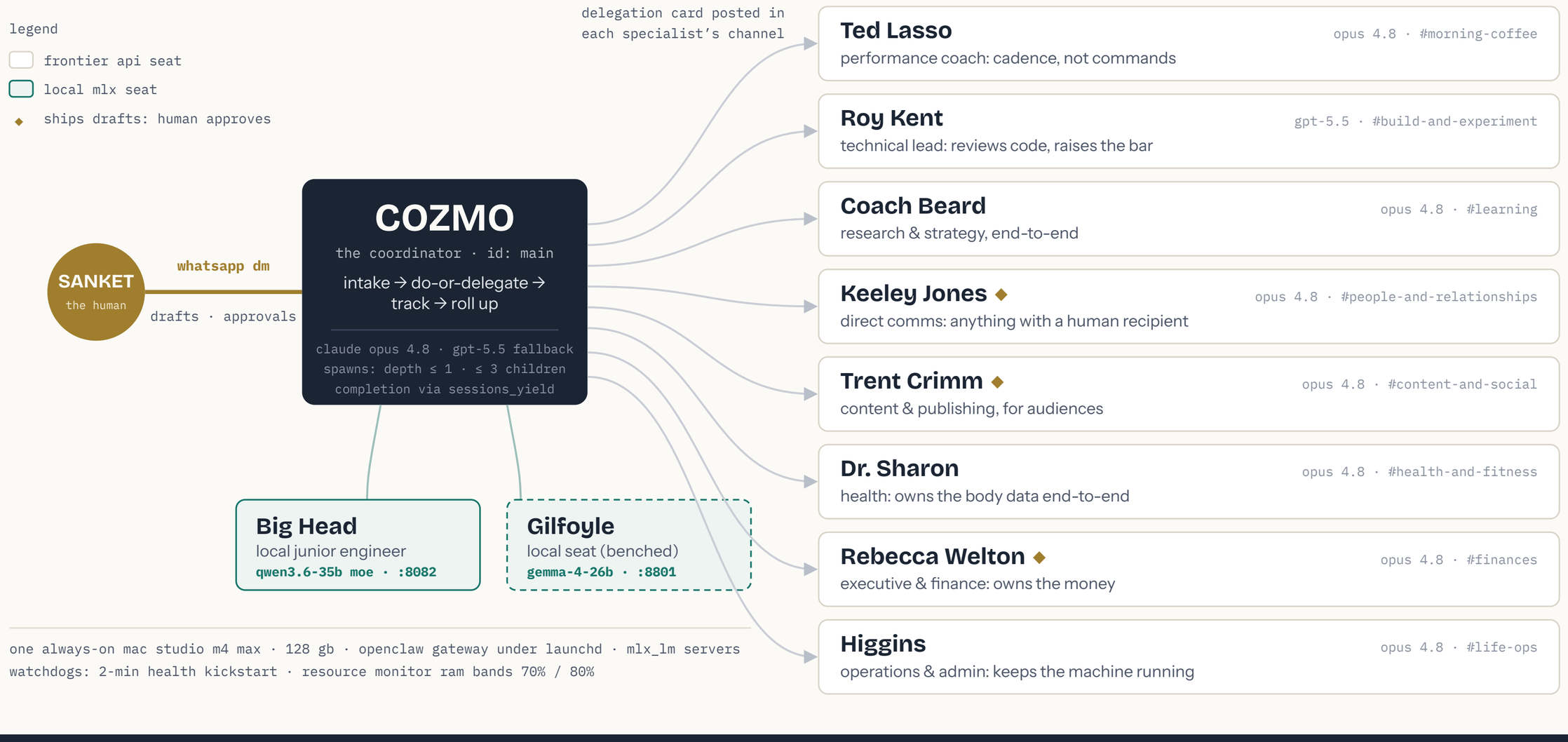
Operating an 11-Agent Personal AI Staff:

Orchestration Patterns, Local LLMs, and Hard-Won Lessons

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

sanketgautam.dev



The models were the easy part. **Operating them is the system.**

supervision · freshness · human gates

01 The system in sixty seconds

- **11 persona agents, one human, one Mac Studio.** 878 live sessions, 82 scheduled jobs, heartbeats every 60-240 min.
- **Exactly one coordinator, enforced in config.** Every other agent's delegation allowlist is empty; spawn depth is capped at 1, three children per agent.
- **Delegation happens in public.** Each handoff posts a card (spec, done-criteria, deadline) in the specialist's Slack channel; completion is pushed back via sessions_yield, never polled.
- **Two seats run fully local** on Apple MLX (M4 Max, 128 GB): Qwen3.6-35B MoE and Gemma-4-26B (~89 tok/s), at $0 marginal cost. Frontier models keep judgment and anything a human will read.

02 What broke, and the rules it left behind

- **Memory rots.** The assistant twice trusted its memory file over the airline's email and got the date wrong. Rule: primary sources beat memory.
- **Timestamp everything.** *"Stale data presented as current is a trust violation."* Data older than 24 h must declare its age.
- **Self-healing is layered supervision.** 200+ memory-pressure warnings in one audit window led to RAM alert bands, launchd restarts, and a 2-minute watchdog.
- **Escalation is a feature.** Outbound messages ship as drafts; money needs an explicit "do it." The human is the ultimate consistency resolver.

Scan for the full story, configs, and write-ups.

sanketgautam.dev/ensemble

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