Saketh Vadlamudi

Hyderabad | 9704808576 | sakethvadlamudi22@gmail.com | LinkedIn | GitHub | | LeetCode | Website

Summary

AI/ML Engineer with ~4 years of experience building production-grade GenAI, RAG, and ML systems at scale. Designed multi-agent reasoning engines, hybrid RAG search systems, deployed cloud-native ML pipelines. Delivered measurable impact: 60%+ latency cuts, 2× performance gains, and 65+ engineer hours saved monthly. Focused on designing fault-tolerant, scalable systems driving product outcomes.

Technical Skills

- **Programming languages:** Python, SQL
- **AI/ML:** Machine Learning (ML), Natural Language Processing (NLP), Retrieval Augmented Generation (RAG), Transformers, Multi Agent Architectures, LLMs
- **Frameworks:** LangGraph, LangChain, FastAPI, PyTorch
- **Cloud & Infra:** AWS (lambda, DynamoDB, Redshift, SageMaker, S3, EC2, ECR, CloudWatch), GCP (BigQuery, Vertex AI)
- **Distributed Systems:** Redis, OpenSearch, Airflow, Docker, GitHub Actions, Unix
- **Other:** Git, Pydantic, CI/CD, Django

Professional Experience

Tiger Analytics - AIML Senior Associate ***May 2025 – Present (Hyderabad)***

- Architected a **distributed multi-agent reasoning engine with horizontal scaling**, leveraging LangGraph to orchestrate parallel, dependency-aware execution graphs across multiple model agents. **Reducing** end-to-end reporting **latency by over 60%.**
- Designed the engine as a **self-correcting, schema-aware, fault-tolerant** pipeline handling **~2000 daily queries** producing **natural** language **insights**, **structured** tabular **data**, and **interactive Chart**.js configurations.
- Owned the full production lifecycle - architecture, implementation, deployment and monitoring.
- **Tech Stack** : Python, LangGraph, Pydantic, AWS (Lambda, DynamoDB, RedShift, Bedrock, CloudWatch), FastAPI, Docker

ValueLabs - Data Scientist 2 (Senior Software Engineer) ***Mar 2022 – May 2025 (Hyderabad)***

- Developed a Machine Learning (ML) model for **Sentiment Analysis**, which accurately classifies customer reviews on products. Achieved 93% accuracy; insights were adopted by product stakeholders to refine customer messaging.
- Architected an **Advertisement Recommendation System** using collaborative filtering, achieving **MAE of 0.79** and improving ad relevance, leading to a 20% increase in click-through rates.
- Deployed ML models using **AWS** services such as EC2, ECR and SageMaker, leveraging Docker for seamless containerization and scalability, while using reliability through robust cloud Infrastructure.
- Spearheaded **Python processes on Factory Design Mechanism**, reduced run time performance by **2x and 12%** decrease in Campaign drop off rate.
- Built **Python automation pipelines** which eliminated repetitive processing tasks, saving **65+ engineer hours monthly** and improving operational SLAs.
- Shipped **4** production-grade **web applications using Django,** focused on scalability, security, reliability.
- **Tech Stack** : Python, ML, NLP, Django, Unix, AWS (EC2, ECR, Sagemaker, S3), BigQuery, Docker, Jira

Additional Projects

Generative AI :- HybridSeek (Link) – Semantic RAG Search Engine

- Built a **hybrid retrieval RAG pipeline** (BM25 + dense vector + RRF) for searching over 10k+ document chunks.
- Designed a section-aware chunking + 1024-dim embedding pipeline, achieving >90% **semantic recall**@10 during evaluation.
- Reduced repeated-query latency by ~**8**× using Redis caching with exact-match hashing.
- Automated ingestion and re-indexing of **1000+** new documents/day using Airflow DAGs and batching.
- Integrated streaming LLM inference with Ollama, delivering **sub-3s time-to-first-token** latency.
- **Tech Stack**: Python, FastAPI, OpenSearch, Redis, Ollama, Jina AI, Airflow, Docker, Langfuse(Observability), Gradio.

Natural Language Processing (NLP):- Language Modelling (Link)

- Engineered a state-of-the-art language model using pre-trained advanced transformers like DistilBert, Bert2, fine-tuned on custom dataset for accurate prediction of top 10 masked word and seamless contextually rich text generation.
- **Tech Stack**: Python, PyTorch, Hugging Face Transformers, Pandas, scikit-learn.

Education

Executive Post Graduation, Data Science and AI , **IIT Roorkee** **2024 - 2025**

Bachelor of Technology, Computer Science, Narasaraopeta Engineering College (CGPA 8.94) **2018 – 2022**

Awards & Certifications

- ValueLabs CEO Choice Award Winner 2024 (Link)
- Google Cloud Certified Professional Machine Learning Engineer (Link)
- Google Cloud Certified Generative AI Leader (Link)
- AWS Certified Cloud Practitioner (Link)