

DILIP NALLAMASA

Buffalo, NY | +1 716-275-7613 | dilip.nallamasa@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)
AI Engineer | Machine Learning Engineer | LLM Systems

EDUCATION

University at Buffalo, SUNY - M.S. in Artificial Intelligence | Buffalo, NY

Aug 2025 - Dec 2026

Jawaharlal Nehru Technological University - B.Tech in Computer Science & Engineering | Hyderabad, India

Dec 2021 - May 2025

TECHNICAL SKILLS

Programming: Python, C++, SQL, TypeScript/JavaScript, Bash

ML / LLMs: PyTorch, Hugging Face Transformers, TRL, PEFT, Large Language Models (LLMs), Generative AI, Qwen, LLM fine-tuning (QLoRA/LoRA, SFT, DPO), scikit-learn, XGBoost, Retrieval-Augmented Generation (RAG), BGE embeddings, BM25, Reciprocal Rank Fusion (RRF), cross-encoder reranking

Agents / Evaluation: LangChain, LangGraph, Model Context Protocol (MCP), agentic workflows, tool calling, structured outputs, LLM evaluation, benchmark design, regression testing, paired bootstrap, human-in-the-loop, guardrails

Serving / MLOps: vLLM model serving, FastAPI, REST APIs, vector databases (Qdrant), PostgreSQL, Redis, Docker, Kubernetes, AWS (EC2, S3, Lambda, Bedrock, SageMaker), Airflow, MLflow, LLM observability (Langfuse), OpenTelemetry, Evidently, CI/CD (GitHub Actions)

WORK EXPERIENCE

Globant - AI Engineer

Aug 2026 - Present

Remote

- Building enterprise Generative AI services grounded in client data, taking LLM features from prototype through evaluation, REST API integration, testing, production debugging, and deployment in a peer-reviewed codebase.
- Collaborating with engineers and stakeholders on model behavior, failure analysis, code review, and end-to-end delivery of production AI features.

PwC - Software Engineer

Sep 2024 - Jul 2025

Hyderabad, India

- Shipped production Python/LangChain/LangGraph Retrieval-Augmented Generation (RAG) and multi-agent workflows for enterprise clients, reducing manual knowledge-retrieval effort by ~30%.
- Built Python/SQL ETL and validation pipelines across 10K+ records; supported architecture, testing, deployment, and handoff, improving workflow turnaround by ~25%.

RESEARCH EXPERIENCE

University at Buffalo - Graduate Researcher, Chess & LLM Modeling [Demo](#)

Jan 2026 - Present

Prof. Kenneth W. Regan

- Built ChessAssistLens to compare a Regan-inspired statistical human-move model, Qwen3-8B, and Maia-3 on identical chess positions using Stockfish MultiPV, legal-move likelihood scoring, and controlled engine substitutions.
- Designed leakage-safe held-out evaluation with paired source-game bootstrapping (2,000 resamples); Maia-derived signals improved AUC by 0.047 (95% CI: +0.040 to +0.054).

University at Buffalo - Research Assistant, Security & Privacy in Physical AI

Aug 2026 - Present

Prof. Haipeng Cai

- Systematizing security and privacy failure modes of LLM/VLM-driven Physical AI across prompt/tool attacks, adversarial perception, data leakage, and unsafe physical-world actions.
- Prototyping an LLM-assisted vulnerability-repair loop that localizes disclosed flaws, proposes patches, and regression-tests fixes in a sandbox.

PROJECTS

OpsVerse | Qwen3-4B, PyTorch/Transformers, TRL/PEFT, vLLM, Qdrant, Langfuse [GitHub](#) [Demo](#)

- Fine-tuned Qwen3-4B with QLoRA/SFT/DPO and built hybrid BGE+BM25 retrieval with RRF/cross-encoder reranking and grounded citations; evaluated 7,383 chunks with 1,914 graded judgments at MRR@10 0.705.
- Gated releases with 15 automated CI evaluation checks and 245 regression tests; Langfuse profiling isolated serving-layer bottlenecks and reduced cost per response by ~6x.

AxonFDE | LangGraph, FastAPI, PostgreSQL, Redis, OpenTelemetry [GitHub](#) [Demo](#)

- Fused minute-level telemetry, orders, and shipping documents to predict cold-chain breaches 49 minutes earlier than threshold alarms across 40 simulated scenarios; detected 23/23 planted cross-system conflicts at 100% precision.
- Built human approval gates, rule-bounded LLM explanations, adversarial SQL controls, and hash-chained audit logs; 0 unauthorized actions across 50 role-action tests and a 4.92s end-to-end decision cycle.

Pro2Pro | LangGraph, MCP, Claude, Amazon Bedrock, LangSmith, Logfire [GitHub](#) [Demo](#)

- Built a checkpointed LangGraph multi-agent workflow from discovery through deployment with Model Context Protocol (MCP) tool integrations, mandatory human approval, and resumable state.
- Added guardrails, Claude/Bedrock fallback routing, per-call cost tracking, LangSmith/Logfire tracing, and 78 evaluation tests.

FIFA2026MLOps | XGBoost, Airflow, MLflow, Docker, Evidently, OpenTelemetry [GitHub](#) [Demo](#)

- Built an automated forecasting pipeline using Elo ratings, gradient boosting, and Monte Carlo simulation, orchestrated with Airflow and GitHub Actions.
- Tracked experiments and model versions in MLflow with Dockerized deployment, OpenTelemetry tracing, and Evidently data-drift monitoring.

PUBLICATIONS

- [GWO-Based Fed-UNet-CNN Model for Leukocytes Classification Across Developmental Stages](#) - Research Square preprint; 99.29% benchmark accuracy
- [Smart-factory big-data analytics and automation](#) - IEEE Xplore