Yosif Soliman | AI Engineer

+20 109 609 5283 • yosifsoli981@gmail.com • y0sif.github.io • y0sif • y0sif
Languages: Arabic (Native), English (Fluent)

Summary

AI Engineer building production multi-agent systems and AI-powered platforms. Experienced in full-stack AI development with Python, Rust, and TypeScript, spanning agent orchestration, RAG pipelines, LLM evaluation, fine-tuning of open models, and end-to-end automation. Active open-source contributor.

Relevant Experience

Bond **Remote, USA**

AI Engineer *Dec 2025 – Present*

- Built an **agent-driven end-to-end testing framework** orchestrating three LLM sub-agents (test runner, log investigator, copy judge) that drive the live product through **13 real-world campaign scenarios** via Playwright, auto-root-causing failures against **Langfuse** traces and emitting evidence-backed reports.
- Built a **55+ scenario** real-backend agent evaluation framework with rubric-based LLM judging across list-building, enrichment, and copy pipelines, enabling **model-per-sub-agent** selection and prompt/tool fixes that eliminated copy fabrication and ICP-filter leakage.
- Built the **Bond CLI** and its agent **skill layer** – a **/cli-campaign** router skill that runs full campaigns end to end: lead sourcing, an enrichment **column tree**, a 5-row test gate, and export to **Instantly**/**HeyReach**, mirroring the in-app multi-agent system.
- Automated ~**10 DFY clients**' end-to-end onboarding (domain, inbox, warmup, campaign launch) from **30–90 minutes to under 1 minute** via natural-language agent workflows.

AI Intern *Sep 2025 – Dec 2025*

- Enhanced the multi-agent GTM system by implementing agent tools and optimizing agent structure, and evaluated LLMs and prompts with **LangSmith** to improve reliability.
- Migrated the Python/**FastAPI** agent backend to queue-based **Supabase edge functions** (Deno/TypeScript), building integration-tested services for AI-column processing and lead enrichment via **LangChain**, and automated agent-evaluator bug reporting with **n8n**.

Education

Ain-Shams University **Cairo, Egypt**

B.Sc. in Computer Science, ***CGPA: 3.48/4.0*** *Oct 2021 – Jul 2025*

Technical Skills

Languages: Python, Rust, TypeScript, SQL, C/C++

AI/ML: LangChain, LangGraph, LangSmith, OpenAI, Anthropic Claude, Rig, RAG, Prompt Engineering, Fine-tuning (QLoRA, Unsloth), Hugging Face, MCP, Langfuse, ChromaDB, OpenSearch

Frameworks: FastAPI, Next.js, Axum, Leptos, Dioxus, BullMQ, Celery

DevOps: Docker, GitHub Actions (CI/CD), AWS, Prometheus, Grafana

Platforms: PostgreSQL, Supabase, Redis, Linux, GCP

Highlighted Projects

- **Locked-In – AI Calendar Intelligence Platform** [Demo]
  - Built a unified AI assistant (Rig + Claude) with **intent detection**, a **user-intelligence system** that learns preferences over time, and structured question cards for multi-step plan generation.
  - Architected full-stack Rust platform (Axum, Leptos WASM, SQLx) with **SSE streaming**, Google Calendar integration, Stripe subscriptions, and OAuth 2.0, plus **OpenTelemetry**/**Langfuse** observability. Migrated from Python/Next.js for type safety and single-binary deployment.
- **whisrs – Voice-to-Text Dictation for Linux** [GitHub] (50+ GitHub stars)
  - Built multi-backend speech-to-text tool supporting **Groq**, **Deepgram**, **OpenAI** (Realtime WebSocket streaming and REST), local **whisper.cpp**, and a generic **ASR sidecar** for bring-your-own local models, with an LLM-powered command mode and a layout-aware cross-desktop daemon.
- **Fine-Tuned Gemma Models – GutWise & Arcwright** [GutWise] [Arcwright]
  - Engineered **GutWise**, an offline-first IBS health assistant for the **Kaggle Gemma Hackathon** (Health & Sciences): a data pipeline that mines clinical guidelines, synthesizes Q&A with Haiku sub-agents, and gates every sample through medical-fact validators and an LLM judge before **QLoRA** fine-tuning, surpassing base Gemma on a held-out clinical eval.
  - Fine-tuned **Arcwright**, a Gemma model specialized for Rust web/AI frameworks (Leptos, Axum, Rig) through a **compile-verified data pipeline**; on the custom **RustWebBench-15** benchmark it ranked **#1**, outscoring its 6.5×-larger sibling Gemma 26B and Claude Haiku.

Open Source Contributions

Aden/Hive – AI Agent Framework (10k+ GitHub stars) [GitHub]: Contributed **29 MCP tools** that add Attio CRM and Linear integrations, with GraphQL/REST clients, OAuth 2.0, and test coverage.

CodeWhale – Terminal AI Coding Agent (36k+ GitHub stars) [GitHub]: Added an async **Tokio** wakeup channel that resumes the parent agent loop the moment parallel sub-agents finish, removing manual nudges.