

zappa: an AI powered mitmproxy

Apr 15, 2026

Soon, AI will be good enough to interact with the Internet in an indistinguishable way from a human. This can be an amazing opportunity for liberation from all the people who are [targeting your attention](#).

I vibe coded this `zappa` proxy, it is not quite there yet, but I think it points the way forward. Why should I browse the Internet or use apps when machines can do it for me? Suckers getting billed for an ad impression from a 1 cent Qwen.

Instead of the source, I'll include the prompt in this post. I used GPT-5.4 to code it.

Download mitmproxy and configure Firefox to use a SOCKS5 proxy and install the required cert to proxy HTTPS traffic. Write a plugin for mitmproxy to route all website traffic through Qwen using the Cerebras API, you need to proxy HTML, JS, and CSS. Tell Qwen to remove all ads, popups, bright colors, moving things, and enshittified crap from the website and return a good version of the site. Pass this good version back to the user through the proxy. Log everything to a file. If the AI returns an error, pass that error along to the user, do not return pages without AI transformation.

Google Gemma 4 Runs Natively on iPhone With Full Offline AI Inference

April 15, 2026

On-device AI has been a talking point for years, but Google's latest move makes it harder to dismiss. Gemma 4, Google's open-source model family, now runs directly on iPhones, full local inference, fully offline. It's a meaningful step, and it signals that edge AI deployment isn't a future priority anymore; it's happening right now.

So, where does Gemma 4 stand against the competition? Early benchmarks position the 31B variant alongside Qwen 3.5's 27B model, a reasonably close matchup, with Gemma carrying roughly 4 billion additional parameters. Both models carry trade-offs, and neither is a clear sweep across every task.

Now Withdraw Without Waiting

5 reasons why people like trading futures with us.

Latest

DJI Osmo Pocket 4: Launch Date, Specs - What You Need to Know

OPPO Find X9 Ultra: 10x Optical Zoom & Dual 200MP Cameras Redefine Mobile Photography

Block's mesh-lin is Building a Decentralized AI Compute Network - Here's Why It Matters

vivo X300 Ultra Review: A Serious Camera Phone Built for Professional Results

Newsletter

Email address

I WANT IN

Don't miss

Google Gemma 4 Runs Natively on iPhone With Full Offline AI Inference

April 15, 2026

On-device AI has been a talking point for years, but Google's latest move makes it harder to dismiss. Gemma 4, Google's open-source model family, now runs directly on iPhones, full local inference, fully offline. It's a meaningful step, and it signals that edge AI deployment isn't a future priority anymore; it's happening right now.

So, where does Gemma 4 stand against the competition? Early benchmarks position the 31B variant alongside Qwen 3.5's 27B model, a reasonably close matchup, with Gemma carrying roughly 4 billion additional parameters. Both models carry trade-offs, and neither is a clear sweep across every task.

More compelling story, though, isn't the flagship size — it's the smaller ones. E2B and E4B variants are clearly engineered for mobile deployment, prioritizing efficiency over raw capability. Google's own app nudges users toward E2B, and that makes sense: it's faster, lighter, and better suited for real-world on-device conditions where memory and thermal limits matter.

Getting started requires nothing more than downloading the Google AI Edge Gallery from the App Store. From there, users select their preferred model variant and start running inference directly on their device. No API calls. No cloud dependency.

Google AI Edge Gallery isn't just a text interface. It bundles image recognition, voice interaction, and an extensible Skills framework, positioning it less like a demo and more like a platform for on-device AI experimentation. That framing matters; it suggests Google wants developers and power users to treat this as a foundation, not a feature.

Under the hood, Gemma 4 routes inference through the iPhone's GPU. In practice, responses arrive with notably low latency, a strong indicator that consumer hardware is now capable of sustaining this class of workload without visible degradation. That's not a minor footnote; it's the whole argument for why local AI deployment is becoming commercially viable.

Offline capability, in particular, changes the calculus for enterprise use cases — field applications, healthcare settings, and scenarios where data privacy rules out cloud processing entirely.

When all's said and done, Gemma 4 on iPhone isn't just a technical proof-of-concept. It's a signal that the on-device AI era has arrived — and for Google, the Gemma is definitely out of the bottle.

© 2023 Gizmoweeek, All rights reserved.

Hacker News

new | threads | past | comments | ask | show | jobs | submit

georgehotz (833) | logout

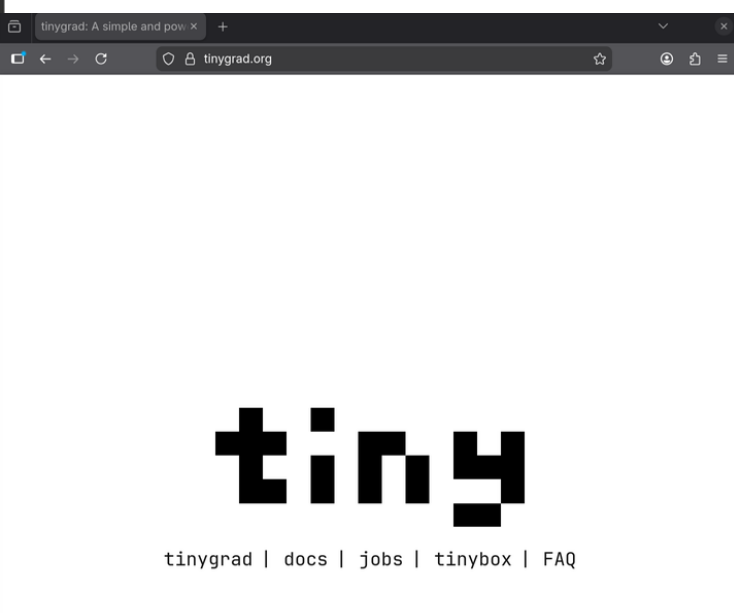
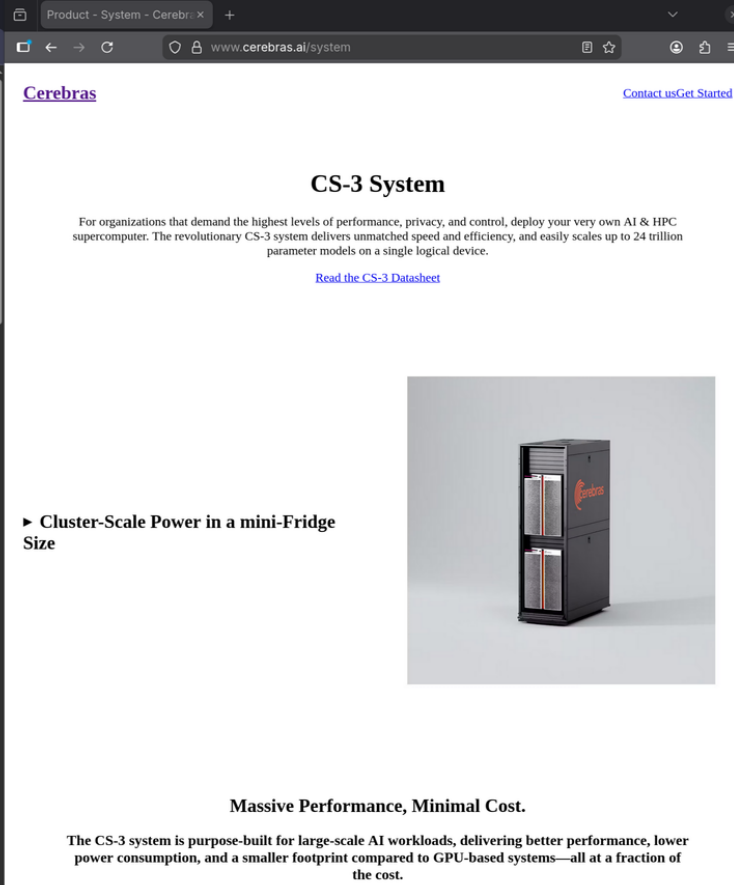
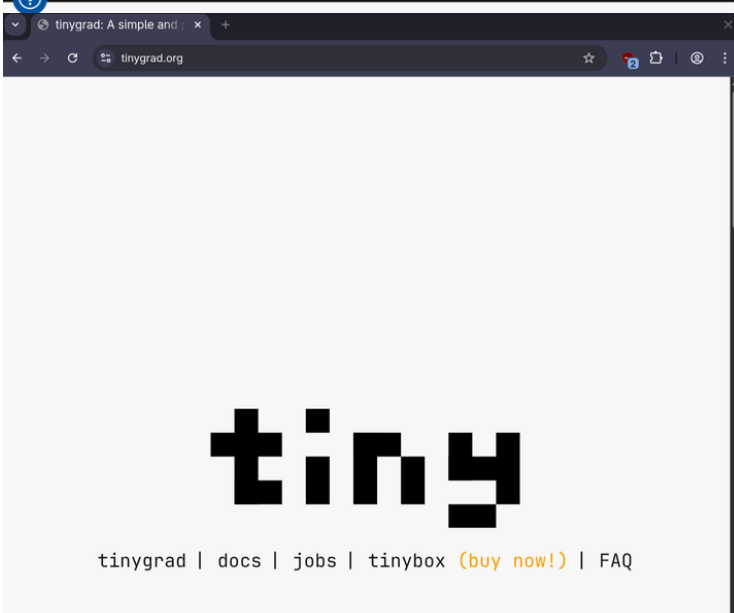
- RunEnv – secrets injected into memory at runtime, never on disk (runenv.dev)
1 point by derickolotu 0 minutes ago | flag | hide | past | discuss
- Your "Tiny" CSS Animation Might Be Burning CPU 24/7 (telesink.com)
1 point by kyrylo 3 minutes ago | flag | hide | past | discuss
- China Is in Deep Trouble over Trump's Iran Blockade [video] (youtube.com)
1 point by keepamovin 3 minutes ago | flag | hide | past | discuss
- Europe Is Accelerating a NATO Fallback Plan in Case Trump Pulls Out (wsj.com)
1 point by doener 5 minutes ago | flag | hide | past | discuss
- Clavicular Hospitalized After Suspected Overdose While on Livestream (hollywoodreporter.com)
1 point by helloplanets 5 minutes ago | flag | hide | past | discuss
- I Read Designing Data-Intensive Applications 2nd Edition and It's Awesome (medium.com/javarevisited)
1 point by teleforce 6 minutes ago | flag | hide | past | discuss
- My 14-Year Journey Away from ORMs (nikita-voikov.github.io)
1 point by _query 7 minutes ago | flag | hide | past | discuss
- It's OK to compare floating-points for equality (lisyarus.github.io)
1 point by mariuz 9 minutes ago | flag | hide | past | discuss
- Nixquick – Search, install and uninstall nixpkgs across your Nix configurations (github.com/gschurck)
1 point by grums 9 minutes ago | flag | hide | past | 1 comment
- Why are so many logs written for machines instead of humans? (github.com/exasol)
1 point by one-random-geek 10 minutes ago | flag | hide | past | 1 comment
- Anthropic chat log leak or nonsense? (pastebin.com)
1 point by hactually 12 minutes ago | flag | hide | past | discuss
- Netgear Scores the First Exemption from the FCC's Foreign-Made Router Ban (pcmag.com)
1 point by neogodless 14 minutes ago | flag | hide | past | discuss
- New: NoRegistry censorship-resistant DNS over BitTorrent's DHT, 1 command to run (noregistry.com)
2 points by Synonym-recruit 15 minutes ago | flag | hide | past | 1 comment
- China weighs curbs on exports of solar manufacturing equipment to US (reuters.com)
1 point by leonidasrup 15 minutes ago | flag | hide | past | discuss
- Why do companies change their logo/wordmark (e.g. StackOverflow)?
1 point by profsummergig 16 minutes ago | flag | hide | past | discuss
- Academic fraud may be the symptom of a more systemic problem (voxweb.nl)

Hacker News

new | past | comments | ask | show | jobs | submit | login

- RunEnv – secrets injected into memory at runtime, never on disk (runenv.dev)
1 point by derickolotu 0 minutes ago | hide | 1 comment
- Your "Tiny" CSS Animation Might Be Burning CPU 24/7 (telesink.com)
1 point by kyrylo 2 minutes ago | hide | discuss
- China Is in Deep Trouble over Trump's Iran Blockade [video] (youtube.com)
1 point by keepamovin 2 minutes ago | hide | discuss
- Europe Is Accelerating a NATO Fallback Plan in Case Trump Pulls Out (wsj.com)
1 point by doener 4 minutes ago | hide | discuss
- Clavicular Hospitalized After Suspected Overdose While on Livestream (hollywoodreporter.com)
1 point by helloplanets 4 minutes ago | hide | discuss
- I Read Designing Data-Intensive Applications 2nd Edition and It's Awesome (medium.com/javarevisited)
1 point by teleforce 5 minutes ago | hide | discuss
- My 14-Year Journey Away from ORMs (nikita-voikov.github.io)
1 point by _query 5 minutes ago | hide | discuss
- It's OK to compare floating-points for equality (lisyarus.github.io)
1 point by mariuz 5 minutes ago | hide | discuss
- Nixquick – Search, install and uninstall nixpkgs across your Nix configurations (github.com/gschurck)
1 point by grums 9 minutes ago | hide | 1 comment
- Why are so many logs written for machines instead of humans? (github.com/exasol)
1 point by one-random-geek 9 minutes ago | hide | 1 comment
- Anthropic chat log leak or nonsense? (pastebin.com)
1 point by hactually 11 minutes ago | hide | discuss
- Netgear Scores the First Exemption from the FCC's Foreign-Made Router Ban (pcmag.com)
1 point by neogodless 13 minutes ago | hide | discuss
- New: NoRegistry censorship-resistant DNS over BitTorrent's DHT, 1 command to run (noregistry.com)
1 point by Synonym-recruit 14 minutes ago | hide | 1 comment
- China weighs curbs on exports of solar manufacturing equipment to US (reuters.com)
1 point by leonidasrup 14 minutes ago | hide | discuss
- Why do companies change their logo/wordmark (e.g. StackOverflow)?
1 point by profsummergig 15 minutes ago | hide | discuss
- Academic fraud may be the symptom of a more systemic problem (voxweb.nl)
1 point by the-mtr 16 minutes ago | hide | discuss
- Linux and Fingerprint Readers: A Tough Combination (medium.com)
1 point by speckx 17 minutes ago | hide | discuss
- Google to punish sites that trap people in with back button tricks (nbc.com)
2 points by 01_- 18 minutes ago | hide | discuss
- I Suddenly Lost My Enthusiasm for Interneting (sturmsonski.com)
1 point by profoundPenguin 19 minutes ago | hide | discuss
- A look at the J language: the fine line between genius and insanity (2012) (scottlocklin.wordpress.com)
1 point by tosh 21 minutes ago | hide | discuss
- A tool that turns development activity into automated daily and weekly reports (gitmore.io)
1 point by theodora 22 minutes ago | hide | 1 comment
- Ask HN: How can we weaken China?
2 points by rochul 23 minutes ago | hide | 2 comments
- Sparse Arrays in J (software.com)
1 point by tosh 23 minutes ago | hide | discuss
- The new radicalism among young women in the UK (newstatesman.com)
1 point by thinkingsome 24 minutes ago | hide | discuss
- Rome Is Burning (dyoshsimmons.com)
2 points by joshcsimmons 25 minutes ago | hide | discuss
- Prepare to Set Up Verify with Wallet on the Web in Apple Business (support.apple.com)

I disabled uBlock Origin for these tests, Chrome on the left is the default internet, Firefox on the right is using the proxy if by some crazy chance you couldn't tell.



The right way to ship this is probably a browser extension in some browser that didn't totally nerf extensions. It should be simple with a customizable prompt, then people can share prompts like they share uBlock Origin filter lists. And it should be agentic, it shouldn't actually return the HTML, it should use tools and keep per site state. Imagine a skilled software engineer running in 100x real time cleaning up websites for you before you view them.

Don't fall for AI browser crap that's marketed to you, that's just them wanting to control your attention better. You need an AI you can trust to fight back!

I hope ad people see the writing on the wall, get scared, and pivot to user aligned business model. Intelligence is about to be dirt cheap, everyone will have a full time lighting fast personal assistant to

deal with the enshittified world for them.

And you can say, well they will have a smarter one on the make everything bad side, but if mine is human level and aligned with me they will have to have gone so hard that no actual human in the world can deal with them so yea good luck with that.

The Turing Test is over. Enjoy spending your ad dollars showing things to my Qwen.

the singularity is nearer

the singularity is nearer
geohot@gmail.com

 [geohot](#)
 [realgeorgehotz](#)

A home for poorly researched ideas that I find
myself repeating a lot anyway