

Yadidiah Kanaparthi

AI Engineer · Full-stack LLM systems · Production deployment

Dubai, UAE / yadidiah.wrk@gmail.com / linkedin.com/in/yadidiah-kanaparthi / github.com/Yadidiah-k / yadidiah.vercel.app

§01 SUMMARY

AI engineer who builds features from idea to full production deployment — owning the lifecycle from scoping and architecture through model integration, infra, and shipping to real users. Built an agent harness with an agentic loop (LangGraph) powering a live legal-translation platform, including a 5B-parameter Gemma model fine-tuned (QLoRA) on a 13K-example domain dataset; built an MCP-compatible gateway integrating Claude, Cursor, ChatGPT, and custom agents; built a production-shaped RAG pipeline end to end (chunking, embeddings, vector store, grounded generation with citation validation). Comfortable scoping ambiguous problems and moving fast across Python/FastAPI, NestJS/Next.js, and cloud infra (GCP/AWS/Azure). 3 peer-reviewed publications (Springer/IEEE).

§02 CORE TECHNICAL SKILLS

FULL-STACK & DELIVERY	FastAPI · NestJS · Express · Next.js · React · TypeScript · Python — production ownership from architecture to deployment
LLMS & AGENTIC SYSTEMS	LangChain · LangGraph · DSPy · MCP tool integration · RAG pipelines · vector DBs (FAISS, Pinecone, MongoDB Atlas) · fine-tuning (LoRA/QLoRA) · prompt engineering
VOICE AI	LiveKit · Pipecat · Deepgram · Whisper · Silero VAD · real-time STT/TTS
ML & DEEP LEARNING	PyTorch · TensorFlow · Hugging Face Transformers · scikit-learn · XGBoost · CNN · RNN/LSTM · BERT · GPT
CLOUD & MLOPS	GCP (Cloud Run, GKE, Vertex AI) · AWS (SageMaker, Lambda, S3) · Azure (ML, Container Apps, Document Intelligence) · Docker · Kubernetes · GitHub Actions · MLflow

§03 PROFESSIONAL EXPERIENCE

AI Engineer — Humai

SEPT 2025 — PRESENT

Owens features end to end — from idea to production — across voice, matching, and a verified-work platform.

- Verified-work platform (Mairit):** Contributed to an MCP-compatible platform (NestJS, Next.js, Python/LangGraph) live in production on Azure Container Apps, using Azure Document Intelligence and Cognitive Services for the legal-translation vertical. The MCP gateway integrates Claude, Cursor, ChatGPT, Lovable, and custom Node agents — coordinating across multiple external client integration surfaces.
- Agent harness architecture:** Built the legal-translation pipeline as an agent harness with an agentic loop (LangGraph) rather than a single linear flow — multi-step tool use and control flow for document translation and validation.
- Domain fine-tuning:** Fine-tuned Gemma-4 (5B params, QLoRA via Unsloth) on a 13K-example English–Arabic legal-translation dataset: chat-template formatting, response-only loss masking, inference validation — owned end to end from data prep through deployment.
- Voice agent pipeline:** Built a sales voice agent on Pipecat (Deepgram STT, GPT-4, Silero VAD) with a parallel Whisper-validation layer that cross-checks transcription in real time without adding response latency.
- Matching engine:** Designed a 6-stage compatibility-matching pipeline (FastAPI + MongoDB Atlas Vector Search, Sentence-BERT and CLIP scoring, auction-based assignment) processing 10K profiles in ~8 minutes at a 70–80% match rate.
- Recruiting platform:** Shipped services for an AI recruiting product — LiveKit interview agents and a CV-analysis pipeline scoring 200+ candidates — across full-stack and infra boundaries under shifting requirements.

Enterprise RAG System — AskMyDocs · FAISS

MAY 2025

- Retrieval-augmented generation over 10 GB+ document collections with OCR ingestion, sub-3-second query latency, and ~89% relevance on internal eval.

Custom GPT language model · PyTorch

AUG 2025

- Decoder-only transformer from scratch — pre-norm architecture, multi-head causal self-attention, weight-tied embeddings (~2M parameters).
- Gradient checkpointing and Flash Attention cut GPU memory ~40%, enabling longer context on commodity hardware.
- Custom BPE tokenizer and training loop (AdamW, cosine LR schedule, top-k / temperature sampling).

Customer churn prediction · XGBoost · FastAPI

JUL 2025

- Churn-prediction service (92.3% accuracy) with SHAP explainability, containerized with Docker/FastAPI for sub-100 ms inference.

3D medical image segmentation — comparative study · TensorFlow / Keras

MAY 2025

- Implemented and benchmarked 5 segmentation architectures (U-Net, V-Net, SA-Net, E1D3, HDC) on 3D MRI brain-tumor data — bachelor's thesis contributing to a peer-reviewed publication.
- Streamlit interface for real-time inference in a clinical-demo setting.

§05 OPEN SOURCE

- **Merged:** documentation fix in **pydantic** (28k+ ★) — PR #13375.
- **Pending review:** bugfix PRs against Hugging Face *huggingface_hub*, Microsoft *VS Code*, Anthropic's *anthropic-sdk-python*, and Google's *python-fire*.

§06 EDUCATION, PUBLICATIONS & CERTIFICATIONS**B.Sc. Computer Science, Magna Cum Laude — Canadian University Dubai**

AUG 2021 — MAY 2025

- **Publications:** 3 peer-reviewed papers, Springer & IEEE (2024–2025) — IoT–ML integration, security threat detection, systematized log analysis.
- **Certifications:** Oracle GenAI Professional · Google MLOps for GenAI · Microsoft GenAI · KNIME Data Science.
- **Job simulation:** BCG GenAI Consulting (Forage, Jan–Mar 2025) — SEC-filings chatbot, NLP pipeline, deployment automation.