Sai Kogilathota

+1(934)799-0867 | saiakhilkogilathota@gmail.com | Stony Brook, New York, United States | LinkedIn

SUMMARY

Software Engineer with 3+ years of experience integrating ML solutions into production systems across fintech and telecom. Skilled in NLP, predictive modeling, and API development using Python, Flask, and cloud platforms like AWS and Azure. Proven track record of building scalable pipelines and delivering data-driven features that improve decision-making and user engagement.

WORK EXPERIENCE

ML Researcher, Stony Brook University, Stony Brook, United States **Aug 2024 — Present**

- **Multimodal Research:** Introducing a lightweight probing framework that predicts hallucinations in **VLM** image captioning by analyzing hidden-states pre-generation; achieved strong results on **4 VLMs**, culminating in a **first-authored EMNLP 2025 submission.**
- **Clinical RAG Assistant:** Built and deployed a **clinical multimodal RAG system** using **LangChain, Pinecone, and AWS (EC2, S3)** to extract patient data and augment diagnostic processes, enabling faster and more accurate clinical decision support.
- **Self-Finetuning LLMs:** Developing a novel framework for self-finetuning LLMs using Agent-to-Agent (A2A) protocols orchestrated with **LangGraph** in controlled environments, enabling adaptive model improvement without manual intervention.

Software Engineer - AI/ML, Lendingkart Finance Limited, Bengaluru, India **Jan 2023 — Aug 2023**

- **LLM Fine-tuning with BERT:** Fine-tuned a BERT-Base model on internal search data using **Hugging Face & PyTorch**, improving document retrieval accuracy by **70%**. Deployed the model on AWS SageMaker for scalable, real-time search capabilities.
- **NLP Analytics for Sales:** Increased sales conversion by **15%** by building and deploying an end-to-end NLP pipeline using **Google STT API** and sentiment models to analyze sales conversations.
- **KYC Automation:** Developed and deployed an **OCR-based and face-matching system** to accelerate onboarding, reducing onboarding time by **80%**; leveraged **PyTorch**, **YOLOv8** for real-time face detection, with end-to-end hosting and scaling on **AWS**.
- **Loan Eligibility Automation:** Automated loan eligibility workflows by embedding predictive analytics using **Scikit-learn and XGBoost** models, reducing manual processing time by **20hrs/week**, deploying this on **AWS EC2**.
- **Engagement Optimization:** Built & deployed a personalized discount engine for **10,000+ users**, leveraging **Kafka** and **AWS SNS**, increasing loan completion rates by **35%**; conducted **A/B testing** on user behavior to further optimize engagement and conversions.

Software Engineer - AI/ML, Comviva Technologies, Bengaluru, India **Aug 2020 — Dec 2022**

- **ML Tool:** Developed and deployed a **Django based no-code ML tool** on **AWS EC2**, enabling business teams to build models there by improving operational efficiency; implemented **CI/CD pipelines** for seamless updates and faster feature releases.
- **Recommendation System:** Implemented **collaborative filtering** algorithms to personalize mobile campaign offers, increasing ARPU (Average Revenue Per User) by **10%**.
- **ML Model Deployment:** Trained and tuned **XGBoost, Ensemble models** and deployed them on **AWS SageMaker** for churn and multi sim user prediction, enabling proactive retention strategies that **reduced churn by 18%** and increased revenue by $150K+/ quarter.
- **Real-Time Data Processing Pipelines:** Built and deployed real-time **ETL workflows** on **AWS** in **Python** using **Pandas** for ingesting telecom usage data into ML pipelines, **reducing processing lag by 40%** and supporting predictive analytics.
- **A/B Testing Framework:** Collaborated with marketing and product teams to develop an experimentation framework to **A/B test ML-driven campaign** features, improving **conversion rates by 17%**.
- **Sentiment Analysis:** Applied sentiment analysis on multilingual customer feedback using **BERT Models** and **Google Translate API,** surfacing key pain points and enabling UX/UI optimizations across 3 product lines.

PROJECTS

Revize | React.js, Django, SQL, AWS | EMNLP 2025 Submission, Link

- Developed a smart web application for **spaced repetition learning**, enabling users to add topics, auto-schedule revision dates, track daily tasks, and review progress with flashcard-based content.
- Integrated **OpenAI SDK** for automated summarization into flashcards, and deployed the application on **AWS EC2** with a **Postgres** database for scalable, reliable performance.

Stock Recommendation System | NLP, LSTM, Pytorch, Galformer, Link

- Developed a novel **PyTorch-based architecture** leveraging **Galformer**, a specialized transformer for stock systems and NLP, to analyze stock sentiment and generate buy/sell recommendations based on purchase price and quantity.
- Deployed the model with a **Streamlit** front end, achieving **15% better prediction accuracy** compared to a baseline LSTM model.

TrialMatch | JavaScript, Python, React, MySQL, KMeans, Link

- A clinical trial recruitment platform with a swipe-based matching interface and integrated **recommendation system**, improving match accuracy for **rare disease trials**, **reducing** participant search time **by ~40%**, and **accelerating** trial enrollment by **up to 30%**.

EDUCATION

Stony Brook University, **Stony Brook, United States** - *Master's, Data Science* **Aug 2023 — May 2025**

Ramaiah Institute of Technology, **Bengaluru, India** - *Bachelor's, Electronics and Communication* **Aug 2016 — Jul 2020**

SKILLS

Programming Languages	:	Python, Java, R, MySQL
ML Frameworks	:	Scikit-learn, XGBoost, TensorFlow (basic), Rasa, spaCy, NLTK
Model Development	:	Supervised & Unsupervised Learning, Cross-Validation, Hyperparameter Tuning, A/B Testing
Data Processing & Pipelines	:	ETL Pipelines, Data Wrangling, Feature Engineering, Pandas, NumPy
APIs & Deployment	:	RESTful API Development, Flask, FastAPI, Docker, CI/CD
Cloud Platforms	:	AWS (SageMaker, EC2, S3), Azure
Version Control & Dev Tools	:	Git, JIRA, VS Code, Postman