LIKITH VISHAL BODDEDA

(408) 752-6185 | likithvishal20@gmail.com | GITHUB |

SKILLS

Languages: Java, C#, C/C++, Python, JavaScrip, React, TypeScript, HTML/CSS

Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, NumPy, Pandas, Matplotlib, Seaborn, NVIDIA TensorRT, NVIDIA Triton Inference Server, NVIDIA DeepStream SDK, NVIDIA Jetson Platforms, OpenCV

Operating Systems: Windows, Ubuntu(Linux), MacOS, Redhat

Collaboration Tools: Git, Jira, Slack

Other Tools:Docker, MLOps Pipelines, Model Monitoring, Data Pipelines, Experimental Design & Evaluation, A/B Testing, Hyperparameter Tuning

PROFESSIONAL EXPERIENCE

VSD Infotech Remote

ML Engineer **Oct 2024 – Present**

- Pioneered the development and seamless integration of cutting-edge multimodal AI solutions, meticulously blending real-time video, lidar, and diverse sensor data to forge a superior understanding of complex autonomous environments.
- Achieved a remarkable 40% boost in model inference performance by expertly leveraging NVIDIA TensorRT, ensuring robust, high-throughput, and low-latency deployments via NVIDIA Triton Inference Server.
- Served as a pivotal contributor to critical team initiatives, consistently offering profound technical insights and propelling key advancements in AI model conceptualization and operationalization.
- Authored and contributed to the evolution of a groundbreaking attention-based model, significantly enhancing predictive analytics capabilities within intricate video streams.
- Championed and successfully integrated Retrieval-Augmented Generation (RAG) paradigms, transforming intelligent video content into concise summaries and highly accurate query responses.
- Designed and implemented scalable MLOps pipelines, automating model training, validation, and deployment, reducing iteration cycles by 35%.
- Collaborated closely with product teams to translate complex business requirements into tangible AI features, ensuring alignment with strategic objectives.
- Optimized resource utilization across GPU clusters, leading to a 20% reduction in operational costs for AI inference workloads.
- Developed custom performance monitoring dashboards for deployed models, enabling proactive identification and resolution of anomalies.
- Participated actively in architectural discussions for future AI system enhancements, contributing to the long-term technology roadmap.

ACADEMIC PROJECTS

Image-model-application-using-LLMS-and-Gemini-Pro

- Developed an end-to-end multimodal AI application leveraging Gemini Pro, integrating both text and image analysis for comprehensive generative capabilities in Q&A, summarization, and image description.Implemented a scalable, interactive interface using Streamlit for seamless user interaction with the AI model, achieving real-time response for 60 queries per minute.Built a robust back-end architecture to handle API integration with Google Generative AI, optimizing the system for efficiency in text and image processing tasks.Enhanced model deployment efficiency by setting up environment management and API security through .env configurations, streamlining the generative AI workflow

CalorieAdvisor GENAI Doctor

- Developed an AI-powered health application leveraging Google Gemini Pro Vision API to analyze real-time meal nutrition. The app enables users to upload food images and instantly receive a detailed breakdown of calories, macronutrients (carbohydrates, proteins, fats), fiber, and sugar content. It provides actionable insights into meal balance, highlights nutritional imbalances, and offers personalized health recommendations. Implemented using Streamlit for UI, integrated API calls for real-time image-based analysis, and ensured secure API key handling through .env configurations. This project showcases expertise in AI integrations, API utilization, and creating impactful user-centric applications.

Sentiment Analysis with ChatGPT and Google Gemini APIs

- Implemented a robust sentiment analysis pipeline leveraging OpenAI's GPT-4 (ChatGPT) and Google Gemini APIs for classifying sentiments in over 10,000 textual samples. Utilized Zero-Shot and Few-Shot Learning techniques, achieving 90% accuracy and 88% precision in sentiment classification. Enhanced performance by 25% through optimized preprocessing, including lowercasing and removing non-alphanumeric characters. Successfully integrated APIs for scalable sentiment classification, leveraging free API credits. Evaluated model performance using metrics like accuracy, precision, recall, and F1 score, showcasing expertise in LLM integration and real-time text analysis workflows.

Blog Generation LLM

- Implemented a blog generation app using the LLaMA 2 model, achieving 90% content relevance and reducing response times to under 2 seconds. Integrated user customization features, increasing engagement by 25%. Managed the entire lifecycle, from training on a 10,000-sample dataset to deployment.

Text Summarization and Classification

- Developed a news summarization system using TextRank and K-Means Clustering, achieving 85% precision and 82% F1-score on 1,000+ articles from BBC News and Reuters. Applied Latent Semantic Analysis (LSA) for topic modeling and tested with Logistic Regression, Random Forest, and SVM. Utilized NLTK, Scikit-learn, and Gensim for implementation.

FUTURE WORKS

AI-Driven Healthcare Documentation & Knowledge System Developing an end-to-end intelligent medical documentation system integrating speech recognition, natural language processing, and large language models to enhance clinical workflows and patient-provider communication.

Core Technical Contributions:

Engineering real-time audio transcription and summarization system using transformer-based models (Whisper, T5) with domain-specific vocabulary enhancement for improved medical terminology accuracy. Building an intelligent Q&A system combining vector database retrieval (FAISS) with fine-tuned LLMs (Llama 3, Med-PaLM) for contextual, fact-grounded medical information retrieval. Creating modular AI architecture coordinating retrieval, reasoning, and explanation components for clinically validated responses. Ensuring HIPAA and GDPR compliance through secure data architecture, encrypted storage, and real-time EMR integration capabilities. Developing bias detection pipelines and human-in-the-loop feedback systems for continual model improvement and evidence-based output verification. Deploying web-based platforms enabling interactive querying of AI-generated medical documentation. Creating a scalable, privacy-compliant AI system demonstrating practical applications of generative AI in healthcare, with emphasis on transparency, safety, and clinical utility.

EDUCATION

Westcliff University — CA, USA
Master of Science in Computer Science | *GPA 3.81/ 4.0* — Mar 2022 - Feb 2024

Gandhi Institute of Technology — Andhra Pradesh, India
Bachelor of Technology in Computer Science & Engineering — Jun 2017 - May 2021

PUBLICATIONS

- Gupta, Savyasachi, and Likith Vishal Boddeda. "Deep Learning approach to Diffusion-based Image Classification with Timestep Conditioning." (Under review at 1st IEEE International Conference on Recent Trends in Computing and Smart Mobility hosted by Maulana Azad National Institute of Technology (MANIT) in Bhopal, India)
- Pasam, J.R., Kasumurthy, S.R., Boddeda, L.V., Mandava, V., Sana, V.V. (2022). Vision for Eyes. In: Misra, R., Kesswani, N., Rajarajan, M., Veeravalli, B., Patel, A. (eds) Internet of Things and Connected Technologies. ICIoTCT 2021. Lecture Notes in Networks and Systems, vol 340. Springer, Cham
- Likith Vishal B, Subhadra K, Sivalaya G.A Comparative Study of Classification Algorithms over Images using Machine Learning And Tensorflow
- Boddeda, Likith & Pragathi, M & Amulya, P & Kumar, N. Suresh. (2020). Image Classification Using Neural Networks and Tensor-flow. Test Engineering and Management. 83.