Mohamed Yousry

mohamedyousry.work.dev@gmail.com • +201003204931 • Mansoura, Dakahlia, Egypt

SUMMARY

AI Engineer and Backend Developer with experience leading B2B and B2C platforms and shipping production AI systems, including agentic and multi-agent architectures (A2A, AG2), LLM deployments, and on-device AI. Strong background in Python, FastAPI, Docker, Redis, and PostgreSQL, with a focus on microservices, automation, and real-time data processing.

EXPERIENCE

Drive-Nova, Lead Engineer (Hybrid · UAE) *Apr 2026 - Present*

- Lead engineering across B2B and B2C platforms (including managed Magento storefronts), owning architecture, delivery, and ongoing platform operations.
- Architecting and managing upcoming AI initiatives, including Agent-to-Agent (A2A) architecture and multi-agent frameworks such as AG2.
- Managing integrations and technical relationships across multiple partner platforms in the tyre industry, spanning Europe, Egypt, and the UAE.

Pillar Construction Group, Fractional CTO & Lead AI/Backend Engineer *May 2025 - Present*

- Designed and delivered a full-stack AI procurement platform aggregating 50,000+ federal (SAM.gov) and SLED (HigherGov) opportunities, with multi-schema PostgreSQL, semantic search via embeddings, and an event-driven processing system.
- Built ELI, a conversational AI assistant with 24 specialized procurement tools (opportunity search, document Q&A, board management) powered by GPT-5/GPT-4o, plus a 4-tier matching pipeline (geographic, metadata, documents, LLM) and a PaddleOCR pipeline for scanned government forms.
- Architected microservices with FastAPI, Celery (6 specialized workers), Redis, and Supabase; deployed full production infrastructure with Docker, Coolify, Nginx, Cloudflare, GitHub Actions, and Sentry monitoring.

MyAly.AI, AI Engineer & Automation Developer *Feb 2025 - Apr 2025*

- Built and maintained Stripe payment systems, including subscriptions, webhooks, and secure checkout, enhancing payment processing efficiency.
- Developed and integrated frontend flows with Backend APIs, improving website performance and user experience.
- Set up and managed infrastructure and deployments using domains, DNS, Nginx, and Git workflows, ensuring robust and fault-tolerant applications.

TotallyNot, Co-founder & AI Engineer *Sep 2024 - Mar 2025*

- Led the development of AI-driven features for the Local on-device AI & Flutter, enhancing system performance and user experience.
- Designed and deployed scalable backend architectures using FastAPI, PostgreSQL, and Docker, improving data processing efficiency.
- Implemented CI/CD pipelines for seamless deployment, improving system stability.

Freelance, Freelance Developer *Jan 2020 - Present*

- Architected and deployed large-scale production systems, including AI pipelines, OCR systems, mobile applications, web platforms, and automation bots, enhancing efficiency and scalability.
- Designed and implemented OCR and document-processing systems, handling image preprocessing, text extraction, validation, and structured data output, improving data processing.
- Managed full infrastructure lifecycle, including cloud setup, server provisioning, Dockerized deployments, CI/CD pipelines, domain/DNS configuration, and production monitoring, ensuring robust and fault-tolerant applications.

SELECTED PROJECTS

ileterate — AI-Powered Multilingual Grammar Checker *Open Source*

- Built a privacy-first, self-hostable grammar and rewriting platform supporting Dutch, English, German, French, and Spanish, combining LanguageTool's rule-based detection with LLM corrections and a unique validation loop that re-checks LLM output to prevent AI-introduced errors.
- Cross-platform Flutter app (iOS, Android, Web, Desktop) with FastAPI backend, RSA + AES end-to-end encryption, API key authentication, and one-command Docker Compose deployment.
- Stack: Python, FastAPI, Flutter, LanguageTool, Docker, OpenAI-compatible LLMs.

CSphere — Multi-Agent AI IDE *In Development*

- Architected and authored full system documentation for a VSCode-based IDE with an 11-agent system (orchestrator-driven, MCP-based) covering autonomous coding, debugging, testing, refactoring, and deployment, plus a free community marketplace for tools, agents, and extensions.
- Designed local-first infrastructure: SQLite plus ONNX Runtime embeddings running directly in Electron with no middleware, optional Supabase sync for auth and usage, and hosted inference for Llama 4, Qwen, and GPT OSS models.
- Stack: Electron, TypeScript, Python, FastAPI, SQLite, ONNX Runtime, Supabase, MCP.

strix-halo-unsloth — ML Infrastructure for AMD Strix Halo *Open Source*

- Authored the working setup guide and prebuilt Podman containers for running PyTorch and Unsloth fine-tuning on AMD Ryzen AI Max+ 395 (gfx1151) with 128 GB unified VRAM; identified the kernel regression and ROCm version constraints that block other setups.
- Stack: PyTorch (ROCm 7.1), Unsloth, LoRA, Podman/Docker, Linux kernel tuning.

EDUCATION

New Mansoura University
Bachelor's • AI Engineering • 2021 - 2026
GPA 3.3/4

SKILLS

Languages: Python • TypeScript • JavaScript • C++ • C • C#
Backend & APIs: FastAPI • Flask • Django • REST APIs • Microservices • Celery • Stripe
AI & ML: Large Language Models (LLMs) • Multi-Agent Frameworks (AG2, MCP) • LangChain • Llama.cpp • PyTorch • TensorFlow • Fine-tuning (LoRA, Unsloth) • PaddleOCR • NLP • Embeddings (pgvector) • ONNX Runtime
Mobile & Frontend: Flutter • React • Next.js • Electron
Databases: PostgreSQL • Supabase • Redis • MySQL • SQLite
DevOps & Automation: Docker • Nginx • Linux • Cloudflare • CI/CD (GitHub Actions) • Git • Sentry • n8n • ETL Pipelines