04 · PERSONAL R&D · MULTI-AGENT · LOCAL-FIRST

Yousef HQ

A one-person “AI office” on a local GPU: 16 agents scout and verify Gulf B2B leads and produce bilingual Arabic/English AI content, while a human approves every outward action.

ROLE Sole builder — architecture, safety rules, approvals, models and local infra; code written with Claude Code under my rules (Oct–Nov 2026)

STACK Claude Code (subagents, skills, hooks, MCP) · Hermes Agent · local Qwen models on LM Studio · Python 3.13 · SQLite FTS5/BM25 · Whisper · Playwright · ffmpeg · Remotion · Three.js

16	$0	−82%	1/6 → 6/6	1,825	16/16
role agents in 4 wings	local inference per run	prompt tokens (17,349 → 3,036)	claims-gate pass rate	automated tests	safeguard breaks caught

THE PROBLEM

One engineer can’t watch Gulf news and tender sources for commercial signals by hand, and producing correct bilingual AI teaching content is slow and easy to get wrong: invented numbers, generic “AI slop” and broken right-to-left rendering.

WHAT I BUILT

- **Agent office.** 16 role agents (Chief of Staff, Scout, Researcher, Reviewer, Offer Strategist, Copywriter, Art Director, Publisher, Analyst…) with least-privilege tools, file-based handoffs, a mailbox and heartbeats, shown live on a 3D dashboard.
- **“Python computes, the model phrases.”** Every number comes from code into a capped fact sheet; the local Qwen model only rewrites it, and a claims validator rejects any number, date or status word without proof, falling back to rule-written text.
- **Lead radar.** Scouting, Arabic/English normalisation, fuzzy dedup, verbatim-quote verification on the live page and explainable 10-dimension scoring: 465 URLs → 44 candidates → 5/5 records verified.
- **Humans approve everything.** Fail-closed Claude Code hooks, approval only from an interactive terminal, a PIN-gated dashboard (scrypt, lockout, CSRF), and no send code at all; mutation tests catch 16/16 deliberate safeguard breaks.



Live HQ dashboard: 16 agents, post pipeline and alerts