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

- Migrated Lambda agent to Kubernetes (EKS), enabling 6-hour stateful workflows vs. 15-min timeout limit across 200 SOC clients.
- Built Redis session manager with Claude-powered summarization, enabling 5+ iterative edits per session and reducing hallucinations 40%.
- Designed validated JSON schema with CI/CD pipeline, cutting deployment errors 70% and enabling weekly releases vs. monthly.
- Improved widget selection accuracy from 70% to 90% via query-based logic, validated with 15 SOC analysts across 50 dashboards.

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

- Built LangGraph troubleshooting system with Claude 3.5 Sonnet, auto-resolving 66% of 2,800 monthly tickets (1,850/month).
- Built 4-agent architecture (router, executor, validator, escalator), improving resolution accuracy 55% → 66% via prompt engineering.
- Implemented LangSmith tracing, achieving 95% LLM call observability and reducing debug time 50%.
- Added Redis/S3 state persistence, cutting repeat customer interactions 30% across multi-step flows.
- Integrated 8 diagnostic APIs (modem health, logs, account status) with 99.2% uptime and sub-200ms P95 across 120k daily requests.

Software Engineer May 2020 – Jul 2022
Dentsu Hyderabad, India

- Optimized Flask APIs with query restructuring and indexing, improving throughput 30% and reducing load times from 12s to 8s.
- Built XGBoost forecasting model on 4.2M records, improving ROI prediction accuracy 15% for $2M+ budget optimization.
- Developed PySpark ETL pipelines processing 300k daily records, reducing feature engineering time 40%.
- Deployed SageMaker endpoints with sub-120ms P95 latency and 30% cost reduction via autoscaling.

Projects

RAG-based PDF Summariser | Live Demo Jan 2024

- Built RAG pipeline with FAISS top-12 retrieval across 4 query variants (1000-token chunks, 200-overlap), reducing review time 60%.

AI Meeting Preparation Agent | Live Demo Dec 2023

- Designed 4-agent LangGraph pipeline with Tavily search and Llama-4-Maverick, reducing meeting prep from 30min to 5min.

BERT Fine-tuning for Next-Word Prediction | GitHub Mar 2023

- Fine-tuned BERT with custom data collator and FP16 training, achieving Top-5 accuracy on 10k+ examples.

Seq2Seq Transliteration with Attention | GitHub Jan 2023

- Built GRU encoder-decoder with Bahdanau attention for English→Hindi, achieving 85% accuracy on 10k samples.

Education

Montclair State University Montclair, NJ
Master of Science in Computer Science, GPA: 3.65 Aug 2022 – May 2024

Technical Skills

Languages: Python, C++, SQL, JavaScript
ML & NLP: TensorFlow, PyTorch, Scikit-learn, XGBoost, Transformers, Hugging Face
LLM & Agents: LangChain, LangGraph, Langsmith, OpenAI API, RAG, MCP, Claude
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, Claude Code