BHARGAV PRASAD KALICHETTI

bhargavprasad9814@gmail.com | www.linkedin.com/in/bhargav-prasad-kalichetti-183552257

Experienced AI Engineer with 5+ years of experience in Generative AI, LLM fine-tuning, RAG/Graph RAG, and multi-agent systems, with strong programming expertise in Python, C, and C++. Skilled in building and deploying enterprise-scale AI solutions using Azure, AWS, Kubernetes, and GPU optimization. Passionate about applying MLOps/LLMOps best practices to deliver scalable, secure, and high-performance AI systems that drive business transformation.

Work Experience

Cisco System | Vzure Solutions

***AI developer*, Texas, USA, (Aug 2024 – Present)**

- Built and Led **Python-based no-code/low-code chatbot platform using Django REST + React**, enabling enterprises to create custom assistants with **OpenAI GPT and local NVIDIA Llama models**; featured real-time chat, drag-and-drop workflow builder, model switching, and validation.
- **Designed and deployed multi-agent systems with LangGraph, CrewAI, AutoGen, and LangChain**, where agents collaborated on domain-specific workflows (manufacturing, analytics, reporting). Implemented an **Agent-to-Agent (A2A) communication protocol to enable seamless coordination**, message passing, and distributed decision-making across agent networks.
- **Integrated persistent memory using Mem0, pgvector, and ChromaDB (alongside Weaviate & Pinecone)** into Python multi-agent systems, enabling **long-term context retention** and **multilingual interactions** for better decision-making.
- Integrated Mem0 for persistent memory management in multi-agent systems, enabling context retention and improved decision-making across agent interactions.
- Integrated **MCP (Model Context Protocol)** into multi-agent systems to enable dynamic tool registration and runtime tool addition, eliminating dependency on OpenAI function calling while providing flexible, protocol-based agent interactions and seamless tool management across distributed agent architectures.
- **Developed a database agent from scratch with LangChain + LangGraph** to interact with **manufacturing relational data across PostgreSQL, Oracle, and SQL Server**. Leveraged local Llama, Anthropic, and OpenAI models to enable natural language querying, **Azure SQL generation, data analysis, and automated reporting** without external API dependencies.
- **Fine-tuned Llama-3B (Hugging Face) with LoRA and QLoRA** on proprietary **manufacturing domain data and SQL query logs, enabling** the model to generate optimized SQL queries for manufacturing databases. Delivered 35% higher domain-task accuracy and 50% lower inference latency, improving analytics and reporting efficiency for enterprise customers.
- Implemented a **custom RAG system for the database agent** that stores and indexes database schemas in **ChromaDB and pgvector**, enabling intelligent schema retrieval and context-aware SQL query generation for improved accuracy and performance in manufacturing data analysis.
- **Constructed an Azure model optimization pipeline** using **ONNX**, **NVIDIA TensorRT**, **TensorFlow Lite** (for IoT Edge), and **Apache TVM** for efficient, cross-hardware deployment—resulting in 3× faster inference and 50% lower compute cost.
- **Incorporated GPU partitioning techniques** where NVIDIA GPU time-slicing used for shared workloads and **MIG (Multi-Instance GPU)** when hardware allowed—optimizing GPU utilization and reducing cost per job.
- **Implemented smart GPU-aware autoscaling** in AKS using the Luna autoscaler—optimizing instance choices per GPU slice allocation, reducing GPU provisioning costs and improving utilization.
- Enabled observability with **eBPF-enhanced tooling, along with Prometheus + Grafana dashboards**, delivering real-time metrics, alerts, and insights with low performance impact.

University of North Texas

***Research and Instructional Assistant*, Texas, USA, (March 2023 –Aug 2024)**

- Researched and developed a cybersecurity knowledge graph, training custom Llama 3 models on structured domain data, improving expertise classification accuracy by 30% through graph-enhanced LLM training methods.
- Designed and implemented a **Graph RAG system with LangChain and Neo4j** to retrieve knowledge from cybersecurity interviews; enabled multi-hop reasoning and complex relationship queries, achieving a **40% improvement in context-aware response generation**.
- Built a **Graph RAG system using LangChain and Neo4j** to enhance knowledge retrieval from cybersecurity interview data, enabling complex relationship queries and multi-hop reasoning across connected entities, resulting in 40% improvement in context-aware response generation and enhanced semantic understanding of cybersecurity expertise networks
- Conducted comparative research on transformer architecture using PyTorch and TensorFlow, developing and evaluating custom BERT variants for cybersecurity entity extraction, achieving 15% performance improvement over baseline models
- **Deployed the Graph RAG system on Azure Kubernetes Service (AKS)** with Dockerized microservices, CI/CD pipelines, and monitoring via Grafana, enabling scalable, secure, and production-ready knowledge retrieval for cybersecurity datasets.

Wipro

***Data and Network Analyst* | Remote, India, (Feb 2022 - Dec 2022)**

- Designed and deployed an **AI-assisted risk assessment framework** for secure online banking systems by combining **traditional ML models with network security protocols (NMAP scans, firewall policies, load balancing rules)**; delivered a **40% reduction in exploitable vulnerabilities** and strengthened regulatory compliance (PCI-DSS, ISO 27001).
- Built and operationalized **machine learning threat detection pipelines** on **AWS SageMaker**, leveraging **Random Cut Forest for anomaly detection** and **XGBoost for fraud pattern recognition**; achieved a **25% improvement in early threat detection**

rates and reduced false positives in banking security alerts.

- Integrated ML-driven insights with **SQL Server databases and SIEM systems (Splunk, ELK)** to provide real-time dashboards for security teams, improving incident response times by 30%.

Amara Raja Power Systems

***Research and Development Engineer*, India, (Aug 2020 - Dec 2021)**

- Developed and optimized firmware for ARM and STM32 microcontrollers using Embedded C/C++, creating lightweight graphical user interfaces (GUIs) and implementing advanced communication protocols (SPI, I2C, UART, CAN); these enhancements increased battery life by 20 % and improved user interaction and energy efficiency in battery-management systems
- Implemented **predictive battery lifespan analysis and real-time fault detection algorithms** using Python and Embedded C; leveraged **classical ML models (Random Forests, Regression, and SVM)** for anomaly detection on sensor data streams, reducing downtime and operational costs.
- Designed a **cloud-based IoT monitoring system on AWS IoT Core and Azure IoT Hub (2020-era platforms)** to collect telemetry from power electronics devices, enabling **cloud-hosted analytics and visualization dashboards** for proactive maintenance.
- Applied **time-series ML models (ARIMA, LSTM prototypes in TensorFlow 2.0)** on battery usage data to predict failures and optimize charging cycles, improving reliability in field deployments.
- Developed lightweight graphical user interfaces in **C++ for ARM/STM32-based battery-management systems**, integrating real-time data via **SPI/I²C/UART protocols**. These GUIs improved usability and enabled operators to monitor and adjust system parameters without compromising the 20 % battery-life gain achieved through firmware optimization.

Projects

- **Phone Recommendation System**: Developed a **personalized recommendation engine using Flask**, implementing collaborative and content-based filtering algorithms. Integrated machine learning models that use scikit-learn libraries to analyses user preferences and usage patterns, enhancing the accuracy and relevance of phone recommendations.
- **Advanced Multimodal Image Captioning System**: Built a production-scale image captioning system **using SWIN Transformer for visual features and GPT-3 for text generation, trained on 100K+ images**. Achieved an 89% BLEU score improvement with custom attention mechanisms and deployed a scalable inference pipeline for generating context-aware captions.

Skills and Areas of Expertise

- **Programming Languages**: Python, C/C++, SQL, Embedded C, TypeScript, ReactJS
- **Frameworks & Libraries:** PyTorch, TensorFlow, Keras, Scikit-learn, Spacy, Hugging Face Transformers, Pandas, NumPy, LangChain, CrewAI, LangGraph, AutoGen,Django ,OpenCV, MaterialUI, Plotly, Streamlit, Chainlit
- **Database Technologies**: PostgreSQL, pgvector, MongoDB, ChromaDB, Neo4j, Vector Databases
- **MLOps, LLMOps & Model Optimization:** LLM Fine-tuning (LoRA, QLoRA, PEFT), RAG & Graph RAG, RLHF, Agentic RAG, Multi-Agent Systems, MCP (Model Context Protocol), Prompt Engineering, Chain-of-Thought, GPU Slicing (MIG, Time-Slicing), ONNX, NVIDIA TensorRT, Apache TVM, TensorFlow Lite
- **DevOps, Deployment& APIs:** Docker, Kubernetes (AKS, GPU-aware autoscaling, Helm), NVIDIA NIM Containers, Triton Inference Server, CI/CD Pipelines, Container Orchestration, Azure Arc, Azure Functions, AWS SageMaker, API Development & Integration (REST, GraphQL, gRPC)
- **Version Control & Tools:** Git, GitHub, Bash,GitLab DVC for Data Versioning
- **Software & IDEs**: Linux, MATLAB, Jupyter Notebook, Visual Studio, STMCube-IDE, N8N, MLFlow

Certifications

- Udemy - Java 17 Master Class
- Udemy – Machine Learning: Natural Language Processing in Python
- IBM- Deep Learning with Tensor Flow (2021) Online Course
- Intel- Deep Learning with Multimodal RAG: Chat with Videos
- Simplilearn- Deep Learning (2021) Online course
- Simplilearn- Al Capstone (2021) Online course
- Unlock the Future: Mastering Generative AI, MLOps, AIOps - LLMOps with Open AI and Hugging Face Models Deploy to Prod

Education

- **University of North Texas,** *Master's in Artificial Intelligence* | Texas, USA | GPA: 3.65 / 4.0, (December 2024)
- **Annamacharya Institute of Technology and Science**, *B. Tech in Electronics and Communication Engineering* | Andhra Pradesh, India