SAI SAMBHU PRASAD KALAGA

+ 1 737-420-4595 | linkedin.com/in/sai-sambhu-prasad-kalaga | saisambhuprasadkalaga@gmail.com | United States

EDUCATION

- **Master of Science** in Computer Science, Southern Methodist University, Dallas, Texas, USA ***GPA: 3.9/4.0 | August 2023 - May 2025***
- **Bachelor of Technology** in Computer Science, Anurag Group of Institutions, India ***GPA: 3.6/4.0 | August 2019 - May 2023***

EXPERIENCE

AI Engineer Intern, iLink Digital, Remote, USA ***Oct 2025 – Present***

- Built an intelligence platform with multiple specialized agents orchestrated via LangGraph supervisor pattern, handling complex queries through AI-driven routing and response synthesis through OpenAI and Gemini. Implemented dual-pipeline system combining structured CSV queries with FAISS-based vector search over unstructured PDF documents, ensuring zero-hallucination responses through strict tool-based data retrieval.
- Conducted comprehensive performance analysis across Llama4 vs Qwen3, and Llama 3.2 (3B) vs Qwen 2.5 (7B) models, measuring TPS (tokens-per-second), GPU VRAM utilization (4/8/16-bit quantization), and multi-query batch processing latency. Architected scalable FastAPI backend with Streamlit frontend (STT/TTS), Docker containerization, and OpenAI Evals integration for automated response quality assessment.

Data Science Researcher- AI and ML, Southern Methodist University, Dallas, TX, USA ***March 2024 – November 2025***

- Derived 50+ features from VR surgical training data; trained a Random Forest reaching 95% validation accuracy; deployed with Streamlit; benchmarked XGBoost and LSTM; shipped instructor dashboards that shortened review time by 50%; analysis revealed that students who reflected about every 2 days outperformed peers, an insight the professor adopted across classes to boost engagement and outcomes.
- Automated storage and analysis of a large municipal Police Department VR training data (Python, cloud storage) with schema checks, metadata versioning, and scheduled ETL; decreased manual handling by 70% and cut time to insight from 2 days to 2 hours.

Full Stack, NLP and ML Engineering Intern, Blue Clay Health, Remote, USA ***August 2025 – October 2025***

- Designed and deployed enterprise-grade RAG and semantic-search pipelines using Python, FastAPI, FAISS, and Azure Blob Storage, optimizing ingestion and retrieval workflows to cut high-percentile latency by 35% and boost clinical insight accuracy by 28%.
- Scaled microservice architecture on GCP (Cloud Run, Artifact Registry, Cloud Build) and led Azure-to-GCP migration, enabling cross-team access, faster delivery, and full-stack integration with LLM-driven workflows, automated draft generation, and intelligent summarization.

Data Science Research Intern, Indian School of Business (ISB), Hyderabad, India ***January 2022 – July 2022***

- Automated table extraction (pandas, docx2python) from corporate reports (Thousands of Fortune 500 companies' annual reports) into structured datasets; built Streamlit ML/NLP demos; presented to 50+ business students, improving comprehension 20%.

Software and Data Engineer, Founder, TextHappen (a digital marketing startup), Hyderabad, India ***January 2022 – May 2023***

- Owned full stack feature delivery for a marketing platform, collaborating with frontend engineers and analysts; shipped HTML, CSS, JavaScript, and Python services that enhanced UX. Modelled user behavior with Python, ML and scikit-learn to guide campaigns, lifting click-through rate by 25% and enabling data-driven decisions.

Google Developer Student Club Lead, Google, Hyderabad, India ***August 2021 – July 2023***

- Spearheaded a 1,200+ member community; organized workshops and hands-on sessions across Web, Data, Android, Google Cloud, AI/ML, Flutter, and Dialogflow CX.

PROJECTS

CUDA Accelerated Explainable AI for Heart and Chest X-ray Diagnostics ***Summer 2025***

- Implemented a CUDA accelerated system for dual diagnostics using CNNs on chest X-rays and MLP on tabular indicators, reaching 98.7% validation accuracy. Integrated SHAP and Gemini to generate natural-language rationale in a Streamlit interface for Users.

Detecting Out-of-Tune Instruments with Audio Deep Learning (VGGish) ***Spring 2025***

- Trained CNN and MLP models with VGGish embeddings and mel-spectrograms in TensorFlow, delivering 81% test accuracy on a custom NSynth dataset. Synthesized 300K+ augmented audio clips with Librosa.

SMUBot - a RAG-based Chatbot for International Student Assistance ***Fall 2024***

- Developed a RAG chatbot using Gemini LLM and FAISS semantic search, shortening response time by 70% and attaining 98% answer accuracy on FAQs. Optimized PDF ingestion and indexing with Python and Transformers, reducing retrieval latency by 80%.

SKILLS

- **Programming Languages:** Python, C++, R, C, SQL, Java
- **Frameworks/Technologies:** HTML, CSS, JavaScript, Spring Boot
- **Databases:** MySQL, Oracle, Google BigQuery
- **Mathematics**: Linear Algebra, Calculus, Statistics, Probability
- **Soft Skills**: Initiative, Leadership, Adaptability, Critical Thinking
- **Product Management:** Product Metrics, KPIs, A/B Testing, Causal Inference, Empathy, Hypothesis Testing, Bias Detection
- **Application Environments:** Power BI, Tableau, Excel, Looker, Anaconda, Flask, Airflow, Databricks, Hadoop, FastAPI
- **Software and Cloud:** AWS, GCP, SDLC, Agile, Git, GitHub, CI/CD, SageMaker, Snowflake, Docker, Spark
- **AI and ML Toolkit:** Scikit-Learn, Keras, NumPy, Pandas, Matplotlib, StreamLit, TensorFlow, FAISS, PyTorch, MLP, CNN, NLP, chatbot, LLM, computer vision, HuggingFace, BERT, Tuning

ACHIEVEMENTS

- Published ML-based object detection research in a UGC-approved journal (2023).
- Secured SMU Lyle Graduate Scholarship.
- Awarded 1st place in IBM's AI/ML Hackathon among 50+ teams.