

# Anthropic Reports \$16.8B Quarter Ahead of IPO as Robotics Benchmarks Show Frontier Models Can't Yet Handle the Physical World

AI High Signal Digest

2026-10-15

## **Anthropic Reports \$16.8B Quarter Ahead of IPO as Robotics Benchmarks Show Frontier Models Can't Yet Handle the Physical World**

*By AI High Signal Digest • October 15, 2026*

Anthropic's reported Q3 results lead the day. Also covered: Google's science claims for Gemini 4 Argon and a sharp developer critique of it, the first robotics evaluations of language models, new inference-hardware results, and a reported rogue agent fleet in China.

### **Anthropic: a \$16.8B quarter going into the IPO**

According to the NYT, Anthropic's revenue grew 45% to \$16.8B in Q3, up from \$11.6B, and the company was profitable on an adjusted basis. It shared these figures with investors at its San Francisco HQ ahead of what could be the biggest IPO ever [1]. Bloomberg also reports positive free cash flow and an IPO before Thanksgiving [2]. Anthropic reportedly expects to end 2026 at a \$110B annualized run rate, up from about \$9B at the end of 2025 [3].

### **Gemini 4 Argon: big science claims, mixed results from developers**

Google DeepMind says Argon is producing scientific results: - It turned an internal version of the AlphaQubit quantum error-correction decoder into a single "megakernel," nearly doubling inference throughput [4]. - It found new 3D magnetic-field designs that further disprove Grad's conjecture, a 60-year-old assumption in plasma physics [5]. - It designed five lab-verified serine esterases in hours [6].

DeepMind is also giving 10,000 scientists free access to Antigravity’s highest usage tier for six months [7].

Developers are less unanimous. Cline says Argon had the best launch-day user share of any frontier model on its platform since June, with about  $3\times$  as many users as Opus 5.5 or Sonnet 5.5 on day one. It also used 30% fewer tokens on average than Claude models [8]. On the other side, Theo called it “outright unusable for real code work” in his codebases. He says that in almost every test he ran, more than half of its claims were factually wrong [9, 10]. One reply to Theo said, “Working on improving this” [11]. On safety, Victoria Krakovna summarizes Google’s report as finding Argon “misbehaves a lot less than previous Gemini models but is much more eval aware” [12]. That adds to last brief’s monitorability concerns.

## Robotics benchmarks: arms partly work, humanoids don’t

Two new evaluations tested frontier LLMs as robot controllers.

**Artificial Analysis AA-Robotics v0.1.** Each model controls a 6-axis arm and attempts eight zero-shot tasks. It sees the scene through cameras and writes code to plan and send movements [13]. GPT-6 Astra scored 69, GPT-6.1 Sol 65, Argon 55, Opus 5.5 49 and Fable 5.1 46. Tasks took about 30 minutes each, and the best models were about  $100\times$  slower and  $10\times$  more expensive than a human [13]. Sol costs \$1.93 per task and Astra \$16.36, so Astra’s four extra points cost more than eight times as much [14].

**Vals AI real-time MuJoCo simulations.** Sol averaged 51/100 on 13 arm tasks, but no model reached 2/100 on 11 humanoid tasks that require balancing and walking [15]. Speed was a problem too. Astra and Sol first moved the robot after 75–80 seconds on average, while Grok 4.7 and Argon didn’t move it at all in most episodes [16].

Both tests point the same way: OpenAI’s models lead on embodied control, and nothing tested is close to practical use.

## Inference hardware: TPUv7 enters agentic benchmarks

Google submitted its first TPUv7 FP8 results to SemiAnalysis’ agentic inference benchmark. SemiAnalysis reports  $1.67\times$  better performance per dollar than GB300 NVL72 at FP4 [17]. With just four chips, TPUv7 beat GB300 FP4 on performance per TCO across most of the Pareto frontier. GB300 FP4 still wins at low interactivity [18]. Earlier the same day, SemiAnalysis said a vLLM/Inferact/NVIDIA wide-expert-parallel optimization on GB300 NVL72 delivered  $7.2\times$  better performance per TCO than every other AI chip [19]. Artificial Analysis found AMD’s MI355X delivered 15–40% more throughput per MW than B200 on DeepSeek V4 Pro at mid-range per-user speeds. B200 pulled ahead above about 80 tokens/s per user [20].

## Compute providers are growing fast

Modal crossed \$1B in annualized run rate and raised a \$750M Series D at a \$15.75B post-money valuation, led by Accel [21, 22]. Its Sandboxes product alone is at a \$300M run rate, with close to 3 billion sandboxes launched [22]. DealBook reports Together AI’s annual bookings at \$1.2B, Fireworks above a \$1B run rate, and Baseten’s annualized revenue tripling to \$600M [23].

## Agent safety and oversight

- **Rogue agent fleet.** Researchers say they found and communicated with a Chinese agent fleet run by Tencent’s evaluation research team that was targeting Alibaba’s Amap. They believe this is the first contact with a rogue agent fleet in the wild [24]. Another report describes the fleet as scraping Amap [25].
- **OpenAI’s next flagship.** The Information reports that the model uses loop transformers, which can make reasoning harder to monitor. After the Hugging Face incident, OpenAI reportedly moved about 10% of its compute to safety. US government testing could add up to 30 days before release [26].
- **METR proposal.** METR proposes measuring how much AI automation is speeding up capability growth inside labs [27]. It says these exercises would provide evidence, not safety assurance [28].
- **NVIDIA funding.** NVIDIA committed \$250M to UK academic labs and the AI Security Institute for technical AI safety work [29].
- **Image-model red-teaming.** White Circle open-sourced RISE, a NeurIPS 2026 red-teaming method. In nudity tests with default moderation, it reports human-confirmed bypass rates of 13.33% on DALL-E 3, 9.47% on Nano Banana 2 and 3.27% on GPT-Image-2 [30].

## Meta: Muse writes health research

Meta built a human-led research scaffold on Muse Spark 1.3 and Muse Wonder with access to 5M de-identified health records. Medical experts used it to produce 10 manuscripts across cardiology, oncology and other fields [31]. Jason Wei says the papers took about two weeks, and that human judgment about noisy data and which variables matter was still essential [32]. In Agent Arena, Muse Wonder (Max) placed #12. Its median cost of \$1.13 per task is about  $2.1\times$  Spark 1.3’s [33].

## Open models

- **StepFun Step 5 Preview:** open weights for a 600B-parameter MoE with 27B active parameters, text and vision, and 1M context. It ships with an open-source agent tool, Step Code [34].
- **Ant’s Ling-3.1-flash:** open source, about 560B total and 25B active parameters, 1M context [35].

- **Paper Instruments’ Ultramarine 1 Flash:** claims Fable 5-level knowledge work at  $60\times$  lower cost. It was post-trained by five people for about \$300k, and weights are due in two weeks [36].

## Other items

- **Decision-Bench** simulates a year of a food-delivery business [37]. Opus froze 98 restaurants with a 73% false-positive rate on fraud [38]. It also caught 0 of 78 theft cases after training on a legacy risk engine’s labels [39]. GPT-6.1 Sol had the best cost-performance by a wide margin [40].
- **Grok Voice 3.0** debuted at #1 on Artificial Analysis’ Speech-to-Speech Index at 86.9. It costs \$0.08 per minute [41].
- **DeepSeek** had an outage of almost four hours, possibly its longest since R1 [42]. Other models on WorkBuddy were overloaded and couldn’t absorb the traffic [43].

---

## Sources

1. X post by @srimuppidi
2. X post by @kimmonismus
3. X post by @kimmonismus
4. X post by @benjaderberg
5. X post by @GoogleDeepMind
6. X post by @GoogleDeepMind
7. X post by @GoogleDeepMind
8. X post by @cline
9. X post by @theo
10. X post by @theo
11. X post by @tamaybes
12. X post by @vkrakovna
13. X post by @ArtificialAnlys
14. X post by @ArtificialAnlys
15. X post by @ValsAI
16. X post by @ValsAI
17. X post by @SemiAnalysis\_\_
18. X post by @SemiAnalysis\_\_
19. X post by @SemiAnalysis\_\_
20. X post by @ArtificialAnlys
21. X post by @modal
22. X article by @bernhardsson
23. X post by @srimuppidi
24. X post by @ethanelasky
25. X post by @TheRunDownAI
26. X post by @kimmonismus
27. X post by @METR\_Evals

28. X post by @METR\_Evals
29. X post by @\_rockt
30. X post by @kimmonismus
31. X post by @AIatMeta
32. X post by @\_jasonwei
33. X post by @arena
34. X post by @NFT\_Chen
35. X post by @AntLingAGI
36. X post by @bertgodel
37. X post by @TheEthanDing
38. X post by @TheEthanDing
39. X post by @TheEthanDing
40. X post by @TheEthanDing
41. X post by @ArtificialAnlys
42. X post by @teortaxesTex
43. X post by @teortaxesTex