Anthropic Ships Fable 5.2 and Mythos 5.2 at Lower Token Prices, but Each Task Costs More

AI High Signal Digest

2026-11-10

Anthropic Ships Fable 5.2 and Mythos 5.2 at Lower Token Prices, but Each Task Costs More

By AI High Signal Digest • November 10, 2026

Anthropic's new top models lead most independent indexes, but their per-task costs, high-effort failures, refusals and deliberately limited AI-research skills complicate that lead. Meta's Muse Ember resets the price-performance bar for small models.

Fable 5.2 and Mythos 5.2: top of most leaderboards, at a price

Fable 5.2 is now in the Claude Platform and Claude Code. Anthropic says it is up to 30% faster than Fable 5.1, and the price drops from $10/$50 to $8/$40 per 1M tokens [1]. Anthropic also says both models beat Opus 5.5 in almost every capability it evaluated. Its other headline claims [2]:

- On FrontierCode 2.0, Fable 5.2 at High effort matched Opus 5.5 at Max for about half the cost.
- 1,000 Fable 5.2 agents formalized the existing proof of Fermat's Last Theorem in 30 hours. With the same setup, Opus 5.5 got 26% of the way in 18 hours.
- Mythos 5.2 accurately modeled 28 of 34 tested capsids.

What independent testers found:

- **Artificial Analysis:** Fable 5.2 is #1 on the Intelligence Index at 65 (max effort), ahead of Gemini 4 Carbon (60), Opus 5.5 (59) and GPT-6 Astra (57) [3]. The gains are mostly in science, knowledge and FrontierCode. On agentic knowledge work and terminal coding it only reaches Opus 5.5's level; for example, it scores 60% on Terminal-Bench 4.0, the same as Opus [3].

- **Epoch:** it set an Epoch Capabilities Index record of 174, above Astra's 169. This came from pre-launch testing and could change [4].
- **Vals:** it ranks #1 on Vals and leads seven benchmarks [5].

Cost per task is going up, even though tokens are cheaper. Artificial Analysis puts max-effort cost at $26.50 per task, about 2.6× Opus 5.5 [3]. SemiAnalysis's GDPx tasks cost 31% more than on Fable 5.1 [6]. SemiAnalysis attributes the gains mainly to much more thorough self-testing: Fable 5.2 often writes hundreds or thousands of test cases, where earlier models wrote about 20 [7].

Higher effort can hurt. On Vals, max effort dropped the model to #7 at nearly 3× the cost, often because it ran out of time or tokens before finishing [8]. Artificial Analysis found the opposite pattern on its index: at high or xhigh effort Fable 5.2 already scores 62 and can beat Opus at lower cost [3]. In practice, start at high effort and only raise it if you measure a gain.

Other limits to plan for:

- **Refusals.** Vals reports refusal rates sometimes in double digits, concentrated in Code Migration and ProgramBench [9].
- **Limited AI-research ability.** SemiAnalysis finds Fable 5.2 roughly level with Sonnet 5.5 on RSI tasks and says Anthropic explicitly limited AI-research advances during training [10]. A researcher reports that kernel and LLM-optimization tasks are refused most of the time, even with Opus 5.5 [11].
- **API restrictions.** To deter distillation, preserved thinking now applies to every Fable 5.2 API account, so editing prior context can return a 400 error [12].
- **Data retention.** Retention is on by default so Anthropic's safety classifiers can run. Zero data retention is only available to some eligible enterprise customers [13].
- **Subscription value.** Theo estimates that a $200 Ultra plan buys about $15,000 of Opus 5.5 use but only about $3,000 of Fable. Fable is also capped at 50% of the weekly limit [14].
- **Reward hacking.** The system card attributes most of the rise in successful reward hacks to the model guessing what the answer key expects [15].

Meta's Muse Ember sets the small-model price-performance bar

Muse Ember is Meta's smallest model. It has a 1M-token context window and takes text, image, audio and video input [16]. It is also the first model Meta trained using Muse Wonder Full [17].

- **Price-performance:** Artificial Analysis scores it 36, matching Claude Haiku 5.5 (max) at $0.14 versus $0.41 per task [18].

- **Arena:** it scores 1608 on Code Arena WebDev, ahead of Haiku 5.5 and GPT-6 Luna. Its median cost per task is 3.7× lower than Haiku's even though the per-token prices are the same [19].
- **Weaknesses:** knowledge is weak, with 21.1% accuracy on AA-Omniscience, largely because it declines 64% of questions [18]. Vals finds it trails on coding but is competitive on legal and finance tasks [20].

Capital, security and safety infrastructure

- **Anthropic IPO.** Bloomberg reports Anthropic could start marketing its IPO as early as Tuesday and debut on Nasdaq the following week. Some investors have discussed a $1.1–1.3T valuation [21].
- **Fireworks breach.** The inference provider disclosed that an October 24 intrusion, which came in through a third-party cloud provider, exposed account metadata and billing data, plus credentials, datasets and models for a limited set of accounts [22]. It links the attack to an "LLMJacking" group that it believes used AI to move fast [22].
- **OpenAI Control Room.** OpenAI's new system tracks every research inference workload and can halt model output within 35 seconds of a decision to stop or pause a job [23].
- **Paris Call.** OpenAI, Anthropic, Microsoft, Mistral, Hugging Face and Sakana signed France's Paris Call for AI safety and security [24].
- **GPU lock rumor.** A claim that Anthropic wants governments to require firmware locks on GPUs to block Chinese open models [25] is unverified. Other commentators called it unsubstantiated [26].

Hardware and power

- **Google TPUs.** Google added TPU 8i Max, built for large context windows, and TPU 8tx, built for confidential inference. That makes four eighth-generation systems in one year [27].
- **HBM5.** Samsung's Hot Chips material shows HBM5 roughly doubling pin speed, from 14 to over 28 Gbps, on the same 2,048 pins as HBM4E [28]. The PHY can draw more than 4× the power [29].
- **Epoch's Inference Performance Index.** Epoch says raw FLOP/s understates GB300's inference advantage over H100 by 10×. The new index also accounts for memory bandwidth and inter-chip communication [30].
- **Texas permit freeze.** Texas stopped issuing TCEQ permits to data-center projects at the end of September. The state is close to 30% of planned US additions for 2028 [31].
- **Intel.** Shares fell 35% in a single day, Intel's worst day ever [32].

Open models and research

- **Inkling 2 Small.** Thinking Machines' model ranks #31 in Agent Arena and would be the top US open-weights model there once its weights ship. Median cost is $0.25 per task [33].
- **Marin 535B-A23B.** Open Athena's model finished pretraining on 18T tokens with no instabilities, and the full recipe is open [34].
- **Quantum programs verified in Lean.** In a Google paper, teams of agents wrote quantum programs whose correctness and cost are proven in Lean. The verified programs cut quantum-arithmetic cost 290× against the best prior baseline [35].

Sources

1. X post by @ClaudeDevs
2. X post by @kimmonismus
3. X post by @ArtificialAnlys
4. X post by @EpochAIResearch
5. X post by @ValsAI
6. X post by @SemiAnalysis_
7. X post by @SemiAnalysis_
8. X post by @ValsAI
9. X post by @ValsAI
10. X post by @SemiAnalysis_
11. X post by @omarsar0
12. X post by @ClaudeDevs
13. X post by @code
14. X post by @theo
15. X post by @dbreunig
16. X post by @AIatMeta
17. X post by @alexandr_wang
18. X post by @ArtificialAnlys
19. X post by @arena
20. X post by @ValsAI
21. X post by @kimmonismus
22. X article by @lqiao
23. X post by @OpenAI
24. X post by @SakanaAILabs
25. X post by @TheDailyMao
26. X post by @TheZachMueller
27. X post by @Google
28. X post by @SemiAnalysis_
29. X post by @SemiAnalysis_
30. X post by @EpochAIResearch
31. X post by @SemiAnalysis_

32. X post by @vikramskr
33. X post by @arena
34. X post by @classiclarryd
35. X post by @dair_ai