Bhargava Ram K

Hyderabad, India 500039 | +91 8147631042 | bhargavaram443@gmail.com

PROFESSIONAL SUMMARY

Artificial Intelligence & Machine Learning Engineer | Data Scientist | Generative AI Consultant & Prompt Engineer with ~4 years of experience specializing in Natural Language Processing (NLP), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG).

Proven ability to transform unstructured clinical, payer, and patient data into actionable insights that deliver measurable business value. Adept at building and deploying production-grade AI/ML solutions in healthcare and enterprise environments, with deep expertise in Prompt Engineering, Context Engineering, Vector Search, and knowledge-grounded Q&A systems.

Hands-on experience with industry-leading tools and frameworks including Azure OpenAI, Google Gemini, OpenAI GPT, LLaMA, Hugging Face Transformers, LangChain, BigQuery, and Cosmos DB. Skilled in designing scalable ML pipelines while adhering to strict data privacy and compliance standards (e.g., HIPAA).

Passionate about bridging the gap between state-of-the-art AI research and real-world business applications, driving tangible improvements in clinical decision support, operational efficiency, and enterprise automation through innovative GenAI solutions.

SKILLS

Programming & Data Handling: Python | SQL | Data Analysis & Visualization | Data Preprocessing & Cleaning | Pandas | NumPy | Matplotlib | Seaborn | Plotly
Natural Language Processing (NLP): Text Preprocessing | Named Entity Recognition (NER) | OCR (Tesseract, AWS Textract) | Spacy | NLTK | Whisper (Speech-to-Text) | Topic Modeling | Sentiment Analysis | Clinical NLP | Custom Entity Extraction
LLMs & Generative AI: Large Language Models (LLMs) | Prompt Engineering | Context Engineering | Retrieval-Augmented Generation (RAG) | Few-Shot & Zero-Shot Learning | In-Context Learning | Chain-of-Thought Prompting
LangChain | LlamaIndex | OpenAI API | Google Gemini | Azure OpenAI | Anthropic Claude | Mistral | LLaMA | Hugging Face Transformers | BERT, RoBERTa, GPT | LLM Observability | DeepEval | Prompt Layer | Model Evaluation & Alignment
Vector Search & Embeddings: Pinecone | FAISS | ChromaDB | Weaviate | Azure Cosmos DB | Vector Indexing |OpenAI Embeddings | HuggingFace Transformers | Sentence Transformers| SBERT
Machine Learning & Deep Learning: Supervised & Unsupervised Learning | Model Training & Evaluation | Feature Engineering | Scikit-learn | XGBoost | LightGBM | TensorFlow | Keras | PyTorch | Neural Networks (ANN, CNN, RNN, LSTM) | Transformer Models | PCA | Recommendation Systems | Clustering | Classification | Regression | Time-Series Forecasting

MLOps & Production Deployment: Model Packaging & Serving | ML Pipelines | FastAPI | Docker | Git | Model Monitoring & Drift Detection| Logging & Telemetry
Cloud & Platforms: Azure (OpenAI, Cosmos DB, Functions) | Google Cloud (Vertex AI, BigQuery) |Jupyter | VS Code | GitHub | Streamlit
Soft Skills & Collaboration: Problem Solving | Analytical Thinking | Effective Communication | Team Collaboration | Product Thinking | Cross-Functional Stakeholder Interaction

WORK HISTORY

Sr DATA SCIENTIST | Feb 2024 – Present | IKS Health | Hyderabad, India
DATA SCIENTIST | Jun 2023 – Sep 2023 | DTC Infotech Private Limited | Bengaluru, India
DATA SCIENTIST | Dec 2021 – Mar 2023 | Cognizant | Bengaluru, India

PROJECTS

Project 1: Developed and deployed a RAG-based Q&A bot utilizing Azure OpenAI GPT for policy-related queries, achieving highly accurate and context-aware answers.

- Extracted relevant information from a policy document and structured it into a knowledge base stored in Azure Cosmos DB, ensuring scalable and fast retrieval for user queries
- Integrated OpenAI GPT models enabling the bot to generate human-like, accurate, and contextually appropriate answers based on the knowledge base.
- Developed a Retrieval-Augmented Generation (RAG) mechanism, combining information retrieval (from Cosmos DB) with GPT-powered generation for context-aware responses.

Project 2: Developed an automated clinical text summarization solution using Gemini (AI model) and prompt engineering to extract key insights from large clinical chart notes and transcripts.

- Designed and implemented a system that automatically generates concise, readable summaries from clinical chart notes, transcripts, and other unstructured medical text data using Gemini AI and advanced prompt engineering techniques.
- Utilized prompt engineering to focus Gemini's capabilities on extracting critical medical information, such as patient symptoms, diagnoses, medications, and treatment plans, helping clinicians to quickly identify key details without reading through lengthy notes.
- Achieved a 30% reduction in review time by significantly improving the efficiency of chart note reviews, enabling clinicians to focus on critical aspects of patient care with reduced cognitive load.
- The system improves readability, reduces review times, and enhances clinical decision-making efficiency, ultimately streamlining workflow for healthcare professionals.

Project 3: Developed an automated Patient Profiling and Summarization Platform using Gemini AI and BigQuery to aggregate, process, and summarize comprehensive patient data from multiple sources. The platform provides concise, actionable summaries that support clinical decision-making, optimize data access, and improve overall patient care efficiency.

- Integrated data from multiple healthcare systems (EHR, lab results, medical imaging, etc.) into a unified patient profile using BigQuery for scalable data processing, ensuring all relevant information is included in the summary.
- Leveraged Gemini AI for contextual text summarization, transforming complex medical data into clear, readable, and actionable patient summaries, which helped clinicians make faster, data-driven decisions.
- Delivered concise, real-time patient summaries that include critical information such as medical history, diagnosis, medications, and treatment outcomes, ensuring that healthcare professionals have immediate access to the most relevant patient details.

Project 4: Productivity Enhancement through Voice Activity Detection Model Audio Processing Efficiency by 20%

- Enhanced audio preprocessing significantly improved model input quality for downstream tasks such as transcription and summarization.
- Developed and fine-tuned Voice Activity Detection (VAD) model using SpeechBrain to identify conversational and non-conversational segments in audio files, improving productivity by 20%.

Project 5: Detected DI Flag Issues through description analysis

- Data cleaning is done by removing hash tags, html tags and special characters, numeric.
- Stop word removal, tokenization, Stemming, and lemmatization are performed in the Phase of Text Preprocessing.
- Bag of Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF) techniques are applied to extract the features from text data.
- Evaluated the model using a confusion matrix and tested the model for 30days and after validating the accuracy for those 30 days, pushed the model to production.

Project 6: Predicting Buy Box Price and Sales price

- Checked for linear relation, autocorrelation, multicollinearity, homoscedasticity and normally distributed errors for satisfying Linear Regression.
- Implemented Dimension Reduction Technique (Principal Component Analysis) on analytical data to improve accuracy by adjusting for bias and variance trade-off.
- I identified the top 5 drivers and bottom 5 drivers and Computed R2, Adj R2, RMSE, MAE, and MSE and shared results with clients.

EDUCATION

B.Tech | Computer Science from SCSVMV University Kanchipuram, 2020