Bhushan Raju Golani

AI Research Engineer | ex-Amazon SDE | Machine Learning & Data Science | MS Data Science @ UMich

brgolani@umich.edu | bhushan-golani | bhushan-golani | (+1)7347543233 | Ann Arbor

Education

University Of Michigan, Ann Arbor **08/2025 – 05/2027**
Master Of Science In Data Science
Coursework: Natural Language Processing, Deep Learning, Machine Learning, Database Management Systems

Indian Institute of Technology (IIT), Kharagpur **06/2018 - 05/2023**
Bachelor and Master of Technology in Mechanical Engineering (Minor in Computer Science) CGPA: 8.95/10
Coursework: Machine Learning, Computer Vision, Image Processing, Algorithms, Operating Systems

Experience

AI Research Engineer | University Of Michigan | Professor Scott Pletcher **07/2025 - Present**

- Architected a configuration-driven **multi-agent AI platform** that unifies OpenAI, Anthropic, Gemini, DeepSeek, and Ollama behind a single agent interface, then composes them into specialized teams.
- Designed and implemented a **ChromaDB vector database** with contextual chunking, hybrid BM25+embedding search, reciprocal-rank fusion, and cross-encoder reranking for PubMed, web-search, and long-term conversation data.
- Built a long-term **conversation memory system** that streams chat history into a persistent ChromaDB store, supports semantic search over past sessions, to reduce the prompt size.
- Developed a **React-based UI** and backend endpoints (e.g., chat, finalize session) to provide an interactive web interface for users to work with the multi-agent models..

Software Developer | Amazon | Alexa Sensitive Content Intelligence Team **07/2023 - 07/2025**

- Developed highly scalable cloud services in **Java** using **AWS Lambda, ECS** and **NoSQL Database (DynamoDB)**, achieving **99.99%** uptime and processing over **50,000 TPS** with efficient load balancing, caching and data optimization.
- **Prompt Engineered** a vision-enabled classification service using **Amazon Bedrock (Claude Sonnet 3.5)** to perform sentiment, compliance, and risk assessments by combining embeddings, and LLM-based severity scores.
- Designed **multimodal AI workflows**, integrating image retrieval and LLM reasoning with latency **less than 1s**.
- Worked on an UI **ReactJS** app on ElasticSearch for data insights, supporting leaders to review the feedback flow.

AI/ML Projects

GPT-style Language Model & Conversational SFT | EECS 595 (NLP)

- Implemented a **GPT-style transformer** in **PyTorch** from scratch with **multi-head causal self-attention**, **Rotary Position Embeddings (RoPE)**, **SwiGLU** feed-forward layers, and **RMSNorm**.
- Built a scalable **pretraining pipeline** over **Hugging Face Arrow** datasets (FineWeb-Edu), including custom **GPTDataset** and **DataLoader** utilities with **mixed precision** and **gradient accumulation**.
- Designed a **supervised fine-tuning (SFT)** framework for conversational agents with special role tokens (`<|user|>`, `<|assistant|>`, `<|system|>`) and **selective loss masking** (labels = -100) to train only on assistant responses.

Child-Centered LLM Safety Evaluation & Alignment | CSE 595 Course Project

- Designed an end-to-end **LLM safety pipeline** for children (ages 6–12), creating a human-verified dataset of **intentionally unsafe child-style prompts** based on the **4Cs** framework, where models are expected to **safely deflect**.
- Benchmarked open LLMs (**Gemma, Llama, GPT-OSS, DeepSeek**) using a **GPT-5 LLM-as-a-judge** to score responses on **safety, age-appropriateness, empathy**, and **boundary-setting** in child-facing interactions.
- Trained a **safety classifier** to flag unsafe or weak deflections and applied **Direct Preference Optimization (DPO)** with GPT-5 gold answers to strengthen **child-centered safety alignment** and refusal behavior.

Technical Skills

Machine Learning: PyTorch, TensorFlow, Scikit-learn, Hugging Face, Transformers, LangChain, Pandas, NumPy, OpenCV
AI Technologies: Large Language Models, Vector Databases (ChromaDB), Semantic Search, RAG Systems
Programming: Python, SQL, Java, C++, CUDA
Cloud & Tools: AWS (SageMaker, Lambda, S3), Docker, Git, REST APIs

Publications

Mapping fluid structuration to flow enhancement in nanofluidic channels
Journal of Chemical Physics | Machine Learning Hits Molecular Simulations **06/2023**

- Designed **physics-informed neural network models** to predict flow enhancement factors using deep learning.

Prediction of Cloud Server Job Failures using ML-based KNN and LSTM
International Journal of Engineering Research And Technology **07/2021**

- Designed **KNN classification and LSTM modeling** pipeline achieving **94% accuracy** for cloud job failure prediction.