

YOGVID WANKHEDE

AI Engineer — LLM Systems, Agents & Production ML

+1(909) 705-0369 | yogvidwankhede@gmail.com | [LinkedIn](#) | [GitHub](#) | yogvidwankhede.com

EDUCATION

Master of Science, Data Analytics & Statistics August 2024 - May 2026
Washington University in St. Louis

Bachelor of Technology, Artificial Intelligence August 2020 - June 2024
G.H. Rasoni College of Engineering

Certifications: [AI Fluency](#), [Claude 101](#) (Anthropic)

Research: [Springer Nature Singapore](#) book chapter (conference-presented), sign-language interpretation with Dynamic Time Warping; [IJSRD](#) journal article, cyber-attack forecasting

SKILLS

Programming	Python, TypeScript, JavaScript, SQL, MATLAB
LLM & Agentic AI	LangGraph, LangChain, Claude/OpenAI APIs, Multi-Agent Orchestration, RAG, Prompt Engineering, Evals
ML & Data Science	Machine Learning, PyTorch, scikit-learn, NLP, Computer Vision, Time Series
Cloud & MLOps	GCP (Cloud Run, Cloud Scheduler, Cloud Tasks), AWS (EC2), Docker, MLOps, REST APIs, Agile
Data & Retrieval	PostgreSQL, MySQL, Pinecone, ChromaDB, Embeddings & Semantic Search, Tableau, Power BI

EXPERIENCE

Software Engineering Intern (AI/ML), EUNO (The Epiphany Company, LLC) June 2026 - Present

- Build the 6-stage LLM agent pipeline behind the product's core feature (LangGraph.js, NestJS, TypeScript), turning user conversations into structured records the product acts on, ~98% correct first-pass
- Engineer cost-aware routing across GPT-4o-mini and Claude Haiku 4.5 with per-call token budgeting, cutting the cost of running the feature ~40%, to under \$0.002/cycle
- Ship and own the autonomous worker in production on Google Cloud (Cloud Scheduler, Cloud Run, Cloud Tasks) with a Redis/BullMQ queue, monitoring 2 daily cycles so the feature runs unattended
- Author the runbooks and operational documentation that let teammates across the organization own the system after handoff; provide technical support and troubleshoot failures in an Agile team

Research Assistant, Washington University School of Medicine January 2026 - June 2026

- Implemented the Method of Fundamental Solutions in Python and MATLAB, replacing the Boundary Element Method in a noninvasive imaging pipeline to cut computation time 40% and lift correlation 0.85 to 0.92
- Validated the reconstruction across 118 test episodes at median correlation 0.83, stable under 10% noise, and presented findings to the multidisciplinary clinical research team behind a noninvasive preterm labor test

Teaching Assistant, Washington University in St. Louis August 2025 - January 2026

- Taught weekly labs and recitations for 40+ undergraduates and mentored students one-on-one in office hours, translating abstract theory into working intuition for a non-expert audience

Data Science Intern, ASTERISC December 2023 - June 2024

- Developed a cyber-attack forecasting system in Python (ARIMA, LSTM, Random Forest) on 494K+ imbalanced records, hitting 95.4% accuracy at a 0.6% false-alarm rate a small team could review
- Deployed a Dynamic Time Warping classifier matching emerging attack sequences to known patterns at 88.3% accuracy with a 4-8 day lead time, and led the research initiative behind the published findings

Data Analyst, SAFETECHNO SERVICES January 2021 - December 2023

- Collaborated with engineering, maintenance, supply chain, and finance teams: consolidated data in SQL across 15+ Oil & Gas projects and delivered Tableau and Power BI dashboards, cutting report preparation ~30%

PROJECTS

VentureOS: Multi-Agent AI Co-Founder — LangChain, Gemini, Flask, Server-Sent Events

- Orchestrated 8 domain-specialist agents on LangChain turning a one-sentence startup idea into market research, competitor analysis, 3-year financials, investor matches, and a landing page
- Streamed agent progress over Server-Sent Events so users can follow long asynchronous runs; [shipped live](#), top 10 at OpenAI's Handshake competition

HealthMate-AI: RAG System with Domain-Adapted Embeddings for Medical Decision Support

- Designed a Python retrieval pipeline over 40K+ medical text chunks on Pinecone at 93.3% accuracy, grounding every answer in a cited source rather than model memory
- Fine-tuned [sentence-transformer embeddings](#) with LoRA under 3-fold cross-validation (+18.31% Spearman) and benchmarked GPT-4o against Mistral to justify the production choice

Rubricron: Evaluation Harness for LLM Quality Claims — Python, applied statistics

- Created a Python evaluation harness that blocks model-quality claims the evidence cannot support, running 11 pre-registered checks before any comparison ships as a win
- Wrote the statistical core from scratch (Krippendorff's alpha with bootstrap intervals, Fleiss' kappa, Gwet's AC1), validated against published reference values under 225 tests