Local AI Infrastructure Recommendation — Tan Republic

Run ID: `064ce7dc-ef0e-4237-b88d-2a06e0e19785` **Target:** `https://www.tanrepublic.com/`
Author: Athena (Scientific Researcher) **Task:** DEM-45 (S2-R3) **Date:** 2026-05-01

> **Rubric note.** `skills/local-ai-rubric.md` is a placeholder per the v2 spec (co-design with Josh pending). This recommendation is shipped best-effort and structured around the four required outputs: stack, workloads, fit, cost band. Re-issue once the rubric lands.

1. Recommended stack — HQ Compliance Copilot

A single corporate-HQ workstation in Salem, OR running an open-weights LLM behind a small retrieval index. Customer-facing chat and voice are already provisioned in this run via Hestia (Astro `/api/chat`) and ElevenLabs (S2-Deploy-D2.5) — they are **not** local AI and are out of scope for this recommendation. This stack covers the **internal**, regulator-sensitive workloads.

Layer	Recommendation	Why this choice
Hardware	Mac Mini M4 Pro, 24 GB unified memory, 512 GB SSD	Fits a 7–8B Q5 model with headroom; fanless-quiet for HQ desk; one-line replacement under warranty. ~$1,800.
Alt hardware	Lenovo ThinkStation P3 Tiny, RTX 4060 8 GB, 32 GB DDR5	Pick if Tan Republic prefers Windows / Active Directory. ~$2,000.
Inference runtime	**Ollama** (preferred) or LM Studio	One-binary install; manages model pulls; OpenAI-compatible HTTP API for downstream tools.
Primary model	**Llama 3.1 8B Instruct** (Q5_K_M)	Strong instruction-following at 5–6 GB; permissively licensed for commercial use; runs comfortably in 16 GB unified RAM.
Secondary model	**Qwen 2.5 7B Instruct** (Q5_K_M)	Stronger structured-output and table reasoning; swap in for FDD / state-statute extraction.
Front end	**Open WebUI** or **AnythingLLM**	Multi-user chat UI with workspace separation (HQ marketing / ops /

Layer	Recommendation	Why this choice
		legal); both ship a per-workspace document index.
Retrieval corpus	Smart Tan training PDFs; FDA tanning/sunless guidance; state minor statutes (CA, ID, NV, OR, UT, WA); current FDD; brand voice + claim guardrails	The corpus is the value — without it the model is a generic chatbot.
Backups	Time Machine to network share + offsite weekly	Standard small-business pattern; protects the curated corpus and chat history.

Network: stays on the HQ LAN. No inbound exposure. Model and chat data never leave the building.

---

2. Workloads it runs

1. **Marketing copy compliance pre-flight.** Before HQ ships hero copy, social posts, franchise-pack collateral, or new-service announcements, paste it into the copilot and ask: Does any line in this draft contradict 21 CFR 1040.20, the FTC vitamin-D enforcement pattern, or Smart Tan claim guardrails? The model flags risky language with the specific rule it implicates.
2. **State-rule lookup for the 6-state footprint.** “Can a 16-year-old book a UV bed in Boise this weekend?” → CA / NV / OR / WA: no (full bans); ID: only with written, in-person parental consent under §18-1523; UT: parental consent regime, **verify current statute** before publication. The retrieval index keeps this answer current as states amend.
3. **Red light therapy claim line-drawing.** Open lane (cosmetic — appearance, glow, recovery feel) vs. closed lane (medical — acne, hair, wound healing) per FDA 510(k) clearance scope. The corpus includes the FDA red-light warning-letter precedent so the model can cite it back to staff.
4. **Sunless-tanning / DHA language guardrails.** “FDA-approved DHA color additive applied externally” is acceptable; “FDA-approved spray tan” is not. PPE language (eyewear, nose filters) gets a green checkmark.
5. **Franchisee onboarding Q&A.** New owner asks how Smart Tan certification works, how the 10% federal excise tax applies (UV only), or what the 7-year initial term covers. Standardized answers, sourced from the FDD and Smart Tan corpus, no copying-and-pasting between staff.
6. **FDD / SOP search.** Natural-language search across the franchise documents, staff handbook, and equipment manuals. Replaces “ask Lance” for routine questions.

---

3. Why it fits Tan Republic

- **Compliance surface is the load-bearing risk.** Tan Republic operates in two of the strictest UV regimes in the country (CA, NV under-18 bans, OR/WA near-bans) and markets a **Class II medical-cleared modality** (red light) and an **FDA color-additive** (DHA) that have well-

documented, well-enforced claim limits. A single FTC warning letter or state-AG action against an HQ-issued claim has more cost than a decade of model hardware.

- **Lean franchisor economics.** Per business_profile.md, the franchisor entity reports $1.7 M revenue, $679 K total assets, ~10–30 corporate staff. A $2 K one-time / ~$50/mo recurring spend is tractable; a $20–50 K/year SaaS contract is not.
- **Operational data should not flow to public LLM APIs.** Marketing drafts, FDD content, franchisee performance Q&A — this is the franchisor's IP and contains material business information. Local inference keeps it inside the building, full stop.
- **Predictable cost as franchisee adoption scales.** If 65 locations use the copilot for staff Q&A, cloud-API per-token billing turns into an operational tail risk. Local inference: same monthly bill at 1 user or 100.
- **Customer-facing chat / voice already exists in this run.** Hestia's chat persona and the ElevenLabs voice agent serve consumers from a hosted stack. The **internal** workloads (compliance, FDD, franchisee onboarding) are precisely where local AI earns its keep — no consumer concurrency to engineer for.
- **Wellness positioning lines up with "we take this seriously."** A compliance-first internal tool is on-brand: Smart Tan Certified salons, FDA-aware copy, state-aware membership flows. The mature, plainspoken-Western voice from `synthesis/brief.md` carries through the staff-facing surface, not just the customer-facing one.

---

4. Cost band

Item	One-time	Recurring
Hardware (Mac Mini M4 Pro 24 GB) or ThinkStation P3 + RTX 4060	**$1,800–$2,200**	—
Setup + corpus curation (one engineer-week)	**$1,000–$2,500**	—
Inference runtime (Ollama / LM Studio)	$0	$0
Models (Llama 3.1 8B, Qwen 2.5 7B; permissive licenses)	$0	$0
Front end (Open WebUI or AnythingLLM open-source)	$0	$0–$20/mo (optional cloud sync)
Electricity (idle 10 W, peak 60 W; mixed-use)	—	**~$5–$15/mo**
Periodic corpus refresh (state-law updates, new FDD, new Smart Tan releases)	—	**~$200/quarter** (Athena's heartbeat handles literature monitoring; this is a budget line, not a new vendor)
Year-1 total	**$2,800–$4,700**	**~$70–$200/mo**

Reference point. A single FTC consent decree in indoor-tanning advertising would dwarf the lifetime cost of this stack. The recommendation is conservative — Tan Republic could adopt the lighter Mac Mini option and still cover 80 % of the value.

---

5. Edge constraints + compliance specifics

These are guardrails the corpus must encode and the copilot must surface in answers:

- **21 CFR 1040.20** — federal sunlamp performance standard; warning-label text is mandatory and cannot be diluted in ad copy.
- **21 CFR 878.4635** — sunlamp products are **Class II** (since 2014); 510(k) premarket required.
- **IRC §5000B** — 10 % federal excise tax on UV services (not spray, not red light). Pricing/package logic must respect this.
- **21 CFR 73.2150** — DHA approved as **external** color additive only; **never** for inhalation, eye, lip, or mucous-membrane exposure. PPE provisioning is part of the service description.
- **Red light = FDA-cleared, not approved.** Cosmetic claims permitted; medical/disease claims (acne, hair, wound healing) are misbranding without device-specific 510(k).
- **IARC Group 1 (2009).** UV from tanning devices is classified with tobacco and asbestos. No “safe”, “anti-cancer”, “vitamin D therapy” claims survive scrutiny — these are bright-line FTC enforcement targets.
- **State minor regimes (Tan Republic’s 6 states).** CA / NV — full under-18 ban. OR / WA — near-complete under-18 ban with medical exception. ID — Idaho Code §18-1523, under-14 ban + 14–17 in-person parental consent. UT — parental-consent regime (verify current statute before publication; advocacy reporting suggests tightening).
- **FDA red-light warning-letter precedent.** Smart Tan documented an FDA letter to industry about red-light marketing claims; the corpus should include it so the model can cite the precedent rather than improvise.
- **Bright lines for the copilot to enforce.** No “America’s largest” (false vs. Palm Beach Tan). No “FDA-approved salon” (no such concept). No “vitamin D from tanning” (FDA/FTC enforcement target). No “tan anywhere, any age” (state minors).

---

6. Open items / re-issue triggers

1. Co-design `skills/local-ai-rubric.md` with Josh; re-render this recommendation against the rubric once it lands.
2. If Tan Republic wants per-location (vs. HQ-only) deployment, the math changes — recommend a federated read-only endpoint pattern instead of 65 boxes; spec on request.
3. If RAG accuracy targets are added to the rubric (e.g., ≥ 95 % grounded retrieval on state-rule queries), the corpus engineering scope grows from one-engineer-week to two.
4. Verify the Utah minor-access statute current state with Themis (S2-R4) before publishing any age-related staff training material drawn from this stack.

`deliverable.local_ai_recommendation.done` will fire on KB write + PDF upload.