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*

Charotar University of Science and Technology (CHARUSAT) — **Sept 2020 - May 2024**
Bachelor of Technology in Information Technology — *Gujarat, India*

SKILLS

Languages and Databases: Python, C/C++, JavaScript, SQL, MySQL, MongoDB, Pinecone, Qdrant, PostgreSQL
Frameworks and Libraries: PyTorch, TensorFlow, Keras, HuggingFace, Scikit-learn, NumPy, Pandas, Transformers
Generative AI: Large Language Models (LLMs), AI Agents, RAG, Prompt Engineering, Fine-tuning
Tools & DevOps: Git, GitHub, Bitbucket, Docker, Kubernetes, AWS, GCP, Azure, FastAPI, Flask
Data Science: Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, Cloud Computing
Agentic AI: Tool Calling, Google ADK, A2A, LangGraph, CrewAI, Autogen, and Open AI Agentic SDK

EXPERIENCE

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

- Architected **Retrieval-Augmented Generation (RAG)** document parsing pipeline, 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**
- Conducted **model testing, monitoring, and failure analysis** (15+ cases), improving response quality by **12%**
- Collaborated with cross-functional engineering teams to integrate **AI solutions** into production systems

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%**
- Applied advanced **AI techniques** in medical imaging, demonstrating the impact of AI professionals in healthcare

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 MLOps pipeline** with **PyTorch DDP** across 60K+ training steps, implementing automated checkpointing, 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 benchmarking scripts** to load-test **LLM inference servers**, measuring throughput (TPS/RPS) and latency percentiles using NumPy
- Deployed **GPU-accelerated inference** on AWS EC2 (NVIDIA A10G) with **Docker and CUDA**
- Conducted comparative analysis of GPU memory and concurrency strategies, evaluating **vLLM vs SGLang** using parallel request orchestration and semaphore-based controls

Expense Tracker using MCP

- Deployed a custom **MCP server** on FastMCP Cloud, enabling MCP Client to manage finances through natural language
- Integrated **Claude Desktop MCP Client**, allowing users to query, summarize, 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