

Abdelrahman Sheta

Product Engineer | React, TypeScript & Node.js | AI-Enabled Products

Cairo, Egypt | Open to relocation | abdelrahman.sheta1@gmail.com

LinkedIn: [linkedin.com/in/abdelrahman-m-sheta](https://www.linkedin.com/in/abdelrahman-m-sheta)

SUMMARY

Backend engineer who led the re-platforming of a NestJS/Node.js monolith to event-driven Python/FastAPI, backed by 6,500+ automated tests, cutting API response time 43%, average query time 61%, and database CPU use 38% while sustaining production availability. Cut per-conversation inference cost by over 70% with answer quality held flat through multi-tier model routing with automatic fallback. Builds payment and data systems with HMAC-authenticated webhooks, idempotency, reconciliation, and audit trails, and ships Docker-based exact-commit releases with health checks and rollback.

EXPERIENCE

NLP Engineer, Ozzie (myozzie.app)

Remote | Aug 2026 - Present | Part-time

- Own the recitation model end to end, from annotating children's Quran recitation data through fine-tuning and production deployment; the model has reached 85% accuracy on children's recitations, with a target of 98%.

Founding Engineer, Cravit

Cairo, Egypt | Jun 2025 - Present

- Led the full re-platform from a NestJS monolith to an event-driven Python/FastAPI monorepo (async SQLAlchemy, Celery, Redis Pub/Sub, PostgreSQL): 2,100+ REST endpoints across 50+ business modules, 600+ ORM models, 350+ Alembic migrations, and 6,500+ automated tests, cutting API response times 43%, average query execution 61%, and database CPU usage 38% while sustaining production availability.
- Built a central AI gateway with model routing, fallbacks, and circuit breakers, Redis LLM caching, 4 Qdrant vector collections for vector search (vector databases), a Two-Tower deep-learning recommendation service, and a dbt analytics warehouse.
- Cut per-conversation inference cost by over 70% with answer quality held flat through multi-tier model routing with automatic fallback, prompt caching against byte-stable system prefixes, per-tenant token budgeting, response-format trimming, and a cost-per-resolution metric instrumented alongside latency and quality.
- Shipped LangGraph marketing and operations agents with per-tenant budgets, human approval queues, guardrails, audit logging, fallbacks, circuit breakers, and an outage watchdog.
- Integrated Paymob buy now, pay later (BNPL) across eligibility, checkout, installment schedules, authenticated webhooks using HMAC, idempotency, reconciliation, refunds, and reporting. Built monthly customer subscriptions offered by restaurants and SaaS subscriptions paid by restaurants.
- Owned concept-to-production delivery for a multi-tenant restaurant platform serving 14 restaurants, 22,000+ customer accounts, and 38,000+ orders; shipped 4 Next.js/React/TypeScript applications with real-time updates and 15+ external integrations.
- Fine-tuned transformer classifiers for Arabic, Egyptian dialect, Arabizi, and mixed Arabic/English messages across intent, language, and urgency; extended the experience to real-time voice with Deepgram STT and Azure TTS over WebSockets.

Product/GTM Engineer (Founding Team), Rafeeq

Remote | Jul 2026 - Present | Contract

- Architected the full FastAPI backend for a personal development platform: 144 REST endpoints, 29 SQLAlchemy models, and 44 service modules, with multi-provider LLM routing and automatic fallback, Qdrant vector search, Redis-backed session state, and long-term user memory.
- Built and led the FastAPI AI platform, integrating OpenAI, Anthropic, Gemini, and open-source models with routing, fallbacks, Qdrant retrieval, Redis session state, long-term memory, usage metering, and auditable tenant controls.
- Built Docker-based exact-commit releases, health checks, authenticated backup, restore, and rollback workflows; preserved 4,573 data assertions through a production-copy CRM migration and covered the WhatsApp workspace with 703 Python and 71 TypeScript/Node unit tests.
- Engineered the WhatsApp Business workspace and consent-first enrichment pipeline with suppression controls, encrypted contact data, PII isolation, recorded approvals, spend and usage ledgers, idempotency, reconciliation, and auditable fail-closed behavior.

Full Stack Engineer, Style Protocol

Remote | Jun 2024 - Jun 2025

- Built a Next.js frontend and server APIs, increasing engagement 30%, reducing page-load time 20%, and cutting server response time 25%; deployed 6 smart contracts across Ethereum, Arbitrum, and Solana supporting 1,000+ transactions without operational issues.

Blockchain Developer, Nethermind

Remote | Mar 2024 - Jun 2024

- Identified and helped remediate 2 high-severity DeFi vulnerabilities that could have caused multi-million-dollar losses; audited 7 Ethereum and Starknet projects and led secure-coding workshops.

Web3 Developer, Rhodes Academy for the Creative Arts, Inc.

Remote | Nov 2023 - Mar 2024

- Shipped a decentralized identity system that reduced onboarding administration 30%, a DeFi tuition-payment flow, and a blockchain-based virtual gallery.

Full Stack Developer, EJADA

Cairo, Egypt | Feb 2023 - Oct 2023

- Shipped real-time chat for an e-learning platform, increasing engagement 20% and satisfaction 15%; improved query speed 30%, cut database use 40%, and containerized 4 services to reduce deployment time 50% and downtime 60%.

SKILLS

Product Engineering: Python, TypeScript, FastAPI, Node.js, NestJS, REST APIs, PostgreSQL, Redis, SQLAlchemy, WebSockets, multi-tenant and event-driven architecture, React, Next.js

AI and LLM Tooling: Qdrant, model routing, fallbacks, circuit breakers, LangGraph, OpenAI, Anthropic Claude, Google Gemini, agent orchestration, RAG, prompt caching, cost and latency instrumentation, evaluation gates, human approval, audit logs

ML, NLP, Speech and Vision: vector search, Azure TTS, Deepgram STT, transformer fine-tuning, real-time voice over WebSockets, Two-Tower recommender

Platform, Data and Mobile: Celery, Redis Pub/Sub, Alembic, dbt

Fintech Payments and Controls: authenticated webhooks using HMAC, idempotency, reconciliation, audit trails, tenant isolation, Paymob, recurring billing

Delivery and GTM: Docker, Automated testing, CI/CD, health monitoring, backup and rollback, performance tuning

Web3 and Security: secure coding, smart contract auditing

EDUCATION

Bachelor of Science in Computer Engineering, Alexandria University (2019 - 2023)

CERTIFICATES

IBM AI Engineering Professional Certificate, IBM (Coursera)

LANGUAGES

Arabic (Native) | English (Proficient)