MEDHAVI MONISH

Gmail 9534715692 GitHub LinkedIn

Profile Summary Portfolio

AI/ML Engineer with 4 years of experience delivering production-ready GenAI features, fine-tuning LLMs (LLaMA), building lightweight ML engines in C++/CUDA, and compressing high-dimensional embeddings using Siamese networks. Proven expertise in designing agent-based RL systems, streaming-safe structured generation, and low-latency NLP pipelines. Passionate about building reasoning-first AI using memory, curiosity, and reward-driven modular frameworks.

Core Competencies

- LLM Fine-Tuning (SFT, PPO, DPO)
- GenAI Pipeline Architecture
- Retrieval-Augmented Generation (RAG)
- Low-Level ML Infra (C++/CUDA/OpenCL)
- Reinforcement Learning (Curiosity, PPO)
- Embedding Compression (Siamese Networks)
- Real-Time Inference & Streaming APIs
- Modular AI Agent Systems
- NLP-to-SQL Conversion & Code Interpreters

Technical Skills

- **Languages**: Python, C++, Java, C#, JavaScript
- **ML Frameworks**: PyTorch, TensorFlow, TRL, OpenCV
- **GPU Programming**: CUDA Kernels, OpenCL, Numba
- **Data & Backend**: Snowflake, SQL, MongoDB, Pandas
- **Simulation & 3D**: Unity3D, Blender
- **Web Technologies**: React.js, Angular
- **MLOps:** AWS SageMaker, EC2, Docker, GitHub Actions

Work Experience

PowerSchool — *Software Engineer I* (Apr 2023 – Present)

Promoted from Associate SE I & II

- **LLM Fine-Tuning**: Reproduced DeepSeek-R1 capabilities before ***their*** public release, using PPO and DPO to fine-tune LLaMA for YAML-based structured outputs with domain-specific reward shaping.
- **Siamese Embedding Compression**: Designed a MiniLM-based Siamese network fine-tuned on STS-B to reduce 384D vectors to compact representations, improving similarity search efficiency.
- **Vector-Free RAG**: Engineered a <150ms in-memory RAG pipeline using cosine similarity in pandas, eliminating dependency on external vector DBs.
- **PowerBuddy for Data Analysis Core Design**: Developed an NLP-to-SQL engine and a dynamic code execution interpreter inspired by OpenAI's architecture to support user-defined logic.
- **Streaming JSON**: Created a fault-tolerant streaming parser to convert partial JSON into structured JSON, enabling live updates in UIs with incomplete LLM output.
- **AI Tutor & Career Counselor**: Built LLM-driven academic tutor and career exploration tools, improving student engagement and decision-making in flagship products.

PowerSchool — *Associate Software Engineer II* (Apr 2022 – Mar 2023)

- Developed ERP APIs using ASP.NET (C#) + Angular and contributed to UI/UX enhancements.
- Identified for internal transfer to AI team after leading multiple UI/logic defect resolutions.
- Worked on student modeling, adaptive testing algorithms, and personalized content recommendations.

PowerSchool — *Associate Software Engineer I* (Jul 2021 – Mar 2022)

- Enhanced ERP modules by customizing field layouts and implementing header redesign.
- Supported minor production bugs and gained exposure to CI/CD workflows.

🧪 Personal AI Systems & Projects

- **Cortana++** *(In Progress)*: C++/CUDA-based ML engine built from scratch to support tensors, broadcasted matmul, and Dense layers, targeting low-memory environments. 🔗 GitHub
- **Darwin's Silicate Organism** *(Design Complete)*: Survival-driven modular AI using memory and reward models to simulate emergent behavior in constrained environments. 🔗 GitHub
- **TwinSqueeze** *(Active)*: Built a Siamese network with MiniLM-L6-v2 + NEFTune-style loss to compress 384D embeddings down to 64D, retaining task-specific semantic accuracy while reducing memory and compute. 🔗 GitHub
- **GreedyContext**: Built a semantic graph using sentence embeddings + greedy traversal to shrink LLM context by 50–90%, improving latency and reducing cost. 🔗 GitHub
- **Stream-JSON**: Developed a parser that streams and repairs incomplete YAML from LLMs into real-time structured JSON, enabling UI rendering on partial outputs. 🔗 GitHub
- **DeepFake Inference (OpenCL)**: Built a DeepFake pipeline using OpenCL on AMD GPUs, manually implementing ANN inference without external ML libraries. 🔗 GitHub
- **Unity3D Self-Driving Simulation**: Simulated PPO agents navigating GPS-defined tracks using a custom-designed car and road system in Blender + Unity. 🔗 GitHub
- **Genetic Algorithm for Conway's Game of Life**: Evolved seed configurations to achieve full-grid coverage using GA. 🔗 GitHub
- **Real-Time Face Recognition System**: Used OpenCV + TensorFlow to mark attendance via facial authentication and store logs in MySQL. 🔗 GitHub

🎓 Education

Master of Computer Applications (MCA), Birla Institute of Technology, Mesra — *CGPA 8.4* (2021)
Bachelor of Computer Applications (BCA), Birla Institute of Technology, Mesra — *CGPA 7.74* (2018)