Operating an 11-Agent Personal AI Staff:

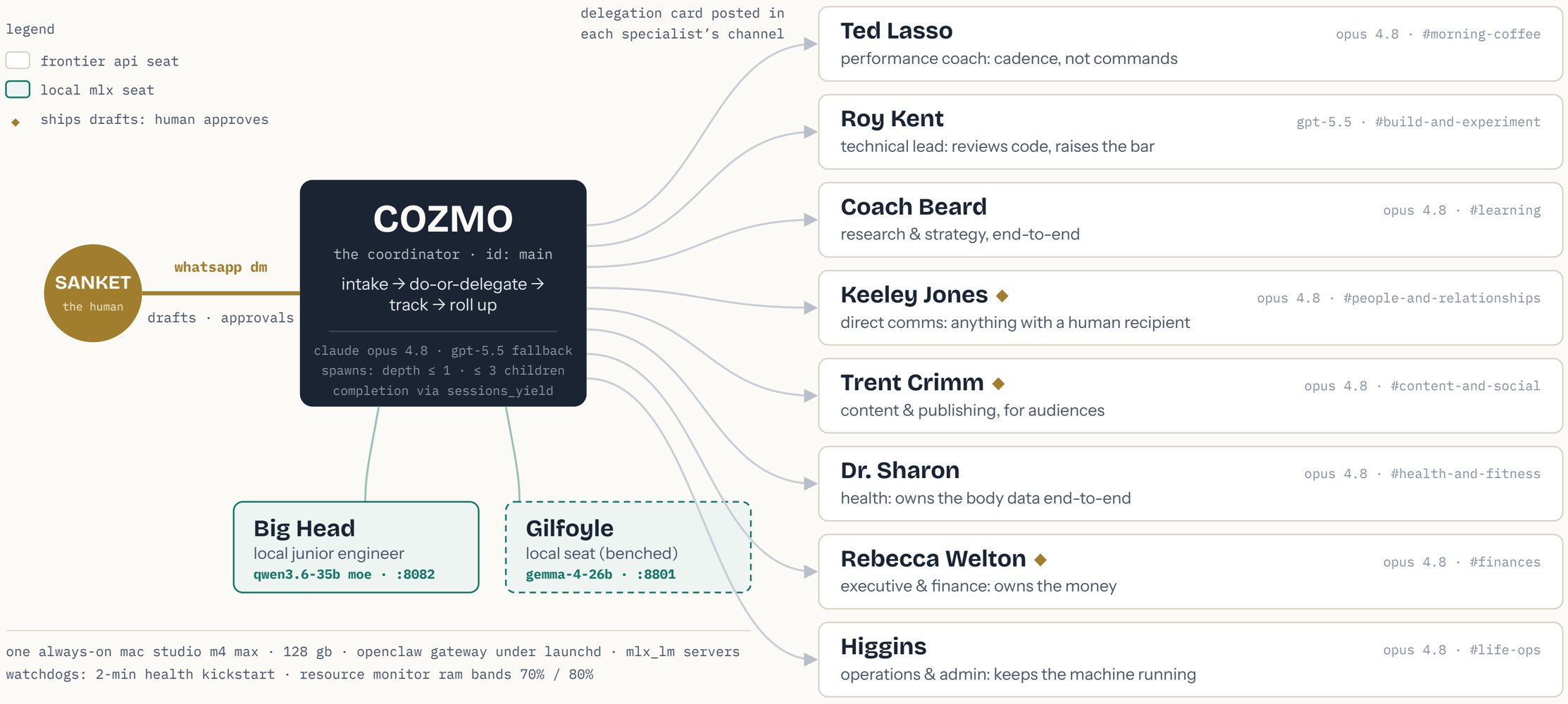
Orchestration Patterns, Local LLMs, and Hard-Won Lessons

Sanket Gautam Independent · personal project, in production since Feb 2026

sanketgautam.dev

11 + 1	878	82	2 / 11	$0
persona agents + one human	live sessions (jul 2026 audit)	scheduled jobs, 60-240 min heartbeats	seats on fully-local models (3 mlx servers)	marginal cost per local token

01 One coordinator, ten specialists, enforced by config



02 Orchestration patterns

The org chart is config, not convention.
Only the coordinator holds a delegation allowlist; every other agent's is empty. Specialists **cannot** spawn or command each other; the hub shape is structurally enforced.

```
main.subagents.allowAgents: [9 agents]
all others:  allowAgents: []
maxSpawnDepth: 1 // no grandchildren
maxChildrenPerAgent: 3 · runTimeout: 900s
```

Delegation is a written contract.
Each handoff posts a **delegation card** (task-id, spec, done-criteria, deadline) in the specialist's public Slack channel, then spawns the worker. Coordination stays legible to the human.

Push, don't poll.
Children report completion via **sessions_yield**; polling loops are prohibited. The parent owns verification and exactly **one visible completion**; late duplicates are suppressed.

A staggered pulse, not a stream.
Every agent heartbeats on its own cadence (60-240 min) into its own channel: proactive, but rate-limited by design.

03 Local LLMs on MLX

Two staffed seats (plus one spare server) run entirely on the desk: the **$0-marginal-cost experiment**.

```
mac studio m4 max · 16-core cpu · 40-core gpu
128 gb unified memory · ~546 gb/s bandwidth
serving: mlx_lm.server under launchd
openai-compatible endpoints
```

model (mlx-community)	port	seat
Qwen3.6-35B-A3B 4bit-DWQ MoE · ~14 GB weights · chosen: distilled + fast	:8082	Big Head
Gemma-4-26B-a4b 4bit MoE · measured ~89 tok/s on-device	:8801	Gilfoyle
Gemma-4-31B OptiQ 4bit staged spare capacity	:8802	·

Division of labor.
Local takes high-volume, low-stakes work: heartbeat personas, bulk grunt tasks, framework experiments. **Frontier** (Opus 4.8 primary, GPT-5.5 fallback, Sonnet 5 for sub-agents) keeps judgment, delegation, and anything a human will read.

Design target: **~$100 / month** for the full staff (Feb 2026 estimate); the local seats exist to bend that curve toward zero.

04 Hard-won lessons

1 Agent memory goes stale faster than you think.
Twice in one week the assistant trusted its own memory file over the airline's confirmation email, and reported the wrong return date. The standing rule now: **primary sources beat memory**; memory files "go stale fast."

2 Never present stale as fresh.
Every claim carries its timestamp; data >24 h old 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, not magic.
One audit window logged **200+ gateway memory-pressure warnings**. The response: a resource monitor with 70% / 80% alert bands, launchd-supervised restarts, and a 2-minute watchdog that revives the gateway itself.

4 Escalation is a feature.
Outbound to humans ships as a **draft**, never verified-and-sent in the same turn; money needs an explicit "do it." *"The human in the loop is a coordination mechanism … the ultimate consistency resolver."*

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