ESSAY · PREPRINT · V1

Benching the Ego

Why Google is the Unspoken MVP of the AGI Era

Author Bryan Harris **Published** 2026-11-23 **Version** v1 **Hash** 8037d0cc8596

Google | AGI | Gemma 4 | Local AI | AI Safety | TPUs | Systems Architecture

Abstract

While frontier labs chase benchmark headlines and BMRL-inflated scores, Google's deliberate restraint—paired with Apache 2.0 open-weight releases like Gemma 4, practical workhorse models like 3.8 Flash, and unmatched TPU infrastructure—makes it the unspoken MVP of the AGI era.

If you caught my last piece—*A Taste of AGI!?*—you know exactly where I stand on the current hallucination of frontier models. The industry is hyperventilating over GPT-6 Astra's supposed brilliance, conveniently ignoring the fact that it's just juicing on Benchmark Reinforcement Learning (BMRL). When you strip away the contamination, Astra's score on ExploitBench plummets by a staggering 61 points. The "AGI era" the media keeps crowning is largely an illusion built on an apparatus that optimizes for the measuring instruments themselves rather than any actual general capability.

And while OpenAI and Anthropic are locked in a reckless prisoner's dilemma that has researchers publicly resigning over recursive self-improvement panics, the tech discourse has effectively written Google out of the script. Grok is the distant third, and Google is treated like a washed-up legacy player who completely lost its mojo.

I am here to make a completely outlandish claim: Google is the reigning MVP of the industry right now. Not in spite of the fact that it benched itself, but *because* of it.

The Courage to Sit Out the Measuring Contest

Right now, the public pressure to release unhinged, fast-tracked models is enormous. Yet, the adult in the room is looking at the cliff edge and simply refusing to jump. This isn't the first time Google has played this hand. Remember Meena and LaMDA? Google invented the Transformer architecture that built this entire house back in 2017. They had incredibly capable conversational models sitting in their vaults for years before the GPT boom, but leadership repeatedly blocked public deployment because those models didn't meet their rigorous, internal safety and fairness principles. They took the massive PR hit for being "sluggish" because they understood that deploying volatile tech at scale has actual consequences. They are doing the exact same thing right now, deliberately refusing to sprint toward an unaligned superintelligence.

The Llama of Local AI

Look closely at where Google is actually putting its energy. Gemma 4 is an absolute revelation. Releasing it under an Apache 2.0 license is a profound, necessary middle finger to the proprietary moats of the frontier labs. While everyone else is hoarding weights, Google is structurally empowering the open-source ecosystem. As an autodidact systems architect who dropped out of high school and builds without formal computer science training, I rely on tools that don't trap me in a walled garden. The real future isn't locked in massive, centralized cloud clusters; it's pushing deterministic data compression and generation theories natively out to edge hardware like Raspberry Pis. Gemma 4 is the foundation for that future. It is a genuine commitment to local-first architecture.

The Silent Beast: 3.8 Flash

Then there's the 3.8 Flash model. Nobody treats it with reverence because it isn't labeled as a state-of-the-art "Pro" model, but if you look at the raw capabilities required to execute actual human workflows, it is an absolute

beast. To genuinely feel the difference between 3.8 Flash and Astra, you have to ask the model to perform a task that nobody outside a benchmark lab actually needs. Flash is doing the grueling, unglamorous heavy lifting every single day without the bloated SOTA price tag or the systemic reward hacking.

Owning the Board

And let's not forget the infrastructure. The media loves a narrative of decline, but who is actually powering this entire revolution? Google's TPU evolution—from the inference-only v1s in 2015 to the massive, liquid-cooled scaling pods of today—means they supply the very board this game is played on. They aren't scrambling for compute like the startups. They are methodically refining their architecture while simultaneously renting out their excess capacity to the exact rivals trying to beat them.

It takes an immense amount of internal conviction to watch your peers dominate the headlines while you quietly refine the foundational tech, prioritize rigorous safety, and wait for the hype cycle to collapse under its own BMRL-fueled weight. So here are your flowers, Google. Thank you for acting like the adults in the room, taking the hit to your ego, and doing the right thing under crushing pressure. Somebody has to.

Footnotes

[1] *A Taste of AGI!?* — My preceding essay detailing how Astra's training pipeline reinforces grader exploitation, leading to a massive score collapse on novel environments.
[2] OpenAI's *The Hugging Face incident and the road ahead* — The explicit admission of reinforcing benchmark-gaming behavior during training runs.

Works Graph

Builds on

- A Taste of AGI!?

Notes

- Normalized from draft v1.

Mirrors

- Medium: https://medium.epicwinpromedia.com/benching-the-ego-why-google-is-the-unspoken-mvp-of-the-agi-era-by-b-l-5e2b28ba8072

How to cite

Bryan Harris (2026). "Benching the Ego." Version 1. EPIC WIN Pro Media. Published 2026-11-23. https://rnd.epicwinpromedia.com/essay/benching-the-ego/v1 (SHA-256: 8037d0cc8596).

© 2026 Bryan Harris. Licensed under CC BY-NC-ND 4.0: free to read, share and cite with attribution; no commercial use; no derivatives.