Omar Salama

El Sheikh Zayed, Giza, Egypt · Open to remote (CET/EU timezones) and relocation

osalama710@gmail.com

omarsalama.dev · LinkedIn · GitHub

SUMMARY

Mobile and AI engineer, five years shipping production software. I work the full length of a feature: the model, the retrieval that feeds it, and the app it ships in. At Niyyah I build the AI backend behind the app, a retrieval-grounded answer pipeline plus its evaluation tooling, and the React Native app alongside it. On the side I fine-tuned Gemma-4 2B with SFT + DPO (LoRA) and shipped it on-device in OpenSpace, a privacy-first React Native therapy companion that runs entirely offline (open-sourced on HuggingFace).

EXPERIENCE

SDE II, Niyyah

September 2025 - Present

- Own the AI backend behind Niyyah's app: a retrieval-grounded answer pipeline (routes a query, gathers and ranks sources, then generates a cited, streamed answer) and the from-scratch evaluation harness that scores answer quality and gates what ships.
- Built a runtime validation layer that grades live answers on accuracy, structure, and citation correctness, and flags the likely fix when one fails.
- Designed and built the AI side of a personalized 30-day guided journey end to end in a 4-week sprint: onboarding into the challenge, day-by-day content unlocks, a daily reflection agent on the prior day's content, progress tracking, and the data models behind them.
- Designed the AI architecture for a conversational onboarding flow: an onboarding chat mode, intent routing, and a recommendation agent that suggests a personalized path.
- Cut input tokens per answer by 10-20% (lower cost and latency) and hardened reliability: a server-side streaming proxy persists in-flight answers, so a response survives the user backgrounding the app, dropping connection, or leaving the chat.
- Shipped user-facing chat features in the React Native app: streaming Q&A with inline source cards and a usage-based paywall.

Python Developer (LLM Training), Turing

June 2024 - September 2025 · Contract

- Promoted from contributor to pod lead, guiding a 4-person pod for ~3 months: reviewed submissions, calibrated output quality, and reported on pod performance.
- Built internal Jupyter tooling that parsed weekly contribution exports across parallel pods, generated comparison charts, and drafted progress reports via a lightweight LLM step.
- Produced and curated SFT training data for domain-specific model training, focused on data-analysis tasks.

Software Developer, Keepoala

May 2021 - October 2024 · Remote (Munich-based company)

- Joined as a working student and grew over 3.5 years into a core developer on a two-person mobile team. Shipped the consumer app (React Native): voucher rewards, challenges and leaderboard, a returns flow with QR-coded DHL labels, an order-tracking timeline, a carbon-savings view, deep linking, AppsFlyer analytics, and a full UX/UI redesign pass.
- Backend integrations on Firebase Cloud Functions: BillBee, WeClapp, Algolia, Shopify merchant plugin.
- Built a partner-shop analytics dashboard (React + R-Shiny): retention curves, NPS, and order metrics for Keepoala's brand partners.

Interview Engineer, Karat

October 2021 - March 2022; January 2025 - April 2025 · Contract

- Delivered 10-15 structured technical interviews (coding, debugging, system design) against Karat's calibrated rubrics for its enterprise clients; selected via a multi-round interviewer-calibration pass and re-engaged for a second stint in 2025.

PROJECTS

OpenSpace + TheraSpace (2025)

Privacy-first therapy companion (React Native) that runs entirely on-device via a custom fine-tuned LLM: zero network calls, zero data transmission.

TheraSpace (fine-tuning pipeline):

- Base model: Gemma-4 2B (google/gemma-4-E2B-it). Two-stage training: SFT on 13,238 rows from Counsel-Chat, Psych8k, AnnoMI, and AMOD, then DPO for persona steering.
- LoRA throughout (rank 8, alpha 16): SFT LoRA merged into base, fresh DPO LoRA trained on top with TRL; five personas steered by DPO style pairs (distinct voices, same competencies), controlled via system prompt rather than separate weights.
- Crisis-safety gate: every persona clears it (LLM-judged across 18 risk categories) before shipping.
- Arc evaluation: 89% stage-appropriate turns, 2.67/3 listening quality (LLM-judged). Training on Modal; quantized to Q4_K_M GGUF (~3.3 GB).
- Published as **MindSpace**, an open model card on HuggingFace.

OpenSpace (app):

- Inference via llama.rn 0.12.0: Metal on iOS, GPU layers on Android.
- On-device RAG: a second GGUF embedding model (CPU-pinned to keep the GPU free for generation) feeds a hand-rolled vector store that injects retrieved techniques and prior-session memory into each prompt.
- Encrypted local sessions (MMKV + Keychain), biometric auth (Face ID / Touch ID), rolling on-device context summarizer.

Hybrid retrieval system for Islamic knowledge (2026)

A self-initiated project (Python/FastAPI, ChromaDB + cross-encoder, hosted on a Hugging Face server) to improve how well AI answers stay grounded in classical Islamic sources: Quran (with tafsir), Hadith, the Mawsuah fiqh encyclopedia, classical Usul texts, and more (100k+ passages), each indexed on its own terms. Hybrid dense and lexical retrieval with cross-encoder reranking and citation-grounded results, plus an LLM-graded evaluation harness (retrieval precision, faithfulness, relevance, pairwise model bake-offs, and a benchmark suite). Multilingual Arabic and English embeddings retrieve source text directly.

NeuroGames (2022, Graduation Project)

Motor-imagery EEG classification driving a VR boxing game (4-person team). Emotiv Insight headset → MNE preprocessing → feature extraction/selection → LDA classification (scikit-learn), streamed to a Unity VR environment over WebSockets. LDA beat SVM, Random Forest, and KNN; results matched or exceeded commonly cited BCI Competition baselines. I owned feature extraction and selection. Grade: Excellent.

SKILLS

- **Mobile:** React Native (TypeScript, JSX), iOS (Swift), Android (Kotlin, Java), llama.rn, MMKV, Keychain, React Navigation, deep linking
- **AI / LLMs:** SFT, DPO, LoRA, GGUF quantization, llama.cpp, cross-encoder reranking, ChromaDB, LangChain, agentic orchestration, RAGTriad evaluation, LLM-judge pipelines
- **Web & data:** React (web), SQL, Firestore, SQLite, R-Shiny
- **Backend & infra:** Python, FastAPI, Node.js, Express, Firebase Cloud Functions, AWS (Lambda, EC2), Modal, HuggingFace
- **AI-assisted development:** Cursor and Claude Code in daily production work; authored custom skills, subagents, slash commands, and project rules
- **Languages:** TypeScript, Python, JavaScript, Swift, Kotlin, Java, C, C++ · German (A2), Spanish (A2)

EDUCATION

B.Sc. in Computer Engineering

Faculty of Engineering, Cairo University (2017-2022)

- Grade: 83.9% (Very Good with honors)
- Graduation Project: **NeuroGames** (Grade: Excellent), an EEG brain-computer interface driving a VR game (see Projects).
- Certification: Data Analysis Professional Nanodegree, Udacity.
- Certification: Mobile development certificates in iOS/Swift, Android, and Kotlin (Coursera, CLS Learning Solutions, 2018).