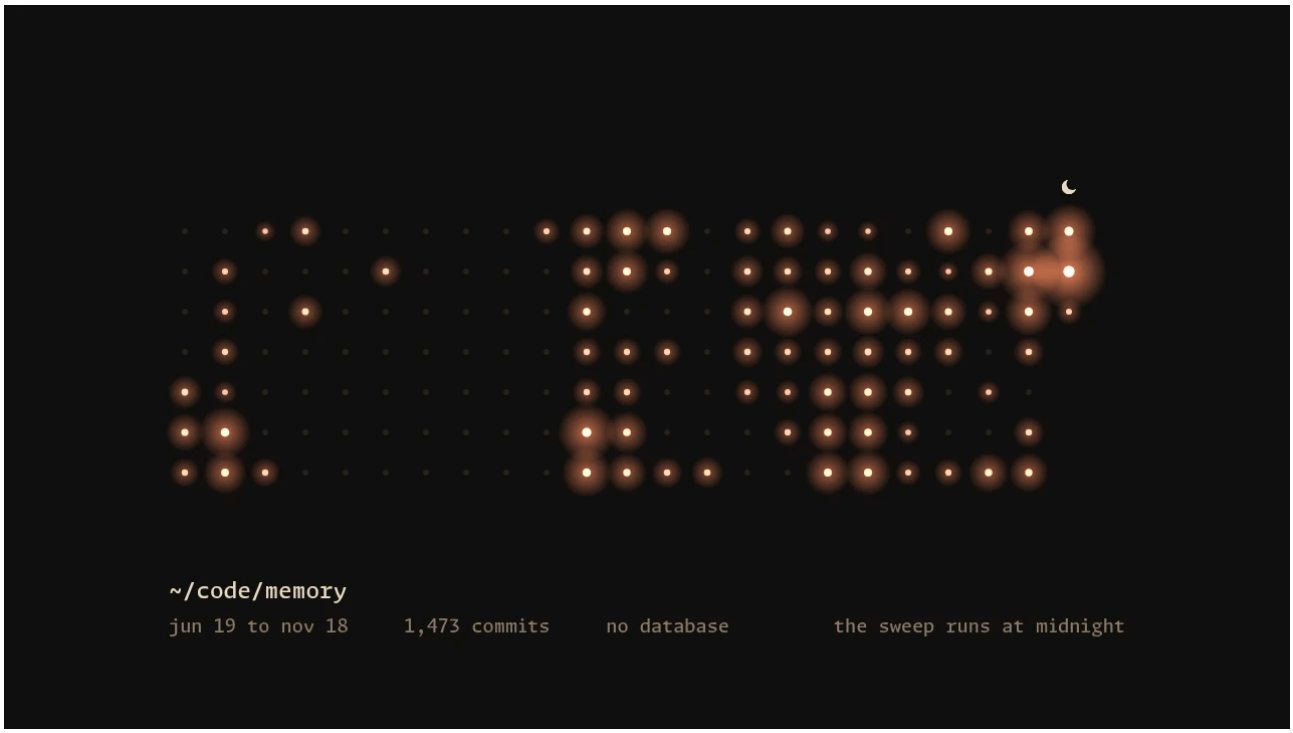
Roll your own LLM memory

Isaac Bythewood · 2026-11-18 · 3 min read

ai sysadmin



My Claude Code memory is a folder of Markdown in git that hooks write to and a nightly timer cleans up.

Since June I've had Claude Code keep a memory vault for me at `~/code/memory`. It's plain Markdown files in a local git repo with no database and no remote, and it holds everything, code and personal stuff alike. It's been running great and it's the main reason Claude feels like it knows what I'm working on.

The layout is simple. A `journal/` folder with a file per day, `projects/`, `decisions/` and `topics/` folders for notes that stay current, an `index.md` with one line per note, an `open.md` for whatever's still outstanding, and a `CLAUDE.md` with the rules for keeping it. My `~/CLAUDE.md` tells Claude to read `index.md` at the start of a task and open what it needs.

I didn't read up on any of this when I set it up, plain files in git just seemed like the obvious way to do it. The other day I had Claude compare it to the recent stuff like Karpathy's LLM wiki and Letta's sleep-time agents and it's mostly the same shape, so I was doing some of it before it was cool, or at least before I knew it was.

Getting it going

Writing to it is all hooks in `~/.claude/settings.json`. Stop, PreCompact and SessionEnd run `vault-capture.sh`, which every eight messages kicks off a background `claude -p` on Sonnet to read what's new in the transcript, update the notes and commit. The session never waits on it. SessionStart runs `vault-digest.sh`, which puts `open.md` and the latest journal headings in front of Claude.

Then a systemd timer on the host runs a sweep every night:

```
[Timer]
OnCalendar=*-*-* 00:00 America/New_York
Persistent=true
```

It catches any session the hooks missed and tidies up the notes that changed that day, dropping stale facts, merging duplicates and clearing finished items out of `open.md`.

All the scripts are in taproot under `dotfiles/claude/hooks/` with a writeup in `docs/vault.md`.

The new stuff

Up until this week the session itself stopped every few messages to write the vault, which worked fine but was slow and made way too many commits. The background writer and the nightly cleanup are new and I like them a lot so far. Sessions are faster, the index is smaller and the notes actually read like the current state of things instead of a pile of appends.

Roll your own

Don't trust your LLM harness to handle memory for you. Claude Code has a built-in auto memory and I ran it next to the vault for a couple weeks before turning it off with `"autoMemoryEnabled": false`. It wasn't in git, I couldn't see what it was doing, and it drifted until it was telling Claude things that weren't true anymore.

I think everyone should roll their own and know how it works. Use plain text files, or SQLite if you want to query it, and keep it somewhere you own. Mine doesn't care what model reads it, so if I switch providers or run a local model when Claude is down it all still works.