

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

persona agents + one human

878

live sessions (jul 2026 audit)

82

scheduled jobs, 60–240 min heartbeats

2 / 11

seats on fully-local models (3 mlx servers)

\$0

marginal cost per local token

01 One coordinator, ten specialists, enforced by config

LEGEND

- frontier api seat
- local mlx seat
- ◆ ships drafts: human approves

SANKET
the human

whatsapp dm +
drafts · approvals

COZMO

the coordinator · id: main
intake → do-or-delegate → track →
roll up
claude opus 4.8 · gpt-5.5 fallback
spawns: depth ≤ 1 · ≤ 3 children
completion via sessions_yield

Coach

performance coach: cadence, not commands

opus 4.8 · #morning-coffee

Tech Lead

reviews code, raises the bar

gpt-5.5 · #build-and-experiment

Research Lead

research & strategy, end-to-end

opus 4.8 · #learning

Comms ◆

direct comms: anything with a human recipient

opus 4.8 · #people-and-relationships

Content ◆

content & publishing, for audiences

opus 4.8 · #content-and-social

Health

owns the body data end-to-end

opus 4.8 · #health-and-fitness

Finance ◆

executive & finance: owns the money

opus 4.8 · #finances

Life Ops

operations & admin: keeps the machine running

opus 4.8 · #life-ops

Local Junior

local junior engineer · qwen3.6–35b moe · :8082

Local Spare

local seat (benched) · gemma–4–26b · :8801

one always-on mac studio m4 max · 128 gb · openclaw gateway under launchd · mlx_lm servers

watchdogs: 2-min health kickstart · resource monitor ram bands 70% / 80% · delegation card posted in each specialist's channel

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, so the hub shape is structurally enforced.

```
main allowAgents: [8 agents]
others allowAgents: []
maxSpawnDepth: 1
maxChildren: 3
completion: sessions_yield
```

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.

Qwen3.6–35B–A3B 4bit DWQ :8082
MoE · ~14 GB weights · distilled + fast Local Junior

Gemma–4–26B–a4b 4bit :8801
MoE · measured ~89 tok/s on-device Local Spare

Gemma–4–31B OptiQ 4bit :8802
staged spare capacity

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.

2 Never present stale as fresh.

Every claim carries its timestamp; data over 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.

