

The Model on Your Mac

A private, no-cost AI that answers questions about your live web apps — running on your laptop, not in the cloud.

JEFF FRANZEN · JUNE 23, 2026 · FOR DISCUSSION — INDEPENDENT RESEARCH

It already works. The live chapter-intelligence site at `chapters.jbf.com` can be opened, and a question — "What's the one-sentence takeaway for this chapter?" — answered in about half a second by a language model running entirely on a MacBook. No API key. No cloud. No tokens billed. Nothing left the machine.

The point of this brief is the idea that makes it possible, because it is reusable on every app I have shipped.

The one idea

A web page does not reach a model through the server it was downloaded from. It reaches the model through *the browser*. And my browser runs on my Mac — which is also where the model runs. So when I open my own live site, the page asks `localhost` for an answer, and `localhost` is this laptop.

That is the whole trick. The site can sit on Vercel, on a custom domain, anywhere — it makes no difference. The hosting is just where the HTML came from. The intelligence is local, summoned by a page that happens to be live. It feels like the website got smart. Really, the laptop did.

Why it answers in half a second, not five

SMALL MODEL	A 3-billion-parameter model (qwen2.5:3b) is plenty to read a report and answer from it. Bigger models are on tap when quality needs to climb.
PINNED	The model is held in memory so it never reloads between questions.
WARMED	The page's text is fed to the model the moment a report loads — so the <i>first</i> question is already cached, not read fresh.
TRIMMED	Only the text that answers the question is sent. Less to read is less to wait for.
MEASURED	0.14s to read the report, 0.42s to write the answer — ~0.57s total, on an M4 Max.

The slow part of local AI was never the thinking. It was the model loading off disk and re-reading a long prompt. Remove both, and it is instant.

Wire it into any app

The same four moves bolt this onto anything I have already built. It is a drop-in panel — a retractable sidebar that reads the page's own text and chats with the local model.

- 1

PICK THE MODEL

Choose the smallest one that answers well

qwen2.5:3b for speed; step up to 7b or a Gemma model when the answer needs more nuance. Switchable live, mid-demo.
- 2

OPEN THE DOOR

Whitelist the app's domain

The local model only answers pages it trusts. Add the app's address to the one setting (OLLAMA_ORIGINS) so the live site is allowed to ask. One time, per app.
- 3

DROP THE PANEL

Add one line to the page

A single script tag adds the Ask sidebar. Gated behind ?ask in the URL, so ordinary visitors never see it — only I do, when I want it.
- 4

MAKE IT INSTANT

Warm and pin

Load the model into memory and pre-read the page before the first question. Now every answer lands in about a second.

Local versus cloud — the honest trade

	LOCAL MODEL (THIS)	CLOUD API
Cost	Free. No tokens, no bill, ever.	Pay per question.
Privacy	Nothing leaves the Mac.	Data goes to a vendor.
Who can use it	Only me, on my Mac.	Anyone who opens the app.
Internet	Works with the Wi-Fi off.	Needs a connection.
Quality ceiling	Good; a small local model.	Highest; frontier models.

"The website didn't get smart. The laptop did — and the live page just gave it something to talk about."

– THE WHOLE BRIEF, IN ONE LINE

What it can't do — and the fix

This is a private tool, not a public feature. Other people opening the same live site get nothing, because their `localhost` is their machine, with no model on it. It also wants Chrome (Safari blocks the local call) and the Mac awake with the model running.

When the audience needs to be everyone, the page stops asking the laptop and asks a hosted model behind the app instead — that costs tokens but reaches the world. Or, for a short window, my Mac's model can be exposed through a tunnel and the page pointed at it. The local trick is for me: demos, private data, zero cost, no internet.

Independent research by a Red Cross volunteer. Not an official American Red Cross or FEMA application or position.

SOURCES: [CHAPTERS.JBF.COM](https://chapters.jbf.com) · [OLLAMA](https://ollama.com) · [APPLE](https://apple.com)
M4 MAX | 2 / 2