

Sai Kumar Nune

New Haven, CT – 06511 — nunesaikumarnsk@gmail.com — +1 (636) 891-2315 — [LinkedIn](#)

Professional Summary

AI/ML Engineer with **5+ years of experience** developing and deploying **LLM-driven systems**, **multi-agent RAG architectures**, and **semantic search applications** for enterprises across healthcare, finance, and telecom. Skilled in designing **LangChain** and **LangGraph**-based pipelines that integrate **GPT-4**, **LLaMA2**, and **vector databases** (FAISS, ChromaDB) for context-aware automation and knowledge retrieval. Experienced in building **FastAPI**-powered AI microservices, implementing **MLOps pipelines** with Docker, Kubernetes, and CI/CD, and fine-tuning models using **Hugging Face Transformers** and **TensorFlow**. Adept at translating business goals into scalable AI workflows, reducing latency, and improving accuracy through **prompt optimization**, **metadata retrieval**, and **responsible AI practices**.

Category	Tools & Technologies
AI/ML & NLP	GPT-3/4, LLaMA2, T5, BERT, SBERT, LangChain, LangGraph, Hugging Face, RAG, Prompt Engineering, TensorFlow, PyTorch, Scikit-learn, XGBoost, SpaCy, NLTK, Semantic Search, NER, Summarization, Tool-using Agents, Transformers, ANN, CNN, Ensemble Methods, Recommender Systems, Few-shot Learning
Data & Engineering	Python (Pandas, NumPy, Matplotlib), SQL, NoSQL, MySQL, PostgreSQL, MongoDB, Airflow, REST APIs, FAISS, ChromaDB, Neo4j, Pinecone, Spark, Flask
Cloud & DevOps	AWS (SageMaker, Lambda, EC2, S3), Azure, Docker, Kubernetes, OpenShift, Git/GitHub, GitHub Actions, Bitbucket, Harness, CI/CD
Visualization	Power BI, Tableau, VS Code, Excel
Computer Vision	OpenCV, Image Classification, Object Detection, Vision Transformers

Educational Details

Sacred Heart University, Fairfield, ConnecticutJan 2024 – Mar 2025
Master of Science (M.S.) in Computer Science – CGPA: 3.76/4

Lakireddy Bali Reddy College of Engineering, IndiaDec 2021
Bachelor of Technology in Electronics and Communication Engineering – CGPA: 7.59

Professional Experience:

Responsive – Remote (AI Engineer)July 2024 – Present

Roles & Responsibilities:

- Designed and deployed enterprise-grade **LangGraph-powered GPT-4 chatbots** for internal automation, improving service response time by **35%** through **multi-agent orchestration** and intelligent workflow management.
- Built and integrated **memory-aware agents** and **multi-tool integrations** using **LangChain React Agent** framework, orchestrating multi-step reasoning workflows for real-time, personalized, **context-driven user interactions**.
- Engineered robust **Retrieval-Augmented Generation (RAG)** systems with **LangChain**, **FAISS**, **ChromaDB**, **Pinecone**, and **SQL** databases to enhance enterprise knowledge access and deliver **context-aware, factually accurate responses**.
- Increased ingestion throughput by **40%** through **token-optimized chunking pipelines** for PDFs and CSVs with intelligent data preprocessing using **Python**, **Pandas**, and **NumPy**.
- Designed **graph-based Q&A agents** using **Neo4j** graph database, implementing **semantic search** and improving real-time contextual retrieval for complex query workflows across distributed knowledge bases.
- Developed and optimized **prompt engineering strategies** for **GPT-4** and **LLaMA2**, enabling domain-specific automation, personalized content generation, and improved model response quality.
- Applied **few-shot learning** techniques and advanced **NLP pipelines** using **Transformers**, **BERT**, and **Hugging Face** libraries for **NER**, **summarization**, and classification tasks.

- Created and maintained **CI/CD pipelines** using **Docker**, **Kubernetes**, **OpenShift**, **Git/GitHub**, **GitHub Actions**, **Bitbucket**, and **Harness** for streamlined model deployment, version control, and continuous monitoring.
- Deployed and managed **NLP and CV models** in production environments on **AWS SageMaker**, **Lambda**, **EC2**, and **S3** with high availability and scalability.
- Built **REST APIs** using **Flask** for model serving and integration with front-end applications, ensuring low-latency responses.
- Incorporated **responsible AI practices** by implementing model explainability, fairness checks, bias detection, and compliance measures in deployed systems.
- Collaborated in cross-functional Agile teams using tools like **VS Code** for development, ensuring alignment with business goals and measurable impact.

Tech Stack: GPT-3/4, LLaMA2, LangChain, LangGraph, LangChain React Agent, FAISS, ChromaDB, Neo4j, Pinecone, SQL, NoSQL, MongoDB, PostgreSQL, MySQL, OpenAI API, Python, Pandas, NumPy, Matplotlib, Flask, REST APIs, Docker, Kubernetes, OpenShift, Git, GitHub, GitHub Actions, Bitbucket, Harness, AWS (SageMaker, Lambda, EC2, S3), CI/CD, RAG, Prompt Engineering, Few-shot Learning, Vector Databases, Semantic Search, Multi-Agent Orchestration, Tool-using Agents, Transformers, BERT, Hugging Face, NER, Summarization, Real-time LLM Applications.

Wipro – Hyderabad, India (*GenAI / ML Engineer*)

Dec 2021 – Mar 2024

Roles & Responsibilities:

- Designed and deployed **Retrieval-Augmented Generation (RAG) systems** using **LangChain**, **FAISS**, and **ChromaDB**, integrating LLMs (**GPT-4**, **LLaMA2**, **T5**) to support healthcare claims automation, medical code extraction, and intelligent document summarization.
- Reduced manual triage by **35%** through optimized RAG workflows, improved metadata retrieval pipelines, and **semantic search** capabilities for claim fraud detection and evidence retrieval.
- **Fine-tuned LLMs** (GPT-4, LLaMA2, T5) for summarization and document intelligence pipelines using **prompt engineering**, **few-shot learning**, and transfer learning techniques.
- Built advanced **NLP pipelines** using **AWS SageMaker**, **BERT**, **T5**, **SBERT**, **spaCy**, and **NLTK** for clinical code recommendations, **NER**, textual entailment, and intelligent chatbot responses.
- Developed and deployed **deep learning** and **neural network models** (ANN, CNN) using **TensorFlow** and **PyTorch** with hyperparameter tuning, regularization, and backpropagation to improve model accuracy.
- Trained **computer vision models** with **PyTorch**, **TensorFlow**, and **OpenCV** for document-based visual information extraction, dental image analysis, document verification, and **object detection** tasks.
- Applied **machine learning techniques** including **Regression**, **Classification**, **Clustering**, **Ensemble Methods**, **XGBoost**, and **Scikit-learn** for predictive analytics, behavioral modeling, and segmentation.
- Led deployment optimization, reducing inference latency by **20%** through fine-tuning deployment strategies on **AWS Lambda**, **EC2**, and **S3** in collaboration with DevOps teams.
- Replicated and extended features of **WiNow**, Wipro's internal AI chatbot platform—built chatbot workflows for timesheet entry, leave requests, approvals, and ticket status tracking using **conversational AI**, demonstrating up to **80%** reduction in manual effort and **50%** faster turnaround.
- Engineered and optimized **ETL workflows** using **Apache Airflow** and **Spark** to ingest, transform, and process large-scale structured and unstructured datasets from healthcare repositories and **SQL** databases.
- Built and maintained **CI/CD pipelines** with **Docker**, **Kubernetes**, **Git**, and **Azure** cloud-native tools to automate retraining, testing, and production deployment of AI models.

- Developed automated **data pipelines** with **Python**, **NumPy**, **Pandas**, and **Matplotlib** for data preprocessing, feature engineering, and visualization.
- Created interactive **dashboards** and **data visualizations** using **Tableau**, **Power BI**, and **Excel** to communicate model insights and business metrics to stakeholders.

Tech Stack: LangChain, FAISS, ChromaDB, GPT-4, LLaMA2, T5, BERT, SBERT, spaCy, NLTK, TensorFlow, PyTorch, Scikit-learn, XGBoost, OpenCV, AWS (SageMaker, Lambda, EC2, S3), Azure, Docker, Kubernetes, Git, Apache Airflow, Spark, Python, Pandas, NumPy, Matplotlib, SQL, MySQL, PostgreSQL, Tableau, Power BI, Excel, Prompt Engineering, Few-shot Learning, NER, Summarization, ANN, CNN, Ensemble Methods, Regression, Classification, Clustering, Object Detection, Image Classification, Vision Transformers, Semantic Search, RAG.

Zensar Technologies – Hyderabad, India (*Associate AI Engineer*)

Dec 2020 – Nov 2021

Roles & Responsibilities:

- Created **classification**, **regression**, and **anomaly detection models** using **ensemble methods**, **XGBoost**, and **Scikit-learn** with retraining workflows using **Apache Spark** and **Airflow**.
- Deployed **transformer-based NER** and classification models built with **TensorFlow**, **PyTorch**, and **Hugging Face Transformers** on **AWS** and **Azure** cloud platforms for finance and telecom domains.
- Developed automated **data pipelines** and **ETL workflows** with **Apache Airflow**, **Python**, **NumPy**, and **Pandas** to process large-scale structured and unstructured datasets from **SQL**, **NoSQL**, **MySQL**, **PostgreSQL**, and **MongoDB** databases.
- Applied **NLP techniques** using **spaCy**, **NLTK**, and **BERT** for text analytics, **NER**, sentiment analysis, and **summarization** tasks.
- Built executive **dashboards** and interactive **data visualizations** using **Power BI**, **Tableau**, and **Excel** to visualize model performance and business metrics, communicating insights to stakeholders.
- Implemented **CI/CD pipelines** using **Git/GitHub** for version control and automated model deployment.
- Utilized **VS Code** for development and collaborated with cross-functional teams to ensure model explainability, alignment with business goals, and measurable impact of AI solutions.
- Designed and built **recommender systems** using collaborative filtering and content-based approaches for personalized user experiences.

Tech Stack: Python, TensorFlow, PyTorch, Scikit-learn, XGBoost, Hugging Face, Transformers, Apache Spark, Airflow, NumPy, Pandas, Matplotlib, spaCy, NLTK, BERT, AWS, Azure, Git, GitHub, Power BI, Tableau, Excel, VS Code, SQL, NoSQL, MySQL, PostgreSQL, MongoDB, Regression, Classification, Clustering, Ensemble Methods, NER, Summarization, Recommender Systems.

Key Professional Projects

Healthcare RAG Bot:

Built **GPT-4 LangChain Q&A** system with **FAISS** vector database for healthcare triage automation, implementing **semantic search** and **prompt engineering**, cutting response time by **40%**.

Multi-Agent RAG System for Claims Processing:

Engineered **LangGraph-based orchestration system** integrating **GPT-4** and **LLaMA2** with **FAISS** and **ChromaDB** for automated healthcare claims fraud detection.

Reduced manual review time by **35%** and improved detection accuracy by **22%** through optimized RAG workflows and metadata retrieval pipelines.

Auto-Form Extraction:

Created **Vision Transformer-based** document extractor using **PyTorch** and **OpenCV**, improving accuracy from 79% to **92%** through **deep learning** techniques.

Chat Agent Orchestration:

Built **LangGraph**-based multi-agent system with tool-using agents and memory-aware capabilities, reducing SLA violations by **30%**.

Awards and Certifications:

- **NVIDIA-Certified Associate: Generative AI LLMs**
- **DeepLearning.AI – Generative AI with LLMs (Coursera) – 2025**
- **Udemy – Artificial Intelligence A-Z – 2025**
- **Udemy – Certified Machine Learning Engineer – 2024**