Vineeth Sai **Varikuntla**

vineethsai4444 — Portfolio — vineeth-sai — vineeth-sai — vineeth-sai

Education

Master of Science in Data Science — **University of the Pacific, San Francisco**
Advanced ML/Deep Learning | NLP | Computer Vision — *Aug 2023 – May 2025*

Bachelor of Technology in Computer Science & Engineering — **Jawaharlal Nehru Technological University**
Machine Learning, Data Mining, Artificial Intelligence — *July 2018 – July 2022*

Professional Experience

Google (Sunnyvale, CA)
Data Scientist *— AI/ML Engineer | AI Agents & RAG | Automation & Analytics* — *May 2025 – Jan 2026 (Contract)*

- Engineered RL-optimized AI agents and fine-tuning pipelines on AgentSpace to automate LLM data labeling, eliminating ∼200 weekly engineering hours and reducing labeling costs by 35%.
- Designed SQL dashboards to monitor MCP reliability metrics and ran A/B tests on agent configurations, improving task completion rate by 22% and reducing agent failure incidents by 18%.
- Architected Agentic RAG workflows for multimodal classification with prompt optimization, LLM-as-a-Judge evaluation, and Responsible AI guardrails before production release.

Foundation AI (Hyderabad, India)
Data Scientist *— ML Engineer | NLP | Computer Vision | Document Processing* — *Aug 2022 – May 2023*

- Productionized high-volume document extraction pipelines with CI/CD and MLOps monitoring; fine-tuned BERT, RoBERTa, and LayoutLM v3 for document unitization, applying quantization to cut inference latency by 40% and OCR misclassification by 28%.
- Built model evaluation dashboards to track drift, precision, and recall across document classes in real time, enabling proactive retraining and achieving 95%+ accuracy in a cloud-native production environment.

Aion Labs (San Francisco, CA)
Research Scientist/ AI Engineer *— Fine-tuning diffusion models | Multimodal RAG | Agentic RAG* — *Jan 2025 – May 2025*

- Delivered a production RAG chatbot for a FinTech client from prototype to Dockerized deployment with sub-100 ms inference (Azure OpenAI, Pinecone); built agentic RAG workflows and diffusion model fine-tuning pipelines.
- Fine-tuned Stable Diffusion on proprietary art/media datasets and implemented LLM-based candidate screening with spider-chart explainability reports, delivering interpretable client-facing insights from black-box model decisions.

Internships

Nurjana Technologies (SF, CA)
Data Science Intern *| SNN | Akida Neuromorphic Processor* — *May 2024 – Aug 2024*

- Pushed the frontier of ultra-low-power AI for space: optimized SNN object detection on the Akida chip, slashing inference latency by 50% and energy consumption by 40%, enabling real-time vision in power-constrained satellite environments.

BambiHealth (SF, CA)
Data Scientist – SDE Intern *| Speech Recognition | Backend* — *Feb 2024 – May 2024*

- Engineered a fault-tolerant multi-cloud speech-to-text pipeline for clinical use, boosting transcription accuracy by 25% and cutting latency by 30%, directly improving the speed and reliability of patient-facing healthcare workflows.

Zummit Info. Labs (Bangalore, India)
Data Scientist Intern *| Computer Vision | Medical Imaging* — *May 2022 – Sept 2022*

- Designed an end-to-end SageMaker medical imaging pipeline (U-Net segmentation + PyRadiomics feature extraction), boosting diagnostic accuracy by 30% and processing throughput by 2.5×, accelerating radiologist workflows at scale.

Projects

Pixel-to-Action: Fine-Tuning a Multimodal LLM for Autonomous Decision Making — **Ministral-8B, GRPO, QLoRA, PyTorch**
Transformed an 8.4B-param vision-language model into an autonomous action-taking agent that learns directly from raw screen pixels with no hand-crafted features, achieving a ***95.6% hit rate*** *on a live game environment using GRPO reinforcement learning, QLoRA parameter-efficient fine-tuning, and a Gaussian policy head for continuous action output.*

AgenticRAG Research Assistant — **RAG, LangChain, Ollama, Qwen2.5, Llama3.1**
Built a multi-agent RAG assistant with autonomous retrieval, reranking, and context-aware Q&A over large document corpora; integrated RAGAS-based LLM evaluation to benchmark answer faithfulness and reduce hallucination rate.

Real-Time Conversational Voice Agent — **Deepgram, HuggingFace, FAISS, Streamlit**
Built a low-latency conversational voice assistant using Deepgram for real-time transcription, FAISS for personalized semantic retrieval, and HuggingFace models for response generation; optimized the end-to-end inference pipeline to achieve sub-200 ms response time.

Technical Skills

- **Programming & Tools:** Python, SQL (Snowflake/BigQuery), Snowpark, Cortex AI, Git, Docker, Kubernetes, CI/CD, FastAPI, Streamlit, AWS (Bedrock, SageMaker), GCP (Vertex AI), Azure AI, Terraform, GoogleADK
- **AI/ML Technologies:** PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace Transformers, LangChain, LangGraph, LlamaIndex, CrewAI, FAISS, Pinecone, ChromaDB, Weaviate, Milvus, MLflow, Kubeflow, vLLM, TensorRT, DSPy, RAGAS, Arize, PydanticAI, SGLang, LiteLLM, Mem0, GraphRAG, LangSmith, OpenTelemetry, Ollama, n8n
- **Key Concepts:** LLMs, Generative AI, RAG, Fine-Tuning, Quantization, Inference Optimization, Prompt Engineering, Agentic RAG, RLHF, Diffusion Models, Computer Vision, NLP, Evaluation, A/B Testing, MCP, LLM-as-a-Judge, GraphRAG, MLOps, Multi-Agent Systems, Retrieval Optimization, Knowledge Distillation, Embeddings, Vector Search, Context Window Management