BasedAI

BasedAI: Proposed Plan and Technical Brief

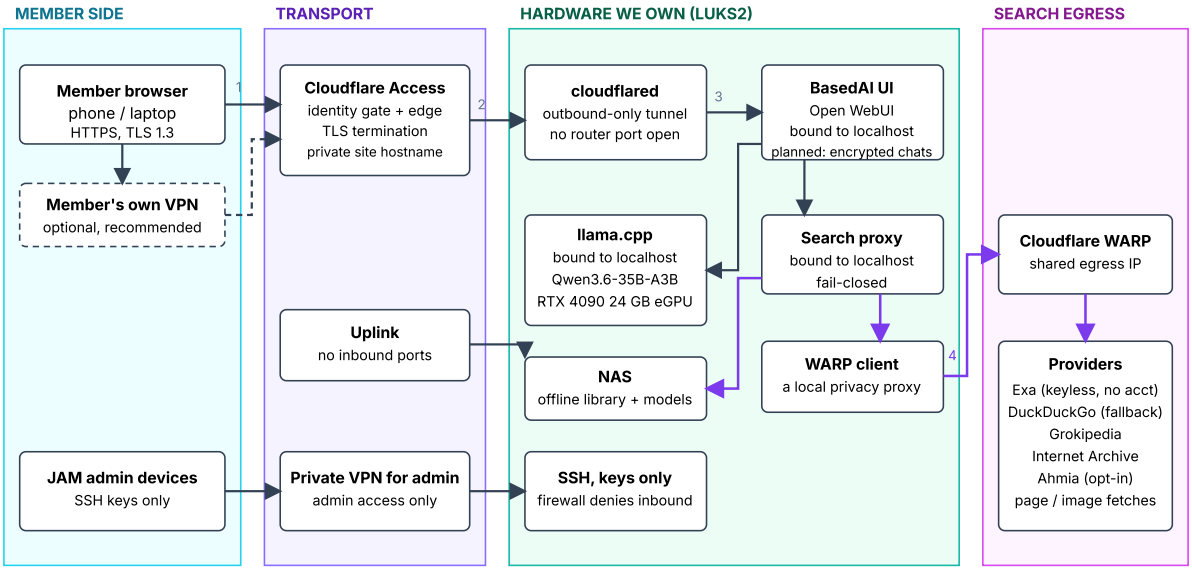
A proposed plan for a private, self-hosted assistant. Figures marked EST are estimates, not measurements. Author: JAM Solutions LLC.

1. The ask

We propose a private chat assistant for a small group of members, run on hardware we own. No third-party AI provider sees prompts. Planned Chats are encrypted per member; even the admin cannot read or export them. Planned Anon mode writes nothing. Searches leave our hardware through Cloudflare WARP with no member account or email attached. The plan serves 8 to 10 members at once.

Funding target: about $4,500 (about 9 members at $500 one-time) for lifetime access. No profit: every dollar goes to hardware, receipts are shared, and contributors are refunded if the goal is missed.

2. Proposed architecture



(1) Member HTTPS to Cloudflare Access, optionally through the member's own VPN. (2) Outbound-only cloudflared tunnel; no router ports open. (3) Open WebUI and the model (RTX 4090 eGPU) bound to localhost on hardware we own; NAS holds the offline library and model shelf. (4) Search and image fetches go through WARP (fail closed if WARP is down). Admin path: a private VPN for admin, plus SSH keys.

3. Hosting

- BasedAI runs on a dedicated GPU server we own and operate, not a big-tech AI cloud.
- Members reach it through an encrypted, outbound-only Cloudflare Tunnel with a login, and no inbound ports are open.
- The nofilters.si site is hosted on Cloudflare Pages.
- Planned An RTX 4090 upgrade to serve more members.

4. Proposed hardware

Item	Proposed spec
Host	A dedicated GPU workstation we already own, with full-disk encryption (LUKS2) and Thunderbolt. A 24 GB card is added so the model stays in VRAM.
GPU + enclosure	Used RTX 4090 24 GB in a Thunderbolt eGPU enclosure with a suitable power supply. Weights fully in VRAM (Q4_K 21.71 GB or IQ4_XS 18.73 GB), locked with `--mlock`.
Backup power	Two 1500 VA / 1000 W pure sine UPS units: one for the uplink and networking equipment, one for the NAS and eGPU enclosure.
Uplink	Private uplink with no inbound ports. No home-ISP account tie. Off-grid backup when local lines are down, cut or censored. Plan price is not in the $4,500.
NAS	Offline reference library and model shelf, so chat and lookup keep working without the public internet.
Electricity	~4 h/day at ~350 W + 20 h idle at ~30 W, about 2 kWh/day, at $0.17/kWh. About $10/month, up to about $20 with heavy use EST.

The funding target of about $4,500 covers the GPU upgrade, enclosure, backup power, shipping, and a buffer (about $4,100 bare minimum). Receipts will be shared. The Thunderbolt link mainly affects the one-time model load (tens of seconds for ~20 GB, EST). Once weights are in VRAM, decode should be close to a desktop 4090, somewhat lower in practice.

5. Proposed capacity and speed

Members generating at the same instant	Aggregate t/s	Per member t/s	Context per slot (-c 131072)
1	~160	~160	131k
4	~250 to 350	~60 to 90	~32k
8	~300 to 420	~35 to 55	~16k
10	~320 to 450	~30 to 45	~13k

All figures are EST, from desktop 4090 llama.cpp numbers, not measurements on this enclosure. Default: Qwen3.6-35B-A3B, thinking off (first word ~0.25 s); thinking on is opt-in (~5 to 15 s). Method: `--parallel N`; aggregate assumed 1.6x to 2.8x of 160 t/s; per-member = aggregate / N; context is split across slots. People read ~5 to 8 words/s, so 30+ t/s feels instant. "Members generating" means simultaneous generations, not logged-in users. A bake-off switches only on a clear win and a member vote.

6. Security and privacy

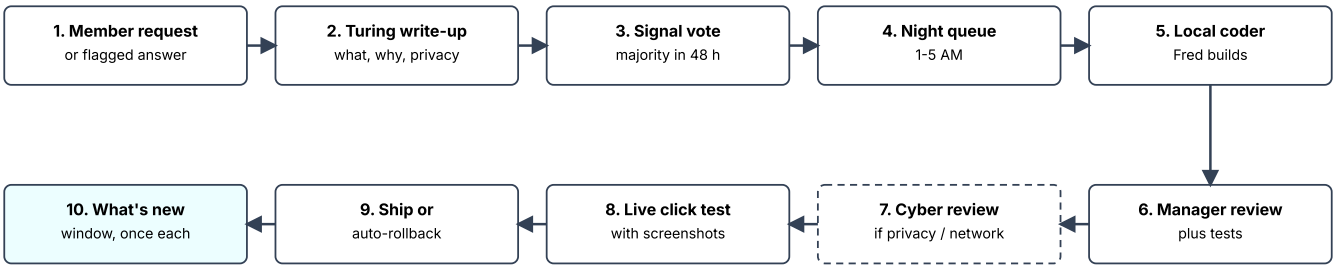
Layer	Proposed controls
Network	No open router ports. Cloudflare Access, then an outbound-only Cloudflare Tunnel. Search and images egress only through WARP (a local privacy proxy, fail closed) to a keyless Exa endpoint (no account, key or email; DuckDuckGo fallback). DNS over TLS. App services bound to localhost. Admin access via a private VPN only. Images in answers are proxied, never fetched by the member's browser.
Host	LUKS2 full-disk encryption (passphrase at boot). SSH is public-key only, over a private VPN only. sudo requires a password.
App	Access, then a BasedAI login. Open signup off (accounts created by JAM). Planned Admin cannot view or export other members' chats. Security headers (CSP, HSTS, nosniff, no-referrer). No third-party AI API: prompts never leave the server except as search queries.
Data	Planned Per-member encryption at rest: a random data key, wrapped by an Argon2id key from the member's password, unwrapped in RAM only during the session, plus a recovery code and no admin escrow. Planned Anon mode: temporary chat, nothing written to logs or the database, images proxied; a unique query must grep to zero hits in every log and DB dump (permanent test). Default logs hold time and status only, short retention, no query text.
Ops	Staging, automated tests, live click test with screenshots, auto-rollback. Signal votes on changes. No vote or agent can weaken privacy guarantees or add censorship.

7. Who can see your searches?

Observer	Without your own VPN	With your own VPN
Your ISP / Wi-Fi	That you connected to Cloudflare / the site hostname, when, and how much data. Not the content (TLS).	Only encrypted traffic to your VPN. Not even that you use BasedAI.
Cloudflare edge	Your IP and, because it terminates TLS, technically the request content. Residual risk of any Cloudflare-fronted site.	The VPN exit IP instead of yours; content exposure unchanged.
WARP, search providers, fetched sites	The server, not your device, runs the search. Providers see the query text from a shared Cloudflare WARP IP, with no account, cookie or email, and cannot tell which member asked. Fetched sites see a WARP IP and no member cookies. Your VPN does not change this path.	
JAM / a subpoena	Planned Account names and login times, not chat contents (no admin view, no escrow key). A demand can get encrypted blobs without keys, plus time-and-status logs with no query text. Planned Anon and temporary chats leave nothing. Your VPN does not change this.	

Residual risks. (1) Cloudflare terminates TLS at the edge, so it can technically see request content in transit. (2) Search providers see the query text, unlinked to a person or account. (3) Physical access to an unlocked machine on hardware we own, or a compromised member device, can expose a live session.

8. Agents and change management



Dashed box: Cyber reviews only when the change touches privacy, login or the network.

- **Turing** is the AI dev manager and change control board: intakes member requests and flagged bad answers, writes plain-English cards (what, why, privacy impact), posts them to Signal, and keeps the approved list.
- **Fred**, on the server's own local model, writes the code. Member requests never go to an outside AI company for coding. Daytime chats come first and pause agent work mid-task.
- **Backup managers** (Claude Code, then Grok Build) plan and review hard jobs and unstick the coder. They manage; the local model codes. Review prompts contain code and tickets, not member chats. Work falls over automatically on usage limits.
- **Cyber** reviews any change that touches privacy, login or the network. **Signal vote:** thumbs-up majority of all members within 48 hours; thumbs-down and no reaction do not count; anonymous members may vote privately through Turing.
- **Night queue** runs 1-5 AM. A stuck detector (the same error twice, or 3 minutes with no progress) wakes a manager for a hint.
- **Gates:** automated tests, the anon-mode zero-hits privacy test, a live click test with screenshots, and auto-rollback to the last good version. A morning What's new window shows the version, NEW and BUG FIXES, only what passed, once per member per version. Each coder mistake becomes a written rule and a test. Typical speed (not a guarantee): small fixes overnight; bigger features a few nights.

9. Rollout plan

Phase	What we propose to do
1. Fund	Raise the ~$4,500 target (about 9 members at $500 one-time). Share receipts. Refund if the goal is missed.
2. Buy and install	Purchase the 4090, enclosure, PSU and two UPS units. Confirm Thunderbolt fit and power. Install on hardware we own, enable the uplink, and load the default model fully into VRAM.
3. Bake-off models	Start on Qwen3.6-35B-A3B with thinking off (opt-in thinking). Compare shelf models on real BasedAI questions. Switch only on a clear win and a member vote.
4. Enable features	Turn on member features (Access, per-member encryption, anon mode, parallel slots, What's new). Members vote on further changes.