

AVANIGADDA SAI DIVYA SREE

Hyderabad, India | +91 8074040794 | [Portfolio](#) | [Github](#) | divyastansbts13@gmail.com

SUMMARY

AI Engineer specializing in end-to-end Agentic AI development, multi-agent orchestration, and production deployment of LLM-powered systems. Proven track record building RAG pipelines, function calling workflows, and context-engineered prompting strategies. Experienced deploying fault-tolerant AI agents integrated with enterprise APIs and cloud infrastructure (GCP, AWS, Azure-compatible), with strong Python skills and a forward-deployment mindset for solving real-world business problems.

EDUCATION

B.Tech, Computer Science Engineering (AI-ML) • SREENIDHI INSTITUTE OF SCIENCE AND TECHNOLOGY (SNIST), Hyderabad
GPA: 8.8

TECHNICAL SKILLS

Languages: Python, Java, SQL

Agentic AI & LLM: LangGraph, LangChain, Vertex AI SDK, RAG Pipelines, Function Calling, Structured Outputs, Prompt Engineering, Context Engineering, Embeddings, Vector Search, Hallucination Reduction, Model Evaluation Loops.

AI/ML Frameworks: Transformers, TensorFlow, Keras, CNN/CRNN, Azure OpenAI API compatible LLM Proxies.

Backend & Cloud DevOps: FastAPI, Flask, Django, GCP (Cloud Run, Cloud SQL), AWS, Docker, Kubernetes-compatible deployments, RESTful APIs & Enterprise System Integration, ERP/Database Integration.

Data Engineering: High-Throughput Ingestion Pipelines, Feature Extraction, Data Validation, Batch Processing (100k+ records), Real-time Stream Preprocessing. **Workflows & Tools:** Semantic Search Indexing, Context Caching Optimization, Git, Agile Methodologies, Automated CI/CD Testing, Evaluation & Testing Frameworks.

WORK EXPERIENCE

PEcAn Ecosystem Modeling Project

Research Contributor – LLM-Assisted Scientific Data Extraction

Ongoing Contribution

- Collaborating with cross-functional researchers to architect and evaluate LLM-based workflows for automating ecological metadata and management event extraction from unstructured scientific publications.
- Investigating advanced temporal reasoning, entity extraction, and normalization techniques to transform complex narratives into highly structured representations for downstream ecosystem modeling tools.
- Constructing robust validation benchmark datasets from published ecological studies to evaluate model extraction accuracy, map failure modes, and engineer mitigation strategies for data-curation automation.

Social Summer of Code 2025 — Software Contributor

July 2025 – October 2025

- **Diagnosed and resolved backend architectural bugs** in complex open-source repositories, integrating fixes across cross-service APIs and enhancing system reliability, directly analogous to forward deployment debugging in production environments.
- **Expanded evaluation and testing frameworks** by authoring modular unit tests and automated CI/CD configurations via Pytest, ensuring zero regression across core execution pathways and continuously improving production system quality.
- **Collaborated with distributed global maintainers** via rigorous pull request reviews, architectural design discussions, and iterative system optimization — demonstrating stakeholder communication and cross-functional collaboration skills.

PROJECTS

Hiro.ai — Recruitment Intelligence Platform

[live demo](#)

Python, FastAPI, LangGraph, LangChain, GCP Cloud Run, Cloud SQL, PostgreSQL, Vertex AI, RAG, Function Calling, React

- **Architected end-to-end agentic AI development** on GCP Cloud Run — designed multi-step workflow agents with tool usage, memory, and stateful orchestration, reducing candidate screening latency from 4 hours to 12 minutes (~95% efficiency gain).
- **Engineered LangGraph DAG-based multi-agent orchestration** with structured function calling and strict execution boundaries at $O(K \times N)$, enabling complex iterative reflection loops while eliminating infinite token expenditure.
- **Developed RAG pipelines and context management strategies:** integrated native Vertex AI Context Caching via the google-genai SDK, slashing repetitive token ingestion overhead and loop TTFT by ~80% across sequential agent turns.
- **Implemented structured outputs via function calling** by binding Google Gemini endpoints to concrete response schemas, achieving 100% deterministic output compliance and drastically reducing hallucinations.
- **Integrated with enterprise databases and APIs:** externalized session persistence via transaction-isolated PostgresSaver checkpoints on Cloud SQL PostgreSQL, solving container ephemerality in stateless cloud runtimes.

Infant Cry Classification System — Audio Diagnosis Platform

[link](#)

TensorFlow, PyTorch, CRNN, GCP Cloud Run, Streamlit

- **Engineered and deployed an end-to-end audio classification AI system** on GCP Cloud Run achieving 94% diagnostic accuracy on real-time acoustic streams, demonstrating production AI deployment experience.
- **Designed scalable preprocessing and feature extraction pipelines** using Librosa to automate high-throughput acoustic feature extraction (MFCCs, Mel-spectrograms) from unstructured input data.
- **Built evaluation and reliability frameworks** — introduced a confidence-scoring layer to optimize model serving reliability and reduce false-positive inferences, improving response quality and consistency.

Real-Time High-Throughput Streaming & Data Processing Pipeline

[link](#)

Python, FastAPI, PostgreSQL, TCP Sockets, In-Memory Caching, Apache Beam / Dataflow Concepts, Locust Distributed Load Testing Framework

- **Large-Scale System Design:** Engineered and deployed a cloud-native, decoupled real-time streaming data pipeline using Python and an asynchronous FastAPI gateway to ingest high-concurrency client telemetry data, simulating massive live-user spikes (1,000+ RPS).
- **Data Processing & Optimization:** Mitigated critical database write-storms and transaction connection starvation by designing a stateful compute worker implementing 10-second fixed tumbling windows.
- **Distributed Computing & Infrastructure:** Architected a thread-safe, volatile TCP network socket queue to fully decouple the stateless ingestion layer from the storage tier, establishing horizontal scalability ready for Google Kubernetes Engine (GKE) and Cloud Pub/Sub scaling.
- **System Testing & Debugging:** Conducted rigorous distributed infrastructure stress testing using the Locust framework to triage performance bottlenecks, analyze backpressure under heavy network loads, and validate pipeline fault tolerance.

ACHIEVEMENTS

- Solved 300+ LeetCode problems (55 hard-level), demonstrating strong algorithmic problem-solving and distributed systems thinking [\[link\]](#)
- Open-source contributor focused on reliability and testing improvements

Published Research: Co-authored reviewed papers: *Building a CV Model with TensorFlow* and *AI-Based Cry Detection and Reason Classification*. [\[Paper\]](#) [\[Paper\]](#)