TAHA-YASSEN SHALABY

Full-Stack AI Engineer
tahashalaby93@gmail.com | +20 1010081699 | U.S. Citizen | Open to Remote & Relocation
github.com/TahaSha | linkedin.com/in/tahashalaby

PROFESSIONAL SUMMARY

Full-Stack AI Engineer with 6+ years building, shipping, and scaling AI-powered products in production. Co-founded an AI/Data Science startup that delivered 12+ ML systems for Fortune 500 clients, powering 2M+ predictions per month. Currently building agentic LLM applications for cybersecurity — on-prem SOC automation with tool-calling agents, retrieval-grounded responses over security logs, and modular SIEM integration, hardened against prompt injection and excessive agency (OWASP LLM Top 10). Owns the full stack: React/Next.js front ends, FastAPI/Node.js back ends, GCP/AWS infrastructure, and production ML pipelines.

PROFESSIONAL EXPERIENCE

AI Engineer / Consultant
May 2025 – Present

Snappers (AI Cybersecurity) | Cairo, EG (Remote for US Client)

- Architected and built a full-stack AI-powered SOC analyst dashboard integrating with Netwitness API, designed modularly for multi-SIEM compatibility
- Built a tool-calling LLM agent (on-prem Qwen via Ollama) that grounds responses in live log data, letting Tier-1 SOC analysts replace hand-written Netwitness/SIEM queries with plain-English requests (e.g., filtering IP traffic by source, destination, and time window) — cutting query construction from minutes of manual syntax to a single natural-language step
- Hardened the tool-calling agent against OWASP LLM Top 10 risks — validating LLM-generated queries and scoping execution to read-only, least-privilege access to limit prompt injection and excessive agency, with prompt/response logging for auditability
- Built automated report generation engine for PCI compliance and incident reports with modular, customizable templates
- Deployed on-prem using Ollama with open-source Qwen models to meet strict data residency requirements
- Tech stack: React, FastAPI, PostgreSQL, Ollama/Qwen, LangChain, Docker Compose, CI/CD pipeline

Tech Lead
Feb 2023 – Apr 2025

Singularity Finance | Cairo, EG

- Led a small dev team building and operating price-feed algorithms for the Flare FTSO oracle network (Songbird canary), consistently landing submissions within the reward band on ~90% of price windows
- Architected and operated the supporting infrastructure on GCP Managed Instance Groups (MIG), sustaining ~99% node uptime across a 3-node validator fleet
- Implemented Node.js/TypeScript backend services integrated with FastAPI Python endpoints for ML algorithms within an Ethereum-based ecosystem
- Managed confidential VMs, automated snapshot/backup processes, and ERC token node infrastructure across the validator fleet

Co-Founder & CTO
Dec 2019 – Dec 2022

Iaccuna (AI/Data Science Startup) | Toronto, CA

- Co-founded and set technical direction for an AI startup that built and shipped production ML for SMEs and Fortune 500 clients
- Delivered 12+ production ML projects for Fortune 500 companies, powering 2M+ predictions/month across manufacturing, media, finance, and telecommunications
- Led a consulting engagement with UCI graduate engineers for a multi-million-dollar California manufacturing company
- Built and deployed production ML systems spanning NLP, time-series forecasting, classification, deep learning (PyTorch, transformers), and distributed computing
- Owned full backend architecture, API development, cloud deployment (GCP), and client delivery pipeline

Mathematics & Computing Educator *2017 – Present*

Cambridge International Schools | Cairo, EG

- Delivered Cambridge Mathematics and Computing curriculum across multiple international schools while maintaining concurrent technical career
- Developed automated reporting systems and custom gradebook tools; founded a robotics club introducing students to coding and STEM

SELECT PROJECTS

AI-Powered SOC Dashboard with Agent Chat — Snappers, 2025

- Full-stack cybersecurity dashboard with natural language AI agent for querying security logs, automated PCI/incident reporting, and modular SIEM integration. On-prem LLM deployment with Ollama/Qwen. React + FastAPI + PostgreSQL + Docker.

BERT-Based Article Relevance Classifier — Iaccuna, 2021

- NLP model using BERT transformer (PyTorch/HuggingFace) achieving 82% F1-score. Reduced manual analyst headcount from 100 to 15 for a US media company.

Distributed Time-Series Forecasting System — Iaccuna, 2022

- Distributed computing pipeline using Ray, PyTorch, and statsmodels for client-specific label forecasting. Deployed on GCP with Plotly dashboards for a US media company.

Customer Churn Prediction Model — Iaccuna, 2022

- Gradient boosting model with hyperparameter tuning that achieved 15% churn reduction for a telecom company. Deployed on GCP and integrated into retention workflows.

Scalable ML Workflow Engine on Kubernetes — Iaccuna, 2022

- MLOps workflow engine (microservices) using FastAPI, Celery/RabbitMQ, and Nginx. Deployed on GKE with auto-scaling, ingress, secrets management, and configurable Dockerfiles.

TECHNICAL SKILLS

Languages	Python, JavaScript/TypeScript, SQL, Bash
AI / ML	PyTorch, HuggingFace Transformers, scikit-learn, XGBoost, LLM agents & tool calling, RAG, prompt engineering, LLM security (OWASP LLM Top 10), LangChain, Ollama, NLP (spaCy, NLTK)
Frontend	React, Next.js, React Router, Tailwind CSS
Backend	FastAPI, Node.js, Celery, RabbitMQ, Nginx
Cloud / DevOps	GCP (GKE, Cloud Run, MIG), AWS, Docker, Docker Compose, Kubernetes, MLOps, CI/CD, Git, Linux
Data	PostgreSQL, Pandas, Ray, Plotly, Seaborn, Matplotlib
AI Tooling	Claude Code, Cursor, MCP (Model Context Protocol), AI-assisted development workflows

EDUCATION

B.Sc. Mechatronics Engineering 2014 – 2019

Arab Academy for Science & Technology (AAST), Cairo, EG

Graduation Project: Autonomous Attendance Robot using Facial Recognition

Computer Science Coursework 2011 – 2012

New Jersey Institute of Technology (NJIT), NJ, USA