

Tejesh Mattaparthi

+1 (913) 225 0842 | tejesh.m@cvmailhub.com | [LinkedIn](#) | [GitHub](#) | Open to Relocate

SUMMARY

AI Engineer with 3 years of experience designing, building, and deploying enterprise AI solutions using Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), AI Agents, and scalable machine learning pipelines across cloud environments. Experienced in developing intelligent assistants, semantic search platforms, document intelligence systems, and production AI services using Python, LangChain, LangGraph, FastAPI, PyTorch, TensorFlow, AWS, Azure OpenAI, and Google Vertex AI. Skilled in prompt engineering, vector databases, MLOps, model deployment, API development, and integrating AI solutions into enterprise applications to improve productivity, automation, and operational efficiency.

SKILLS

Programming Languages: Python, SQL, TypeScript, JavaScript, Bash

Generative AI & LLMs: Large Language Models (LLMs), LangChain, LangGraph, LlamaIndex, Retrieval-Augmented Generation (RAG), Prompt Engineering, AI Agents, Multi-Agent Systems, Function Calling, Tool Calling, MCP (Model Context Protocol), AI Workflows, Agent Orchestration, Context Engineering

Machine Learning & Deep Learning: Scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM, Random Forest, Deep Learning, Transfer Learning, Fine-Tuning (LoRA, PEFT), Feature Engineering, Hyperparameter Tuning, Model Evaluation, SHAP

NLP & Vector Search: Hugging Face Transformers, Sentence Transformers, BERT, RoBERTa, Named Entity Recognition (NER), Semantic Search, Vector Embeddings, Hybrid Search, FAISS, Pinecone, ChromaDB, Weaviate

Cloud, Data Engineering & AI Platforms: AWS (S3, Lambda, SageMaker, Bedrock), Azure OpenAI, Google Vertex AI, Apache Spark, Databricks, Apache Airflow, Snowflake, ETL/ELT Pipelines

Backend, APIs & MLOps: FastAPI, Flask, REST APIs, Docker, Kubernetes, Git, GitHub Actions, CI/CD, MLflow, Terraform, Model Deployment, Model Monitoring, Drift Detection, LLM Evaluation

Databases & Storage: PostgreSQL, MongoDB, Redis

Visualization & Analytics: Power BI, Tableau, Streamlit

Software Engineering: Microservices Architecture, Distributed Systems, API Integration, Unit Testing, Integration Testing

EXPERIENCE

AI Engineer JPMorgan Chase & Co., USA	Aug 2025 – Present
- Designed enterprise AI applications using Large Language Models, LangChain, LangGraph, and Retrieval-Augmented Generation (RAG), improving response accuracy by 36% across internal knowledge management and customer support platforms. - Developed scalable AI pipelines with Python, FastAPI, AWS Bedrock, and SageMaker, processing 80M+ records monthly while reducing inference latency by 41% across production workloads. - Built multi-agent AI workflows integrating function calling, MCP, enterprise APIs, and vector databases, reducing manual operations by 38% while improving workflow automation across business functions. - Implemented semantic search solutions using Sentence Transformers, FAISS, Pinecone, and OpenAI embeddings, decreasing document retrieval time by 63% while improving search relevance across enterprise repositories. - Optimized LLM performance through prompt engineering, retrieval tuning, evaluation frameworks, and response validation, reducing hallucinations by 31% while improving answer consistency for production AI services. - Deployed production AI services using Docker, Kubernetes, MLflow, GitHub Actions, and CI/CD pipelines, shortening model deployment cycles by 46% while improving release reliability and version control. - Integrated monitoring, drift detection, and LLM evaluation frameworks using CloudWatch, MLflow, and custom metrics dashboards, improving production model availability to 99.9% while accelerating issue resolution. - Collaborated with Product Managers, Data Engineers, ML Engineers, and Security teams to deliver enterprise AI capabilities supporting fraud detection, compliance automation, and customer intelligence, accelerating feature delivery by 29%.	
AI/ML Engineer Zensar Technologies, India	May 2021 – Jun 2023
- Built predictive machine learning solutions using Python, SQL, XGBoost, Random Forest, and Logistic Regression, improving business prediction accuracy by 24% across customer analytics and financial datasets. - Engineered scalable ETL pipelines using PySpark, Apache Spark, SQL, dbt, and Apache Airflow, processing 25M+ records while reducing pipeline execution time by 40%. - Constructed recommendation and forecasting models using ARIMA, Prophet, feature engineering, and Bayesian modeling, improving demand planning and operational forecasting across enterprise business functions. - Applied Isolation Forest, clustering algorithms, and statistical analysis to detect anomalous customer behavior, reducing manual investigation effort by 35% while strengthening operational risk monitoring. - Created NLP applications using LangChain, Hugging Face Transformers, BERT, and Retrieval-Augmented Generation, enabling enterprise document search, summarization, and knowledge retrieval across unstructured datasets. - Automated machine learning experimentation using MLflow, Docker, GitHub Actions, and CI/CD, improving reproducibility while shortening model validation and deployment preparation by 33%. - Delivered interactive Power BI dashboards tracking predictive model performance, customer behavior, and operational KPIs, reducing manual reporting effort by 45% while improving executive decision support. - Partnered with cross-functional engineering and business teams to validate AI models through SHAP explainability, hypothesis testing, and A/B experimentation, improving stakeholder confidence and production adoption.	

PROJECTS

Enterprise AI Copilot Platform <i>Python, LangChain, LangGraph, FastAPI, OpenAI, Vertex AI, FAISS, Docker</i>	
- Built an enterprise AI copilot using LangChain, LangGraph, and Retrieval-Augmented Generation (RAG) to automate knowledge retrieval, document summarization, and question answering across 100K+ internal documents, improving response accuracy by 35%. - Developed scalable FastAPI microservices with Docker and Vertex AI, integrating vector search, AI agents, and LLM APIs to process over 750K monthly inference requests with low-latency performance. - Optimized prompt engineering, retrieval strategies, and response evaluation workflows, reducing hallucinated responses by 28% while improving answer consistency and user satisfaction across enterprise AI applications.	
Intelligent Document Processing Platform <i>Python, FastAPI, Hugging Face, AWS, PostgreSQL, ChromaDB</i>	
- Designed an AI-powered document processing platform using transformer models, semantic search, and vector databases to automate extraction, classification, and validation of structured and unstructured documents, reducing manual processing time by 65%. - Implemented automated evaluation, monitoring, and CI/CD pipelines using MLflow, GitHub Actions, and Docker, improving deployment reliability while reducing release cycles by 40%. - Developed OCR and NLP pipelines to extract, normalize, and validate information from PDF and DOCX files, improving document processing accuracy by 32% while accelerating downstream business workflows.	

EDUCATION

Masters in Computer Science <i>University of Central Missouri, USA</i>	May 2026
--	----------