Opus 5.5 and GPT-6 Sol/Luna Put Work per Dollar in Focus

AI High Signal Digest

2026-09-23

Opus 5.5 and GPT-6 Sol/Luna Put Work per Dollar in Focus

By AI High Signal Digest • September 23, 2026

Anthropic and OpenAI pair frontier-model releases with lower prices, but independent evaluations show that token use and task-specific regressions complicate the savings story. New agent-training methods, RL infrastructure, commerce integrations, and serving updates round out the period.

Top Stories

Why it matters: Buyers need workload-level cost and quality comparisons; lower token rates do not guarantee cheaper completed work. [1, 2]

Anthropic's Opus 5.5 is described as matching Claude Fable 5.1 on most tasks, running about 30% faster, and costing about 40% less per task than Opus 5. API rates fell 20% to \$4/\$20 per million input/output tokens; cache reads are 60% cheaper. Artificial Analysis scored it 58 at max effort, its highest measured score, with leading results on six of ten evaluations. But it used 1.6× Opus 5's output tokens, and its max-effort Intelligence Index task cost was nearly unchanged (\$5.98 versus \$5.86). ValsAI found regressions against Opus 5 on medical coding, public benefits, legal research, tax, and HLAB, concentrated in retrieval, fidelity, and rubric-following. Anthropic also says it added Fable 5.1-class safeguards for cyber, bio, and frontier-LLM development, with flagged requests falling back to another model. [3, 4, 5, 6, 1, 7, 8]

OpenAI's GPT-6 Sol and Luna carry much of Astra's capability into faster, more affordable models. Listed API rates are \$2/\$10 and \$0.10/\$0.50 per million input/output tokens, respectively; they are available via API and rolling out in Work and Codex, while Free and Go users can access Luna in the desktop app, not Chat. Artificial Analysis found their Intelligence and Coding Agent Index scores broadly level with GPT-5.6, but cost per Intelligence Index task fell about

50% for Sol and 60% for Luna. Sol gained two coding points, Luna lost two, and both regressed on GDPval as shorter deliverables more often omitted rubric elements. [9, 10, 2, 11]

Research & Innovation

Why it matters: Agent progress increasingly depends on feedback loops that teach systems from real failures and reward good process, not just correct answers. [12, 13]

Perplexity's Computer agent is post-trained on real user sessions after synthetic-environment reinforcement learning. The company says it imitates useful steps and distills validated hints about tool errors into hint-free behavior, while excluding sessions with personally identifiable information and from users who opted out. In a live A/B test, tool-call failures fell from 2.24% to 1.77% without inference-time hints; the annotation process checks hints against information available before the mistake to reduce hindsight bias. [14, 15, 12, 16, 17]

A follow-up on Xiaomi's MiMo-V2.6 describes RL rewards for code quality, exploration, and testing—not only correctness—alongside an SFT-trained agent that filters reward-hacking trajectories and penalties for length and tool-call or format errors. ValsAI ranks Pro and Flash first and second among open-weight models, nearly tied at 59.73 and 59.69, with reported costs of $0.39 and $0.20 per task. Both remain below the Code Migration frontier, and Flash trails Pro by nine points in Legal Research. [13, 18, 19, 20, 21]

ValsAI also reports that ten Opus 5.5 agents devised and Lean-verified a faster shortest-path algorithm in 15 hours, describing it as an improvement over published bounds. [22]

Products & Launches

Why it matters: Consumer agents are being connected to transactions, while vendors package browser and orchestration workflows for everyday use. [23, 24]

Meta's Muse announced commerce integrations: PayPal says Muse agents will let customers shop and check out across PayPal merchants worldwide; Expedia and Instacart describe forthcoming connections for trip planning and grocery orders. [23, 24, 25]

Kimi's browser extension is available now: from a sidebar it can navigate websites and fill forms, while recorded routines can be saved as reusable skills. [26]

DigitalOcean's Managed Agents entered public preview, letting users run Claude Code, Codex, or LangGraph agents in runtimes that pause when idle, with tools behind one governed endpoint and access to 75+ models. [27]

Industry Moves

Why it matters: Scaling agents requires more efficient execution infrastructure and new routes into the AI talent pipeline. [28, 29]

A post quoting DeepSeek's **Elastic Compute** paper describes one production-scale unit spanning about 160 nodes, serving roughly 3 million sandboxes a day, supporting over 380,000 concurrent sandboxes, and creating more than 5,000 per second. The poster says a typical sandbox uses only around 5% of provisioned CPU time and that DSec aims to raise utilization. [28, 30]

The **Horowitz Andreessen Academy** raised $42 million led by a16z to build a selective San Francisco school for young builders. Its ten founding partners include Anthropic, Google, Meta, NVIDIA, and OpenAI; the initial one-year fellowship is tuition-free and begins in Fall 2027, with project work in place of traditional grades and tests. [29]

Quick Takes

Why it matters: Serving, speech, and coding updates offer concrete checks on progress beyond the headline model launches.

- **Serving:** vLLM v0.30.0 reports 5.2–7.7% higher end-to-end throughput for Kimi K3 mixed batches and a 3.23× speedup for Gemma 4 multimodal prefixes on RTX PRO 6000. [31]
- **Speech:** StepAudio 3 ASR reaches 1.7% WER, tied at the top of Artificial Analysis's index; it runs at 88× real time but costs $0.40 per hour and trails on long earnings calls. [32]
- **Coding:** Arena ranks Grok 4.7 #10 on Code Arena: WebDev at 1,632 points, 16 points and six places above Grok 4.6 (the comparison uses xHigh versus High effort). [33]

Sources

1. X post by @ArtificialAnlys
2. X post by @ArtificialAnlys
3. X post by @claudeai
4. X post by @ClaudeDevs
5. X post by @ClaudeDevs
6. X post by @ArtificialAnlys
7. X post by @ValsAI
8. X post by @ClaudeDevs
9. X post by @OpenAI
10. Introducing GPT-6 Sol and Luna | OpenAI
11. X post by @ArtificialAnlys
12. X post by @perplexity_ai

13. X post by @nrehiew_
14. X post by @perplexity_ai
15. X post by @perplexity_ai
16. X post by @perplexity_ai
17. X post by @perplexity_ai
18. X post by @ValsAI
19. X post by @ValsAI
20. X post by @ValsAI
21. X post by @ValsAI
22. X post by @ValsAI
23. X post by @PayPal
24. X post by @Expedia
25. X post by @Instacart
26. X post by @Kimi_Moonshot
27. X post by @digitalocean
28. X post by @teortaxesTex
29. X article by @gaganbiyani
30. X post by @teortaxesTex
31. X post by @vllm_project
32. X post by @ArtificialAnlys
33. X post by @arena