

ARMAND JUNIOR DONGMO NOTUE

ASSOCIATE AI/ML ENGINEER | PYTHON | PYTORCH | GENAI | FASTAPI

Arlington, TX | notuearmand250@gmail.com | armandnotue.com | linkedin.com/in/notue250 | github.com/ARMAND-cod-eng

PROFESSIONAL SUMMARY

Early-career AI/ML Engineer with an M.S. in Data Science and hands-on experience designing, evaluating, and deploying machine learning and generative AI applications. Strong Python, SQL, pandas, NumPy, PyTorch, TensorFlow, FastAPI, REST APIs, vector databases, RAG, LLMs, prompt engineering, Docker, AWS, Git, and CI/CD. Built and operate an end-to-end AI product in production, a multi-agent research platform, and machine learning systems with rigorous evaluation. U.S. Permanent Resident, authorized to work without sponsorship.

TECHNICAL SKILLS

Programming / Data: Python, SQL, TypeScript, pandas, NumPy, PostgreSQL, SQLite

Machine Learning / Deep Learning: PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, CatBoost, BERT, LSTM, Autoencoder, Isolation Forest, SHAP

Generative AI / NLP: LLMs, prompt engineering, LangChain, LangGraph, Hugging Face Transformers, RAG, RAGAS, Claude, GPT-4o, LLaMA-3, NLP, agentic AI

Retrieval / Databases: pgvector, ChromaDB, Pinecone, BGE rerankers, embeddings, vector search, ETL, feature engineering

Backend / Cloud / Dev: FastAPI, REST APIs, Pydantic, asynchronous processing, Docker, AWS EC2, SageMaker, Git, GitHub Actions, MLflow, Grafana, CI/CD

PROFESSIONAL EXPERIENCE

AI Engineer & Founder | IntelliCast AI | Remote | Aug 2025 - Present

- Built and operate a personalized AI news agent, live in early access at intellicast.news: ingestion from 14 outlets, story clustering with local embeddings and pgvector, cross-outlet claim verification, and goal-based ranking (Python, FastAPI, React/TypeScript).
- Developed a streaming script → fact-check → voice pipeline with Llama 3.3 70B and two-host Cartesia voices in English and French, plus grounded follow-up Q&A with citations; about 15 s from tap to first audio in internal testing.
- Ship it as one Docker image on AWS EC2 provisioned with Terraform; GitHub Actions CI runs type checks, linting, and 180 offline backend tests before every image build.

AI Training Specialist, RLHF / SFT | Outlier / Scale AI | Remote | May 2024 - Jan 2026

- Evaluated LLM outputs using structured rubrics for factuality, reasoning, safety, instruction adherence, and code correctness across technical and coding tasks.
- Produced RLHF preference data, SFT examples, adversarial prompts, and evaluation datasets used in model post-training and quality-assurance workflows.
- Verified model-generated code, investigated failure modes, and red-teamed hallucination and jailbreak behaviors to improve model reliability and responsible AI practices.

Data Analyst Intern | The African Think Tank | Dallas, TX | Aug 2024 - Jan 2025

- Cleaned, joined, and analyzed survey data using Python and SQL, built Tableau dashboards, and presented data-driven findings and recommendations to nontechnical leadership.

SELECTED AI / ML PROJECTS

RAG Document Intelligence API (in progress) | Python | FastAPI | LangChain | ChromaDB | Sentence-Transformers

- Building a document question-answering service in FastAPI: PDF ingestion, chunking, sentence-transformer embeddings in ChromaDB, top-k retrieval, and answers grounded in retrieved passages (Llama 3 70B via Together AI).
- Configuration-driven pipeline: models, chunk size, and retrieval depth live in one typed settings module so retrieval experiments stay reproducible.
- Next: BGE reranking, RAGAS evaluation on a fixed question set, streamed responses over SSE, and Docker deployment.

Real-Time Fraud Detection Ensemble | Python | PyTorch | scikit-learn | XGBoost | pandas

- Designed a stacked ensemble: a PyTorch autoencoder trained on legitimate transactions and an Isolation Forest produce anomaly scores that feed an XGBoost classifier over 44 engineered features.
- Achieved 0.9834 ROC-AUC and 85.71% recall at a 0.16% false-positive rate on a held-out test split of the 284,807-transaction ULB dataset, with 86 ms mean scoring latency in a notebook benchmark.
- Analyzed precision–recall trade-offs and feature importance (the Isolation Forest score ranks second) and modeled cost-benefit by decision threshold.

Atlas, Autonomous Multi-Agent Platform | Python | LangGraph | FastAPI | arq | Redis | PostgreSQL | Next.js | Docker

- Built a multi-agent research platform: a LangGraph planner fans out up to five parallel researchers, a critic checks findings against retrieved sources and re-runs only faulted branches, and a synthesizer writes cited answers.
- Runs execute on an arq worker with a Redis event stream and resumable SSE, so a dropped client loses nothing; per-run token budgets cap cost, with Claude Sonnet 5 and GPT-4o failover.
- Measured locally: five parallel branches finish in 23.3 s vs 93.4 s back to back; 168 backend and 34 frontend tests. Now building a graded regression eval suite with a CI gate.

Transformer Sentiment Analysis | Python | PyTorch | Hugging Face Transformers | BERT | FastAPI | Docker

- Fine-tuned BERT with PyTorch and Hugging Face Transformers on more than 50,000 text samples for sentiment classification.
- Built a FastAPI inference service with approximately 120 ms response time and packaged the application with Docker for reproducible deployment.
- Applied NLP preprocessing, model evaluation, and production-oriented API design to move the model from experimentation to a deployable service.

EDUCATION

Master of Science in Data Science | Grand Canyon University | Jan 2024 - Oct 2025 | GPA: 3.43