## Sai Kiran Reddy Soma

+1 (682) 498-3904 ◊ [somasaikiranreddy98@gmail.com](mailto:somasaikiranreddy98@gmail.com)

# OBJECTIVE

Experienced ML/ AI Engineering, with a strong background in systems designing, automation, and deployment. Hands-on expertise in building intelligent systems using LangGraph, Python, and advanced AI / ML architectures like multi-agent systems and RAG pipelines. Holding Masters of Science in Computer Science focused on ML. Expert in developing real-time AI systems for decision making, personalization, and multi-modal applications. Skilled in developing POC through research papers. Thrives in fast-paced, research-driven environments.

# EDUCATION

**Master’s in Computer Science, University of North Texas, Denton, Texas *August 2022 – May 2024***

* **Focus**: Fundamentals of AI, Machine Learning, Data Mining, Information Retrieval, software development for AI, Fundamentals of DB, Big data & data science.

# EXPERIENCE

**AI Engineer, Saiana Technologies Inc, NJ *August 2024 - present***

* + Engineered multi-agent systems using LangGraph to perform complex task delegation, personalized workflows, and dynamic user adaptation.
  + Designed end-to-end RAG pipelines using LangChain and vector databases (Pinecone, Weaviate) for personalized document and memory retrieval.
  + Fine-tuned open-source LLMs (LLaMA 3, Mistral) and integrated Hugging Face Transformers into real-time decision-making systems.
  + Developed semantic memory systems using Redis + vector DBs to store user history and context for behavior modeling and strategic feedback loops.
  + Built scalable APIs with FastAPI, served via Docker and deployed on Kubernetes clusters for production inference and data workflows.
  + Automated training workflows and model deployment using GitHub Actions and Terraform for reproducible infrastructure and version-controlled pipelines.
  + Conducted performance benchmarking across OpenAI, Claude, and LLaMA APIs, balancing cost, latency, and accuracy for different use cases.
  + Collaborated with cross-disciplinary teams (product, backend, data engineering) to embed AI agents into existing infrastructure and UX flows.
  + Integrated observability tools to monitor inference latency, token usage, and retrieval quality using custom metrics and alerting pipelines.

**Machine Learning Engineer, Smart Squads Digital, Hyderabad, India *June 2021– June 2022***

* + Built and deployed production ML models for real-time customer behavior prediction, improving churn detection by 17% and upsell success rate by 22%.
  + Developed customer segmentation pipelines using clustering techniques (KMeans, DBSCAN), integrated into marketing and recommendation platforms.
  + Automated batch feature engineering pipelines using Bash + Python, deployed in Dockerized containers with AWS S3 data feeds.
  + Maintained model versioning and rollback strategies with CI/CD workflows using GitHub Actions and custom shell scripts.
  + Designed FastAPI microservices to serve trained models at scale, with Redis used for request caching and user session memory.
  + Built time-series forecasting models for inventory demand using Prophet and XGBoost, enabling smarter supply chain decisions.
  + Monitored production model health using Prometheus + Grafana, tracking real-time accuracy decay and retraining thresholds.
  + Wrote robust data ingestion and validation scripts to ensure input quality, using custom exception handling and schema validation routines.
  + Conducted extensive A/B tests to evaluate the impact of ML models on business KPIs, documenting results and iteration plans.

**Machine Learning Engineer intern, Smart Squads Digital, Hyderabad, India *Jan 2021 – May 2021***

* + Built a sentiment classification engine using NLP pipelines in spaCy and NLTK, applied to customer reviews and support transcripts.
  + Engineered features and trained models using scikit-learn (SVM, Logistic Regression), achieving over 85% precision on test sets.
  + Designed data visualization dashboards using Tableau to communicate insights from customer sentiment trends.
  + Collaborated with senior engineers via Git workflows, maintaining reproducibility and continuous integration in model development.
  + Presented actionable insights to product managers, helping prioritize user feedback features for next release.- Mentored junior interns and led knowledge transfer sessions for new analytics tools and practices.
  + Analyzed e-commerce funnel metrics and presented insights to product managers to drive UI/UX improvements.
  + Contributed to AB testing design and result interpretation for product feature rollouts.

# RESEARCH PAPER IMPLEMENTATIONS

## Attention is all you need

* + Implemented mini GPT model from scratch using Pytorch, Numpy.
  + Implemented this to learn ins and outs of Large Language Models and how they work.

## In-Context-Learning over Fine Tuning

* + Read the paper, implemented the concept presented in the paper.
  + ICL outperformed FT on tasks involving implicit patterns (e.g., detecting redundant code, simplifying boolean expressions).
  + Across four reasoning tasks (Expression, Code Reading, Boolean, Relation), ICL often beat FT by 2–30+% depending on the model and task.
  + In misleading data tests, ICL stayed relatively close to clean performance, while FT models dropped sharply.
  + ICL achieved high performance with far fewer examples (e.g., 32-shot prompts) compared to thousands of FT training samples

# SKILLS

## AI/ML & LLMs:

LangGraph · LangChain · Retrieval-Augmented Generation (RAG) · Prompt Engineering · Hugging Face Transformers · Fine-tuning (LLaMA 3, Mistral) · Embeddings · OpenAI API · In-Context Learning · NLP (spaCy, NLTK) · scikit-learn

## Multi-Agent & Personalization Systems:

Multi-Agent Architectures · Agent-Based Task Delegation · Strategic Reasoning · Context-Aware Workflows · User Behavior Modeling Semantic Memory Systems

## DevOps & Infrastructure:

Docker · Kubernetes · GitHub Actions · Terraform · FastAPI · Redis · CI/CD Pipelines · AWS (EC2, S3) · Linux Scripting

## Programming & Tools:

Python · Bash · SQL · PyTorch · NumPy · Vector Databases (Pinecone, Weaviate, FAISS) · Postgres · MySQL · Tableau

# CERTIFICATIONS

* + Neural Networks and Deep Learning (by Andrew Ng)
  + TensorFlow for AI (by Mo Rebaie)