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, and deployed cloud-native ML pipelines. Delivered measurable impact: 60%+ latency cuts, 2× performance gains, and 65+ engineer hours saved monthly. Focused on designing robust, scalable AI systems driving product outcomes.

Technical Skills

- **Programming languages:** Python, SQL
- **AI/ML:** Generative AI, Machine Learning (ML), Natural Language Processing (NLP), Retrieval Augmented Generation (RAG), Transformers, Agentic AI, LangGraph, LangChain, Pydantic
- **Cloud Technologies:** AWS (lambda, DynamoDB, Redshift, SageMaker, S3, EC2, ECR), GCP (BigQuery, Vertex AI)
- **Tools:** Git, GitHub Actions, Docker, Fast API, Unix, Django, Redis

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**

Professional Experience

Tiger Analytics   **May 2025 – Present (Hyderabad)**

AIML Senior Associate
- Architected an autonomous, **multi-agent** reasoning engine using **LangGraph** to decompose ambiguous user requests into parallel, dependency-aware execution graphs, **reducing reporting latency by over 60%.**
- Engineered the engine as a **self-correcting, schema-aware** pipeline handling **~2000 daily queries** where orchestrated agents autonomously produce **natural** language **insights**, **structured** tabular **data**, and **interactive Chart**.js configurations from a single user query.
- **Tech Stack** : Python, LangGraph, Pydantic, AWS (Lambda, DynamoDB, RedShift, Bedrock), Fast API, PostgreSQL

ValueLabs   **Mar 2022 – May 2025 (Hyderabad)**

Data Scientist 2 (Senior Software Engineer)
- 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.
- Implemented several **Python-based automations** for handling manual/regular tasks resulted in significant **reductions** of over **65** working **hours** monthly with notable recognition from stakeholders.
- Delivered **4** high-quality, scalable, and secure **Web applications using Django** framework by creating innovative, user-centric solutions that have consistently exceeded client expectations.
- **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 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 with Airflow DAGs, processing 1k+ new documents/day.
- Integrated streaming LLM inference with Ollama, delivering sub-3s time-to-first-token latency.
- **Tech Stack**: Python, FastAPI, OpenSearch, Redis, Langfuse, Ollama, Jina AI, Airflow, Docker, 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.

Awards & Certifications

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