Eswar Sai Korrapati

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

Professional Summary

AI/ML Engineer with 3+ years building production LLM and multi-agent systems. Delivered automation workflows cutting dashboard setup time 45% and auto-resolving 2,800+ support tickets monthly across enterprise clients. Expertise in RAG pipelines, agent orchestration, and AWS deployments with SageMaker, Kubernetes, and PyTorch.

Experience

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

- Re-architected multi-agent platform from Lambda to Kubernetes (EKS), enabling 6-hour stateful workflows for 200 enterprise SOC clients.
- Built session-aware context management with automatic summarization at 80% context threshold, enabling 5+ iterative dashboard edits per session and cutting hallucination errors 40%.
- Cut SOC analyst dashboard configuration time from 4 hours to 15 minutes via agent-driven automation across 200 clients.
- Designed unified JSON schema with 150+ validated fields and CI/CD validation pipeline, reducing config errors 70% and enabling weekly releases.

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

- Built multi-agent troubleshooting system with LangGraph and Claude 3.5 Sonnet (Bedrock), auto-resolving 66% of 2,800 Tier-1 tickets monthly.
- Designed supervisor-worker agent architecture with A/B-tested prompts, improving auto-resolution accuracy from 55% to 66%.
- Added persistent state via Redis and S3, reducing repeat customer interactions 30%.
- Integrated 8 diagnostic APIs with sub-200ms P95 latency, boosting first-contact resolution from 48% to 66%.

Software Engineer May 2020 – Jul 2022
Dentsu Hyderabad, India

- Optimized Flask APIs and PostgreSQL queries, improving throughput 30% and reducing report load times 35%.
- Built XGBoost forecasting model on 4.2M records, improving ROI prediction accuracy 15% and enabling $2M+ budget optimization.
- Developed PySpark ETL pipelines processing 300k daily records, reducing feature engineering time 40%.
- Deployed SageMaker endpoints with sub-120ms latency and 30% cost reduction via autoscaling.

Projects

RAG-based PDF Summariser — Github

- Built RAG summarization tool with LangChain, FAISS, and OpenAI embeddings, cutting document review time 60%.

Next Word Prediction with BERT — GitHub

- Fine-tuned BERT on custom corpus using Hugging Face transformers.

Machine Transliteration System — GitHub

- Built Seq2Seq model with attention for multilingual script conversion, achieving 85% accuracy on 10k samples.

AI Meeting Preparation Agent — Github

- Designed multi-agent pipeline generating structured meeting briefs, reducing prep time from 30 min to 5 min.

Education

Montclair State University Montclair, NJ
M.S. in Computer Science, GPA: 3.65 Aug 2022 – May 2024

Technical Skills

Languages: Python, C++, SQL, JavaScript
ML/NLP: TensorFlow, PyTorch, XGBoost, Transformers, Hugging Face
LLM/Agents: LangChain, LangGraph, RAG, Prompt Engineering, OpenAI API, Claude
Data: Spark, PySpark, Pandas, NumPy, PostgreSQL, MongoDB
MLOps/Cloud: AWS (EKS, Bedrock, SageMaker, EC2, S3), Docker, Kubernetes, MLflow, DVC