Claude Fable 5.1 Workflow Breakdown

How to actually run long, unattended Claude Code sessions now that the model holds focus and costs less to do it.

Why this release changes how I use Claude Code

Most AI models get worse the longer a task runs. Context piles up, they lose the thread, and by hour two you're babysitting a session that started out sharp. Claude Fable 5.1, the model Anthropic released in September 2026, is built to do the opposite. It held focus through a documented 38-hour unattended run, and it costs up to 45% less to run on agentic tasks because of a 75% cut to cache read pricing.

That combination is what makes this worth setting up properly instead of just switching the model dropdown. A cheaper model that still drifts after an hour doesn't change your workflow. A model that stays coherent for a day and a half at a lower cost does.

01

How it actually works

1. Fable 5.1 keeps its own internal record of what it has tried, what worked, and what's left, instead of relying purely on the raw conversation history staying in context.
2. When it hits an error or a dead end, it can diagnose the cause, correct course, and reprioritize its own task list without a human re-explaining the goal.
3. Because cache reads are 75% cheaper, a long session that keeps re-reading the same project files and prior tool output costs a fraction of what it used to, which is why agentic workloads see the biggest savings (up to 45%).

02

The one thing almost everyone gets wrong

People assume a stronger model means they can hand off a task with a vague prompt and walk away. That's backwards.

> **WATCH OUT**
>
> Fable 5.1's autonomy is only as good as the goal you give it. In Anthropic's own case study, the 38-hour run worked because the engineer gave the model a clear problem, a way to check its own results, and permission to run experiments in parallel. Set it loose with a fuzzy goal and no way to verify progress, and it will confidently run in the wrong direction for just as long.

03

The numbers worth knowing before you start

METRIC	FABLE 5.1	FABLE 5
Cache read cost (per million tokens)	$0.25	$1.00
Typical workload cost	~25% lower	baseline
Agentic workload cost	~45% lower	baseline
Standard input / output pricing	$10 / $50 per million tokens	$10 / $50 per million tokens
Terminal-Bench 4.0 (agentic coding)	55.8%	42.0%
Terminal-Bench-Science 0.1 (agentic research)	52.6%	24.7%

SOURCE

- Anthropic: Claude Fable and Mythos 5.1
  https://www.anthropic.com/claude-fable-and-mythos-5-1

Set it up in Claude Code

1. Open Claude Code and select Claude Fable 5.1 as your active model for the session (model picker in settings or the /model command, depending on your Claude Code version).
2. If you're running a long unattended task, give the model a way to verify its own progress, for example a test suite, a benchmark script, or a clear success condition it can check without you.
3. Write the task as a real brief, not a one-liner: the goal, the constraints, what 'done' looks like, and what it's allowed to try on its own (retries, parallel experiments, installing dependencies).
4. If the task will run for hours, use a background or scheduled approach so you're not holding the terminal open, then check back instead of watching it live.
5. Review what it did before you trust the output. Fable 5.1 will tell you what it tried and why, so read that record instead of just checking the final result.

💡 QUICK TIPS

- Start with a task you'd normally give a junior engineer and a full afternoon, that's the range where the 38-hour-style autonomy actually pays off.
- The cache savings only kick in when the model is re-reading the same context repeatedly, so a long session with a stable project structure benefits more than a series of one-off unrelated prompts.

---

04 The commands I actually run

KICK OFF A LONG UNATTENDED RUN

```
I want you to work on [PROBLEM] until it's solved or you hit a genuine blocker. Check your own progress against [SUCCESS CONDITION / TEST]. If a result looks wrong, diagnose why before trying again. You can run multiple approaches in parallel if that's faster than one at a time. Keep a running log of what you tried and why.
```

ASK FOR A STATUS CHECK WITHOUT INTERRUPTING THE RUN

```
Without stopping the current work, summarize what you've tried so far, what worked, what didn't, and what you're doing next.
```

HAVE IT DIAGNOSE ITS OWN ERROR

```
This result looks off: [DESCRIBE WHAT LOOKS WRONG]. Before trying again, figure out whether this is a real problem with the approach or an artifact of [DATA SOURCE / TEST SETUP], and explain your reasoning.
```

SET UP PARALLEL EXPERIMENTS

```
Instead of one attempt at [TASK], run [NUMBER] different approaches in parallel: [APPROACH 1], [APPROACH 2], [APPROACH 3]. Compare the results and tell me which one to keep.
```

When it breaks

ERROR	FIX
The model isn't available as an option in your Claude Code model picker	Confirm your Claude Code version supports Fable 5.1 and that your account or API key has access; Fable 5.1 is generally available but new versions roll out over a short window.
A long unattended run drifted off task	Check whether you gave it a concrete success condition. Without one, even a strong model has nothing to check itself against, so it can wander confidently.
Costs were higher than expected on a long session	The cache savings depend on re-reading stable context. If your session keeps loading unrelated files or restarting from scratch, you're not getting the cache benefit, restructure the task so related work stays together.
You can't tell what the model actually did during a long run	Ask it directly for a running log or a summary of attempts, don't rely on scrollback in a session that's been going for hours.
Mythos 5.1 isn't available to you	That's expected. Mythos 5.1 is restricted to vetted cybersecurity and life sciences professionals through Anthropic's trusted access programs. Fable 5.1 is the generally available version and has the same underlying capability outside those restricted domains.

Ready to go deeper?

This playbook is the short version. Inside the hub you get the full build walkthroughs, the prompt library, and the systems I use to ship this in a weekend.

Join the Build →

https://hub.digicuratoragency.com/join