Dasari Sandhya Rani

AI Engineer: LLM Agents, RAG, Evals
Autonomy AI

sandydasari977@gmail.com +91 8074816864
Portfolio | LinkedIn | GitHub
Indian Institute of Technology, Bombay

ACCOLADES & ACHIEVEMENTS

- Winner of the **ANZ GENAI Hackfest** and awarded **2k Australian dollars** in prize money. ['23]
- Awarded **100%** Scholarship in **NMMS** Scholarship Exam. ['21]
- Secured State **1st Rank** in Intermediate Board Examination and honored with **30k** cash prize. ['16]
- **Centum** holder in all subjects in State Board Examination 2015. ['15]

PROFESSIONAL EXPERIENCE

Autonomy AI | ***AI Researcher*** [Sep'26-present]

Building the agents behind AutonomyAI Studio, a software factory for product teams, usable inside Claude Code, Cursor and MCP.

- Built the **multi-agent setup** behind Autonomy AI Studio's **Autonomous Product Delivery**: agents turn a PM's ticket, PRD, screenshot or Figma frame into a **review-ready pull request** on the customer's production codebase.
- Agents **implement** the change on the customer's **real codebase**, then **build, render and validate** it in the running app with a **preview link** before opening the PR, so every change is verified, not just generated.
- Work on **context management** and **context engineering**, keeping agents accurate on large, unfamiliar codebases and across **long-running, multi-step workflows** with many tool calls and handoffs between agents.
- Improve agent **reliability** and **scalability** in production: agents ship PRs for **170+ product teams**, with a **human in the loop**: engineers review and approve every merge into their own codebase.
- Work on **evals** for the agent system, measuring output quality and catching regressions as agents, prompts, tools and underlying LLMs change between releases, so improvements are measured.
- **Customer-facing:** work directly with customer product and engineering teams on **deployments**, turning their feedback and real failure cases into concrete **agent improvements** and product priorities.

Leoforce | ***AI Research Engineer*** [Oct'25-Sep'26]

Shipped hybrid search RAG agentic system end-to-end from ideation to production deployment at billion-record scale.

- Architected a production **hybrid search engine** combining **BM25** lexical and **HNSW dense vector** retrieval with **RRF** fusion over **1.6B+ candidate records**, improving matching precision by **40%** at scale.
- Built production **multi-agent systems** on **LangGraph** + **LiteLLM**: **tool-calling agents**, **supervisor patterns**, conditional routing; **Langfuse** prompt tracking and **OpenTelemetry** distributed tracing across microservices.
- Designed comprehensive **RAG evaluation pipeline** using **DeepEval** with metrics including **faithfulness**, **answer relevancy**, **contextual precision/recall**, and **RAGAS scores** to iterate on prompts, retrieval and chunking.
- Fine-tuned **embedding models** with **LoRA/QLoRA** (**65% less memory**) and served them on **vLLM** with **continuous batching** and **tensor parallelism** for **3x GPU throughput** on production inference workloads.
- Built an **NLP pipeline** for resume parsing (**NER**, skill extraction, normalization); tested **semantic chunking** and **dimension reduction** for embedding optimization across the billion-record candidate index.
- Led the **OpenSearch 1.3** $\rightarrow$ **3.0** migration with **FAISS**/**Lucene k-NN**, tuning index mappings, query DSL and cost; built async **ETL pipelines** for embedding generation with **batch processing** and **parallel workers**.
- Built async **FastAPI microservices** (connection pooling, request batching, streaming) and the **observability stack** with custom metrics and alerting for **ML inference endpoints** handling **10k+ QPS**.

ANZ Operations and Technology | ***AI Engineer*** [Jun'23-Sep'25]

Acknowledged as a high-achieving team member from a team of 40; received the highest appraisal among graduates in my cohort.

- **Architected** and **deployed KnowHow Assist**, an enterprise **RAG** system over **60k+ documents** on **Vertex AI**: ingestion, **multi-modal embeddings**, **vector search** at **200ms** query latency across diverse enterprise content.
- Developed a **full-stack RAG solution**: **React.js/TypeScript** front end, **FastAPI** back end and a **real-time streaming chat interface**, with Vertex AI **foundation models** for **document analysis** and **semantic search**.
- **Spearheaded** a team to develop an **automation solution** that significantly reduced business workload by **10x**.
- Established and managed **CI/CD pipelines** with **Codefresh**, **GitLab CI**, and **Azure DevOps**, while orchestrating **Docker** and **Kubernetes** for scalable, efficient deployment of ML and RAG services to production.
- Applied **deep learning** techniques for **document analysis**, achieving **95% accuracy** in automated **data extraction**.

THOUGHTWORKS.PVT.LTD | ***ML & Big Data Analytics Internship*** [Dec'21-May'22]

Awarded full-time internship (1 out of 5 IITB students) acknowledging exemplary performance and sincerity in project work

- Extracted **medical data** from documents and converted it through **OCR** using the **docTR** model for healthcare data processing; recorded observations and indicators against the patient's medical history with **spaCy** and **medCAT**.
- Developed a **Label Studio API** with **docTR** to extract text from **multiple medical data reports**.
- Implemented **semi-supervised learning**, beating **supervised** baselines by **3-5%** accuracy on small labeled data.

BAHMNI Open Source EMR, in collab. with KCDH | *Data Engineer, Prof. Ganesh Ramakrishnan* [Jun'22-Nov'22]

- Enhanced **LayoutLM** deep learning model with **FUNSD dataset** achieving **80% accuracy** in document understanding; developed tool for **key entity recognition** from documents using **SimpleDLM**, improving **F1 by 14.4%**.
- Designed and maintained **data pipelines** supporting data transformation processes, increasing efficiency by **25%**; implemented **data quality checks** and monitoring systems resulting in **15%** reduction in data errors.

PUBLICATIONS

- A. Mohammed, S. R. Dasari, Y. M. Desai, "SCF Prediction using the Finite Element Method Coupled with Sobol Sampling and Bayesian Optimization", on Soft Computing, Machine Learning and Optimisation, Civil-Comp, UK
- A. Mohammed, S. R. Dasari, Y. M. Desai, "Mesh Sensitivity Study of Steel Tubular T-joints for the Computation of Stress Concentration Factors", on AIJR Proceedings

KEY PROJECTS

OpenAcme | *Open-source AI Agent Platform, Creator* [Jun'26]

- Built an **AI workforce platform** in **TypeScript**: role-specialized agents, **multi-provider LLMs**, **MCP**, web + CLI.
- CLI published on npm as **@openacme/cli**; **~90 GitHub stars** and **3,000+ npm downloads** since launch.

Most popular AI agent repos are weekend projects | *Blog post* [Jun'26]

- Reviewed the source of popular agent frameworks (LangChain, LangGraph, CrewAI) and argued production needs what they skip: **context-overflow handling**, **stream recovery**, **OAuth refresh**, **skill supply-chain security**.

Neural Network based approach for SCFs, MTech Thesis | *DL Project, Prof. Yogesh M Desai* [Jun'22-Jun'23]

- Implemented an efficient algorithm in **ANSYS-APDL** with **numpy** and **pandas**, cutting time complexity by **90%**.
- Trained **neural networks** with **Bayesian Optimization** for material-property and geometric-parameter analysis.
- Used **Sobol sampling** and neural networks for structural engineering applications.

Automated 3D Reconstruction from Satellite Images | *Image Processing Course Project* [May'21-Jun'21]

- Developed a **stereo pipeline** including **pushbroom** sensor modeling, **geographic** coordinate systems projection and **localization functions** with error of **0.096 pixels** for **Epipolar Rectification** and **Stereo Matching**.
- Implemented **point cloud** by triangulating correspondences and projecting them on **UTM** reference system.
- Explored clustering algorithms like **K-Means** and **DBSCAN** to convert LiDAR points into optimal building clusters.

YPredict - Real-time Analytics Pipeline | *Big Data Solutions, Self Project* [Nov'21-Dec'21]

- Built a real-time pipeline on **Confluent Kafka** streaming **50,000+** daily transactions from a **Flask** site, **99.9% uptime**.
- Implemented **parallel consumer** processes achieving sub-**200ms** processing speeds for efficient **MySQL** storage.
- Built a **Power BI** dashboard tracking **10+ metrics**, with the **MySQL** database hosted on **aiven.io**.

EDUCATIONAL QUALIFICATIONS

Post Graduation: M.Tech, IIT BOMBAY, CPI: **9.49** [2021-2023]
Graduation: B.Tech, IIT TIRUPATI, CPI: **8.62** [2017-2021]

TECHNICAL SKILLS

- **AI Engineering:** Multi-agent Systems, Agent Harness & Context Engineering, Tool Calling, LLM Evals (DeepEval, RAGAS), LLM Observability (Langfuse, OpenTelemetry), RAG & Hybrid Search (BM25 + HNSW, RRF), MCP, LoRA/QLoRA Fine-tuning, vLLM Serving, Prompt Engineering, Embeddings, Agent Orchestration
- **GenAI & LLM Tools:** LangGraph, LangChain, LiteLLM, OpenAI API, Google Gemini, Google Cloud Vertex AI, Sentence Transformers, Semantic Chunking, DeepEval, RAGAS, Langfuse, Claude Code, MCP
- **ML & Deep Learning:** PyTorch, TensorFlow, scikit-learn, NLP, NER, OCR, spaCy, medCAT, docTR, LayoutLM, Bayesian Optimization, Semi-supervised Learning, Document AI, Computer Vision
- **Analytics:** Machine Learning, Deep Learning, RAG Systems, A/B Testing, Data Modeling, Power BI
- **Languages:** Python, JavaScript, TypeScript, C, C++, R, SQL, HTML, LaTeX
- **Technologies and Frameworks:** React.js, Node.js, FastAPI, RESTful APIs, Kafka, Hadoop, MongoDB, SQLite, PostgreSQL, GCP, OpenSearch, ElasticSearch, FAISS, HNSW, BM25, k-NN Vector Search
- **DevOps:** Docker, Kubernetes, Jenkins, Codefresh, GitLab CI, Azure DevOps, CI/CD Pipelines, Microservices