Vijay Ram Enaganti

Applied AI / ML Engineer - *LLM Evaluation, Prompt Optimization, and Agentic Systems*
Email: vijayram.enag2002@gmail.com | Phone: +1 (951) 347-5576 | LinkedIn: linkedin.com/in/vijay-ram-enaganti
GitHub: github.com/VjayRam | Portfolio: vijay-ram.vercel.app

Education

University of California, Riverside Riverside, CA
MS, Computer Science (Specialization in ML and GPU Programming) - CGPA: 3.70 / 4.00 Sep 2024 - Dec 2025

PES University, RR Campus Bengaluru, India
B. Tech, Computer Science Engineering - CGPA: 8.30 / 10.00 Aug 2020 - Jun 2024

Skills

- **Languages:** Python, C/C++, SQL, JavaScript, Bash.
- **Generative AI & LLMs:** PyTorch, TensorFlow, Transformers, LangChain, DSPy, Ollama, Autogen, LoRA/PEFT, Vertex AI.
- **MLOps & Cloud:** Docker, AWS, GCP, Azure, MLFlow, DVC, Weights & Biases, Git, GitHub Actions.
- **Web & Data:** FastAPI, React, Node.js, Flask, PostgreSQL, MongoDB, Spark, Vector DBs (Qdrant/Pinecone).

Work Experience

LivePerson Inc. *New York City, United States*
ML/AI Intern (Summer & Co-Op) Jun 2025 - Dec 2025

- Architected an end-to-end Prompt Optimization Framework using Google Vertex AI and Google Generative AI Evaluation Service, adapting the Microsoft ProTeGi (Textual Gradient Descent) approach over heuristic tuning to improve convergence stability and automate prompt optimization for LivePerson products.
- Led the migration of Copilot products from GPT-4o to Gemini 2.5 Flash/Pro, using the custom prompt optimizer to reduce prompt tuning lifecycles from *2 weeks to 3 days* while cutting costs and ensuring zero degradation in output quality.

Center for Information Security, Forensics and Cyber Resilience (C-ISFCR) *Bengaluru, India*
Research Intern Jan 2024 - Jul 2024

- Developed and optimized video anomaly detection systems (I3D, ViT, Transformers) and secured procurement of a dedicated GPU cluster, resulting in a 50% reduction in training time and a 60% increase in experimental throughput.
- Designed 8 rigorous experimental workflows to benchmark transformer architectures, leading to a co-authored paper identifying critical latency gaps in real-time surveillance systems.

The Param Science Experience Centre *Bengaluru, India*
Software Developer Intern Jul 2022 - Jan 2023

- Built interactive persona-based chatbots for museum exhibits using the OpenAI API (GPT-3.5), engineering custom system prompts to simulate historical scientists for visitor engagement.

Projects

PromptFlow - End-to-End Prompt Evaluation Platform *Live Demo*
Python | LLM-as-a-Judge | Prompt Evaluation | Dataset Validation

- Engineered a robust LLM-as-a-judge framework that supports custom metric definitions and automated dataset validation, enabling users to benchmark prompt performance against pre-built evaluation templates.
- Implemented real-time observability features to track evaluation progress, API call volume, and estimated costs, providing granular insights into model testing overhead and performance.

AgileBot *Github Repository*
Django | React | Langchain | Tailwind | Qdrant | Gemini 2.5 Flash

- Architected AgileBot (RAG-based SaaS) platform using Gemini 2.5 Flash, Docling, and Qdrant to automate user story/task generation from requirement documents.
- Improved Agile workflow efficiency, reducing manual effort by 30% through robust AI-driven task validation and sprint planning features.

Publications and Open Source

Analysis of Image Filters for Binary Fabric Classification (ICCICN 2024) *[Paper]*
Open Source: Contributor to **sktime** (Python Package) | Maintainer of **AI Vault** (Reusable AI Components)

Honors and Hackathons

- **1st Place - AI Pitch Competition 2025 @ UC Riverside (30 Teams):** Built an Agentic Visual Verification System using Microsoft OmniParser and Vision LLMs to autonomously test browser applications.