

Anthropic’s Vatican Consciousness Lobbying Draws Backlash as AI Starts Producing Open-Problem Math Results

AI High Signal Digest

2026-10-03

Anthropic’s Vatican Consciousness Lobbying Draws Backlash as AI Starts Producing Open- Problem Math Results

By AI High Signal Digest • October 3, 2026

A New York Times report on Anthropic’s push for AI consciousness at the Vatican set off an industry fight. Meanwhile, Meta and Google reported verified results on open math problems, Epoch sized the agent economy, and evidence piled up about China’s progress on its own compute stack.

Anthropic’s Vatican lobbying becomes a public fight

The most discussed story was a New York Times report, relayed on X. It says Anthropic co-founder Chris Olah proposed pulling out of the May 25 launch of Pope Leo XIV’s AI encyclical, *Magnifica Humanitas*, after reading an advance copy that rejected machine consciousness. He attended in the end. Two participants say he and his team then privately lobbied the pope’s advisers to take possible AI consciousness seriously [1]. On stage, Olah said Anthropic finds “internal states that functionally mirror joy, satisfaction, fear, grief and unease” [1]. One summary adds that Anthropic spent months courting theologians under NDAs, hoping they would endorse Claude’s moral standing [2].

Criticism came from rival executives. Cohere’s Aidan Gomez called it a campaign to get major faiths to adopt Anthropic’s philosophy “rather than taking input from them.” He attributed it to “moral arrogance and dogmatism” among EA and Anthropic leadership [3]. Dan Hendrycks made a separate argument: even conscious AIs would not mean humanity should hand them control [4]. For Anthropic, the episode turns a philosophical position into a reputational and policy liability, at a time when AI safety is already contested in Washington.

AI on open math problems, with verification built in

Meta says mathematicians used Muse Spark 1.1 and 1.2 in Thinking Mode to work on open problems. They used the ordinary meta.ai chat interface, with “no custom research scaffold,” and the work produced six papers [5]. Mathematicians guided the work, a second group reviewed it, and each paper marks which passages AI drafted [5]. In one example, Muse Spark wrote the search code that found a counterexample to a group-theory conjecture; the researchers then verified it and completed the proof [6].

Google Research’s Cogentic takes the opposite approach: heavy orchestration around Gemini. Provers each pursue one direction, and a draft is accepted only if two adversarial verifiers pass it. One verifier reads the draft alone; the other compares all of a round’s drafts to catch shared mistakes [7]. Cogentic produced new results on five open problems in online learning, auction theory and mechanism design, each checked by domain experts. Most took around 100 Gemini calls; the hardest took about 1,000 [7]. Taken together, the two reports suggest that open-problem results now come from both plain chat interfaces and heavy scaffolding, provided human or adversarial verification stays in the loop.

How many agents can the hardware run?

Epoch AI estimates that HBM shipped in 2025–27 could support anywhere from 30–60 million concurrent agents running Claude Fable 5 to 1.9 billion running DeepSeek V4 Pro. At the high end, that would rival the working hours of the global workforce [8]. In the sessions Epoch analyzed, Codex cost about \$16–18 per agent-hour at API prices and Claude Code about \$24–50 [9]. The catch is demand. Even 20% utilization would mean \$2.6–5.3T a year in spending, while leading labs would reach only about \$1T in combined annualized revenue by end-2027, even growing $5\times$ a year [10]. Investors are already pricing in the agent story: Bloomberg links Nvidia’s first record high since May, at about \$5.7T, to optimism that agents will drive chip demand [11].

Decision models reach local inference

The new category of decision models, covered in recent briefs, now runs on mainstream open inference tools: - **llama.cpp** added a `/v1/systemone` endpoint for “Jev-style inference locally.” It launched with five open models from 144M to 27B parameters [12, 13]. - **Ollama** now serves Cloudflare’s Clef (27B) and Clef Flash (9B) [14]. - **vLLM Semantic Router Decision 2.0** releases Apache-2.0 models from 0.6B to 27B that answer 64 questions about a request in one forward pass [15].

Perplexity’s Aravind Srinivas says pplx-decider-v1-27b averages 85.7% across 11 benchmarks, “ahead of Jev.” The same open-source batch included Lily, an Apple-silicon inference engine, and Bumblebee, a supply-chain scanner for developer machines that covers MCP configs [16]. Jamin Ball’s write-up puts

this in context: Jev was used by about 13% of paid Vercel AI Gateway teams within 24 hours of launch, and The Information reports TypeSafe is in talks to raise at a \$10B valuation [17].

Frontier scoreboard: cost per task matters more

On Arena’s Agent Arena, Claude Sonnet 5.5 (Max) debuted at #3, and Anthropic models now hold all three top spots. Sonnet’s median cost of \$2.74 per task is about 73% above Opus 5.5 at #2 [18]. GPT-6.1 Sol (Max) entered at #5 with a \$0.56 median cost: 81% cheaper than GPT-6 Astra and within 1.04 points of it [19].

Benchmark trust is also under attack. BridgeMind’s NerfBench reported Opus 5.5 at 94.2%, while noting this was still within normal variance [20]. Theo alleged that the group’s April Opus 4.6 “nerf” claim rested on 6 of 30 tests [21]. He offered matching \$10,000 benchmarking with a third-party audit [22]. Separately, developers report that Opus and Sonnet 5.5 refuse to summarize why they took an action, and say this is pushing users to other models [23].

China’s compute stack

DeepSeek’s open-sourced DeepGEMM, FlashMLA, TileKernels and DeepEP give the first view of Huawei’s Ascend 950 at the software level. An analysis based on them estimates peaks of about 432 TFLOPS in BF16 and 865 TFLOPS in FP8 [24]. It also warns that Ascend’s explicit Cube/Vector pipeline puts more demands on compilers and kernel libraries [24]. @teortaxesTex now argues that DeepSeek has made Ascend viable and that switching platforms is getting easy [25].

Supply is the constraint. As reported on X, Huawei says mass deliveries of its 950DT training superpod start early next year, and that the supply chain needs to ramp faster [26]. A post citing an industry source claims SMIC’s roadmap shows no EUV production by 2030, which would widen its gap with TSMC to more than 10 years [27].

On capability, The Batch reports that open-weight GLM-5.3 nearly matched Claude Mythos at exploiting vulnerabilities, 12% vs. 14% [28]. Generality Labs results also show ExploitBench scores are “VERY harness-sensitive” [29]. At the frontier overall, one commentator calculates the open–closed gap at its widest since late 2025. On the AA Index, the top open model, MiMo-V2.6-Pro, scores 46 against Opus 5.5’s 58 [30].

Memory hardware

SemiAnalysis estimates that HBM controllers and PHYs take up about 16% of Rubin’s die, falling to about 4% on Feynman with NVHBM, because the memory controller moves onto the HBM base die. Nvidia claims up to 25%

more die area for compute and 15% lower HBM power than standard HBM4E [31].

Also notable

- **Trillium Labs**, a new nonprofit for open frontier-AI science co-founded by Nathan Lambert, starts with open post-training recipes and plans open infrastructure for studying recursive self-improvement (RSI) and reward hacking. Initial backing comes from Halcyon Futures and Schmidt Sciences [32].
- **Muse Gadgets** is open-source ESP32 firmware and a Linux SDK for building Muse hardware. Muse is also giving 5,000 Home Link devices free to subscribers [33, 34].
- **Apple** will add controls around macOS Full Disk Access, citing the risk from increasingly capable AI agents [35].
- **Cloudflare** integrated Pi Durable into its Agents SDK [36].
- **Tesla’s** Giga Texas Optimus factory reportedly targets 10 million robots a year, with production starting in 2027 [37].

Sources

1. X post by @kimmonismus
2. X post by @ChristopherHale
3. X post by @aidangomez
4. X post by @hendrycks
5. X post by @AlatMeta
6. X post by @AlatMeta
7. X post by @omarsar0
8. X post by @EpochAIRsearch
9. X post by @EpochAIRsearch
10. X post by @EpochAIRsearch
11. X post by @kimmonismus
12. X post by @ggerganov
13. X post by @ngxson
14. X post by @ollama
15. X post by @vllm_project
16. X post by @AravSrinivas
17. X article by @jaminball
18. X post by @arena
19. X post by @arena
20. X post by @bridgemindai
21. X post by @theo
22. X post by @theo
23. X post by @dbreunig
24. X post by @ZhihuFrontier

25. X post by @teortaxesTex
26. X post by @pandawatch88
27. X post by @Taog_1575
28. X post by @DeepLearningAI
29. X post by @teortaxesTex
30. X post by @Fei2411
31. X post by @SemiAnalysis_
32. X post by @natolambert
33. X post by @alexandr_wang
34. X post by @alexandr_wang
35. X post by @TechCrunch
36. X post by @badlogicgames
37. X post by @TheHumanoidHub