Gemini is Cooked but GCP is Cooking

Gemini is Cooked but GCP is Cooking or why DeepMind's long term failure is GCP's short term gain

AUG 7 https://newsletter.semianalysis.com/p/gemini-is-cooked-but-gcp-is-cooking

View in browser

Gemini is Cooked but GCP is Cooking

GCP YoY rev growth >100%, DeepMind's long term failure is Google Cloud's short term gain

MAX KAN, JOEY BROOKHART, DOUG O'LAUGHLIN, AND DYLAN PATEL

AUG 6 • PAID

QQ:1413221435

READ IN APP

On Wednesday, August 5th, Google announced a complete overhaul of DeepMind leadership. A quick recap:

- Demis Hassabis, DeepMind co-founder and former CEO, is no longer involved in day-to-day operations.
- Jeff Dean, former Google Chief Scientist and Gemini co-lead is leaving to start a neolab called Discovery Loop. Jeff is the undisputed GOAT of Google engineering, co-founded Google Brain, and started the TPU program.
- Joining Jeff are Sanjay Ghemawat, Quoc Le, and Oriol Vinyals. Sanjay and Quoc were both Google Fellows, which is a title reserved for the company's top dozen or so technical contributors.[1] Oriol was a Gemini co-lead.
- Koray Kavukcuoglu, former DeepMind CTO and the one remaining Gemini co-lead, is replacing Demis as the leader of DeepMind/Gemini.

Obviously these are not the actions of people excited about Gemini 4 Pro.

QQ:1413221435

For all intents and purposes, we believe **DeepMind is no longer a frontier lab**. We said as much a few months ago to our Tokenomics clients due to large numbers of

departures from their reinforcement learning teams and poor compute allocation. Google will continue meandering on and releasing models, but their odds of reaching SOTA again have dropped to zero.

Furthermore, the **biggest beneficiary** of today's news is neither Anthropic nor OpenAI—it's **Google Cloud**. Whereas Gemini and GCP used to desperately fight for compute allocation, it's now clear that Thomas Kurian won. **We expect GCP revenue growth to meaningfully accelerate as a result**.

Gemini 3 Pro was the peak

At the end of 2025, Anthropic, OpenAI, and Google were the clear AI big 3. Some still held this opinion as recent as last week, but the downfall of Gemini in 2026 was clear to those paying attention. Here's an excerpt from an institutional note we published to our Tokenomics Model subscribers on **July 9th**:

QQ:1413221435

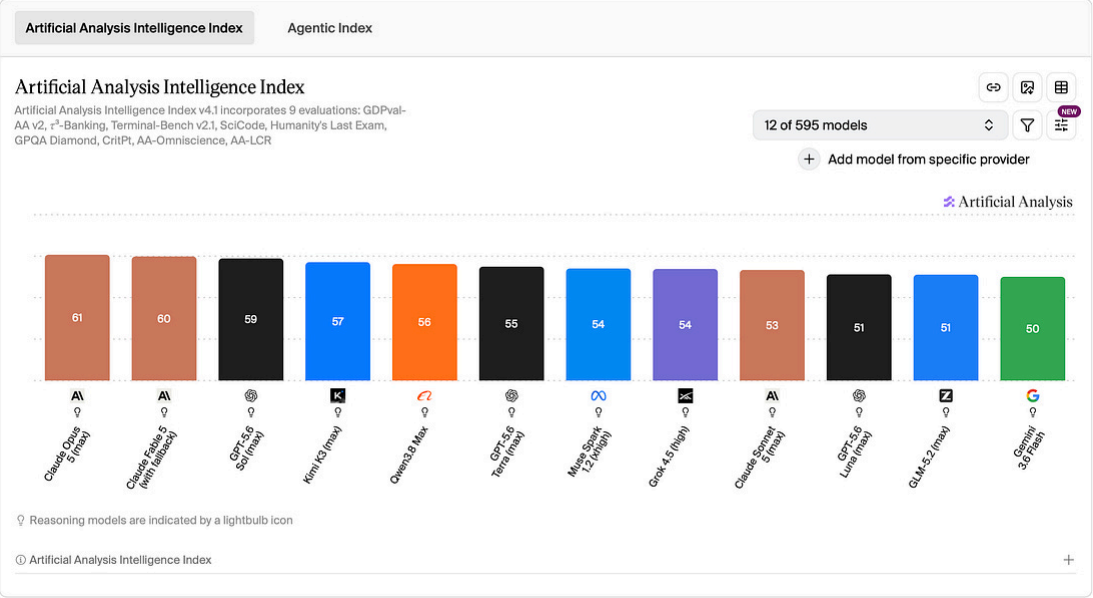
> - In November 2025, Gemini 3 Pro was arguably the best model in the world, forcing Sam Altman to issue a "code red" at OpenAI.
> - Since then, the gap between Deepmind and OpenAI/Anthropic has widened dramatically. Gemini 3.5 Flash was a total flop, and **industry chatter suggests the upcoming Gemini 3.5 Pro is roughly Opus 4.5 level**. This is quite literally worse than GLM 5.2, much less Fable 5 and GPT 5.6.
> - We believe things will only get worse for Gemini, as Google lacks the religious conviction necessary for building RSI. We think it's likely both MSL and SpaceXAI surpass them in model quality by the end of this year.
> - Along with Noam Shazeer and John Jumper, most of the best RL people at Gemini recently left the company.
> - OpenAI has overcome their pre-training issues, and a much larger model code named "Doug" is actively in the works.

The second to last point is worth emphasizing. Jeff, Sanjay, Quoc, and Oriol are just the latest in a long string of high profile departures from DeepMind. Despite discovering most of the foundational breakthroughs behind the current LLM paradigm, Google is now simply unable to retain top AI talent.

Notably, Jeff and co. have followed a playbook to create their neolab, Discovery Loop, that was established by David Silver of Ineffable Intelligence in November last year. To pursue their mission, top AI researchers are now choosing to leave Google, raise billions from outside investors and Google Ventures, and then spend that money on Nvidia GPUs in GCP.

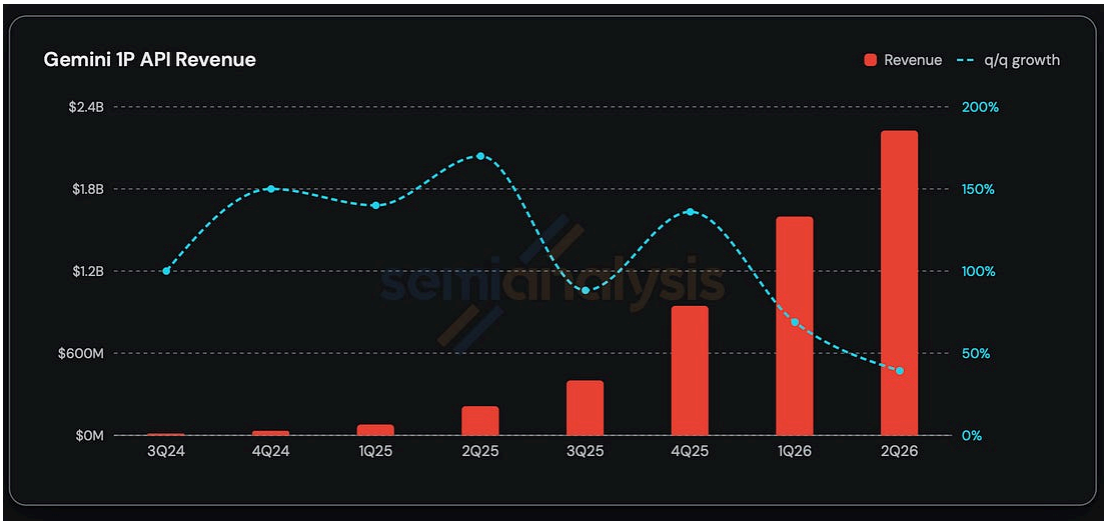
QQ:1413221435

Since our institutional note, Google has silently canceled Gemini 3.5 Pro and is now coping by hyping up Gemini 4 instead. As a bridge model, they released Gemini 3.6 Flash, but it's generally worse than Muse Spark 1.2, Grok 4.5, and the tier 1 Chinese open-source models. Depending on how you count, Gemini is currently in 8th or 9th place, and we don't see Gemini 4 reversing their downfall.



Source: Artificial Analysis

At the same time, Gemini first-party API token growth has famously decelerated. Whereas they grew 60% in 1Q26, going from 10B to 16B tokens per minute, 2Q26 saw just 38% growth to 22B. This has caused a corresponding decline in Gemini 1P API revenue growth.



Source: SemiAnalysis Tokenomics Model

However, **Gemini Enterprise Agent Platform** (formerly Vertex) **has been meaningfully stronger overall** thanks to third-party models like Claude. To see a

QQ:1413221435

full breakdown of Amazon Bedrock, Microsoft Foundry, and Google Vertex revenue separated by model provider, as well as all the implications for OpenAI and Anthropic, see our Tokenomics Model.

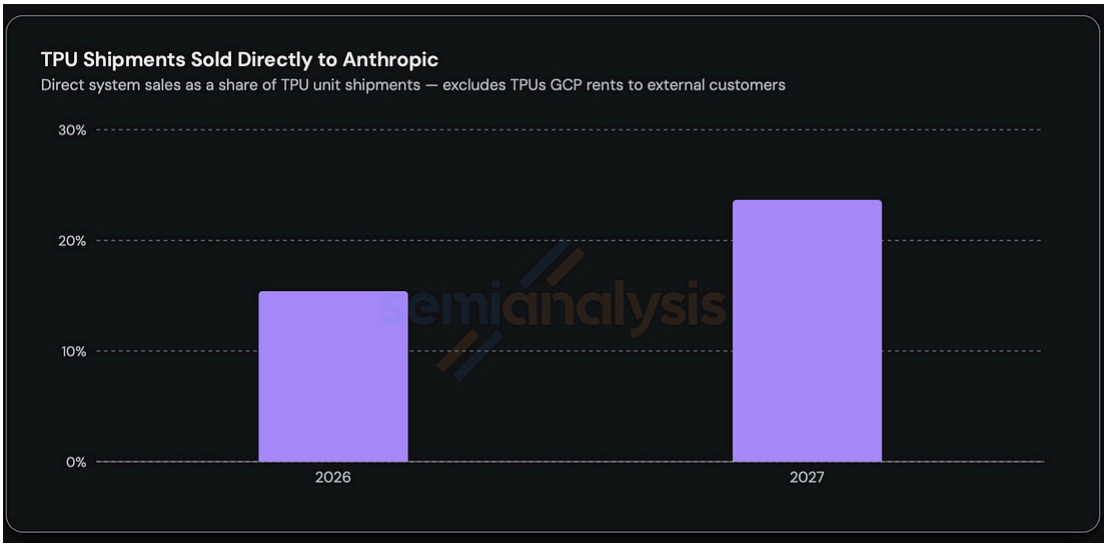
There’s no substitute for conviction

We believe Gemini’s core issue has always been a fundamental lack of conviction. Compute is the lifeblood of AI progress, and all the AGI-pilled labs are desperately trying to acquire as much as possible.

Of course, pre-paying multiple GWs is also very scary and expensive, especially if you don’t have a rapidly growing, high margin API business. However, as we’ve explained in previous newsletters, there are creative ways to overcome this limitation and still monetize excess capacity at very attractive rates **while maintaining the optionality to claw everything back for your research team**. Explicitly, this is what we see happening at Meta and SpaceX.

Google, on the other hand, decided it was totally worth it to sell enormous amounts of compute to Gemini’s fiercest competitors on long term contracts without any hope of ever returning it to DeepMind.

More than 20% of total TPU shipments from 3Q26 to 4Q27 are being sold directly to Anthropic.[2] This is excluding the hundreds of thousands of TPUs GCP already rents to Anthropic today, and the many hundreds of thousands more they’ve committed to rent to Anthropic and Meta over the next 6 quarters.



Source: SemiAnalysis Accelerator Model

For full details on the number of TPUs, GPUs, Trainiums, etc being produced each quarter till 2030 and their end customers, see our Accelerator Model.

QQ:1413221435

If you've ever listened to an interview of Google Cloud CEO Thomas Kurian, you know he is not AGI pilled. In one podcast, for example, he argued that it's great for TPUs to become "general purpose infrastructure" that supports customers like Citadel, the Department of Energy, and generic high performance computing. And when asked why he was selling compute to Anthropic despite them competing with Gemini, he said this was the natural consequence of Google being a "platform company."

This same man likely just won a major internal political battle to have even more control over Google's compute allocation.

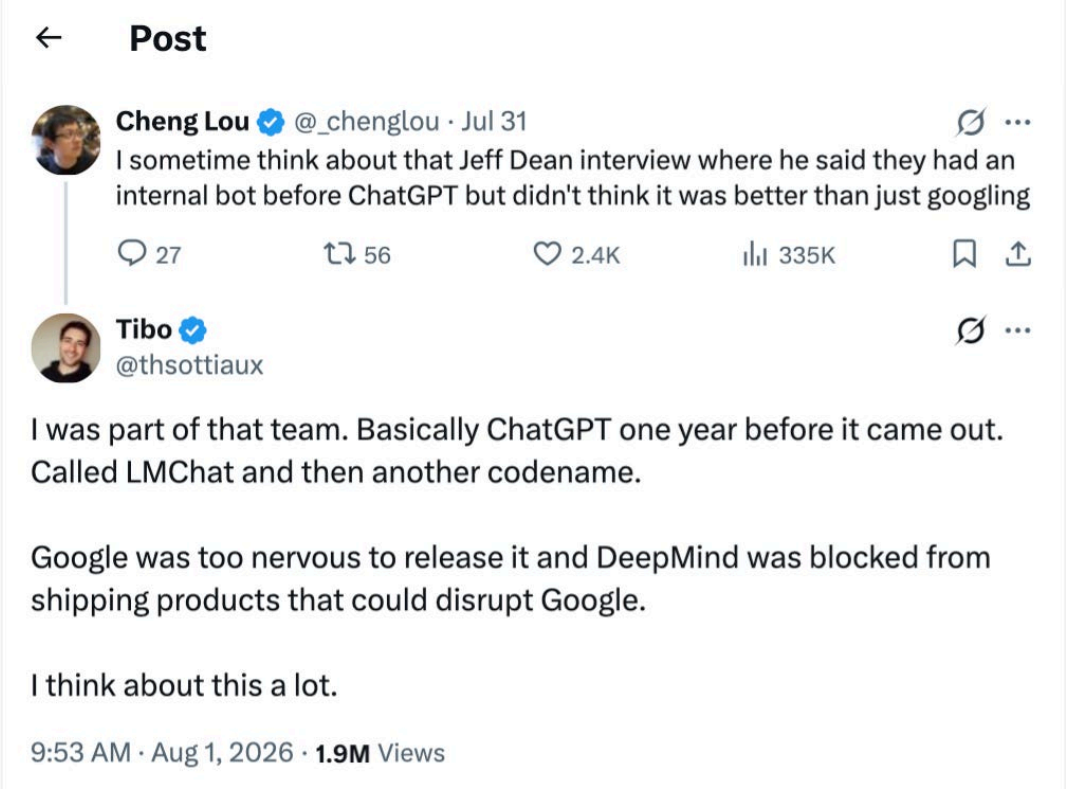
Attempting to steelman DeepMind's frontier ambitions

We've obviously been quite bearish on DeepMind thus far, and if we had to steelman the case for why they'll still be able to train a true SOTA model in the future, it would go something like the following:

- The current setup clearly wasn't working. With the existing leadership team, their odds of catching up to Anthropic/OpenAI looked extremely slim.
- Now that they've cleaned house, the new guys can start from a blank slate. Maybe they'll even acqui-hire a neolab like SSI or Thinking Machines.
- With this new team, their odds of catching up to the frontier actually increase.

Perhaps there's some world in which this happens, but we think the odds are basically zero. The issue with Google was not Jeff Dean nor Noam Shazeer, but rather their extremely bureaucratic, painfully slow, and strategically timid culture. Remember that DeepMind had an AI chatbot 1 year before ChatGPT but was not allowed to release it due to fears of disrupting their core business.

QQ:1413221435
QQ:1413221435



Source: Tibo on X

QQ:1413221435

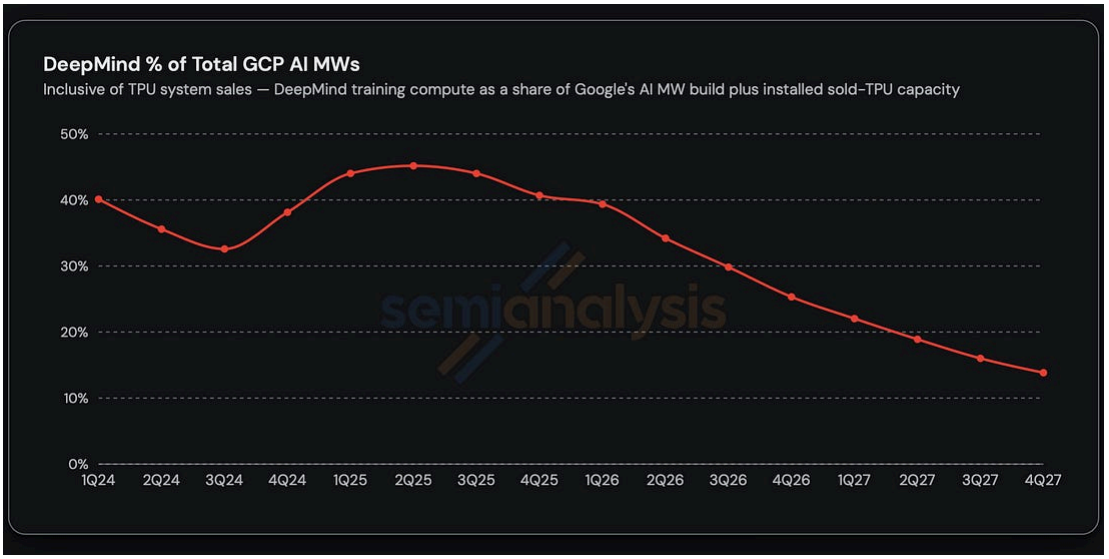
Tibo, funny enough, is another great example of brain drain from Google. A 9 year Google/DeepMind veteran, he left in July 2024 to join OpenAI, co-led the original coding agent team, and is now the head of Codex. Meanwhile, Google “acquired” Windsurf to build its own agentic coding platform in July 2025, and they don’t exactly have much wind in their sails thus far.

Let’s also not forget that DeepMind was the original AI acqui-hire! And with the benefit of 12 years’ hindsight, we now know how that played out. Adding another neolab to the roster would likely just result in more of the same.

The Financialization of GCP

Our Tokenomics Model estimates that Gemini ARR was $12B in 2Q26. In contrast, by the end of 2027, GCP will be doing over $73B in third party AI ARR IaaS/TaaS and another $120B of TPU sales. **$200B of external sales at high 30s EBIT margins vs a first party business generating just $12B today shows where the focus is**. Google management is clear: the short term benefits from focusing GCP on financialization is worth abandoning any long term competitiveness at the frontier.

QQ:1413221435



Source: SemiAnalysis Tokenomics Model

Despite the weaknesses with Gemini, GCP continues to accelerate. This last quarter, GCP growth was 82%. However, that wasn't all from what we would traditionally call the "cloud" business. GCP has a new revenue source in selling full systems of its TPUs to external SPVs that run datacenters for customers such as Anthropic. **These TPU sales are accounted for on a gross basis at ~$35B/GW**. In 2Q26, we estimate that TPU sales were around $1.2B in the quarter, and thus core GCP grew in the low 70s. Given over $150B of TPU systems backlog, **these TPU system sales will drive GCP's 2027 growth rate into the mid 100s vs sellside consensus at 64%**.



Source: SemiAnalysis Tokenomics Model

Although EBIT margins for these system sales are slightly lower than core cloud margins in the low 30% range, we still expect total GCP to deliver mid to high 30s EBIT margins going forward. How investors decide to capitalize the current TPU backlog and any large future sales is an open question. However, given the strong compute demand from labs, we expect more multi-GW deals to be announced soon

QQ:1413221435

adding to this backlog. In all, **we estimate that over $250B of additional TPU Bookings could be added to GCP RPO in coming quarters**. Until then, this eye-popping acceleration in GCP growth coupled with good margins will **add ~$3 to Google EPS in 2027**.

For an even more detailed breakdown of Amazon, Microsoft, and Oracle's cloud businesses, as well as updated Google projections come earnings season, see our Tokenomics Model.

Be Careful What you Wish For

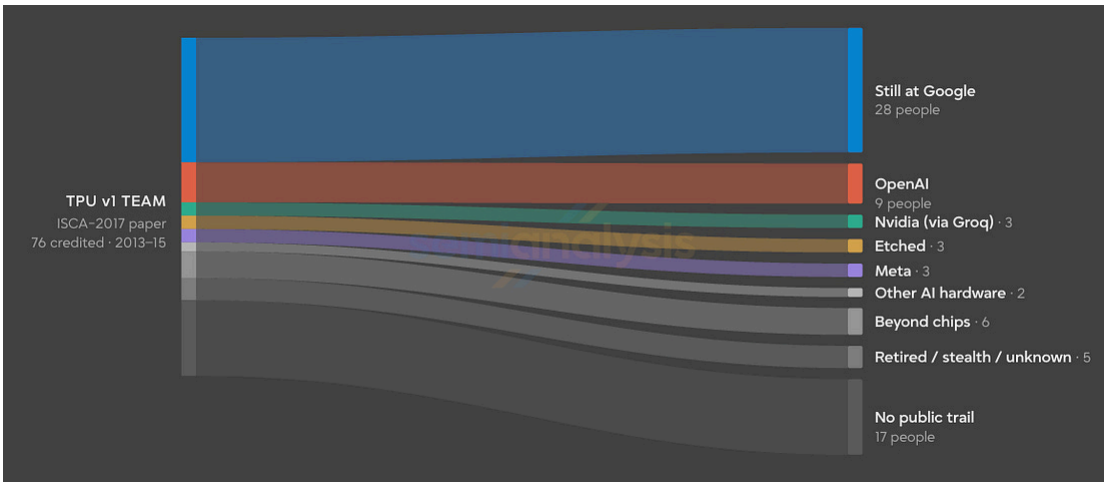
Google losing key talent and hammering external TPU sales will be the pivotal moment in the company's relatively short and storied history. There is an irony in GDM inventing the transformer, but likely losing the eventual AI market. Google will join the ranks of other legendary tech giants like IBM and Intel to give up on the harder thing and do the thing that will make you more money.

One can remember the shift from owning the entire vertical stack of media and distribution to just "the dumb pipes" a la Charter in the cable market. But almost always something that is financially intelligent in the short run often has a longer term strategic consequence. And that is exactly what Google is doing with Gemini dropping out of the serious frontier race and shifting to selling TPUs as a service and system.

Technology is hard, and it is rare that the last winners of the previous regime win the next, as bureaucracy, the innovator's dilemma and the reality of being a big company come to bite. It is no mistake that we don't live in IBM's world, which was once the most dominant technology company of its time. Instead they turned to financial engineering, gave up pursuing high technology, and entrenched themselves at what they were good at, mainframe and mainframe consulting. This same story played out almost exactly at Intel, when they first quit mobile, and then eventually stumbled EUV to lose their core business model itself.

When we look back in the decades to come, there is a chance that Google's all in shift to selling TPU will mark the beginning of the end. Because once they start to post those impressive numbers, it will be literally impossible to stop giving shareholders what they want or shares will tank. The company is going to do just fine in terms of revenue, but it is no better than a glorified neocloud, and while the TPU is the best of the best today, there is no guarantee that chip strength will continue. A meaningful amount of the core team has moved on, but if the remaining talent is strong enough maybe they can keep pace.

QQ:1413221435
QQ:1413221435



Many of the original builders of Google have gone, and by giving up Gemini they are giving up the vision of the future where Google is a real player. We've previously explained how there's a chance TPU's TCO advantage may crumble in the near future, with less aggressive technology choices versus say Nvidia. Maintaining the empire will be a lot of work, and it becomes demoralizing as many of the great technology leads have left. Will Google be able to maintain their TPU edge? Time will tell.

1 Technically Sanjay is a Senior Google Fellow, the only one besides Jeff.

2 Shipment refers to when the accelerator supplier recognizes revenue, which is typically when they ship a baseboard/package to the downstream manufacturer. This is different from when the chips are actually operational.



A guest post by

Doug O'Laughlin

Writing about what I like right now - aka Semiconductors.

Subscribe to Doug

Like Comment Restack

© 2026 Dylan Patel
548 Market Street PMB 72296, San Francisco, CA 94104
Unsubscribe

Start writing

QQ:1413221435

QQ:1413221435

QQ:1413221435