Ahmed Khaled Ali

+20 128 831 1023 | a.khaled.ali.99@gmail.com | Linkedin | Github | Portfolio | Cairo, Egypt

ML Engineer with 3 years of experience deploying generative AI models to production. Currently serving 10K+ users across platforms built with PyTorch, Hugging Face, and cloud GPU infrastructure.

Experience

Full-Stack ML Engineer — Sep 2024 – Present
SongLabAI Inc. (Startup) — *Remote*

- Built AdviceBuddy.ai, an AI mental wellness platform reaching 10K+ users and 50K+ conversations, deploying Llama 3.1-8B on Modal GPUs with therapeutic safety monitoring and content filtering
- Integrated MuseTalk lip-sync model for real-time AI video avatar responses and built text-to-speech pipeline supporting 12 voice configurations (gender × accent) as a premium feature
- Deployed 3 generative AI models into SongLabAI (50+ daily users): Pyramid Flow for video jingle generation, ACE-Step for full song synthesis with lyrics, and fine-tuned MusicGen-Large 3.3B for instrumental generation via Hugging Face Inference Endpoints
- Reduced inference costs across 5 production models through request batching, response caching, endpoint auto-scaling, and A/B testing deployment configurations

Founding Engineer — Apr 2024 – Sep 2024
Connectyed (Contract) — *Remote*

- Built and shipped full MVP for professional matchmaking platform in 5 months as sole engineer, from database design to production deployment on AWS
- Developed recommendation engine for match suggestions using AWS-hosted ML service, with Laravel REST API backend and Vue.js SPA frontend

Freelance ML Engineer & Developer — Mar 2022 – Sep 2024
Self-Employed (Upwork) — *Remote*

- Delivered 15+ client projects: ML model training and deployment (PyTorch, scikit-learn), data analysis dashboards (Plotly, Streamlit), and full-stack web applications (Python, FastAPI)

Projects

Enterprise RAG System | *FastAPI, ChromaDB, Redis, PostgreSQL, Docker* — 2024

- Built async document processing pipeline with multiple chunking strategies (semantic, recursive, fixed-size) achieving 1K+ chunks/min throughput, with semantic search via ChromaDB and query reranking
- Deployed multi-service architecture via Docker Compose (API, PostgreSQL, Redis, ChromaDB, Streamlit) with RBAC (4 roles, 5 permissions), rate limiting via slowapi, and JWT auth with token refresh

ML Training & Serving Pipeline | *PyTorch, Azure ML, FastAPI, Parquet* — 2024

- Implemented DLRM training pipeline on Amazon Reviews data with feature engineering (time-based features, normalization) and Azure ML integration with spot instance scheduling (60–90% cost reduction)
- Built FastAPI serving layer with batch prediction and top-k recommendation endpoints, cold-start fallback scoring, and 80 unit tests covering model architecture and data processing

NLP Analysis Pipeline | *Hugging Face Transformers, BERTopic, Plotly* — 2024

- Built sentiment classification (80%+ F1), BERTopic topic clustering, and NER with entity grouping across automotive review corpus, with 3 interactive Plotly dashboards and 115 unit tests

Education

Obour High Institute for Engineering and Technology — Cairo, Egypt
Bachelor of Engineering in Computer and Control Engineering — *Sep 2019 – Jul 2024*

Technical Skills

Languages: Python, SQL, JavaScript
ML & AI: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, scikit-learn, Pandas, NumPy
Infrastructure: Docker, AWS, Azure ML, Modal, Git, PostgreSQL, MySQL, ChromaDB, Redis, Supabase
Frameworks: FastAPI, Next.js, Streamlit, Gradio
Practices: Model Fine-tuning, Inference Optimization, CI/CD, A/B Testing, Vector Databases