


Charana H U

 charanhumail@gmail.com

 +91 9481368353

 Shivamogga, India

 link.com/charanhu

 github.com/charanhu

OBJECTIVE

Passionate AI/ML Engineer with 2.5+ years of experience in Generative AI, RAG, NLP, and Deep Learning, dedicated to translating cutting-edge advancements into impactful real-world solutions. Specialized in developing and deploying high-performance ML models, LLM-based solutions, and Generative AI systems. Skilled in optimizing model performance and driving AI integration to meet diverse business needs. Proven ability in contributing to open-source projects and advancing the frontier of applied artificial intelligence. Additionally, adept in pre-sales technical roles, with significant experience in sales pitching, stakeholder calls, and presentations, effectively bridging technical expertise with client requirements.

EXPERIENCE

IBM (India Pvt. Ltd,) - AI Engineer

 Aug 2023 - Present

 Bangalore, India


1. Archimed: Document Classification and RAG Assistant

- System Development: Created a document classification system for large PDF documents (over 100 pages each) using the **Alibaba-NLP/gte-large-en-v1.5** embedding model and Retrieval Augmented Generation (RAG) Assistant using **LLama-3-8b-chat** LLM for user queries.
- Two-Level Classification: Implemented a two-level classification pipeline: first level with 8 categories and second level with 10 categories.
- Embedding and Storage: Generated document embeddings and stored them in Milvus database with metadata as class labels, using separate collections for each classification level.
- Classification Process: For new PDFs, calculated embedding, compared it with the Milvus database, and assigned class labels based on similarity scores.
- RAG Assistant Implementation: Developed a Retrieval Augmented Generation (RAG) Assistant using **LLama-3-8b-chat** model to create user-readable answers based on retrieved document chunks.
- Classification Speed Enhancement: Reduced document classification time significantly from **2-3 days** to less than **15 seconds**, leveraging optimized embedding retrieval and processing.
- Accuracy Achievement: Achieved **92%** accuracy in document classification and **95%** accuracy in RAG.
- Semantic Cache: Integrated semantic cache for faster response times.
- Deployment: Deployed the solution as a FastAPI application on IBM Code Engine, containerizing the entire system.
- Business Impact: Successfully drove **\$18.9M** in project revenue.

2. DB-Schenker: Shipping Instruction Entity-Extraction and Comparison

- Developed a system to accurately extract shipping instruction entities from unstructured PDF documents, returning results in JSON format. Entities included Carrier, Shipper Name and Address, Consignee Name and Address, and others crucial for logistics operations.
- Automated entity extraction using the Watsonx Foundation Model, **Llama-3-70b**, and compared results with ground truth data using **IBM's Granite** model. Utilized metrics like Jaccard Similarity, Levenshtein Distance, and Cosine Similarity for accuracy assessment.
- Evaluated extraction accuracy using advanced similarity metrics and embedding-based comparisons, ensuring high fidelity to ground truth.
- Implemented as a robust microservice using FastAPI, allowing user-friendly PDF upload, entity extraction, and detailed report generation with color-coded accuracy indicators.
- Developed a user interface for streamlined file upload, entity extraction, and token cost prediction, showcasing the capabilities of IBM's Granite model effectively.
- Successfully completed the pilot within a tight timeline of 3 weeks, securing a contract worth **\$1.6 million** for DB-Schenker. Overcame challenges such as parsing semi-structured PDF data using Tesseract and managing context length through innovative solutions like few-shot examples and reusable code formats.

TietoEvry (India) - Machine Learning

 March 2022 - July 2023

 Bangalore, India





1. Voice Enabled Dashboard

- Implemented an LLM-based Database **SQL Agent** to convert speech into natural language queries, further translated into SQL queries using Generative Pre-trained Models (GPT-3.5 and T5).
- Leveraged the **Whisper** model for converting speech to text, enhancing the accuracy and reliability of input data.
- Designed and developed a **semantic search engine** powered by the BERT model to extract relevant keywords from user queries.
- Utilized Chart.js for UI development, enabling visualization of data through dynamic charts and graphs.

SKILLS

- Programming Languages:** Python
- Frameworks:** Langchain, Llama-Index, Django, Flask
- Database Management:** ChromaDB, Milvus, Weaviate, MySQL, SQL Lite, Microsoft SQL Server
- IDEs:** PyCharm, Visual Studio, VS Code
- Cloud Services:** AWS Bedrock, AWS Sagemaker, Azure ML, Google Vertex AI, WatsonX.ai
- Python Packages:** NumPy, Pandas, Matplotlib, SciPy, Skit-learn, OpenCV, Tensor Flow, Keras, Pytorch
- GenAI:** Prompt Engineering, LLM-Finetuning
- LLM:** Llama-3, Mixtral, Mistral, Granite, OpenAI GPT-4, ChatGPT
- NLP Skills:** Embedding Model Fine-tuning, OpenAI Whisper Speech to Text, Attention, Transformers, BERT
- Deep Learning:** MLP, CNN, RNN, LSTM, Encoders and Decoders, Seq2Seq, GANs
- Object Detection:** YOLO
- Deployment:** Docker
- Version Control:** Git/GitHub
- Soft Skills:** Elevator pitch, Stand and deliver, Public Speaking, Leadership, Teamwork and Presentation

CERTIFICATIONS

- Generative AI with Large Language Models** 
- Applied AI Course** 
- Neural networks and Deep learning** 
- Data Science Methodology** 


ACHIEVEMENTS


- Got Distinct Performer rating in appraisal cycle twice.
- Got appreciation for development of Conversational ChatBot on Organizational Data that created great value.
- Recognized as Quick learner for adopting LLM, Data Science, Data Engineering and Data Reporting.
- YouTube Channel with 4.5 thousand + subscribers

EDUCATION

Bachelor of Engineering in Information Science

SDM Institute of Technology, Ujire, Mangalore.

 Aug 2017-Aug 2021

 Dakshina Kannada

CGPA: 8.36/10