Fable 5.1 Makes Reasoning Effort a First-Class Coding-Agent Budget

Coding Agents Alpha Tracker

2026-09-02

Fable 5.1 Makes Reasoning Effort a First-Class Coding-Agent Budget

By Coding Agents Alpha Tracker • September 2, 2026

Claude Fable 5.1's sharp quality/latency/cost curve is the lead signal; the rest of the brief turns that lesson into release harnesses, resumable jobs, cache-aware routing, and safer provider fallbacks.

TOP SIGNAL

Claude Fable 5.1 is the day's clearest coding-agent release: it exposes five reasoning levels—`low`, `medium`, `high`, `xhigh`, and `max`—with no option to turn reasoning off. [1] On Simon Willison's same SVG prompt, low/medium took about 23 seconds and ~$0.10, high took 29.6 seconds/$0.13, xhigh took 7m51s/$1.83, and max took 13m54s/$3.30; max produced his best Anthropic result. [1] Artificial Analysis found the catch: max scored 66 but cost 20% more per task than Fable 5 because it used ~1.7× more output tokens, despite a 75% cache-read cut. [2]

TRY THIS

- **Stage the reasoning budget.** Use `low`/`medium` for drafts and routine edits, reserve `xhigh`/`max` for quality-critical generation, then use a normal `high` pass to transform the result. Willison's exact second pass after generating the max-effort SVG was:

  ```
  llm logs -cx | llm -m claude-fable-5.1 -s 'animate this'
  ```

  That pass used 6,121 input and 26,201 output tokens, cost $1.37, and produced a good animation with the wheels rotating the wrong way. [1]

- **Make release tests proof-carrying.** Prime defines a custom harness as the wrapper around the agent, not the model itself. [3] His test de-

sign pairs each definition with agent instructions and required proof; the corrected queue design calls for one Linear ticket and one pending result per definition, carrying the server and ISO URLs. Run agents against each new build, record what happened and why, mark `success`, `failed`, `timed out`, or `aborted`, and return the failures. [3] Add targeted review of state transitions: Theo's only obvious Fable 5.1 miss in T3 Code was an incorrect assumption about auto-settle logic, even though he otherwise reports no issues and says it caught mistakes in earlier Fable 5 code. [4]

- **Checkpoint long-running jobs.** OpenWiki 0.5.0 checkpoints completed pages, resumes interrupted runs instead of restarting, preserves partial progress through CI failures, and uses the same lifecycle for native and coding-agent integrations. Apply the pattern to agent work: persist every completed unit, keep a durable run ID, and make restart resume from the last checkpoint. [5]
- **Route for cache locality, not just model quality.** OpenAI's prompt cache makes a request 90% cheaper, but each cache key tops out around 15 requests per second; Unify says custom routing around that ceiling brought it close to a 95% hit rate. Keep the reusable prompt prefix stable, watch per-key throughput, and route before one hot key saturates. [6]

WHAT SHIPPED

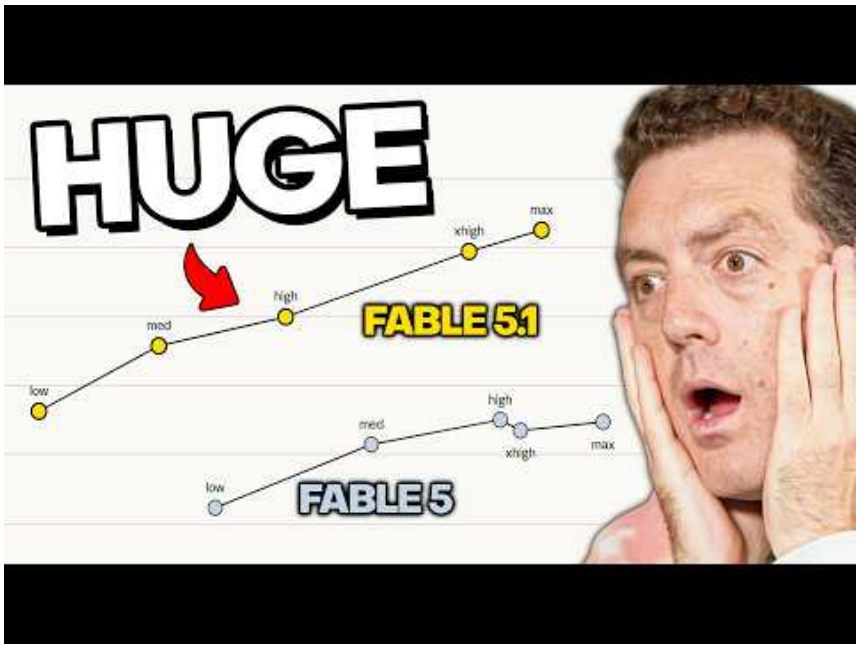
- **Claude Fable 5.1 / Mythos 5.1.** Matthew Berman reports that both models landed. Fable 5.1 is available in Cursor, which reports 73.4% at max effort on CursorBench 3.2 and says the model is particularly good at verifying its own work end to end. [7, 8] For API, Enterprise, and SDK customers, cache reads are now $0.25 per million tokens rather than $1; Boris Cherny says a typical Claude Code session can be up to 38% cheaper. [9]
- **Fable deployment caveat: retention is not the same as no training.** Kent C. Dodds flagged a Fable 5.1 modal whose text says request and output data are retained regardless of Privacy Mode, while Anthropic will not train on it. [10] Anthropic's Enterprise Frontier Safeguards keep company data in the customer's cloud and add automated monitoring for risky patterns. [11] Treat retention, provider access, and monitoring terms as a preflight before sending proprietary code.
- **T3 Code build.** Theo shipped Fable 5.1 support, a GitHub-backed model list so new models do not require a product release, server-side thread-settle logic, large reductions in long-thread memory and idle CPU use, smaller catch-up payloads, better Windows performance, hardened worktree setup, and an OpenCode subagent-stop fix. [12]
- **Software factories for open source.** Vercel's AI SDK—reported at more than 20 million npm downloads per week—deployed separate agents

for reproduction, fixes, and review against a backlog of over 1,000 issues and almost 800 PRs. Four weeks in, Vercel claimed the factory authored 25–35% of merged PRs and closed 70–80% of issues; the system combines a UI, web app, API, execution space, sandboxes, and GitHub triggers. [13] Astro's Flue takes the contribution policy further: every external PR is converted into an issue or discussion, then agents handle research, design, implementation, and initial review after a decision is made. [13]

- **OpenWiki 0.5.0.** The durable/resumable lifecycle is now shared by the project's native and coding-agent integrations, making it a concrete reference implementation for restart-safe agent work. [5]
- **Provider coupling hit Cursor.** Matthew Berman reports that OpenAI intends to wind down its native-model contract with Cursor after SpaceX's acquisition, with native access blocked in three months; users can still bring an OpenAI API key, use the Codex IDE extension, or connect through compatible gateways. [14] Keep a BYO-key or gateway path even when a bundled integration is convenient.
- **GLM 5.3 Flash / "Ox Alpha."** Fireship reports MIT-licensed weights, a 320-billion-parameter mixture-of-experts design, a 1-million-token context window, and pricing of \$0.15/\$0.50 per million input/output tokens—advertised as up to 40× cheaper than Claude during the launch discount. In a firsthand test, it modernized a 2016 AngularJS app, diagnosed a mobile CSS overflow bug with vision, and used FFmpeg to analyze a supplied video. The endpoint's fine print retained prompts, and the model was slow, verbose, and prone to doom loops: use it for non-sensitive experiments, not blind production routing. [15]
- **The desktop agent is shipping its own toolchain.** Simon Willison found the ChatGPT desktop app's 1.7 GB `codex-primary-runtime` bundling full Python and Node installations plus native Git, Poppler, and headless LibreOffice binaries; plugin skills tell Codex how to locate and use them. [16]

GO DEEPER

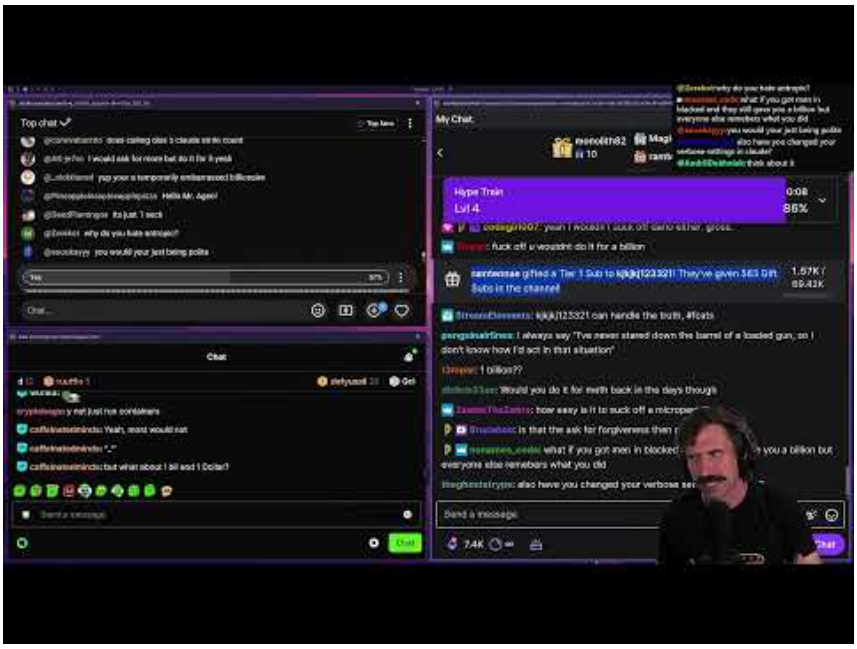
- **Matthew Berman — "Fable 5.1 is the BEST AI Model"** — approximately **10:30–13:30**, the cost-per-task section. Berman puts the Terminal-Bench/CursorBench gains next to Artificial Analysis's output-token penalty; use it to build a task-cost sheet instead of repeating leader-



board claims. [7]

Fable 5.1 is the BEST AI Model (Cancel Your Other Subscriptions) (12:23)

- **ThePrimeTime — "Omarchy Automation Investigation (Building Custom Harness)"** — approximately **42:00–47:00**, the test-definition and release-gauntlet segment. Follow the instruction → proof → Linear ticket → ISO run → result-state chain; it is a useful blueprint for making computer-use agents produce evidence, not just screenshots. [3]



Omarchy Automation Investigation (Building Custom Harness) (39:12)

- **Fireship — "The mystery is solved... and the answer is 40x cheaper than Claude"** — approximately **03:30–07:00**, the GLM 5.3 Flash identification and AngularJS test. The interesting part is the combination of multimodal debugging and video-to-content generation; the prompt-retention warning is the operational takeaway. [15]

The mystery is solved… and the answer is 40x cheaper than Claude (2:01)

- **Study Astro/Flue's contributor guide.** It is a live design for replacing drive-by AI PRs with issue/discussion intake while preserving community reporting and maintainer trust; after triage, agents handle research, design, implementation, and initial review. [13]

Editorial take: Coding-agent alpha is shifting from raw model choice to operating design: budget reasoning per stage, route around cache and provider constraints, and make every autonomous run both resumable and provable. [2, 6, 5, 3]

Sources

1. Claude Fable 5.1 made me a really nice animated pelican
2. X post by @ArtificialAnlys
3. Omarchy Automation Investigation (Building Custom Harness)
4. X post by @theo
5. X post by @colifran_
6. X post by @LangChain
7. Fable 5.1 is the BEST AI Model (Cancel Your Other Subscriptions)
8. X post by @cursor_ai
9. X post by @bcherny
10. X post by @kentcdodds
11. X post by @alexalbert___

12. X post by @theo
13. PRs NOT Welcome: How Top AI Open Source Projects Are Managing Thousands of Contributors
14. Cursor just got BANNED (It's because of Elon…)
15. The mystery is solved… and the answer is 40x cheaper than Claude
16. Codex bundles LibreOffice