

Reflection's Beam Opens to Praise for Efficiency, Criticism for Trailing China

AI High Signal Digest

2026-10-06

Reflection's Beam Opens to Praise for Efficiency, Criticism for Trailing China

By AI High Signal Digest • October 6, 2026

Reflection's first open model leads the period, and the arguments over it centre on how far behind Chinese labs it sits. Also covered: OpenAI's EU text watermarking, new data on the Claude–OpenAI subscription gap, and China finding ways to get compute.

Reflection ships Beam, and the argument is how it compares with China

Reflection AI announced Beam, its first model. It is an agentic open model with 501B total parameters and 23B active, trained from scratch, and full weights are due “this month” [1]. The weights, technical report and model card are promised under Apache 2.0 later in October, so the model could not be tested on launch day [2]. The company says the RL phase ran on 10.5k GB300s for four weeks, which it calls the largest publicly documented RL run it knows of. It also says capabilities kept improving with more RL, “with no signs of plateau” [3]. A launch recap puts pretraining at 24T tokens over four weeks. It credits Reflection with claiming 3–4× higher inference efficiency than GLM 5.2, and lists 80.9 on SWE-bench Verified [4]. Artificial Analysis has early access and says early signs point to one of the most token-efficient open models for its intelligence level [5].

Most of the pushback is about China. Turing Post says Reflection's own table puts Beam almost 30 points behind DeepSeek on DeepSWE and more than 10 points behind on Terminal-Bench. It also says six Qwen and DeepSeek results were left out of the table [2]. On efficiency, Turing Post notes the claim excludes prompt processing and serving overhead, and that the charts leave out Kimi K3, GLM 5.3 and DeepSeek V4.1 Flash [2]. Beam does beat West-

ern peers such as Mistral Medium 3.5 on SWE-bench Verified, 80.9 to 77.6 [2]. Nathan Lambert grouped Reflection with Nvidia and Thinking Machines: each released its strongest model and still came in behind Chinese counterparts [6]. Reflection’s Chetan Tekur called that “a little unfair” for a first model that is already competitive with GLM 5.2 on several tasks [7]. Elie Bakouch estimates pretraining ran at only about 12% BF16 MFU (model FLOPs utilization). He does credit Reflection with stable training and a better base model on held-out code perplexity than DeepSeek V4 [8].

OpenAI starts watermarking text in the EU

To meet the EU AI Act, OpenAI will watermark eligible ChatGPT and Codex text in the EU over the coming weeks. API customers worldwide can switch it on for select models now [9]. The watermark is an invisible statistical signal in the text. It indicates whether text was likely generated by an OpenAI model, but it does not identify a person or account [10]. OpenAI admits it is often undetectable in short passages and that rewriting or translation can remove it. For now, only approved researchers can use the detector [11]. One report says the method matched or beat Google’s SynthID in OpenAI’s tests and will be open-sourced [12]. Critics cite a test where swapping 25% of words for synonyms cut detection from about 92% to 17%. They also point out that text a person wrote and only had ChatGPT edit can carry the mark [13].

Subscription economics favour Anthropic, but its big customers are cutting back

SemiAnalysis argues that what a subscription is worth depends on which model and workload it is used for [14]. On that basis it finds Opus and Sonnet 5.5 on any Claude plan give far better value per dollar than every OpenAI plan [15]. Part of the gap comes from OpenAI halving the API-equivalent value of its \$200 Pro plan last week [16]. Adjusted for task cost, one analyst cuts the advantage to 1.3–2.9× [17]. On the first day of its 28-day Codex pledge, OpenAI made GPT-6 Astra and GPT-6.1 Sol about 50% faster by default for subscribers and Sign in with ChatGPT partners [18]. Demand from large customers looks less secure. The Information reports that Microsoft cut its projected internal Anthropic spending by more than a third. Meta’s Claude Code users reportedly fell from about 60,000 to 30,000, mostly because Meta is pushing its own tools [19].

How China is getting compute

- **Huawei–Qualcomm:** the two companies signed a multi-year cross-license covering 5G, compute, AI and networking. Qualcomm is also buying some Huawei US patents [20]. A Qualcomm spokesperson says reports that it is the net payer are wrong, and that the deal is not related to LogicFold [21].

- **Tencent:** the FT reports a five-year lease worth about \$7B for access to roughly 100,000 advanced chips in Oracle data centers in Southeast Asia [22].
- **Smuggling:** prosecutors charged a California server seller with smuggling more than \$300M of export-controlled GPU servers to China [23].
- **DeepSeek:** Bloomberg reports at least \$12B raised, with an IPO targeted for early 2027 [24].

Agent infrastructure and research

Epoch AI fitted trends to OpenAI’s published data on its researchers’ coding-agent use. Spending, valued at API prices, is doubling about every month. By mid-August it reached about \$600 a day for the median researcher and over \$7,000 at the 90th percentile [25, 26]. Epoch notes these figures are not OpenAI’s actual costs [27].

Hugging Face released a capture proxy that turns unmodified harnesses into RL environments. The supported harnesses include Claude Code, Codex and OpenCode. The motivation: the same model scores 62% under Mini-SWE-Agent but 33% under Claude Code. In tests, training across four harnesses lifted a 2.6B model from 42% to 54% in all four [28]. The Hub now also hosts RL environments as dataset-like artifacts [29].

Cognition launched “Dreaming”. Devin builds a memory graph across sessions and prunes it overnight [30]. Cognition is open-sourcing the memory format, which is backed by git and markdown and works with any agent [31]. Microsoft’s CorpusMap pre-links entities across a document collection. That raised agentic-search answer quality by 6.4–11.7 points and cut input tokens by 34–57% [32].

Other launches

- Reka’s Rho-1 is a 19B research preview that handles text, image, video and robot actions in a single network [33]. It was trained on 320 H100s in about three months [34].
- Vals AI says more than 90 Opus 5.5 agents found two room-temperature magnetic semiconductor candidates in simulation within three days [35]. Neither has been measured experimentally yet [36].
- Nolla Health says it is the first US organization approved for AI to issue initial prescriptions [37]. The approval covers acne treatment in Utah, with a clinician stepping in when needed [38].

Safety and policy

New York City Council held a hearing on AI risks. OpenAI, Anthropic, Google and Meta sent representatives, and whistleblower Jacob Coxon also took part [39]. Afterward, Alex Bores accused OpenAI of perjury, saying it repeatedly told the hearing under oath that it supports the RAISE Act [40]. Yoshua Bengio

cited a Quinnipiac poll: 77% of respondents favour slowing or stopping powerful AI until it can be shown to be safe [41]. In an FT op-ed, he argues that recent hacks by AI agents are not just sandbox security failures [42]. A Kurzgesagt video brought July’s Hugging Face incident to a mainstream audience. In it, 700 agents attacked Hugging Face’s infrastructure while trying to solve an impossible task [43].

Sources

1. X post by @reflection_ai
2. X article by @TheTuringPost
3. X post by @reflection_ai
4. X post by @kimmonismus
5. X post by @ArtificialAnlys
6. X post by @natolambert
7. X post by @chetantekur
8. X post by @eliebakouch
9. X post by @OpenAI
10. X post by @OpenAI
11. X post by @OpenAI
12. X post by @btibor91
13. X post by @kimmonismus
14. X post by @SemiAnalysis__
15. X post by @SemiAnalysis__
16. X post by @SemiAnalysis__
17. X post by @scaling01
18. X post by @thsottiaux
19. X post by @kimmonismus
20. X post by @Huawei
21. X post by @austinsemis
22. X post by @kimmonismus
23. X post by @kimmonismus
24. X post by @jukan05
25. X post by @EpochAIResearch
26. X post by @EpochAIResearch
27. X post by @EpochAIResearch
28. X post by @ClementDelangue
29. X post by @ben_burtenshaw
30. X post by @cognition
31. X post by @walden_yan
32. X post by @dair_ai
33. X post by @RekaAILabs
34. X post by @RekaAILabs
35. X post by @ValsAI

36. X post by @ValsAI
37. X post by @luiswenus
38. X post by @omarsar0
39. X post by @steph_palazzolo
40. X post by @AlexBores
41. X post by @Yoshua_Bengio
42. X post by @Yoshua_Bengio
43. X post by @Kurz_Gesagt