V SaraVivek

+91 9618934336 | vivekvarikuti22@gmail.com | github.com/vivekvari-dl | linkedin.com/in/vivekvar | vivekvari.dev

Technical Skills

GenAI & LLMs: LangChain, LangGraph, RAG Pipelines, Agent Orchestration, Finetuning, VLLM
Backend: Python, FastAPI, Microservices, REST APIs
Databases: MongoDB, PostgreSQL, Redis, Vector Databases(Pinecone)
Cloud & MLOps: Azure, CI/CD, weights biases, Docker, Kubernetes
ML Frameworks: PyTorch, TensorFlow, scikit-learn, XGBoost, Hugging Face, Diffusers

Experience

AI Engineer Aug 2025 – present
AI4AP POLICE *On-site*

- Developed AI-powered legal compliance system for POCSO charge sheet evaluation using RAG and NLP, automatically generating compliance scorecards and evidence gap reports within 5 minutes per case
- Built FastAPI backend services with secure document processing pipeline, handling complaint-to-charge-sheet workflows and delivering AI-generated summaries and section recommendations to investigating officers
- Engineered knowledge base with 1000+ indexed legal documents (POCSO Act, court judgments) using vector embeddings, enabling citation-backed AI responses and reducing prosecutor review time by 60%
- Created React-based dashboards with chat interface and case management features, implementing role-based access control and encryption to ensure compliance with data protection laws
- Implemented human-in-the-loop feedback system for continuous model improvement, reducing AI inaccuracies by 40% through structured validation workflows and curated training datasets
- Collaborated with prosecutors and investigating officers to define evidence requirements and validation rules, delivering technical presentations and maintaining comprehensive system documentation

Machine Learning Intern Dec 2024 – Feb 2025
GGS Information Services *On-site*

- Engineered production-ready 3D compression algorithms achieving 60% model size reduction while maintaining 95% geometric accuracy, deployed across 15+ enterprise clients with 40% rendering speed improvement
- Built end-to-end AI pipeline for 2D-to-3D conversion using custom GANs, processing 500+ STEP files daily with 85% accuracy and reducing manual CAD modeling time by 70% for engineering teams
- Implemented MLOps infrastructure with real-time monitoring dashboards tracking model drift, performance degradation, and inference latency, achieving 99.5% uptime in production
- Led cross-functional collaboration with engineering and product teams to establish AI evaluation frameworks beyond accuracy metrics, implementing A/B testing for model performance optimization
- Optimized deep learning inference pipelines using CUDA kernels and model quantization, achieving 30% latency reduction and 50% memory footprint decrease for real-time deployment

Publications

PPAG: Progressive Pose Attention Generation for Identity-Preserving Image Synthesis

- Published novel zero-shot identity-preserving image generation model using progressive pose attention mechanisms, achieving state-of-the-art performance in pose transfer with 92% identity preservation score

Projects

GSPO-DeepSeek-R1-Distill-Qwen-1.5B | *PyTorch, Transformers, Wandb, Flash Attention*

- Implemented Group Sequence Policy Optimization (GSPO) algorithm from Qwen Team research, achieving superior stability with 50-75% clipping rates vs 0.01-0.02% for baseline PPO/GRPO methods on reasoning tasks
- Engineered complete knowledge distillation pipeline from DeepSeek-R1 to Qwen-1.5B architecture, incorporating 8-bit optimization and gradient checkpointing for memory-efficient training on H100/RTX hardware

- Developed comprehensive benchmark suite with 60.0% accuracy on ZebraLogic and 75.8% on custom math problems, demonstrating -1.4% performance improvement over PPO (-2.9%) and GRPO (-3.8%) baselines with full experiment reproducibility

Dial 112 AI: Emergency Call Intelligence | *Speech Recognition, NLP, Real-time Processing*

- Developed production AI system for Andhra Pradesh Police processing 1000+ emergency calls daily, implementing speech-to-text, sentiment analysis, and priority classification
- Built real-time geospatial analysis and emergency dispatch optimization, reducing average response time by 25% through intelligent resource allocation

CADify: AI-Powered 3D CAD Generation | *PyTorch3D, OpenAI, OpenCV, Computer Vision*

- Developed multimodal AI system transforming 2D engineering diagrams into 3D CAD models with 92% geometric accuracy using computer vision and LLM integration
- Built robust feature extraction pipeline handling technical drawings, sketches, and annotations with advanced OCR and geometric reasoning capabilities
- Created end-to-end production system processing 100+ diagrams daily, reducing CAD modeling time from hours to minutes for engineering teams

DeepRE: Deep Reinforcement Learning for Self-Verification | *PyTorch, VLLM, Ray, Flash Attention*

- Reproduced DeepSeek R1 Zero achieving 90% functional parity, implementing advanced RL techniques for LLM self-verification on complex reasoning tasks
- Engineered distributed training pipeline using VLLM backend and Flash Attention 2, enabling cost-effective training of 3B parameter models on consumer hardware with 40% cost reduction
- Built comprehensive evaluation framework measuring reasoning accuracy, self-correction capabilities, and computational efficiency across mathematical and logical reasoning benchmarks

PPAG: Pose-Guided Image Generation | *PyTorch, ControlNet, Gradio, Diffusers*

- Engineered production-ready pose transfer system integrating 5 ControlNet models, achieving 90% pose accuracy on COCO-Pose benchmark with real-time inference
- Implemented advanced prompt engineering pipeline with dynamic negative prompting and attention guidance, improving generation quality by 40% and reducing artifacts by 65%
- Optimized inference pipeline using torch.compile and mixed precision training, achieving sub-2 second generation time for 512x512 images

Owl CLI: OS-Level AI Agent | *LLMs, System Integration, Windows APIs*

- Built intelligent CLI agent leveraging Google Gemini LLM for natural language to system command translation, enabling conversational OS interaction with 95% command accuracy
- Engineered autonomous security auditing system with real-time monitoring, policy violation detection, and automated threat response capabilities

Achievements

AI & Innovation Excellence

- Won 1st prize in nationwide AI Agent Hackathon conducted by Andhra Pradesh Police (2025)
- Selected for Microsoft for Startups Founders Hub Program (Jan 2025)

Technical Certifications

- Elite Certification in NPTEL Blockchain Technology - May 2024
- Elite Certification in Introduction to Industry 4.0 and IIoT by NPTEL - Dec 2023

Academic Excellence

- All India Rank 2000 in JEE Mains Paper-2 Architecture (80,000+ candidates) - April 2023
- Secured 85th percentile in JEE Mains General Paper - March 2022

Education

Usha Rama College of Engineering — Telaprolu, AP

Bachelor of Technology in AI and ML; — *Dec 2021 - May 2025*