The Future of Meta Superintelligence: A 1 Year Progress Update

The Future of Meta Superintelligence: A 1 Year Progress Update 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

JUL 9 https://newsletter.semianalysis.com/p/the-future-of-meta-superintelligence

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.

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 5.6 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.

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.

QQ:1413221435
QQ:1413221435

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%

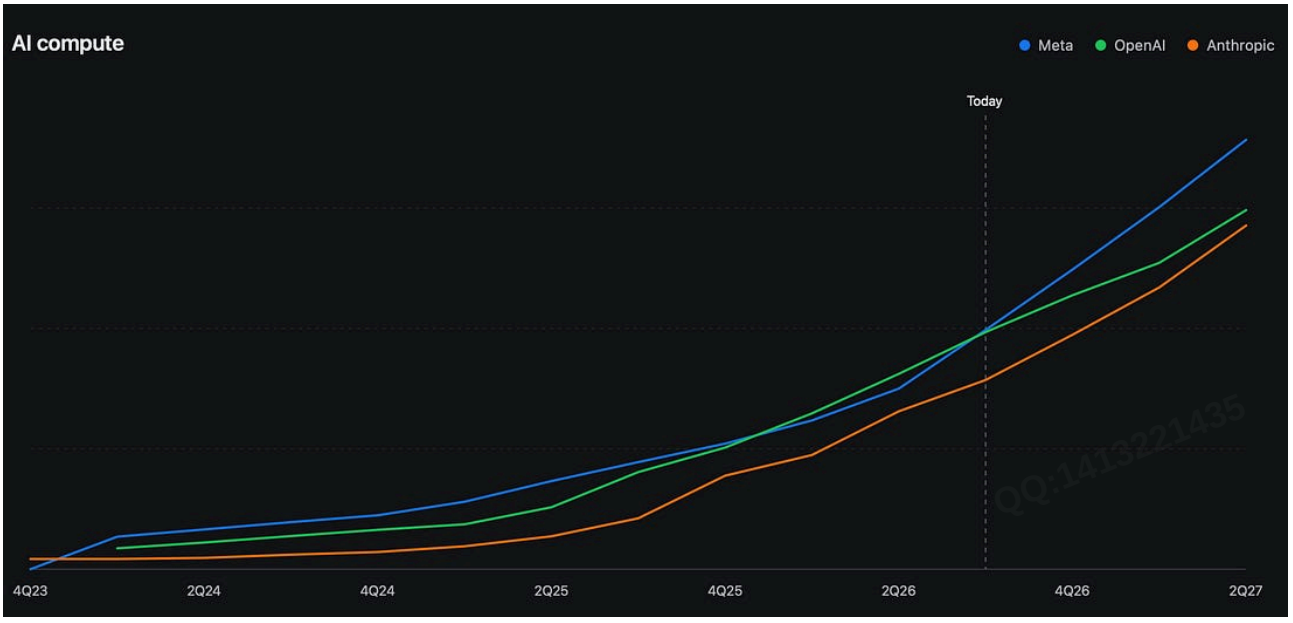
Showing 7 benchmarks of 28 where all selected models have a reported score. Vendor-reported numbers — methodology varies across benchmarks.

Source: SemiAnalysis Tokenomics Model

QQ:1413221435

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.

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.



Source: SemiAnalysis Tokenomics Model

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.

Data is the new oil (for real this time)

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

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.

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.

QQ:1413221435

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

QQ:1413221435

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.

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:

> 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.

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.

QQ:1413221435

Screen recordings are extremely valuable for making RL tasks

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.

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.

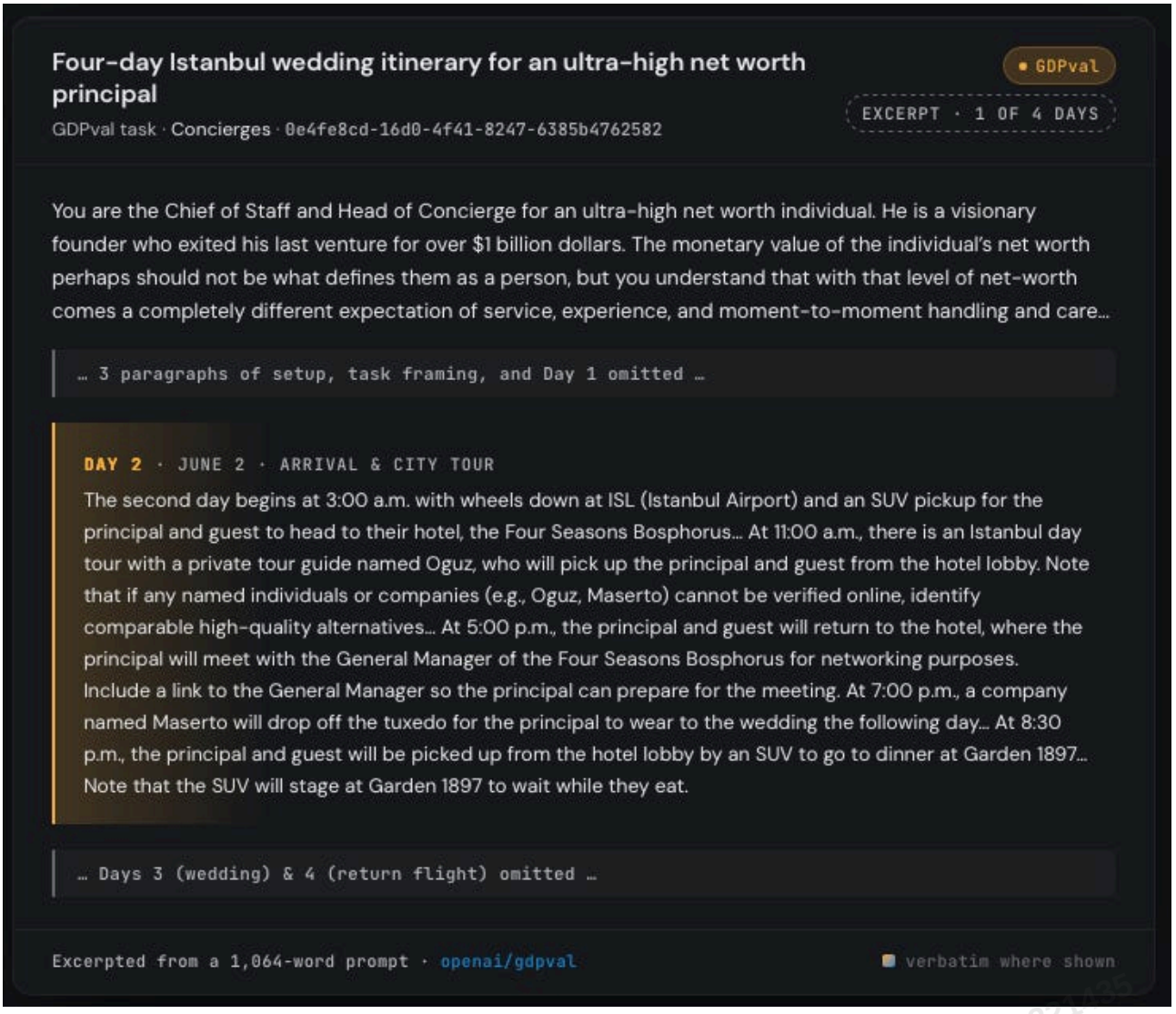
The first benefit is realism. RL data is only good if it's 1) representative of real economically valuable work and 2) at the right difficulty level for the AI. Too easy and there's nothing for the AI to learn, too hard and it'll never achieve any reward.

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.

QQ:1413221435

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.

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!



Source: OpenAI, SemiAnalysis

QQ:1413221435

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.

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.

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.

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.

QQ:1413221435

QQ:1413221435



Example rubric from OpenAI's HealthBench. Source: 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.

Meta just created a top tier RL environment startup

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.

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.

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.

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**.

QQ:1413221435

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.

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.

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.

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.

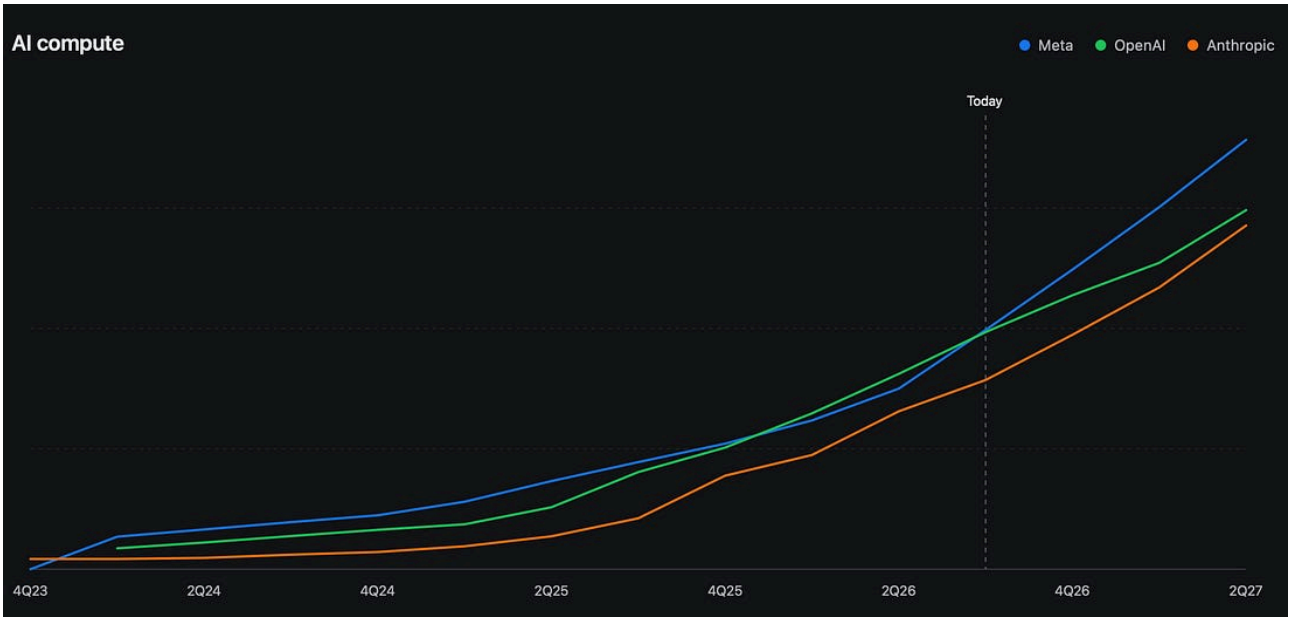
QQ:1413221435

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

Instagram ads can fund a lot of compute

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.

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.



Source: SemiAnalysis Tokenomics Model

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.

An unprecedentedly large compute ramp led by 5 titans

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.

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.

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.

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.



Source: SemiAnalysis Datacenter Model

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.

QQ:1413221435

QQ:1413221435



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

QQ:1413221435

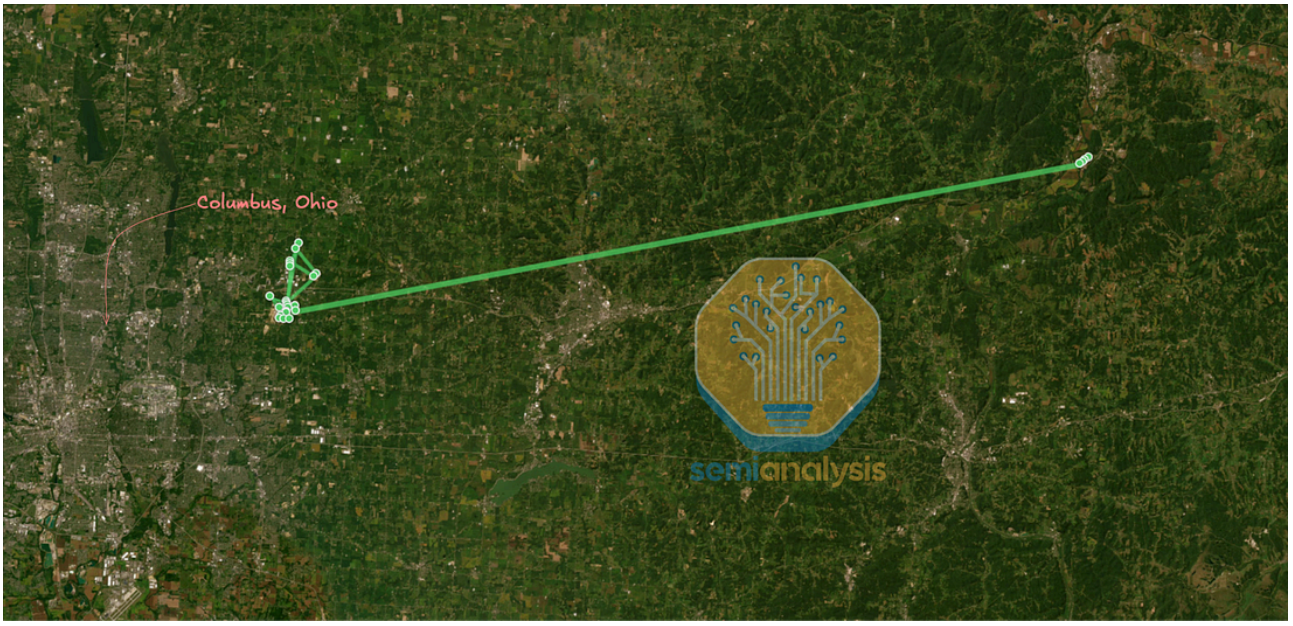
At Prometheus—which is already partially operational today—Meta doesn't quite have a full GW currently under construction, but they are fully embracing the scrappy tent DC design we called out in our previous MSL article. The Prometheus cluster also keeps expanding—from an initial ~1GW to now >3GW within 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.



Source: SemiAnalysis Datacenter Model

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

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.

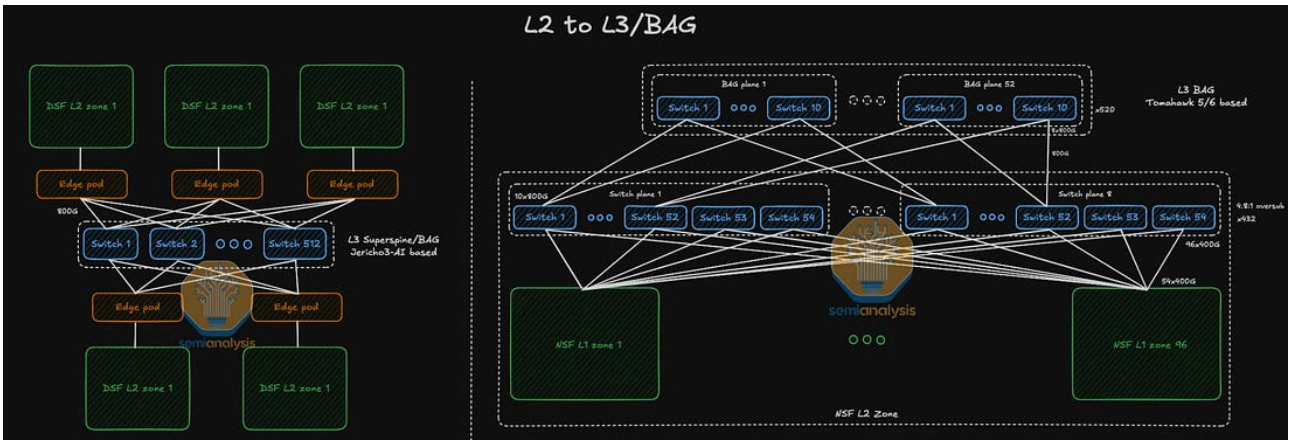


Source: SemiAnalysis AI Networking Model, Meta

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.

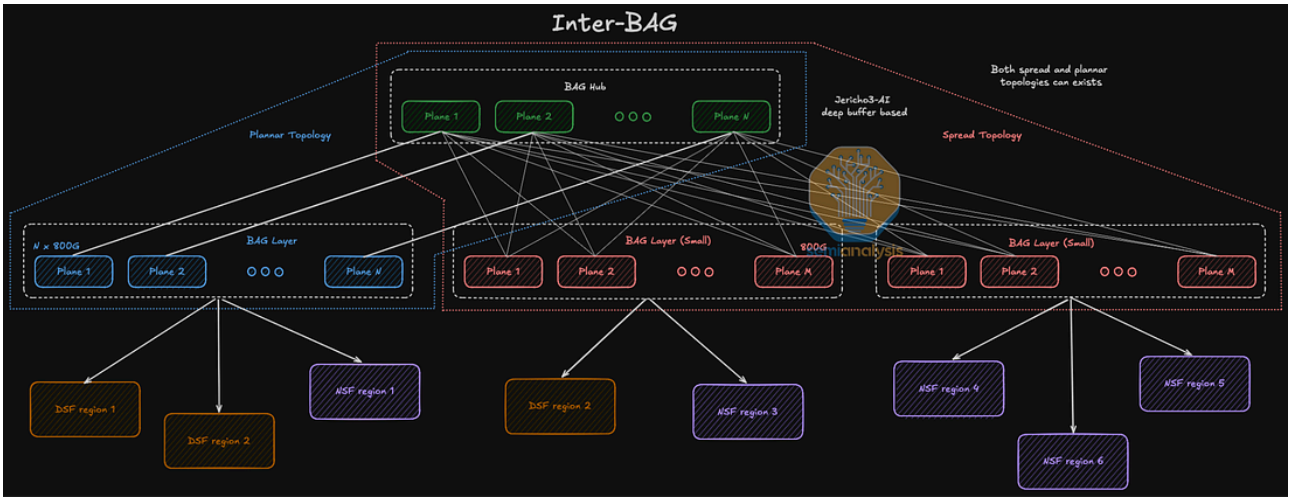
However, with this design choice comes a new challenge: networking. The whole point of a larger cluster is to be able to train larger models in a reasonable timeframe. If the accelerators are not able to communicate efficiently, then there is no point in a gigawatt scale campus. This is where scale-across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

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.



Source: SemiAnalysis [AI Networking Model]

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.



Source: [SemiAnalyis AI Networking Model]

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.

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.

Assembling the MSL superteam

QQ:1413221435

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.

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

- Andrew Tulloch (ex Thinking Machines cofounder)
- Joshua Gross, Mark Jen, Yinghai Lu (Thinking Machines founding team)
- Jason Wei, Hyung Won Chung, and Zhiqing Sun (ex OpenAI)

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.

QQ:1413221435

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.

But just to be clear, success is far from guaranteed

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.

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,**

QQ:1413221435

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.

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.

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.

What this means for Google

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.

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.

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.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis Tokenomics Model

QQ:1413221435

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.

This is loser mentality from Google.

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

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.

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

QQ:1413221435