Yahya Alnwsany

☎ +20 102 257 0742 | ✉ yahyaalnwsany39@gmail.com | Cairo, Egypt
Portfolio | GitHub | LinkedIn | Hugging Face | Docker Hub

SUMMARY

Agentic AI engineer architecting end-to-end ecosystems across Model Context Protocol (MCP), autonomous tool-calling agents, and neural voice surfaces. Delivered 10+ production-grade AI applications, fine-tuned seven transformer architectures, and led AI education programs while earning 20+ advanced certifications. Expert in orchestrating planner-executor loops, real-time speech interfaces, and safety-aligned evaluation pipelines that balance performance, cost, and responsible AI principles.

KEY ACHIEVEMENTS

- Launched 10+ AI platforms across NLP, computer vision, voice, and RAG, including safety-critical moderation and medical ML workloads serving 40K+ end users and internal stakeholders.
- Designed cross-provider MCP ecosystems powering autonomous tool-calling agents, driving a 38% reduction in manual resolution time for developer support and operations teams.
- Fine-tuned and released seven transformer-based models (DistilBERT, GPT-2, LlamaGuard, BLIP, NVIDIA NeMo) achieving up to 18% accuracy gains while keeping inference costs flat through quantization and LoRA adaptation.
- Deployed multilingual voice agents with streaming ASR/TTS handoffs that shortened customer response latency by 44% and enabled 24/7 conversational coverage.
- Completed 20+ advanced AI certifications while mentoring 200+ learners through ambassador, instructor, and coordinator programs that achieved 94% satisfaction scores.

PROFESSIONAL EXPERIENCE

LLM Engineer (Freelance) Aug 2025 – Present
Turing Remote

- Architected CodeGenBot, a retrieval-augmented coding assistant that cut retrieval latency by 35% and lifted pass@1 to 56% on an internal HumanEval-style benchmark using MCP-compliant tool adapters, guardrailed function calling, and adaptive reflection loops.
- Built an evaluation harness and conversational Streamlit IDE that executes code, visualizes outputs, and logs session history, reducing debugging time for pilot developers by 42%.
- Orchestrated autonomous remediation flows with ReAct planning, MCP, and Assistants APIs, enabling shell, ticketing, and documentation agents that resolved 28% of incidents without human escalation.

AI Engineer (Contract) Aug 2025 – Present
Jupiter AI Labs Remote

- Co-developed multi-agent copilots on hybrid cloud infrastructure, generating three proof-of-concept demos that accelerated partner onboarding by 25% and showcased self-healing workflows.
- Led weekly technical design reviews guiding adoption of scalable inference patterns and cost-optimized GPU workloads across distributed teams.
- Implemented observability with LangSmith and OpenTelemetry traces, cutting mean time to detection for hallucination regressions from 4 hours to 45 minutes.
- Piloted a voice-enabled field engineering agent using WebRTC, Riva TTS, and GPT-4o mini that automated 63% of routine service desk interactions.

NLP Engineer Intern Jun 2025 – Aug 2025
Cellula Technologies Remote

- Delivered a toxicity moderation stack combining a DistilBERT LoRA adapter with a Bidirectional LSTM baseline, boosting macro-F1 by 13 points across nine categories.
- Deployed moderation tooling via Streamlit dashboards that trimmed analyst review cycles by 45% and enabled side-by-side model comparison.

- Integrated voice escalation through Twilio Voice and NeMo speech services, providing real-time triage assistants that cut manual routing time by 31%.

Coding Instructor May 2024 – Aug 2025
iSchool Remote

- Delivered 48 live AI and programming workshops to 210+ students, achieving 94% satisfaction scores and 88% project completion.
- Authored modular curricula on ML, LLMs, and Python automation, later syndicated to three partner academies.

Coordinator Jun 2024 – Oct 2024
almentor Remote

- Launched a four-track AI learning program, onboarding 120 participants and improving certification pass rates by 27%.
- Streamlined event logistics and feedback loops, cutting planning cycle time from 10 to 6 days per cohort.

Ambassador Aug 2024 – Oct 2024
Solve Hub Remote

- Advocated responsible AI adoption by hosting six community events for 320 attendees, generating 40+ cross-disciplinary collaborations.
- Built a resource hub of ethical AI tooling that increased member engagement by 33%.

Internship Trainee Oct 2023 – Jan 2024
DeepLearning.AI Remote

- Completed a TensorFlow-focused residency, delivering four deep learning prototypes covering CNNs, RNNs, and emerging LLM workflows.
- Collaborated on curriculum pilots that improved learner assessment accuracy by 18%.

SELECTED PROJECTS

Dual-Stage Toxic Moderation App | *Streamlit, LlamaGuard, DistilBERT, BLIP, MCP*

- Combined LlamaGuard hard-safety gating with DistilBERT+LoRA classifiers, BLIP image captioning, and MCP-governed escalation agents inside Streamlit, reducing moderator false negatives by 22% and processing 12K+ content items in pilot deployments.

CodeGenBot RAG Code Assistant | *RAG, MCP, Streamlit*

- Implemented semantic retrieval over HumanEval-style corpora with planner-executor agents orchestrated through MCP, boosting code suggestion acceptance rate by 18% while providing inline execution, testing, and telemetry dashboards.

Arabic Transcriber Pro | *NVIDIA NeMo, Whisper, Hugging Face, Streamlit*

- Built a Hugging Face-hosted Streamlit ASR app with NVIDIA NeMo and Whisper diarization, supporting multi-format uploads, auto-resampling, and RTL UX; achieved 11% lower WER versus baseline, powered contact-center voice agents, and served 470+ transcripts in first month.

VoyceOps Autonomous Agent | *GPT-4o, Riva TTS, MCP*

- Delivered a multilingual voice concierge leveraging GPT-4o, Riva TTS, and MCP toolchains for scheduling, CRM updates, and knowledge retrieval, automating 61% of inbound calls with CSAT scores above 4.6/5.

Respiration App | *ML, Mobile Development*

- Delivered a mobile ML system with 0.8 breaths/min MAE for real-time respiration monitoring, integrating with wearable sensors for healthcare and fitness pilots.

Arabic Poetry GPT-2 Fine-Tuning | *GPT-2, Hugging Face, Gradio*

- Fine-tuned GPT-2 on a curated Arabic poetry corpus, improving BLEU scores by 24% and releasing interactive demos on Hugging Face and Gradio for creative writing use cases.

EDUCATION

Kafr El-Sheikh University Oct 2022 – Jul 2026
B.Sc. in Artificial Intelligence Egypt

WorldQuant University (Online) Oct 2022 – Jul 2026
Applied Data Science Program Remote

CERTIFICATIONS

TensorFlow Developer Specialization Mar 2024
DeepLearning.AI

Deep Learning & Machine Learning Specializations Oct 2023 & Aug 2023
DeepLearning.AI

NLP Engineer Internship Certificate Aug 2025
Cellula Technologies

TECHNICAL SKILLS

Programming Languages: Python, C++, SQL, Shell Scripting

Frameworks & Libraries: PyTorch, TensorFlow, Transformers, Hugging Face, LangChain, LangGraph, LlamaIndex, NVIDIA NeMo, Whisper, Riva, Ultralytics YOLOv5/v8, ONNX Runtime

AI & ML Domains: Agentic Orchestration & MCP, Retrieval-Augmented Generation, Prompt Engineering & PEFT-LoRA, Voice AI, Speech, NLP, and Computer Vision

Tooling & Deployment: Docker, Docker Compose, Kubernetes, Helm, Streamlit, Gradio, Hugging Face Spaces, WebRTC, Twilio Voice, Vonage