Jaya Krishna

AI/ML Engineer

(405) 916-9097 | jsangoju@saintpeters.edu| Jersey City, NJ (07304)

PROFESSIONAL SUMMARY

Results-driven AI and Machine Learning Engineer with 4 years of experience developing scalable AI and conversational systems using advanced NLP and large language models like GPT-4. Skilled in deploying cloud-based solutions on AWS and GCP, optimizing models with PyTorch and TensorFlow, and integrating vector databases with RAG techniques. Certified in Google Cloud and AWS ML specialties, with strong expertise in Agile environments and data-driven decision-making.

TECHNICAL SKILLS

Programming Languages: Python, TypeScript, JavaScript, SQL
Machine Learning & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Random Forest, LSTM, CNNs, Vision Transformers (ViT), AutoEncoders, Transfer Learning
Natural Language Processing (NLP) & Large Language Models (LLMs): Hugging Face Transformers, GPT-3/4, BERT, LLaMA, Gemini, T5, TF-IDF, NLTK, spaCy, Named Entity Recognition (NER), Semantic Search, Prompt Engineering & Tuning, Instruction Tuning, Few-Shot Learning, Self-Supervised Learning, PEFT (LoRA, QLoRA)
Generative AI & Frameworks: LangChain, LangGraph, LangSmith, OpenAI API, Retrieval-Augmented Generation (RAG), CrewAI, AutoGen
Vector Databases & Search Technologies: Pinecone, FAISS, ChromaDB, AWS OpenSearch, Lucidworks Fusion, Solr, Neo4j
API Development & Integration: REST APIs, GraphQL, OpenAPI, OAuth2.0 Authentication, Function Calling
Data Engineering & Pipelines: Pandas, Hydra, Weights & Biases, Neo4j
Cloud Platforms & Deployment: AWS (Lambda, EC2, SageMaker, S3), GCP (Cloud Functions), Firebase, Databricks, Docker, Kubernetes, Flask, FastAPI
DevOps & CI/CD Tools: Jenkins, GitHub Actions, TeamCity, Git
Monitoring & Observability: LangSmith, Custom Logging, Token-Level Tracing, Prompt Versioning
Project Management Tools: Agile/Scrum, JIRA, GitHub Projects, Jupyter Notebook, VS Code

WORK EXPERIENCE

AI Engineer I

JP Morgan, Jersey City, NJ – USA **October 2024 – Present**

- Apply Agile and Scrum methodologies throughout the Software Development Life Cycle (SDLC) to deliver high-quality data solutions in cross-functional team environments.
- Lead a team of developers in designing and optimizing AI-powered chatbots, improving customer query resolution efficiency by 40% through advanced NLP techniques such as GPT-4, Dialogflow CX, and Kore.ai, along with iterative prompt engineering.
- Manage end-to-end development of a scalable, cloud-based conversational AI system on GCP (Vertex AI, Cloud Functions), handling over 100,000 daily conversations with 99.9% uptime.
- Implement real-time analytics using Power BI and Kore. Ai's reporting tools, enabling data-driven decisions that boost user engagement by 40% and reduce drop-off rates by 15%.
- Design conversational flows—including intents, entities, and dialogues—in Kore.ai to enhance user experience and achieve 20% higher engagement with ChatGPT-generated outbound campaigns.
- Deploy chatbots across web, mobile, and messaging platforms using Kore.ai, ensuring seamless integration through REST APIs and OAuth 2.0 authentication.
- Pioneer Retrieval-Augmented Generation (RAG) solutions using LangChain, Pinecone, and GPT-4 to improve chatbot accuracy and reduce manual intervention by 30%.
- Conduct A/B testing and quality assurance for chatbot performance, leveraging Lang Smith for token-level tracing and custom logging to maintain over 95% intent recognition accuracy.
- Collaborate with cross-functional teams (content, product) to optimize ChatGPT and Dialogflow CX scripts, streamlining deployment using CI/CD workflows with GitHub Actions and Docker.

ML Engineer
WebCode IT, Hyderabad - INDIA **June 2020 – July 2023**

- Led data analysis in Agile and Scrum environments, collaborating with teams to ensure insights aligned with business goals.
- Developed and deployed conversational AI solutions using Dialogflow, Rasa, and APIs, aligning with client requirements and business objectives.
- Collaborated with cross-functional teams to design and implement intelligent voice user interfaces (VUI) for web and mobile platforms, ensuring smooth integration and enhanced user experience.
- Applied NLP libraries such as spaCy and NLTK to preprocess and analyze large textual datasets for sentiment analysis, named entity recognition, and intent classification.
- Trained and optimized deep learning models for speech recognition and synthesis with TensorFlow and PyTorch, achieving accuracy rates exceeding 90%.
- Integrated Google Speech-to-Text and Text-to-Speech APIs to enable real-time speech conversion, reducing response time by 30%.
- Managed and optimized databases including Firebase, Grafana, and SQL to improve query performance, maintain data integrity, and ensure security.
- Enhanced speech recognition accuracy by 15% through fine-tuning TensorFlow deep learning models, resulting in more reliable conversational interactions.
- Developed interactive data visualizations using Matplotlib, Zoho Analytics, and Power BI to deliver actionable insights and support data-driven decision-making.

PROJECTS & RESEARCH HIGHLIGHTS

Generative AI Chatbot for Internal Employee Support **Domain: Gen AI, LLMs, RAG Pipelines, Prompt Engineering**

- Developed an internal-facing Generative AI chatbot using LLMs (GPT-4) and LangChain RAG pipeline, enabling instant policy and documentation Q&A for 10,000+ corporate users.
- Integrated vector database (Pinecone) for semantic document retrieval and used embedding models (OpenAI Ada) to increase relevant response accuracy by ~38%.
- Deployed the solution using FastAPI, Docker, and AWS ECS, reducing support response times by 50% and increasing resolution rates by 42% in pilot testing.

PUBLICATIONS

- **Internal White Paper: "TF-IDF-Based Ranking System for Doctor Recommendations"**
- **Internal Release: "Intent Detection and Classification in NLP Chatbots for Medical Triage"**
- **Google Cloud Certified – Professional Machine Learning Engineer**
- **AWS Certified Machine Learning – Specialty**
- **Natural Language Processing Specialization – DeepLearning.AI (Coursera)**
- **Applied Data Science with Python – University of Michigan**
- **Deploying Machine Learning Models in Production**
- **SQL for Data Science – University of California, Davis**

EDUCATION

Masters in computer and information system | Saint peter’s University, Jersey City, NJ -USA
Bachelor of Technology in Information Technology | JNT University Kakinada – INDIA