S Santosh Kumar

Hyderabad, India | (+91) 8886125285 | sontasantosh5698@gmail.com

Profile Summary

Experienced Digital Specialist Engineer with 2+ years of expertise in Generative AI, LLM fine-tuning, NLP, and AI deployment using Python, TensorFlow, Spacy, Pytorch and deep learning frameworks. Proven ability to develop AI-driven solutions, enhance data processing efficiency, and optimize machine learning models. Skilled in Retrieval-Augmented Generation (RAG), vector databases, prompt engineering, and cloud-based AI deployment.

Key Skills and Technical Expertise

- **Artificial Intelligence & Machine Learning:** Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), NLP, Deep Learning
- **Programming:** Python, Django, Flask
- **AI Frameworks and Tools:** TensorFlow, Hugging Face, Llama Index, Lang Chain
- **Cloud Services:** Azure Key Vault, Azure Blob Storage, Azure Cognitive Services, API Management
-

Education

Master of Computer Applications (MCA)

University College of Engineering, Osmania University, Hyderabad, 2022

Professional Experience

Digital Specialist Engineer

Infosys | Duration: 2 Years 8 months

- Contributed to the R&D team, developing Proof-of-Concepts for cutting-edge AI/ML and Generative AI technologies.
- Specialized in Natural Language Processing (NLP), Deep Learning applications.
- Applied Prompt Engineering to improve model performance and tailor solutions to client needs.
- Worked on various Azure services, including Azure OpenAI, Embeddings, Content Safety, Speech Services, Blob Storage, API Management (APIM), and Key Vault for secure and scalable AI deployment

Projects

Document Embedding and Retrieval with Chroma DB, Llama Index, and FAISS

- Implemented document embedding storage and retrieval using Chroma DB for efficient query handling.
- Optimized the embedding pipeline for improved semantic search performance.
  .
- Embeddings are saved inside Chroma DB, with a parallel implementation using Llama Index vector store and FAISS for efficient retrieval

Document Summarization and Query System with RAG

- Developed a document summarization system using a Retrieval-Augmented Generation (RAG) framework.
- Integrated Falcon 40B-Instruct and Hugging Face embedding models for efficient data embedding and retrieval.
- The system utilizes OpenAI GPT-4 as well as the offline Falcon model for generating summaries and handling complex queries.

Metadata-Driven Document Retrieval System

- Designed a metadata framework for efficient document categorization.
- Integrated semantic search using metadata for advanced query capabilities.

Integration of Azure Key Vault for Secure Key Management

- Replaced traditional .env files with Azure Key Vault to securely store sensitive keys.
- Automated runtime key access, improving security and operational efficiency.

Dynamic Integration of Azure Blob Storage

- Streamlined integration with Azure Speech Service for audio file processing.
- Automated workflows to create containers, upload files, transcribe audio, and clean up post-processing.

Sensitive Information Detection and Masking in Documents

- Developed a solution to detect and mask sensitive information using Presidio analyzer library.
- Automated document processing workflows to ensure compliance with privacy regulations.

Voice-Based Conversational Insights Application

- Built a voice assistant app with real-time transcription using the Web Speech API.
- Integrated ChatGPT-4 for conversation summaries and Q&A generation.
- Added video-to-text processing with the Whisper model for multimedia insights extraction

Azure API Management (APIM) for Secure AI Service Access

- Built a secure API gateway using Azure APIM for controlled access to Azure OpenAI, Embeddings, Content Safety, and Speech Services.
- Implemented OAuth 2.0 and JWT authentication for role-based secure access.