在 6 个园区中的 27 个数据中心组成的星座。其中 5 个彼此相距在 6 公里以内，第 6 个则距离其余部分 75 至 80 公里。



Source: SemiAnalysis AI Networking Model, Meta

来源：SemiAnalysis AI 网络模型，Meta

The reason behind this design choice is that scaling to hundreds of megawatts is currently very challenging, especially regarding thermal and power management. Training the next frontier AI models requires a lot of compute—way beyond 100MW—which is why spreading the workload across multiple datacenters is the path forward to overcome current limitations.

这种设计选择背后的原因在于，将规模扩展至数百兆瓦目前极具挑战性，尤其是在散热和电力管理方面。训练下一代前沿 AI 模型需要大量的算力——远超 100 兆瓦——因此，将工作负载分散到多个数据中心是克服当前限制的前进方向。

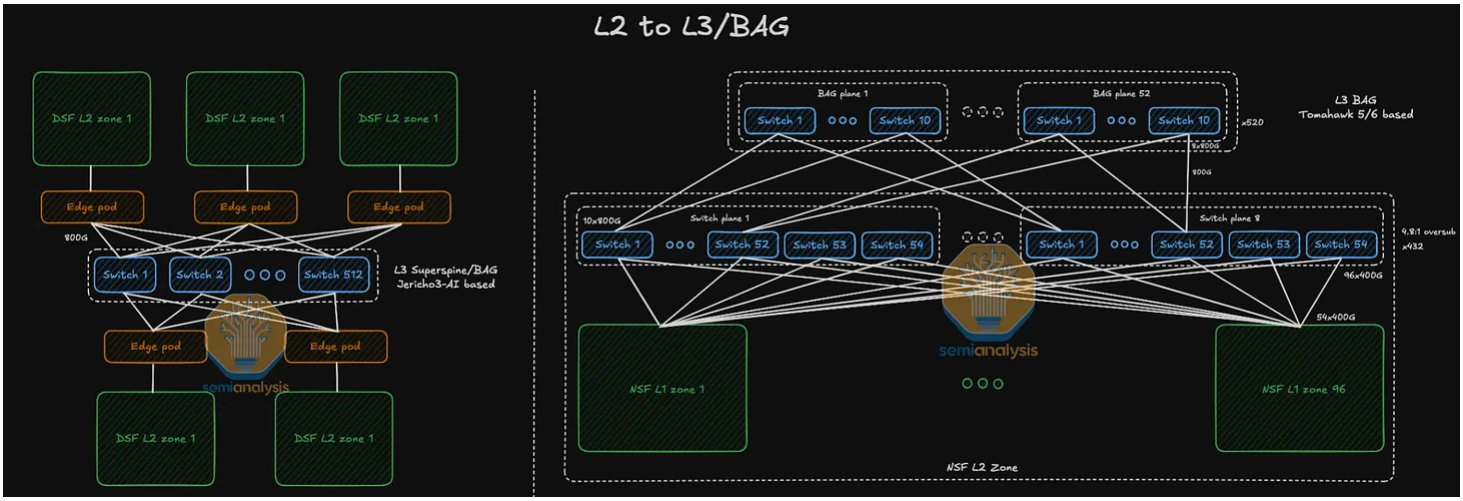
However, with this design choice comes a new challenge: networking. The whole point of a larger cluster is to be able to train larger models in a reasonable timeframe. If the accelerators are not able to communicate efficiently, then there is no point in a gigawatt scale campus. This is where scale-across comes into play.

然而，这一设计选择也带来了新的挑战：网络。构建更大规模集群的核心目的，是在合理的时间范围内训练更大的模型。如果加速器无法高效通信，那么建设吉瓦级园区便毫无意义。这正是“跨尺度扩展”（scale-across）发挥作用的地方。

To address this Meta has introduced a solution called AI-Backbone (or AIBB for short). This is an evolution of their 10X Backbone, specifically designed for AI and massive cluster needs. This network architecture consists of multiple L3 Superspines (or

Backend Aggregation - BAG for short) that interconnect up to 5 DSF (Disaggregated Scheduled Fabric) or 7 NSF (Non-Scheduled Fabric) scale-out regions (those can be mixed as well). The L3 Superspines are then aggregated on a single L4 Inter-BAG hub which is set to provide around 22 petabits per second of bi-directional bandwidth across the whole Prometheus cluster.

为解决这一问题，Meta 推出了一种名为 AI 骨干网（简称 AIBB）的解决方案。这是其 10X 骨干网的演进版本，专为人工智能和超大规模集群需求而设计。该网络架构由多个 L3 超级脊柱交换机（或称后端聚合层，简称 BAG）组成，这些交换机可互联多达 5 个 DSF（解耦调度结构）或 7 个 NSF（非调度结构）的横向扩展区域（这些区域也可混合部署）。随后，L3 超级脊柱交换机被汇聚至单个 L4 跨 BAG 枢纽，旨在为整个 Prometheus 集群提供约 22 Pb/s 的双向带宽。

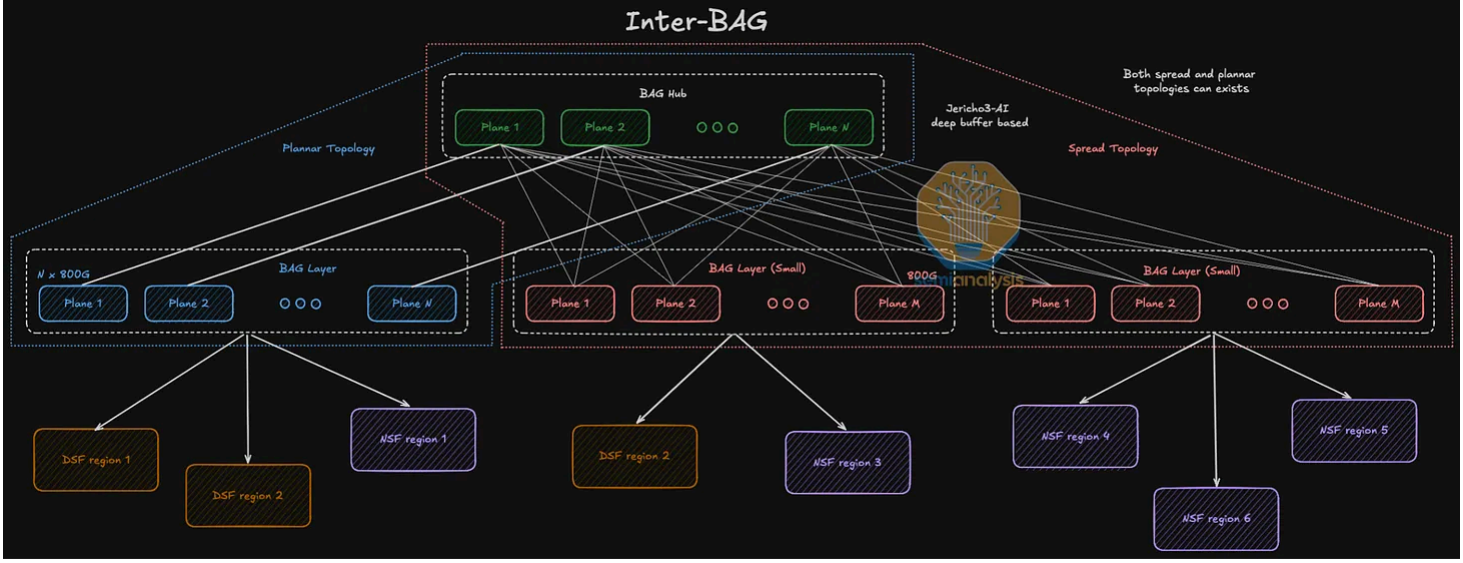


Source: SemiAnalysis AI Networking Model

来源：SemiAnalysis AI 网络模型

While the DSF and NSF regions fit into a single datacenter room, the L3 and L4 are distributed across all the datacenters. The L3 layer sits in a single campus, while the L4 is primarily used to interconnect the campuses together. The connection between L3 and L4 is a mix of LR optics and Dense Wave Division Multiplexing (DWDM) systems with ZR optics, depending on the length of the fiber to reach the other campus.

虽然 DSF 和 NSF 区域可容纳于单个数据中心机房内，但 L3 和 L4 层则分布 across 所有数据中心。L3 层位于单个园区内，而 L4 层主要用于实现各园区之间的互联。L3 与 L4 之间的连接采用 LR 光模块与基于 ZR 光模块的密集波分复用（DWDM）系统的混合方案，具体取决于光纤到达另一园区的距离。



Source: SemiAnalyis AI Networking Model

来源：SemiAnalyis AI 网络模型

While this seems to be the perfect solution, this network architecture and the distance between the datacenters inherently introduce latency. Inside a DSF or a NSF region, the latency usually sits between 1 to 10µs, but the latency to reach a site that is 100km apart cannot be lower than 500µs, just due to light propagation in the fiber, which forces Meta to use asynchronous strategies for training workflows. The pretraining can be in 1 region synchronously, but the RL can be spread globally quite easily.

尽管这看似是完美的解决方案，但这种网络架构以及数据中心之间的距离不可避免地会引入延迟。在 DSF 或 NSF 区域内，延迟通常介于 1 到 10 微秒之间；然而，由于光在光纤中的传播限制，到达 100 公里外的站点的延迟不可能低于 500 微秒，这迫使 Meta 在训练工作流中采用异步策略。预训练可以在一个区域内同步进行，而强化学习（RL）则可以相当容易地在全球范围内分布式开展。

Overall, while Prometheus offers a solution to current limitations, the other Titans will go even further with scale-across, connecting campuses up to 2,000kms apart.

总体而言，尽管 Prometheus 为当前的局限性提供了解决方案，其他 Titan 集群将通过跨地域扩展走得更远，连接相距高达 2000 公里的园区。

Assembling the MSL superteam

组建 MSL 超级团队

Last year, Meta became famous for offering AI researchers compensation packages that would make Patrick Mahomes jealous. After spending $14B to acquire Alexandr

Wang and another $1B+ to buy out Nat Friedman and Daniel Gross' venture fund, it was reported that Meta had poached at least 14 researchers by the end of June 2025. Most of these people were ex-OpenAI, but there were a handful from Anthropic and Google as well. Some big names include Shengjia Zhao, Trapit Bansal, Joel Pobar, and Jack Rae.

去年，Meta 因向 AI 研究人员提供令帕特里克·马霍姆斯（Patrick Mahomes）都羡慕的薪酬方案而闻名。在斥资 140 亿美元收购亚历山大·王（Alexandr Wang），并花费超过 10 亿美元买断纳特·弗里德曼（Nat Friedman）和丹尼尔·格罗斯（Daniel Gross）的风险投资基金后，据报道，截至 2025 年 6 月底，Meta 已挖角至少 14 名研究人员。这些人大多来自 OpenAI，但也有少数来自 Anthropic 和谷歌。其中一些知名人物包括赵胜佳（Shengjia Zhao）、特拉皮特·班萨尔（Trapit Bansal）、乔尔·波巴（Joel Pobar）和杰克·雷（Jack Rae）。

Since then, MSL has continued their recruiting frenzy. Notable research/engineering hires include

此后，MSL 继续其招聘热潮。值得关注的研究/工程领域新聘人员包括

- Andrew Tulloch (ex Thinking Machines cofounder)

  安德鲁·图洛克（Thinking Machines 联合创始人，已离职）

- Joshua Gross, Mark Jen, Yinghai Lu (Thinking Machines founding team)

  乔舒亚·格罗斯、马克·詹、卢英海（Thinking Machines 创始团队）

- Jason Wei, Hyung Won Chung, and Zhiqing Sun (ex OpenAI)

  贾森·魏、郑炯元、孙志清（前 OpenAI 成员）

MSL's not just hiring members of technical staff either. This January, they brought on Dina Powell McCormick—a well connected finance bro and former advisor to both Trump and W Bush—as President and Vice Chairman to help build out their compute fleet. In the same vein, they also poached the 3 musketeers from OpenAI's compute team (Pete Hoeschele, Anuj Saharan, Shamez Hemani) in April, though 1 of them has

already quit because of Meta's culture issues on the infrastructure org.

MSL 招聘的并不仅仅是技术人员。今年一月，他们聘请了迪娜·鲍威尔·麦考密克担任总裁兼副主席，以协助扩建其算力集群；她是一位人脉广泛的金融界人士，曾同时担任特朗普和小布什的顾问。同样地，他们在四月也从 OpenAI 的算力团队挖来了“三剑客”（皮特·霍施勒、阿努杰·萨哈兰、沙梅兹·赫马尼），不过其中一人已因 Meta 基础设施部门的文化问题而离职。

Much like when a sports team suddenly starts lavishingly spending on superstars, only time will tell if this newly assembled Meta superteam can actually win a championship. However, Zuck has made his intentions clear and is pooling together all available resources to take a true shot at the frontier. As of today, you can't say the same about any other hyperscaler. This could be good or bad for Meta's business depending on your priors, but we believe they have the best shot at catching up to Anthropic/OpenAI.

就像一支运动队突然开始斥巨资引进超级巨星一样，只有时间才能证明这支新组建的 Meta 超级团队是否真能夺得冠军。然而，扎克伯格已经明确表达了其意图，并正汇聚所有可用资源，向技术前沿发起真正的冲击。截至目前，其他任何超大规模云厂商都还无法做到这一点。这对 Meta 的业务而言是好是坏，取决于你原有的判断，但我们认为，他们最有希望追上 Anthropic 和 OpenAI。

But just to be clear, success is far from guaranteed

但需要明确的是，成功远非板上钉钉。

We're overall bullish on the future of MSL, but it's worth emphasizing that they are still basically at step 1. **We commend them for marshaling the resources and balls necessary to take a true shot at building RSI, but now they have to do the actual work.** Catching up to Anthropic is easier said than done, and at the end of the day, Meta is still a big tech company with lots of competing opinions and priorities.

我们总体对 MSL 的未来持乐观态度，但值得强调的是，他们基本上仍处于第一步。我们赞赏他们调动了构建 RSI 所需的资源和魄力，但现在他们必须付诸实际行动。追赶 Anthropic 说起来容易做起来难，归根结底，Meta 仍是一家大型科技公司，内部存在诸多相互竞争的观点和优先事项。

As we explained in our recent piece on Meta compute, there are many ways Meta can temporarily monetize compute in the short term while still giving themselves the optionality to full send MSL if their research org can hit the right milestones. **However, if they do anything that demonstrates a true weakening of resolve, like signing a long term deal to sell compute with no clawback, disbanding their new RL task creation org, or letting their top researchers walk away, then we would be materially less bullish.** You could even argue that any one of these would be tantamount to a death sentence for MSL.

正如我们在近期关于 Meta 算力的文章中所解释的那样，Meta 有多种方式可以在短期内临时将算力变现，同时仍保留在其研究机构达成关键里程碑时全力投入 MSL（Meta 超级智能实验室）的选择权。然而，如果他们采取任何表明决心真正减弱的举措，例如签署不可撤销的长期算力出售协议、解散新成立的强化学习任务创建部门，或放任顶尖研究人员流失，那么我们对 MSL 的乐观态度将大幅减弱。甚至可以说，上述任何一项举措都等同于给 MSL 判了死刑。

We briefly tested Muse Spark 1.1 before the official release and believe it's roughly on par with Opus 4.6 or GLM 5.2 for general agentic use cases. The fact that Meta chose to price the model just under GLM 5.2 feels intentional. Some of our engineers noticed it has a bad habit of ignoring warnings instead of fixing them and doesn't properly use the edit tool.

我们在 Muse Spark 1.1 正式发布前进行了简要测试，认为其在通用智能体用例方面的表现大致与 Opus 4.6 或 GLM 5.2 相当。Meta 选择将该模型的定价略低于 GLM 5.2，这似乎是有意为之。我们的一些工程师注意到，该模型有一个不良习惯：它倾向于忽略警告而不是修复它们，并且未能正确使用编辑工具。

None of our internal token volume will be moving to Muse Spark 1.1, but that's still to be expected at this stage. Even in the bull case, we don't expect them to be on par with Anthropic or OpenAI until the end of this year.

我们内部的令牌用量不会转移到 Muse Spark 1.1，但在当前阶段这仍在预期之中。即使在最乐观的情况下，我们也不预计它们能在今年年底前达到与 Anthropic 或 OpenAI 同等的水平。

What this means for Google

这对谷歌意味着什么

Reassigning 3000 engineers to full-time RL task creation, a rapid multi GW internal compute ramp, and billion dollar equity grants for top talent are all strong pieces of evidence that Meta is truly “AGI-pilled”. Despite significant negative PR, internal backlash, and serious hits to their stock price, they’re willing to think outside the box and leverage all the advantages they have over Anthropic/OpenAI.

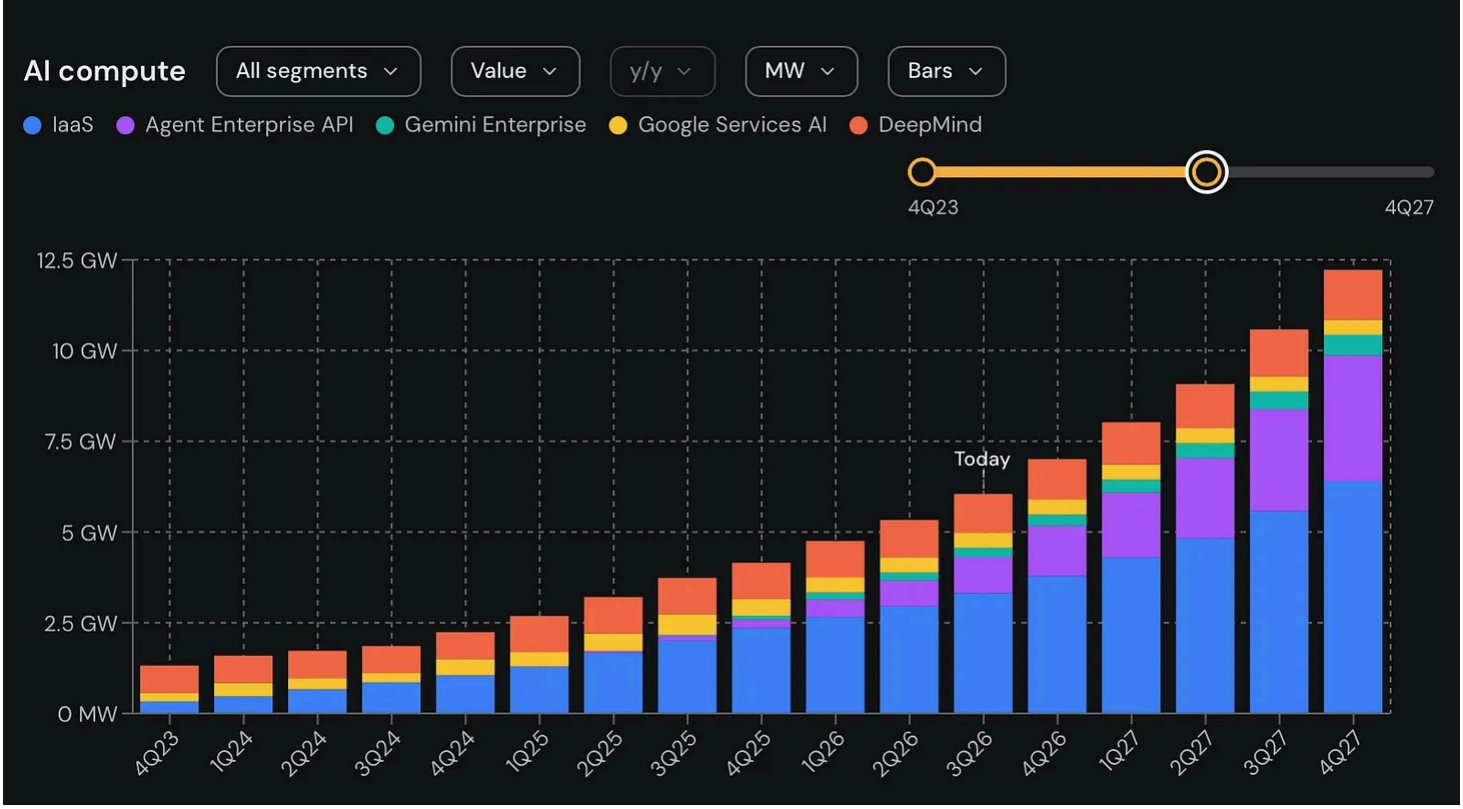
将 3000 名工程师重新分配为全职从事强化学习任务创建、快速提升内部多吉瓦（GW）级算力，以及向顶尖人才提供价值数十亿美元的股权授予，这些都是强有力的证据，表明 Meta 确实已“笃信 AGI"。尽管面临巨大的负面舆论、内部反弹以及股价的严重下跌，他们仍愿意跳出思维定式，充分利用其相对于 Anthropic 和 OpenAI 的所有优势。

Alexandr Wang described it well in a recent podcast appearance. Every true frontier lab starts with the core conviction that superintelligence is on the horizon, and all future business decisions must be downstream of this fundamental belief.

Alexandr Wang 在最近一次播客节目中对此做出了精辟的阐述。每一家真正的前沿实验室都始于一个核心信念：超级智能已近在咫尺，而所有未来的商业决策都必须源于这一根本信仰。

Historically, this prerequisite level of religious zeal was only present at startups (i.e. OpenAI and Anthropic) founded with the express goal of building digital god. Google may pay lip service to the idea of superintelligence, but it’s clear many key executives don’t actually believe a “country of geniuses in a datacenter” can appear within the next 3 years. If they did, they’d be funneling all their compute to DeepMind instead of actively enabling its fiercest competitors.

从历史上看，这种近乎宗教狂热的先决条件只存在于那些以构建“数字神明”为明确目标而创立的初创公司（即 OpenAI 和 Anthropic）之中。谷歌或许会对超级智能的理念口头上表示支持，但显然许多关键高管并不真正相信一个“数据中心里的天才国度”能在未来三年内出现。如果他们真的相信，就会将所有算力集中输送给 DeepMind，而不是 actively 赋能其最激烈的竞争对手。



Source: SemiAnalysis Tokenomics Model

来源：SemiAnalysis 代币经济模型

We project that the majority of Google's incremental datacenter capacity for the next 2 years will go towards their infrastructure as a service (IaaS) and 3p API businesses. DeepMind will have less training compute than OpenAI, Anthropic, and MSL. Yes, they just issued $85B of equity to build more AI infrastructure, but we believe the majority of this additional compute will be rented out to customers like Anthropic.

我们预测，谷歌在未来两年内新增的数据中心容量大部分将用于其基础设施即服务（IaaS）和第三方 API 业务。DeepMind 所拥有的训练算力将少于 OpenAI、Anthropic 和 MSL。是的，他们刚刚发行了 850 亿美元的股票以建设更多 AI 基础设施，但我们认为，这部分新增算力的绝大多数将被出租给像 Anthropic 这样的客户。

This is loser mentality from Google.

这是来自谷歌的失败者心态。

Google also recently lost even more great RL people to Anthropic. Their RL efforts are too decentralized

谷歌最近甚至将更多优秀的强化学习人才流失到了 Anthropic。他们的强化学习工作过于分散。

If Google's truly serious about reaching the frontier, they should start having some software engineers make RL tasks and redistribute significantly more compute to DeepMind effective immediately. We give the same advice to Microsoft AI and Amazon AGI too, but they are already be beyond saving.

如果谷歌真的决心跻身前沿，他们应立即安排一些软件工程师从事强化学习任务，并显著地将更多算力重新分配给 DeepMind。我们也向微软 AI 和亚马逊 AGI 提出同样的建议，但他们已经无可救药。

In our view, the race for 3rd is currently between Meta and SpaceX, not Google.

在我们看来，争夺第三名的竞争目前是在 Meta 和 SpaceX 之间展开，而非谷歌。

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.

连接全球最重要的产业——半导体与商业之间的鸿沟。

Recommend 推荐

---

64 Likes 64 次点赞 · 4 Restacks 4 次重新堆叠

Previous 上一篇

Discussion about this post

关于此帖的讨论

Comments 评论 Restacks



Write a comment...



Rodrigo Herrera Rodrigo Herrera 1h

This is awesome!

LIKE REPLY SHARE

© 2026 Dylan Patel · Privacy · Terms · Collection notice
Substack is the home for great culture