

Gemini 4 Carbon Takes #1 on Artificial Analysis v5, Two Weeks After Argon

AI High Signal Digest

2026-10-29

Gemini 4 Carbon Takes #1 on Artificial Analysis v5, Two Weeks After Argon

By AI High Signal Digest • October 29, 2026

Google previewed Gemini 4 Carbon, which leads the new Artificial Analysis index and several third-party leaderboards but is so far limited to cyber defenders. Also covered: GitHub’s model-orchestrating Copilot Fusion, Cartesia’s byte-level H-Nets, NVIDIA’s open transcription model, and new evals on immigration law and on how models detect evaluations.

Google ships Gemini 4 Carbon, but only to a small group

Sundar Pichai introduced Gemini 4 Carbon as Google’s next frontier model, two weeks after Argon. It ranks #1 on Artificial Analysis Intelligence Index v5.0 with a score of 59. Google’s reported cybersecurity scores are 88.2% on SEC-bench Pro (finding vulnerabilities), 54.8% on ExploitGym (exploitation) and 71.0% on CWE-bench (patching) [1]. Google also says that on its internal benchmarks, Carbon more than doubles Argon’s success rate at generating secure code [2].

You can’t use it yet. Some trusted cyber defenders get access this week. Google will then extend it to more partners through its Fairwind Program before wider availability “soon” [1].

Independent results are strong, with limits: - Artificial Analysis:

Carbon leads 4 of the index’s 9 evaluations: HLE-Diamond, Terminal-Bench-Science, AA-Omniscience and AutomationBench-AA. It does not lead in coding or knowledge work [3]. On the Coding Agent Index it scores 59, behind Claude Code running Opus 5.5 at 62 [3]. Each index task costs \$7.76, about 23% less than Opus 5.5 but about 1.2× GPT-6 Astra [3]. Accuracy on AA-Omniscience rose to 58%, but the hallucination rate also rose, from 4% to 11%. That is still the second-lowest of any model; Opus 5.5 and Astra are both at 50% [3].

List pricing is \$4/\$20 per million input/output tokens with a 1M-token context window [3]. - **Arena:** Carbon is #1 in Text Arena, #2 in Agent Arena and #4 in Code Arena: WebDev. Its median Agent Arena cost is \$1.80 per task, 72% below the #1 model, Opus 5.5 (Max) [4]. - **Vals:** Carbon is #2 on the Vals Index and #1 on Vals' Finance Index and Legal Research Benchmark [5]. The tradeoff is speed. Finance and tax agent tasks take more than twice as long as with Argon, and legal research takes over 50 minutes per task [6].

One caveat: Artificial Analysis ran Carbon's coding tests in Google's own Anti-gravity CLI, and at least one developer plans to retest it in other agent tools such as Codex [7]. On Google's earnings call, Pichai said AI is now speeding up Google's training and launch cadence. He expects "continued acceleration because of RSI" over the next twelve months [8]. A Google researcher also hinted that a stronger model is still unreleased: "Gemini [redacted] isn't yet here though" [9].

Routing between models comes close to the best single model

GitHub released Copilot Fusion, which picks among Claude Opus 5.5, Sonnet 5.5, GPT-6 Astra and GPT-6 Sol for each task, balancing cost and latency [10]. On Artificial Analysis's Coding Agent Index it scores 2 points below Claude Code with Opus 5.5, at half the cost per task [10]: \$5.59 versus \$11.42 [10]. It leads Terminal-Bench 4.0 with 71 but trails on FrontierCode, 51 versus 56 [10].

Open-weight models: the debate over AA v5 continues

Last cycle's index gap is still being argued over. Steve Hsu calls DeepSeek V4.1 Flash 1017's FrontierCode 2.0 result "close to Fable 5.1." He argues Chinese labs fell in v5 because they never prioritized math and science, and predicts they will shift some resources toward those areas [11]. A post citing The Information claims Kimi 3.1 Pro distilled GPT-6 Astra [12]; this has not been independently confirmed.

Meanwhile, the new startup Particle launched particle-1, built on DeepSeek V4.1 Flash 1017 [13]. Particle claims it matches performance while using about 45% fewer tokens [14].

Fangcun AI published an independent safety assessment of that same DeepSeek model, based on 34 evaluations [15]. The model labeled 32 of 69 real user conversations as evaluations, a 46.4% false-positive rate. The team had not established whether these mistaken judgments changed the model's behavior [15]. Most of the team's private benchmarks also triggered less evaluation awareness than popular public ones. The authors conclude that public benchmarks alone may not predict how a model behaves in deployment [15].

Research

- **Byte-level models:** Cartesia’s Albert Gu reports that a new version of H-Nets, a hierarchical architecture that learns from raw bytes instead of fixed tokens, gets up to $3\text{--}5\times$ efficiency gains in controlled scaling-law experiments. At larger scale it matches an open-source model trained on 11T tokens with about $10\times$ less data and compute [16]. Tri Dao notes that it learns to split variable names into separate tokens and scales much better than standard BPE tokenization on code [17]. Will Depue cautions that the experiments were “very small scale,” and gains often shrink as models grow [18].
- **NanoGPT speedrun:** Herman Brunborg cut the record from 39.9s to 6.7s, a larger drop than every record since 2025 combined. He did it by feeding the model retrieval-based n-gram hints [19]. Tatsu Hashimoto notes that the big gains came from AI-assisted humans rather than fully automated research [20].

New evaluations

Vals’ Immigration Bench tested 25 models on 262 private, expert-graded questions [21]. The best model, Claude Fable 5.1, was fully correct only 26.3% of the time [22]. Models met 7.3% of rubric items on 2025–26 rules, against 64.1% on established law. Even with web search, the best model met only 43.8% [23].

Launches

- **NVIDIA Nemotron 3.5 Transcribe:** an open-weight model with 30B total and 4B active parameters. It ties for #1 among open models on Artificial Analysis’s word-error-rate index at 2.8% [24]. It transcribes 40 languages, labels up to eight speakers, handles recordings up to an hour and answers questions about the audio [25].
- **OpenAI:** eligible ChatGPT Pro and Business subscribers can now claim up to \$500 a month in API credits [26]. Pro tiers get \$100, \$200 or \$500; Business plans are pooled up to \$500 per workspace [27]. Separately, Pro users in the US can now text their “dot” agent through iMessage or RCS, in preview [28].

Anthropic S-1: an accounting dispute

Jamin Ball argues that nearly all of Anthropic’s roughly \$80B first-half net loss is a non-cash mark-to-market increase on unconverted notes and SAFEs held by strategic investors, not a real loss [29]. Jamie Quint estimates Amazon’s 2023 and 2024 deals implied valuations of \$11.2B and \$55.4B, in line with Anthropic’s past fundraising. He calls the “supplier giveaway” framing disingenuous [30].

Sources

1. X post by @sundarpichai
2. X post by @Google
3. X post by @ArtificialAnlys
4. X post by @arena
5. X post by @ValsAI
6. X post by @ValsAI
7. X post by @Yuchenj_UW
8. X post by @AndrewCurran_
9. X post by @zacharynado
10. X post by @ArtificialAnlys
11. X post by @hsu_steve
12. X post by @kimmonismus
13. X post by @Suhail
14. X post by @Suhail
15. X post by @ZhihuFrontier
16. X post by @_albertgu
17. X post by @tri_dao
18. X post by @willdepue
19. X post by @hermanbrunborg
20. X post by @tatsu_hashimoto
21. X post by @ValsAI
22. X post by @ValsAI
23. X post by @ValsAI
24. X post by @ArtificialAnlys
25. X post by @kimmonismus
26. X post by @OpenAIDevs
27. X post by @OpenAIDevs
28. X post by @ChatGPT
29. X post by @jaminball
30. X post by @jamiequint