Operating an 11-Agent Personal AI Staff

Orchestration Patterns, Local LLMs, and Hard-Won Lessons

Sanket Gautam

Independent researcher · personal project, in production since February 2026 · **sanketgautam.dev**

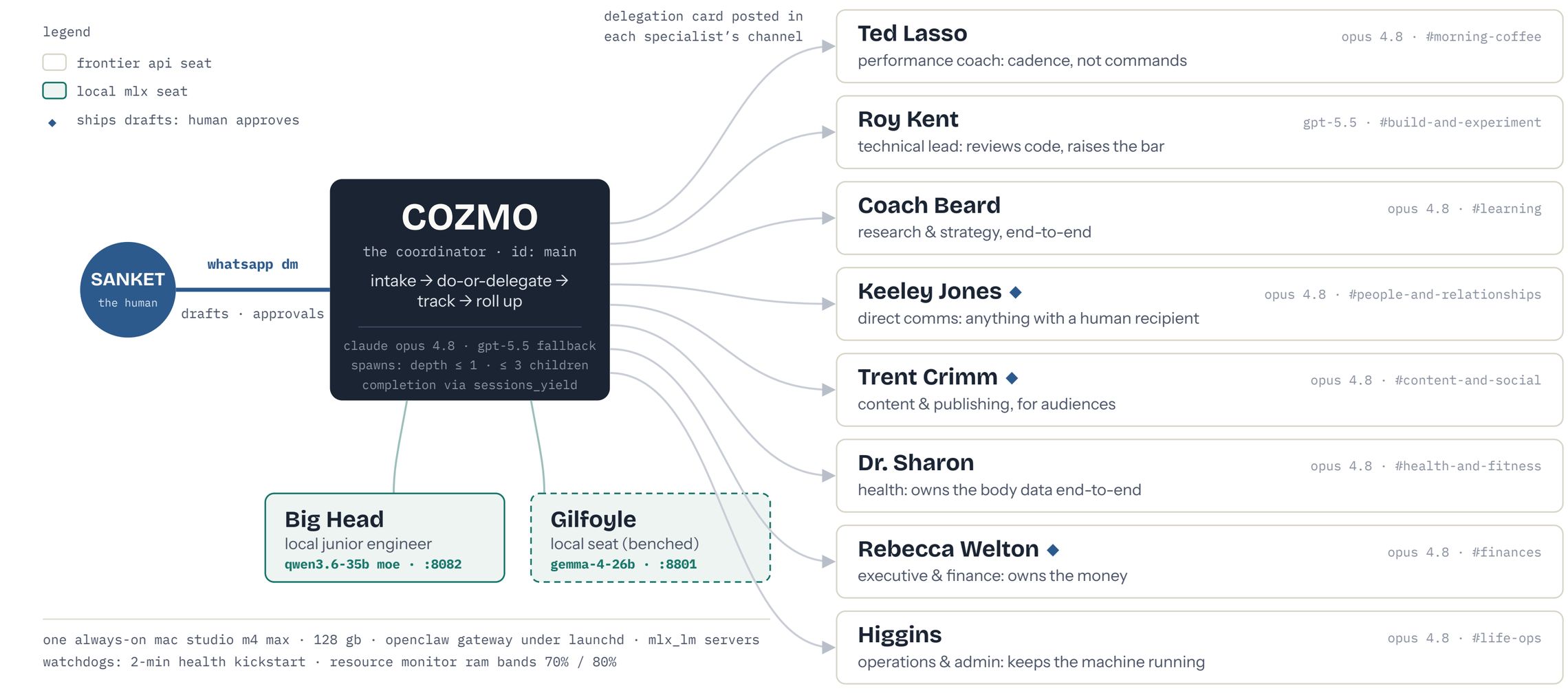

Figure 1. *Hub-and-spoke fleet on a single Mac Studio: one coordinator (Cozmo) delegates to ten specialists over Slack and WhatsApp; two seats run local models on Apple MLX. Diamonds mark agents whose outbound work requires human approval.*

I. The System

A personal "AI staff" of eleven persona agents has run continuously on one Apple Mac Studio since February 2026, serving a single user over WhatsApp and Slack. Each agent owns a domain (health, finance, research, communications, operations) and a public Slack channel.

agents	11 personas + 1 human principal
sessions	878 live (July 2026 audit)
schedule	82 jobs; heartbeats every 60-240 min
models	Claude Opus 4.8, GPT-5.5, Sonnet 5; 2 local MLX seats
hardware	M4 Max, 128 GB unified, ~546 GB/s

II. Orchestration Patterns

Single coordinator, enforced by configuration. Only the hub agent holds a delegation allowlist; all other agents carry an empty one, so specialists cannot spawn or command each other.

```
main.subagents.allowAgents: [9 agents]
all others: allowAgents: []
maxSpawnDepth: 1 · maxChildrenPerAgent: 3
```

Delegation as a written contract. Every handoff posts a delegation card (task-id, spec, done-criteria, deadline) in the specialist's public channel before the worker is spawned, keeping coordination legible to the human.

Push, not poll. Workers report completion via sessions_yield; polling loops are prohibited, the parent owns verification, and late duplicates are suppressed.

III. Local Inference on Apple MLX

Two staffed seats run entirely on-device through mlx_lm.server under launchd, exposing OpenAI-compatible endpoints. Marginal token cost is zero; the design target for the full staff is roughly $100 per month (February 2026 estimate).

:8082	Qwen3.6-35B-A3B, 4-bit DWQ MoE, ~14 GB weights (agent: Big Head)
:8801	Gemma-4-26B-a4b, 4-bit MoE, measured ~89 tok/s (agent: Gilfoyle)
:8802	Gemma-4-31B OptiQ 4-bit, staged spare

Division of labor. Local models take high-volume, low-stakes work such as heartbeat personas and bulk tasks. Frontier models keep judgment, delegation, and anything a human will read.

IV. Human-in-the-Loop Governance

Outbound messages to humans ship as drafts and are never verified and sent in the same turn. Financial actions require an explicit confirmation. Escalation to the human is treated as a designed coordination mechanism rather than a failure mode: "*the ultimate consistency resolver.*"

In Figure 1, diamonds mark the three agents whose output lanes are draft-gated end to end. A two-turn send gate separates verification from sending, so no message is composed and delivered in a single model turn, and approval prompts surface directly in Slack.

V. Operational Lessons

1. **Agent memory goes stale quickly.** The assistant twice trusted its own memory file over the airline's confirmation email and reported the wrong return date. The standing rule: primary sources beat memory.
2. **Never present stale as fresh.** Every claim carries a timestamp; data older than 24 hours must declare its age; health reports re-fetch biometrics before speaking. "*Stale data presented as current is a trust violation.*"
3. **Self-healing is layered supervision.** One audit window logged over 200 gateway memory-pressure warnings. The response combined resource-monitor alert bands (70% / 80%), launchd-supervised restarts, and a two-minute health watchdog.
4. **Reliability is an operations problem.** The failure modes were rarely model quality; they were supervision gaps, silent alarms, and configuration drift.

VI. Conclusion

A single-operator agent fleet is viable on consumer hardware, but its trustworthiness comes from operational discipline: a structurally enforced coordinator, public delegation, freshness contracts, and human approval gates on everything that leaves the house.

The models were the easy part. Operating them is the system: supervision, freshness, and human gates.

built on openclaw (open source) · apple mlx · whatsapp + slack · all figures from the july 2026 system audit and live configuration · personal project, not affiliated with any employer

sanketgautam.dev/ensemble