Wael Fahmy

Senior AI Engineer | LLM Applications and Inference

Ontario, Canada · wael.anwar.fahmy@gmail.com · Authorized to work in Canada
linkedin.com/in/wael-fahmy · github.com/wael-fahmy · waelfahmy.com

SUMMARY

AI engineer with 6+ years of professional software experience, plus internships since 2018. Sole engineer on Charti, an AI medical scribe used in Canadian clinics: live audio streaming, speech-to-text, LLM notes and self-hosted GPU inference. Built Fitlog's AI platform service and lead model calibration and safety research at Lakehead University. MSc in Computer Science (GPA 4.0/4.0, 94% average).

TECHNICAL SKILLS

LLM and speech: vLLM, Qwen, Whisper, Deepgram, AssemblyAI, RAG, pgvector, LangGraph, OpenAI Agents SDK, CrewAI, MCP, Hugging Face, Ollama, PyTorch, LoRA, guardrails, human-in-the-loop, evaluation

Engineering: Python, FastAPI, Pydantic, gRPC, WebSockets, SSE, TypeScript, Node.js (NestJS), React, React Native, Next.js, Django, Celery, REST, SQL

Infrastructure: Docker, OpenShift (Kubernetes), Terraform, Traefik, GitHub Actions, Bitbucket Pipelines, AWS, DigitalOcean GPUs, Slurm, PostgreSQL, MySQL, Redis

Testing and observability: pytest, Vitest, Jest, Playwright, OpenTelemetry, Grafana

Security and privacy: RBAC, MFA, OAuth2/OIDC, controls supporting PHIPA and PIPEDA

PROFESSIONAL EXPERIENCE

Fitlog | Montreal, QC (Remote) | **Dec 2022 to Present**
Senior Full Stack Engineer, AI & Cybersecurity | Sep 2025 to Present
Senior Software Developer, Full Stack & Cybersecurity | Feb to Aug 2025
Senior Full Stack Developer | Dec 2022 to Jan 2025
Flexible hours (20 to 40 per week), Dec 2022 to Aug 2025; full-time since Sep 2025.

- Built the Fitlog AI worker in Python: one gateway for every model call, with role-based model routing, Pydantic-validated outputs, cost and audit logging and fallback models; 391 tests.
- Develop LLM features over client health data with Whisper speech-to-text (since Sep 2025); led the design of MCP tools for cardio, diet and training protocols; security audits and production RBAC.
- Develop Fitlog's React and React Native apps, used by 27,500+ coaches and 350,000+ clients; CI/CD.

Charti Health AI Inc. | Thunder Bay, ON (Remote) | **Feb 2026 to Present**
Founding Engineer & Technical Lead (part-time since Jun 2026)

- Sole engineer on an AI medical scribe used in Canadian clinics (51 clinician sign-ups, 29 generating notes); own the architecture, models, deployment and security.
- Stream browser audio over WebSockets to a gRPC AI worker for transcription, LLM draft notes and clinician review, on FastAPI, PostgreSQL with pgvector, Celery and Redis.
- Built GPU and CPU worker modes: Qwen3.5-27B on vLLM (RTX 6000 Ada and L40 GPUs, DigitalOcean Toronto, moved from RunPod for reliability) or managed APIs; 6,293 AI jobs at 99.7% success, averaging 15.2 seconds.
- Route across 7 AI providers and 13 models (Deepgram, AssemblyAI, Soniox, OpenRouter, faster-whisper) with per-model health checks, automatic failover, spend logging and a cost cap.
- Built a tablet voice intake kiosk, PDF form filling and 20+ specialty templates with strict JSON schemas and prompt-injection guards; benchmarked notes against a competitor.
- Implemented encryption, audit logs, RBAC, MFA and Canadian data residency supporting PHIPA and PIPEDA; about 2,100 backend pytest tests, OpenTelemetry tracing and GitHub Actions deploys.

APPLIED ML RESEARCH

Robot Trust Research Program | Lakehead University | **Nov 2026 to Present**
Lead researcher with Prof. Abedalrhman Alkhateeb (preliminary results)

- Fine-tuned the NVIDIA Cosmos3-Edge video world model with LoRA on 2,500 simulated clips; tip rates had a mean absolute error of 0.035 against simulator rates.
- Tested vision-language safety checks for robot arms: a misleading camera view cut a frontier model's stop rate from 0.96 to 0.43, and carrying earlier observations forward restored it to 1.00.
- Run training and evaluation with Slurm on Digital Research Alliance of Canada H100 clusters; each decision rule is frozen and hashed before results are read.

Preprints: first author of ProtCOVER, conformal prediction for protein variant effect predictors (bioRxiv; submitted to RECOMB 2027; github.com/wael-fahmy/protcover). Two more health preprints on survival analysis and identifiability.

EARLIER EXPERIENCE

UKKERA | Cairo, Egypt (Remote) | **Nov 2022 to Jun 2024**
Senior Software Developer

- Built Node.js, Spring Boot and .NET Core services for education platforms handling 50,000+ requests per day.
- Contributed to the Saudi Ministry of Energy design system and the LCGPA training platform; wrote software requirements and API documentation.

SmartTech Sys | Cairo, Egypt | **Mar 2022 to Dec 2022**
Senior Software Developer

- Added Python OCR, NLP and computer vision steps to UiPath and iRobo bots so they could read scanned business documents across 20+ client workflows.

CARITech | Giza, Egypt | **Sep 2020 to Mar 2022**
Software Developer

- Developed Phoenix core banking modules in .NET and Gupta; investigated recurring production defects.

Earlier internships (2018 to 2019): Teradata (data science), OrdoChao, Egyptian Ambulance Organization.

SELECTED PROJECTS

- **Rennli (independent product, early development):** AI phone receptionist for booking businesses. Built a FastAPI backend, multi-brand configuration and Astro websites in a shared monorepo; real-time call processing with Pipecat is planned.
- **MSc projects:** compared ChatGPT with AutoGPT and AgentGPT on USACO programming problems (group project); serverless LLM sentiment classifier on AWS SageMaker and Lambda; malware image classification with EfficientNet (99.2% accuracy).

EDUCATION

MSc in Computer Science, Lakehead University Sep 2024 to Sep 2025
GPA 4.0/4.0, **94%** overall average; A+ in every graded course.

BSc in Electrical Engineering, Cairo University Sep 2015 to May 2020
Computer and Communication major; WES verified.