Lokale AI-implementatie — techniek & hardware Greenvox

*Interne voorbereiding voor Jamal. Vervolg op het eerdere verslag **"AVG-proof lokale AI — research voor Greenvox"**. Dit deel gaat puur over de **techniek en hardware**: welke machine, welke telefoon, en wat voor app. Dit is research, geen juridisch advies en geen garantie. Alle bedragen zijn **indicatief** en moeten bij aanschaf gecontroleerd worden — chip-tekorten maken de tarieven nu grillig.*

Het Jarvis-gesprek — kort samengevat

Jamal besprak met Jarvis de **technische uitvoering** van de Greenvox-tool. De kern:

- Greenvox wil een **lokale AI-assistent** die klantmanagers ontzorgt en klantcontact schrijft in hun eigen stijl.
- Er spelen **bijzondere persoonsgegevens** (UWV / re-integratie) — dus geen Amerikaanse cloud, alles **lokaal of in de EU**.
- Klantmanagers gebruiken het via een **telefoon-app**.
- De tool moet "alles" kunnen: bij bestanden, WhatsApp sturen, enzovoort.
- Open vraag uit het gesprek: *"zijn Apple-servers wel geschikt voor zo'n lokaal model?"*

Goed nieuws vooraf: die laatste aanname klopt niet helemaal. Apple Silicon (Mac mini / Mac Studio) is **juist sterk** voor lokale AI. Dat leg ik hieronder uit.

De 5 actiepunten zijn nu uitgezocht — met echte bronnen en indicatieve tarieven:

1. Hardware: Apple Silicon vs GPU-server
2. Welke taken bepalen de hardware
3. Kostenvoorstel hardware (3 opties)
4. Goedkoopste + veiligste telefoon onder de 500 euro
5. PWA vs native app

1. Hardware — Apple Silicon vs GPU-server

Waarom Apple Silicon juist goed is

Een lokaal taalmodel is vooral **geheugen-bandbreedte-werk**. Het model moet helemaal in het geheugen passen.

Apple heeft hier een truc: **unified memory**. Het werkgeheugen en het grafische geheugen zijn één geheel. Daardoor kan een Mac een **groot model** laden dat **niet** op een gewone losse videokaart past.

- Een **RTX 4090** (sterke NVIDIA-kaart) heeft maar **24 GB** videogeheugen.
- Een grote taak zoals **Llama 3.3 70B** heeft in Q4 ongeveer **42 GB** nodig.
- Op één 4090 past dat dus **niet** in één keer — hij moet stukken naar het trage systeemgeheugen schuiven, en zakt dan terug naar zo'n **8–15 woorden per seconde**.
- Een **Mac met 64 GB+ unified memory** laadt datzelfde model **wél** in één keer.

Bron: Best GPU for Llama 70B in 2026 (DEV Community) · RTX 4090 local LLM benchmarks (Hardware Corner)

De vergelijking in het kort

(a) Mac Studio / Mac mini (Apple Silicon)

- Mac Studio M4 Max met veel geheugen draait een 70B-model op ongeveer **17–22 woorden per seconde**. Dat is prima voor tekst schrijven.
- Verbruikt heel weinig stroom: zo'n **60–85 watt** onder volle belasting.
- Sterk punt: **groot model per euro** + zuinig + stil + klein kastje.
- Zwak punt: per los antwoord iets **trager** dan een zware NVIDIA-kaart, en **minder geschikt voor heel veel gebruikers tegelijk**.

Bron: Mac Studio M4 Max 128GB — 70B models tested (CraftRigs) · M4 Max AI inference benchmarks (Sean Kim)

Let op de Mac mini: een Mac mini M4 Pro kan een 70B-model technisch laden, maar de bandbreedte is lager, dus dan zak je naar **4–7 woorden per seconde**. Voor een **kleiner model** (bv. een 8B–32B model) is de Mac mini juist weer prima.

Bron: Run Llama 3.1 70B on Apple M4 (SpecPicks)

(b) NVIDIA GPU-server

- Eén **RTX 4090 (24 GB)** is goed voor **kleinere/middelgrote modellen** en is dan supersnel.
- Voor een **70B-model** heb je eigenlijk **2x 4090** nodig (samen 48 GB). Dan haal je **25–30 woorden per seconde** — sneller dan de Mac.
- Zwak punt: **veel meer stroom** (een rig kan **250–300 watt** trekken), warmer, luider, en de bouw is complexer.

Bron: 2x RTX 4090 for LLMs guide (Will It Run AI) · Mac M3 Max vs RTX 4090 showdown (SitePoint)

Welke modellen draaien erop

- **Llama 3.3 70B** (Q4): ± **42 GB** nodig → Mac met **64 GB+**, of **2x 24 GB GPU**.
- **Mistral / Qwen / Gemma** in kleinere maten (8B–32B): passen makkelijk op **24–48 GB**.

Bron: Ollama VRAM requirements table (Local AI Master)

Conclusie actiepunt 1

> De aanname *"Apple is niet geschikt"* is **onjuist**. Voor een EU-/lokaal model met een **beperkt aantal gebruikers** is een **Mac Studio met veel unified memory** vaak de **simpelste, zuinigste en stilste** keuze. Een **GPU-server** is beter zodra je **veel mensen tegelijk** wilt bedienen of maximale snelheid wilt.

---

2. Welke taken + frequentie bepalen de hardware

De juiste machine hangt **niet** af van wat klinkt als "het sterkste", maar van **hoe Greenvox het gebruikt**. Drie vragen bepalen alles:

1. **Hoeveel klantmanagers tegelijk?** Eén tegelijk = licht. Tien tegelijk = zwaar.
2. **Hoe vaak per dag?** Af en toe een mail = licht. De hele dag door = zwaarder.

3. **Wat voor taken?** Een mail of brief schrijven = **licht**. Lange documenten samenvatten of veel mensen tegelijk = **zwaarder**.

Vuistregel: tekst schrijven is licht werk. Het wordt pas zwaar als er **veel mensen op hetzelfde moment** een antwoord vragen. Dan heb je software als **vLLM** nodig (verdeelt het werk slim) en meer rekenkracht.

3 scenario's

SCENARIO	SITUATIE	PASSENDE HARDWARE (INDICATIEF)
Klein	1–3 klantmanagers, af en toe, alleen tekst	**Mac mini M4 Pro** of **1x RTX 4090** + middelgroot model
Midden	3–8 klantmanagers, dagelijks gebruik	**Mac Studio M4 Max (64 GB+)** of **2x RTX 4090** + 70B-model
Groot	8+ tegelijk, hele dag, zware documenten	**GPU-server (2x 24 GB+)** met **vLLM**, of meerdere Macs

> **Voor Jamal:** dit moet je **bij Greenvox uitvragen**. Vraag concreet: *hoeveel mensen, hoe vaak, en wat voor taken?* Zonder dat antwoord koop je of te weinig (traag) of te veel (zonde van het geld).

---

3. Hardwarekosten — voorstel in 3 opties

Bedragen zijn **indicatief** (incl. btw, NL, juni 2026). **Altijd controleren** vlak voor aankoop — de tarieven schommelen door chip-tekorten. Apple heeft het instap-tarief van de Mac mini M4 Pro recent zelfs **verhoogd** om die reden.

Bron prijsbeweging: Apple hikes M4 Pro Mac mini price amid memory costs (MacRumors)

OPTIE	WAT	INDICATIEF TARIEF	GESCHIKT VOOR
Instap	Mac mini M4 Pro (veel geheugen) **óf** 1x RTX 4090-pc	**± 2.000 – 3.000 euro**	klein gebruik, middelgroot model
Midden	Mac Studio M4 Max (64 GB+ unified)	**± 3.000 – 5.000 euro**	dagelijks, 70B-model, zuinig
Krachtig	GPU-server met 2x sterke NVIDIA-kaart	**± 5.000 – 10.000+ euro**	veel gebruikers tegelijk, max snelheid

Bron tarief-indicatie Mac Studio/mini: Apple Mac Studio · Apple Mac mini. Exacte euro-tarieven en geheugenopties wisselen — sommige 128 GB-opties zijn tijdelijk niet leverbaar — dus **op de Apple-site checken** op het moment van kopen.

Het grote voordeel: "stekker eruit"-controle

Alle drie de opties koop je **één keer** en zet je **lokaal** neer (bij Greenvox op kantoor, of bij ons). Dan geldt:

- De data **verlaat het pand niet**.
- Je kunt letterlijk de **netwerkstekker eruit trekken** en het model blijft werken — bewijs dat er niets naar buiten gaat.
- Geen maandelijkse cloud-rekening die oploopt bij meer gebruik.

> Dit is precies het verkoopargument richting Greenvox' IT'er Hamza: **eigen hardware = controleerbaar.**

---

4. Goedkoopste + veiligste telefoon onder de 500 euro

Voor klantmanagers die met gevoelige data werken telt vooral: **lang beveiligingsupdates** + **bedrijf kan de telefoon beheren (MDM)** + sterke beveiligingschip. Drie sterke kandidaten onder de 500 euro:

Aanrader 1 — Google Pixel 9a (± 499 euro, controleren)

- **7 jaar** beveiligingsupdates — de **langste** in deze prijsklasse. Niemand anders biedt dit onder de 500 euro.
- Eigen beveiligingschip **Titan M2**: controleert bij elke start of Android niet gemanipuleerd is.
- Reden: wie met UWV-data werkt wil een toestel dat **jarenlang veilig** blijft.

Bron: Pixel update-beleid (Google Help) · Titan M2 uitleg (Android Authority)

Aanrader 2 — Samsung Galaxy A56 (± 400–500 euro, controleren)

- Samsung biedt op de A-serie meerdere jaren OS-updates en tot **5 jaar** beveiligingspatches.
- Heeft **Samsung Knox**: beveiliging die vanaf de chip is ingebouwd, **plus** sterke zakelijke beheer-software (Knox Suite) waarmee IT de telefoons centraal kan beveiligen en beheren.
- Reden: als Greenvox de telefoons **zelf wil beheren** (apps verplichten, op afstand wissen bij verlies), is Knox de sterkste keuze.

Bron: Samsung Knox vs Android Enterprise (AirDroid) · Knox Platform for Enterprise (Samsung Knox)

Aanrader 3 (budget) — Samsung Galaxy A36 / A35 (± 300 euro, controleren)

- Goedkoper, ook met Knox en meerjarige updates.
- Reden: als budget de doorslag geeft en de eisen niet het allerhoogste hoeven.

Bron: Ranking beste smartphones onder 500 euro (Notebookcheck)

> **Advies:** voor **maximale levensduur + privacy** → **Pixel 9a**. Voor **maximaal zakelijk beheer (MDM/Knox)** → **Samsung Galaxy A56**. Beide ruim onder de 500 euro (tarief controleren bij aankoop).

---

5. PWA vs native app

Greenvox wil een **telefoon-app**. Er zijn twee soorten:

- **Native app** = aparte app voor iPhone én Android, die via de app-store geïnstalleerd wordt.
- **PWA** ("Progressive Web App") = een **website die zich gedraagt als een app**. Je opent 'm in de browser en zet 'm op het beginscherm. Hij krijgt een eigen icoontje, draait full-screen, en voelt als een echte app.

Wat kan een PWA in 2026?

FUNCTIE	PWA
Bij bestanden	**Ja** — file system access wordt nu breed ondersteund
Pushmeldingen	**Ja** op Android (volwaardig). Op iPhone **beperkt** (alleen na "zet op beginscherm")
Offline werken	**Ja** — via service workers
Installeren zonder app-store	**Ja** — dat is juist de kracht

Bron: PWA vs Native App 2026 (MagicBell) · Push notifications in PWAs (MagicBell)

Waarom een PWA de app-store omzeilt

Een native app moet door de **goedkeuring van Apple/Google** heen, kost meer om in twee versies te bouwen, en updates duren langer. Een PWA is **één website** die je **direct** kunt updaten — geen wachten op een store.

Voor een **intern bedrijfstooltje** (niet voor het grote publiek) is de app-store-zichtbaarheid niet nodig. Daarom is een PWA hier logisch.

Bron: PWA vs Native App 2026 (Capslock Agency)

Belangrijk voor beheerde bedrijfstelefoons (MDM)

- Een PWA kan via **MDM** worden uitgerold: op Android via **Managed Google Play**, op iPhone via een **configuratieprofiel**. De app kan dan op de telefoon staan **ook al is de Play Store geblokkeerd**.

- Wel een aandachtspunt: MDM-systemen werken **van oudsher iets soepeler met native apps**. Voor Greenvox is dit te ondervangen omdat de telefoons door het bedrijf beheerd worden.

Bron: App-distributie via MDM (ManageEngine)

Praktisch advies voor Greenvox

> **Begin met een PWA.** Sneller te bouwen, één codebasis, direct te updaten, geen app-store-gedoe, en op door het bedrijf beheerde telefoons prima uit te rollen. Kies vooral **Android-telefoons** (Samsung met Knox), want daar werken pushmeldingen volledig. Pas als er later een harde reden is (zware hardware-functies), kun je naar native.

Hoe werkt veilige telefoon-toegang tot de eigen server?

De telefoon-app praat met de **lokale AI-server** van Greenvox. Die verbinding moet dicht. De bouwstenen:

- **HTTPS / TLS** — alle verkeer tussen telefoon en server is versleuteld. Niemand kan meelezen.
- **VPN of Tailscale** — de server staat **niet open op het internet**. De telefoon komt er alleen bij via een afgeschermd privé-netwerk (zoals wij Tailscale gebruiken voor onze eigen machines). Buitenstaanders zien de server niet eens.
- **Versleuteling op het toestel zelf** — moderne telefoons versleutelen alle data standaard. Bij verlies is de inhoud onbruikbaar.
- **MFA / biometrie** — inloggen met vingerafdruk of gezicht, eventueel met een extra code.
- **MDM** — het bedrijf beheert de telefoons centraal: app verplichten, beveiliging afdwingen, en bij verlies **op afstand wissen**.

> Samengevat: model **lokaal**, verbinding via **VPN/Tailscale + HTTPS**, toegang met **biometrie**, telefoons **beheerd via MDM**. Zo blijft de UWV-data binnen het pand en

> achter slot en grendel.

Belangrijk om eerlijk bij te zeggen

- Alle **tarieven zijn indicatief** en moeten bij aankoop **gecontroleerd** worden (chip-tekorten).
- De **juridische kant** (DPIA, verwerkersovereenkomst, AVG artikel 9) blijft bij een **privacy-jurist / FG** liggen — zie het eerdere verslag.
- Dit verslag geeft **geen garantie** en is bedoeld als voorbereiding, niet als eindadvies.

Bronnen staan als klikbare links door het verslag heen.