Susindra Reddy Bandi

Richmond, TX 77407
+1-346-542-2344 | susindrar@gmail.com | linkedin.com/in/susindra-reddy | susindrareddy.com

Summary

AI/ML Engineer specializing in LLM fine-tuning (SFT, DPO), agentic AI systems, and production deployment. Reduced operational costs by 60% and improved inference throughput by 3.5x using vLLM. Built autonomous multi-agent systems and RAG applications serving 20K+ users. Proficient in PyTorch, LangChain, LangGraph, and modern MLOps tools.

Education

University of Houston — Houston, TX
Master of Science in Data Science and Artificial Intelligence; GPA: 3.3/4.0 — *Aug 2025 – May 2027*

Madanapalle Institute of Technology and Science — Andhra Pradesh, India
Bachelor of Technology in Computer Science and Artificial Intelligence; GPA: 8.21/10.0 — *Aug 2020 – May 2024*

Experience

Pravaah Consulting — Bengaluru, India
AI/ML Developer — *May 2024 – June 2025*

- **LLM Fine-Tuning & Evaluation**: **Led fine-tuning initiative** for **Llama 3.1 8B**, guiding 2 interns through supervised fine-tuning (SFT) with 50K custom examples and **Direct Preference Optimization (DPO)**, achieving **22% improvement** in task-specific accuracy and **18% better alignment scores**.
- **Document Automation Pipeline**: Orchestrated agentic workflow automating document processing utilizing fine-tuned **Tesseract OCR**, reducing manual processing time by **75%** and achieving **94% extraction accuracy** across 10+ document types; monitored agent traces via **LangSmith**.
- **Model Serving & Optimization**: Deployed fine-tuned LLMs leveraging **vLLM** with PagedAttention and continuous batching, increasing inference throughput by **3.5x** and serving 10,000+ requests/day with **200ms P95 latency**.
- **Technologies**: PyTorch, Hugging Face Transformers, vLLM, Tesseract OCR, FastAPI, Docker, GCP, Weights & Biases, LangSmith, Pandas, NumPy

Pravaah Consulting — Bengaluru, India
AI/ML Intern — *Aug 2023 – May 2024*

- **RAG Architecture**: Implemented production-ready **RAG architecture** utilizing LangChain and Pinecone vector database, processing 50,000+ documents; reduced customer support query resolution time by **40%** and achieved **89% answer accuracy**.
- **Recommendation Engine**: Architected collaborative filtering recommendation engine for e-commerce platform utilizing matrix factorization and deep learning embeddings, improving user engagement metrics.
- **Technologies**: PyTorch, LangChain, Pinecone, ChromaDB, Scikit-learn, TensorFlow, REST APIs, Git

Entrepreneurial Experience

Nexus Medical App — Remote
Founder & Lead Developer — *May 2025 – July 2025*

- **Intelligent Medical Platform**: Spearheaded ML-powered exam prep platform serving **200+ students** for NEET PG, INI-CET, and FMGE with **4.9/5.0 average rating** and **85% user retention rate**.
- **Voice Agent System**: Formulated real-time voice agent employing **LiveKit** and **RAG** for medical consultation simulations, achieving **95% transcription accuracy** and **sub-500ms** response latency.
- **Medical Content Development**: Curated 200+ clinical cases and adaptive testing modules with personalized performance analytics for medical aspirants.

BMR BUZZ (BMR Education) — Remote
Co-Founder & Developer — *Sep 2022 – June 2025*

- **Autonomous Content Generation**: Established end-to-end agentic workflow generating **400+ SEO-optimized blog posts daily**: autonomously identifies trending topics, scrapes content, and publishes leveraging **GPT-4o**; achieving **20,000+ monthly impressions**.

- **Content Recommendation Engine**: Designed real-time recommendation solution utilizing semantic embeddings and **Pinecone vector search**, serving personalized blog suggestions as users scroll; improved user engagement and content discovery.
- **Cost Optimization**: Migrated content pipeline from LangChain/LangGraph to custom Selenium-based agents, reducing operational costs to **$200/month** while maintaining 400+ daily posts.

Projects

Autonomous Job Application Agent (Current): Architecting multi-agent framework utilizing **LangGraph** to coordinate specialized sub-agents for job discovery and form completion, targeting **90% reduction** in application time; integrating **Model Context Protocol (MCP)** for standardized tool usage and **Supabase** for state management.
Smart Glasses for Visually Impaired: Devised ML-powered wearable utilizing computer vision for object recognition and navigation assistance; awarded **3rd place** at National Hackathon.

Awards & Honors

Innovator of the Year: Pravaah Consulting (2024)
Best Intern Award: Pravaah Consulting (2023)
3rd Rank - AI Researcher Nationwide: Anveshana National Hackathon, India's largest engineering research competition (2022)

Skills

Agentic AI & Orchestration: LangGraph, AutoGen, LangChain, Multi-Agent Systems, Function Calling, Model Context Protocol (MCP)
Core AI, NLP & Deep Learning: PyTorch, TensorFlow, Transformers, NLP, Prompt Engineering, Computer Vision
LLM Fine-Tuning & GenAI: SFT, DPO, LoRA/QLoRA, Quantization (bitsandbytes), vLLM, Hugging Face, Weights & Biases
RAG & MLOps: Pinecone, ChromaDB, Supabase, Docker, GCP, FastAPI, Git, Pandas, NumPy
Languages & OS: Python, SQL, JavaScript, Linux, Bash