JITEN BHALAVAT

jbha0504@umd.edu | (240) 481-5543 | LinkedIn | GitHub | jitenbhalavat.com

EDUCATION

University of Maryland, College Park **Aug 2024 - Expected May 2026**
Master of Science in Applied Machine Learning *College Park, Maryland*
Coursework: Machine Learning, Deep Learning, Natural Language Processing, Advanced Optimization, Cloud Computing

Charotar University of Science and Technology (CHARUSAT) **Sept 2020 - May 2024**
Bachelor of Technology in Information Technology *Gujarat, India*
Coursework: Data Structures, Full-Stack Development, Database Management System, Network Security

SKILLS

Core Competencies: Cloud Service Development, Distributed and Parallel Systems Computing, Backend Development, Data Science, Web Development, Algorithm Design, Database Management, Cloud Infrastructure (distributed storage), Integration Testing, CI/CD, Microservices, Shell Scripting
Languages and Databases: Python, C/C++, Java, JavaScript, SQL, MySQL, MongoDB, HTML, Node.js, React
Frameworks and Libraries: PyTorch, TensorFlow, Keras, HuggingFace, Scikit-learn, NumPy, Pandas, Transformers
Generative AI: LLMs, AI Agents,Vector Databases, Prompt Engineering, Fine-tuning, LangGraph, CrewAI, Autogen.
Tools & DevOps: Git, GitHub, Docker, Kubernetes, AWS, GCP, Azure, FastAPI, Flask, Tableau, PowerBI, GIT (version control), Hadoop, Kafka, Spark, Operating Systems, Bash, Linux, UBUNTU, Powershell,Postman, DevOps

EXPERIENCE

Machine Learning Engineer Intern **Sept 2023 - Apr 2024**
Plutomen Technologies Pvt. Ltd. *Gujarat, India*

- Built **anomaly detection** using **statistical** and **machine learning techniques** within an **Agile software development lifecycle**, reducing false positives by **25%** and improving system reliability and observability.
- Architected **Retrieval-Augmented Generation (RAG)** data processing pipelines, boosting data extraction accuracy from 45% to 89%, eliminating **15 hours/week** of manual compliance review, using **LlamaIndex and LangChain**
- Improved chatbot answer relevance by **30%** through **optimized chunking** and semantic search using **Qdrant, Pinecone**
- Developed and deployed Flask-based **REST APIs** with **PostgreSQL**, enabling real-time inference under **200ms latency**

Research Assistant **Apr 2023 - Jun 2023**
Charotar University of Science and Technology *Gujarat, India*

- Trained a U-Net CNN architecture for MRI image segmentation, achieving **91% accuracy**
- Evaluated model performance using IoU and accuracy metrics, outperforming baseline CNNs by **15%**

Machine Learning Engineer Intern **May 2022 - Jun 2022**
NXON Pvt. Ltd. *Gujarat, India*

- Developed transformer-based **AI solution** using Parameter-Efficient Fine-Tuning **(PEFT)** with **LoRA** on T5, reducing trainable parameters by **99%** while achieving 60% faster training convergence on **distributed multi-GPU infrastructure**
- Built scalable **data processing pipeline** with **PyTorch DDP** across 60K+ training steps, implementing automated checkpointing, metrics monitoring, and **model evaluation workflows** that reduced debugging cycles by **40%**
- Optimized model performance across 10 programming languages using **CodeBLEU** evaluation metrics and **beam search**, enabling robust multi-language code intelligence

PROJECTS

ClassTopper: Full Stack Multi-modal (Text + Video) RAG Platform

- **Led 3-person team** building **multi-modal RAG** chatbot for videos and documents, improving study efficiency by 40%
- Enhanced content retention by **60%** via mind-mapping feature, visualizing topic relationships for better knowledge recall
- Engineered adaptive quiz system generating personalized assessments from knowledge gap analysis, boosting test performance by 27% and identifying weak areas with **85% accuracy**

Cloud LLM Inference Benchmark

- Implemented **async Python scripts** to load-test **LLM inference**, measuring throughput and latency percentiles
- Deployed **GPU-accelerated inference** on AWS EC2 (NVIDIA A10G) with **Docker and CUDA**
- Benchmarked **vLLM vs SGLang** frameworks for GPU memory and concurrency optimization

Expense Tracker using MCP

- Deployed MCP server on FastMCP Cloud, prototyping agent-based workflow for finances through multi-agent tooling
- Integrated **MCP Client** as **AI-assisted workflow**, allowing users to query, and analyze expenses via conversational AI
- Leveraged a lightweight **SQLite backend** for fast, local, and secure storage with instant retrieval and real-time insights.

CERTIFICATIONS AND ACHIEVEMENTS

Google Cloud Computing - Google Cloud Educational Badges (9 Badges)
AWS Academy Graduate - AWS Academy Machine Learning Foundations and AWS Academy Cloud Developing