Vamana Sesha Sai Siddartha Koppaka

+1 930-333-2483 | vkoppaka@iu.edu | Linkedin-SiddarthaKoppaka | Github-SiddarthaKoppaka | Portfolio

Education

Indiana University — Bloomington, IN
Master of Science in Computer Science — *Aug 2023 - May 2025*

Gurukula Kangri (Deemed to be University) — Haridwar, India
Bachelor of Technology in Computer Science and Engineering — *Aug 2019 - May 2023*

Technical Skills

Programming & Data Processing: Python, SQL, PostgreSQL, PySpark, Apache Spark, Airflow, Kafka, dbt, Snowflake, JS
Machine Learning & AI: Scikit-Learn, TensorFlow, PyTorch, Hugging Face, OpenAI API, LLMs, NLP, Time-Series Analysis, Computer Vision, transformers, LangChain, LangGraph, LangSmith, VertexAI, Ollama, NLTK, Spacy
Cloud & DevOps: AWS (S3, Lambda, ECS, SageMaker), GCP (BigQuery, Cloud Run), Docker, Kubernetes, CI/CD , Redis
Data Visualization: Power BI, Tableau, Matplotlib, Seaborn, Plotly, d3.js
Software Engineering: Large-scale distributed systems, Multithreading, Data processing systems, Pattern Recognition, Natural Language Processing

Experience

AI Engineer — Nov 2024 – Present
Kelley School of Business - Indiana University — *Bloomington, IN*
- Engineered a Retrieval-Augmented Generation (RAG) system using LangGraph and LangChain, achieving 92% top-3 retrieval accuracy over a Qdrant vector index built on 50K+ legal documents.
- Designed asynchronous REST APIs in FastAPI for serving inference, handling 15K+ requests/day with sub-200ms latency.
- Fine-tuned transformer models with PyTorch for sentiment classification on 100K+ business reviews, enhancing few-shot prediction performance to 90%.
- Implemented NLTK and spaCy pipelines for named entity tagging and dependency parsing on over 3M sentences.
- Deployed AI services in containerized Docker environments, tracking inference health metrics with Prometheus and Grafana across cloud-hosted instances.

AI Engineer — Aug 2024 – Dec 2024
Luddy School of Informatics - Indiana University — *Remote*
- Collected and processed 500K+ structured and unstructured data points from online sources, improving data accessibility for analytical studies by 40%.
- Fine-tuned transformer-based models, improving zero-shot classification accuracy by 35% and enabling automated categorization of 100K+ records.

AI/ML Developer — May 2024 – Aug 2024
Hyphenova — *Remote*
- Developed and tested user-facing features for the Hyphenova application using TypeScript and Next.js, enhancing performance and user experience by 10%.
- Implemented unit and integration tests with Jest and React Testing Library, improving frontend test coverage by 85%.
- Designed multi-cloud infrastructure with Terraform on AWS, Azure, and GCP, reducing provisioning time by 40%.
- Automated CI/CD pipelines using Docker and Terraform, cutting deployment time by 40% and streamlining feature delivery
- Developed real-time data pipelines with Apache Spark, Kafka, and AWS Data Lake, enabling continuous analytics.

Projects

AI-Powered Document Search Engine | *FastAPI, LangChain, Qdrant, Llama 3, Docker, AWS, LLMs* Nov 2024 – Mar 2025
- Built an advanced search system integrating Llama 3 using vLLM with Qdrant VectorDB, and Langchain.
- Enhanced document retrieval with FAISS, BM25 and Qdrant, reducing query latency by 60%, while deploying the inference API on AWS using Bedrock & EC2.

User Churn Prediction & Retention Analysis | *Big Data, Databricks, Azure, ML, CI/CD* May 2024 – June 2024
- Processed 50M+ telecom records using Spark on Databricks, automated ETL with Apache Airflow, and stored output in Azure Blob Storage.
- Trained XGBoost and LSTMs, achieving a 92.5% F1-score, and used MLflow for tracking and comparison.

Thinkwise: AI Idea Evaluator | *LangGraph, LangChain, Gemini, FastAPI, MongoDB, ReactJS* Feb 2025 – Apr 2025
- Developed a multi-agent system that ranks and explains business ideas based on ROI and effort using Gemini-powered ReAct agents; reduced evaluation time by 70%.
- Enabled real-time analysis for 100+ concurrent users with session tracking, stores in MongoDB.

Buddy: Your Personal AI Companion | *LangGraph, Redis, LangChain, Node.js, TypeScript, Google Cloud* Apr 2025 – Present
- Created a context-aware AI assistant that automates daily tasks; improved daily productivity tracking by 40%.
- Supports persistent user sessions across 500+ interactions, leveraging Redis for stateful memory and LangGraph.Integrated Twilio based AI text-to-speech for voice based conversation.

Certifications

AWS Cloud Practitioner — Sep 2024 – Sep 2027
Credential ID d8506a288f5b4e4cac01b453f6fece1d

Databricks Academy: Fundamentals Accreditation — Mar 2025 - Mar 2026