SNEHIT VADDI

vaddisnehit@gmail.com | (682)-365-9352 | linkedin.com/snehitvaddi | github.com/snehitvaddi

Publications

- **Can Small Models Reason About Legal Documents? A Comparative Study**, Under Review, [arXiv, 2026]
- **Do Hallucination Neurons Generalize? Evidence from Cross-Domain Transfer in LLMs**, Under Review, 2026.
- **Detecting E. coli Contamination on Leaf Surfaces Using UV-C Fluorescence and Deep Learning** [SPIE, 2025]
- **An Effective Model for Pothole Classification and Admin Alerting System**, [IEEE ICCMST, 2023]

Professional Experience

ModMed
Boca Raton, FL (Remote)
AI Engineer - GenAI Applications & LLM Systems — *Feb 2025 – April 2026*

- **Clinical Ambient AI Scribe:** Shipped clinical AI Scribe serving **15,000+ providers** across 11 specialties (**400K+ daily encounters**), automating 70% of documentation. Built pipeline covering real-time transcription, LLM-powered SOAP generation, and specialty-specific templates.
- **Agentic Document Processing:** Built intelligent document pipeline using OpenAI Agents SDK, OCR, and fine-tuned small vision-language model (Qwen2-VL via HuggingFace) to classify, extract, and route **10M+ clinical pages/month**. Fine-tuned the VLM on domain-specific layouts, achieving superior accuracy over general-purpose models. Replaced $400K/month vendor with $20K/month in-house system (**95% cost reduction**).
- **Text2SQL & Semantic Intelligence Layer:** Built clinical knowledge graph over ModMed's EHR warehouse (200+ tables, 11 specialties) with pgvector embeddings, enabling natural language queries over clinical metrics. Reduced analyst request volume by **60%**.
- **Multi-Agent RAG Architecture:** Architected production RAG with LoRA-finetuned SLMs, hybrid pgvector + BM25 retrieval, document parsing, chunking, vectorization, and cross-encoder reranking, achieving **94% retrieval precision** across 50K+ clinical documents.
- **MEDHALT, Hallucination Detection (Open-Source):** Open-sourced clinical hallucination detection suite (DeBERTa NER + LLM-as-judge validation) achieving **92% accuracy** vs. GPT-4. Built evaluation pipelines with MLflow tracking, golden-set regression testing, and automated quality monitoring.
- **Monitoring & Guardrails:** Developed LangChain + LangGraph + Claude monitoring framework for Scribe quality, cutting incident response to under **5 minutes**. Implemented PHI-safe pipelines with PII/PHI redaction, prompt injection filtering, and audit logging.
- **Databricks & Infrastructure:** Data pipelines on Databricks processing **2M+ records/day** via PySpark and Delta Lake. MLflow for experiment tracking, model registry, and serving. Deployed on Kubernetes (**99.9% uptime**) with CI/CD and A/B experimentation.

GeoSpider AI
USA (Remote)
AI Software Developer Intern — *May 2024 – Jul 2024*

- **Multi-Agent RAG System:** Built LangGraph-based multi-agent RAG system that **autonomously resolved 65%** of customer tickets across a 50K-doc knowledge base, with dynamic routing, LLM-as-judge scoring, and few-shot prompt selection.
- **Intelligent Routing & Evaluation:** Built LLM-as-judge routing layer with dynamic few-shot prompting. Routes to specialized sub-agents and scores quality, improving **helpfulness from 43% to 76%** and relevance by 30%.
- **Hybrid Search & Inference:** Designed FAISS + keyword hybrid search with semantic reranking achieving **92% recall@10**, serving 150+ concurrent users via vLLM with **40% p95 latency reduction**.
- **Agent Infra & Serving:** Implemented Redis-backed agent memory for multi-turn context management. Built FastAPI inference gateway with model fallback logic, request queuing, and structured logging.

University of Florida
Gainesville, FL
Graduate Researcher, AI/ML — *Feb 2023 – Dec 2024*

- **Computer Vision Research (Published):** Developed hybrid YOLOv8-ViT model improving small-object detection by **15%** with Grad-CAM/EigenCAM explainability. Published in **SPIE 2025** and **IEEE 2023**. Presented at both conferences.
- **Interactive Analytics & Multi-Modal:** Built React dashboard with Grad-CAM visualizations replacing static reports

(**adoption: 15% → 85%**). Prototyped CLIP-based multi-modal retrieval between lab images and research reports.

- **ML Pipeline Automation:** Automated model retraining via MLflow + GitHub Actions CI/CD, cutting deployment from **4 hours to 15 minutes**. Built batch inference with configurable PyTorch/TensorFlow backends.

AT&T (via Accenture)
Bangalore, India

Software Data Engineer — *Jun 2021 – Dec 2022*

- **Intelligent Dispatch Optimization:** Engineered BERT + XGBoost intent classifier (88% F1) predicting technician dispatch necessity. Eliminated 12K unnecessary dispatches/year, saving **$2M annually**.
- **Proactive Anomaly Detection:** Built Elasticsearch + Word2Vec anomaly detection system for network telemetry, surfacing incidents and recommending diagnosis paths, **cutting diagnosis time by 40%** and Tier-2 escalations by 25%.
- **Data Engineering at Scale:** Optimized PySpark/Delta Lake pipelines processing **1M+ logs/day** (30% latency reduction). Designed Azure Synapse warehouse with dbt. Built NLP ticket categorization pipeline using fine-tuned BERT (91% accuracy, 500K+ monthly interactions).

Technical Skills

Languages & APIs:	Python, TypeScript, SQL, RESTful APIs, FastAPI, React.js, Streamlit, pandas, scikit-learn
GenAI & LLMOps:	LangChain, LangGraph, DSPy, OpenAI API, Claude/Anthropic API, HuggingFace Transformers, PyTorch, RAG, AI Agents, Multi-Agent Orchestration, LoRA/QLoRA Fine-Tuning, Prompt Engineering, vLLM, Guardrails, Hallucination Detection, LLM-as-Judge Evals
Databricks & Data:	Databricks (Unity Catalog, MLflow, Delta Lake), PySpark, Apache Spark, Snowflake, dbt, Airflow, PostgreSQL/pgvector, FAISS, Pinecone, Qdrant, Elasticsearch, BM25, Cross-Encoder Reranking
Cloud & Infra:	AWS (S3, EC2, Bedrock, Lambda, SageMaker), Azure (Synapse, ML), Docker, Kubernetes, GitHub Actions, CI/CD, Terraform, Datadog, Blue-Green Deployment, Auto-Scaling

Projects

- **MCP BrandForge Agent** (GitHub): Autonomous AI agent using Claude's MCP framework for cross-platform content scheduling (LinkedIn, YouTube, Twitter), visual generation via DALL-E3, and engagement tracking with memory persistence.
- **Resume2Portfolio.com** (Live): AI SaaS with 1,000+ active users. Transforms resumes into portfolio websites in under 60 seconds via GPT-4, Vertex AI, and LangChain multi-agent workflow.

Education

University of Florida
Gainesville, FL

Master of Science in Computer & Information Science — *Jan 2023 – Dec 2024*

GITAM University
Visakhapatnam, India

Bachelor of Technology in Computer Science (GPA: 3.9/4.0) — *May 2017 – Jun 2021*