MEDHAVI MONISH

Gmail 9534715692 GitHub LinkedIn

Summary

Machine Learning Engineer with 4 years of experience building optimized AI systems, including embedding compression using Siamese Networks, custom RAG, CUDA-based ML engines, and reinforcement learning workflows.
Specialized in fine-tuning large language models (e.g., LLaMA) and designing low-level ML infrastructure in C++/CUDA.
Passionate about reasoning-first AI, open-source ML engines, and modular agent frameworks driven by memory, curiosity, and survival behavior.

Technical Skills

Languages & Systems: C++, Python, C#, Java, JavaScript
Frameworks & Libraries: PyTorch, TensorFlow, TRL, OpenCV
GPU Programming Tools: CUDA Kernels, OpenCL, Numba
Simulation & 3D Tools: Unity3D, Blender
Data & Retrieval Tools: Snowflake, SQL, MongoDB
Web Technologies: React.js, Angular

AI Systems & Personal Projects

- **Designing and building Cortana++**, a lightweight ML engine in C++/CUDA with support for tensor operations, broadcasted matmul, reduction, and Dense layers. ***(In Progress)***
- **Architecting Darwin's Silicate Organism**, a modular AI framework inspired by survival-driven intelligence, designed around memory, reward, and curiosity. ***(Design complete; implementation pending)***
- **Developed GreedyContext**, a context-selection solution that reduces LLM token usage and latency by building a semantic trail of relevant messages using sentence embeddings, cosine similarity, and greedy graph traversal.
- **Developed a DeepFake pipeline** on an AMD GPU using OpenCL, manually implementing ANN inference without external ML libraries.
- **Built a self-driving car simulation** in Unity3D using a Blender-designed vehicle and track, with continuous-action PPO agents and GPS integration for navigating new routes.
- **Implemented a genetic algorithm system** to evolve minimal seed configurations for full-grid coverage in Conway's Game of Life.
- **Created a real-time face recognition attendance system** using OpenCV (detection), TensorFlow (recognition), and MySQL (logging).

Professional Experience

Software Engineer I – PowerSchool (2023–Present)

- Solely developed a DeepSeek-R1 replica before its official release, fine-tuning LLaMA using PPO with TRL, structured output enforcement (YAML), and domain-specific reward shaping for reasoning consistency.
- Independently engineered a vector-free RAG system with <150ms latency using pandas and cosine similarity.

- Designed and implemented a Siamese network to compress embeddings while improving similarity accuracy.
- Led core algorithm design for PowerBuddy, including:
  - NLP-to-SQL translation engine
  - A code interpreter system inspired by OpenAI's implementation
- Devised a robust method to stream partial JSON from LLMs, overcoming standard parser failures during incomplete generation.
- Trained LLMs using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), applying structured reward signals and output formatting constraints.
- Developed additional PowerBuddy tools, including an AI Tutor powered by GPT and a student career counselor to assist with academic guidance and exploration.

Associate Software Engineer II – PowerSchool (2022–2023)

- Built ERP module APIs using C# (ASP.NET) and Angular 4.
- Fixed production bugs, updated UI components, and contributed to CI/CD pipelines.
- Mid-year, **recommended by Director of ERP team to join AI team** for advanced product development.
- Contributed to AI-driven education features including:
  - ○ Knowledge tracing for modeling student understanding
  - ○ Debugging learning algorithms in the Learning Nav system
  - ○ Adaptive testing logic that adjusts question difficulty based on student progress
  - ○ Personalized content suggestions to enhance student learning outcomes.

Associate Software Engineer I – PowerSchool (2021–2022)

- Added some fields in a screen, fixed some production bugs.
- Changed Header bar design with new requirements.

Education

Master of Computer Applications (MCA), Birla Institute of Technology, Mesra – CGPA: 8.4