RAVITEJA KUNAPAREDDY

Chicago, IL | ravitejakunapareddy09@gmail.com | +1 815) 909-9283
https://www.linkedin.com/in/ravi-kunapareddy/ | https://github.com/RaviKunapareddy

ABOUT

Generative AI Engineer with 2+ years of experience fine-tuning LLMs (Mistral-7B, QLoRA), building RAG pipelines, and orchestrating agentic systems (LangGraph, CrewAI). Skilled in deploying GenAI APIs via FastAPI on SageMaker, Vertex AI, and AWS Bedrock. Experienced in bias-aware LLM evaluation, hallucination mitigation, and compliance-ready architectures powering real-world NLP applications. Adept in Python and Node.js, with strong grounding in prompt engineering, speech-to-text (Transcribe), and multilingual NLP (Translate).

EDUCATION

Northern Illinois University — **Chicago, IL**
MS in Management Information Systems — **May 2025**

- *Relevant Coursework*: Python Programming, Advanced Predictive Analytics, Applied Business Analytics (SAS), Business Data Visualization, Database Management Systems, Information Technology Project Management.

SRKR Engineering College — **Vizag, India**
B.E. in Electrical and Electronics Engineering — **May 2020**

- *Relevant Coursework*: Data Structures & Algorithms, C Programming, Python Programming, Artificial Intelligence, Natural Language Processing.

TECHNICAL SKILLS

- **LLMs & Fine-Tuning:** GPT-4, Claude, Gemini, Mistral-7B, QLoRA, LoRA, Prompt Engineering, Instruction Tuning.
- **RAG & Agentic AI:** Retrieval-Augmented Generation (RAG), FAISS, Qdrant, LangChain, LangGraph, CrewAI, Context Caching.
- **GenAI Deployment:** Python, FastAPI (Async), Docker, REST APIs, Kubernetes, Vertex AI, GitLab CI/CD.
- **Cloud AI Stack:** AWS SageMaker, Bedrock, Textract, Kendra, Transcribe, Azure ML, Redis.
- **MLOps & Data Pipelines:** Airflow, MLFlow, Weights & Biases, pandas, NumPy, SQL, PostgreSQL, JSON/XML
- **Evaluation & Risk Controls:** Output Auditing, Hallucination Detection, Prompt Risk Controls, Fallback Logic

TECHNICAL PROJECTS

Healthcare RAG Assistant | Gemini + FAISS — **Sep 2024 - Nov 2024**

- Developed and deployed an end-to-end RAG system for clinical QA workflows, using Gemini LLM, FAISS vector search, and CrossEncoder reranking — enabling accurate, real-time answers across 1,500+ parsed MedQuAD documents
- Designed safety-first generation logic with fallback triggers, streaming-compatible outputs, and hallucination truncation — tailored for use in medically sensitive contexts.
- Built a modular CLI-based pipeline with prompt injection defense, context caching, and metadata-aware chunking — emulating AWS Textract and Kendra workflows for structured semantic search.
- Applied domain-tuned sentence-transformer embeddings and retrieval prompts to achieve high-recall, constrained generation, and explainable outputs for clinical support.

Fine-Tuned LLM - Financial Risk Extraction — **Nov 2023 - Mar 2024**

- Fine-tuned Mistral-7B using QLoRA to extract structured risk data from SEC 10-K filings — leveraging custom domain prompts and schema-constrained generation to produce JSON outputs (severity, impact, mitigation) for audit and compliance workflows.
- Built an end-to-end legal-NLP pipeline using SentencePiece for tokenizer customization and domain-specific prompt templates to improve reliability across 5,000+ filings.
- Deployed lightweight 300MB LoRA adapters, reducing inference cost by ~40% and enabling real-time anomaly detection on extracted risk factors.
- Enabled integration with stock modeling tools and audit platforms through structured output design — supporting compliance reporting, regulatory transparency, and valuation analysis

Structured Multi-Agent Task Execution — **Aug 2023 - Oct 2023**

- Led the design of a modular agentic AI system using CrewAI and SentenceTransformers to triage 7,000+ user queries and dynamically route them to memory, knowledge, and tool agents — inspired by enterprise support automation
- Integrated Qdrant as a persistent vector memory layer for long-term context retention, and orchestrated Gemini-based synthesis via prompt-controlled agent delegation
- Implemented production-grade safety mechanisms: prompt validation filters, hallucination suppressors, and fallback flows — emulating LLM safety protocols used in real-time agents
- Built full observability into the agent stack using feedback logs, dynamic context tracing, and per-agent metrics — enabling iterative debugging and performance tuning

PROFESSIONAL EXPERIENCE

TCS (Client : British Telecom) — **Chennai, India**
Machine Learning Engineer — **Jan 2020 - May 2022**

- Developed and deployed ML and GenAI models to analyze telecom usage data, network performance metrics, and IT support logs, improving predictive maintenance accuracy and service quality metrics.

- Built scalable RAG (Retrieval-Augmented Generation) pipelines to extract insights from internal knowledge bases, technical manuals, and customer communication records
- Engineered low-latency inference APIs using FastAPI and WebSockets to power AI-driven chat assistants and ticket classification systems.
- Applied Scikit-Learn and TensorFlow for model selection, evaluation, and retraining in anomaly detection, churn prediction, and fault localization.
- Integrated OpenAI and Azure OpenAI models into secure LangChain-like agents with structured outputs, fallback logic, and context caching to reduce hallucinations.
- Automated feature pipelines and model retraining using Airflow and MLFlow, with continuous monitoring for data drift and model degradation.
- Developed REST and GraphQL APIs with schema validation, observability hooks, and retry logic to support integration with internal telecom platforms.
- Maintained CI/CD workflows using GitLab and Jenkins to ensure reliable deployment and version control across production environments.
- Collaborated with network engineers and service teams to align ML outputs with telecom service design, diagnostics, and regulatory compliance.
- Contributed to Agile sprints, technical documentation, and internal demos to accelerate GenAI adoption across BT's digital services.

CERTIFICATIONS

- **AWS Certified Machine Learning – Associate | Credential ID:** https://www.credly.com/badges/1b1da8df-bfac-4c95-9b46-79dd8c313b2dxxx