Eswar Sai Korrapati

551-344-5356 | eksai0726@gmail.com | LinkedIn | GitHub | Portfolio

Professional Summary

AI/ML Engineer with 3+ years of experience building production LLM and multi-agent systems. I've delivered automation workflows that reduced operational effort 30–45% across enterprise clients, including RAG pipelines, agent orchestration, and AWS-based deployments at scale. Skilled in end-to-end ML development with SageMaker, Spark, PyTorch, and Kubernetes, and experienced partnering with SOC, platform, and support teams to ship reliable AI features quickly.

Experience

Software Engineer – AI Oct 2025 – Present
eSentire Pleasanton, CA

- Partnered with SOC analysts to build a multi-agent system using LangChain and AWS Lambda that automates dashboard setup, cutting configuration time 45% across 200+ clients.
- Designed a unified JSON schema contract that eliminated dashboard misconfigurations and improved data accuracy 30%.
- Deployed a RAG-based retrieval system for schema and metadata lookup, reducing new client onboarding from 3 days to under 1 day.
- Implemented an automated validation pipeline with schema verification, dropping validation errors 70% and enabling weekly production releases.

Software Engineer – AI Aug 2024 – Sep 2025
Comcast San Jose, CA

- Built an agentic troubleshooting system using LangChain, LangGraph, and Claude 3.5 Sonnet (Bedrock) that automated 2,800+ Tier-1 tickets monthly, reducing manual support load 35%.
- Led technical design of a supervisor–worker agent architecture, improving automated resolution accuracy from 55% to 66%.
- Added persistent state using Redis and S3, cutting repeated customer interactions 30% across multi-step troubleshooting flows.
- Integrated diagnostic APIs for modem telemetry, event logs, and device status, improving first-contact resolution rates by 18%.

Software Engineer May 2020 – Jul 2022
Dentsu Hyderabad, India

- Optimized Flask API performance with query restructuring and indexing, improving throughput 30% and reducing report load times from 12s to under 8s for 50+ analysts.
- Developed an XGBoost forecasting model trained on 4.2M ad records, improving weekly ROI prediction accuracy 15%.
- Built PySpark ETL pipelines processing 300k daily records, reducing feature engineering time 40% and enabling daily retraining.
- Deployed SageMaker endpoints with sub-120ms P95 latency and containerized Flask microservices on EC2, with DVC for experiment tracking and reproducibility.

Projects

RAG-based PDF Summariser — Live Demo
- Built a RAG summarisation tool using LangChain, FAISS, and OpenAI embeddings, reducing long-document review workload by 60%.

Next Word Prediction with BERT — GitHub
- Fine-tuned BERT for next-word prediction, demonstrating transformer modelling and training pipeline development.

Machine Transliteration System — GitHub
- Developed a Seq2Seq transliteration model capable of multilingual script conversion.

AI Meeting Preparation Agent — Live Demo
- Designed a multi-agent pipeline generating structured meeting insights using retrieval, summarisation, and role-based agents.

Education

Montclair State University Montclair, NJ
Master of Science in Computer Science, GPA: 3.65 May 2024

Technical Skills

Languages: Python, C++, SQL, JavaScript
ML & NLP: TensorFlow, PyTorch, Scikit-learn, XGBoost, Transformers, Hugging Face
LLM & Agents: LangChain, LangGraph, OpenAI API, RAG, MCP
Data Engineering: Spark, PySpark, Pandas, NumPy, MongoDB, MySQL
MLOps: MLflow, DVC, Airflow, Jenkins, CI/CD
Cloud & DevOps: AWS Lambda, EC2, S3, SageMaker, CloudWatch, Bedrock, Docker, Kubernetes
Tools: Git, Jira, FastAPI, Streamlit, VS Code, Cursor