## OVERVIEW

Excelled with fine-tuning large language models, architecting GenAI pipelines, and developing performance-optimized ML infrastructure across both research and production environments. Targeting opportunities to leverage experience to build scalable, intelligent systems that solve complex real-world problems and drive product innovation.

# Medhavi Monish

**AI/ML ENGINEER | GENAI | LLM FINE-TUNING**

+91 9534715692
monishmedhavi@gmail.com
github/MedhaviMonish
Portfolio_Website
linkedin/medhavi-monish

## CORE COMPETENCIES

- Large Language Model (LLM) Fine-Tuning
- Generative AI (GenAI) Systems Design
- Retrieval-Augmented Generation (RAG)
- Deep Learning & Neural Network Architecture
- Low-Level ML Infrastructure
- Data Preprocessing & Feature Engineering
- Model Evaluation & Optimization
- Reinforcement Learning
- End-to-End ML Pipelines
- Simulation & Agent-Based Modeling
- Cloud & ML Ops
- Natural Language Processing (NLP)
- Cross-Functional Collaboration
- Agile Product-Oriented AI Delivery

## PROFILE SUMMARY

- AI/ML Engineer with **4 years** of experience in designing and deploying **Machine Learning systems**, including large-scale **GenAI models**, **reinforcement learning agents**, and infrastructure-level ML components using **C++/CUDA** and **Python**.
- Currently at PowerSchool as **Software Engineer I**, developing **GenAI-powered education technology** features such as **AI tutors**, **career counselors**, and **knowledge-tracing tools**, while leading **ML design** for flagship products.
- Lead Developer of **PowerBuddy**, an AI-powered analytics platform at PowerSchool, architecting core algorithms for **GPT-based data interpretation**, **NLP-to-SQL pipelines**, and **vector-free retrieval systems** optimized for sub-150ms latency.
- Skilled in **fine-tuning large language models (LLMs)** such as **LLaMA** using **Supervised Fine-Tuning (SFT)**, **PPO**, and **DPO techniques**, with hands-on experience on **AWS SageMaker** focusing on **structured outputs**, **reward shaping**, and **output consistency**.
- Proficient in **ML infrastructure**, building lightweight AI engines from scratch with **C++/CUDA**, implementing **tensor operations**, **matmul kernels**, and **ANN inference** on non-NVIDIA GPUs via **OpenCL**.
- Proven capability to deliver **end-to-end AI projects**, covering **data preprocessing**, **model training**, **evaluation**, and **deployment** for applications like **real-time face recognition**, **semantic context selection**, and **self-driving simulation agents**.
- Strong **cross-functional skills,** combining back-end engineering (**C#**, **SQL**, **Angular**), open-source frameworks (**PyTorch**, **TensorFlow**, **OpenCV**), and data tools (**Snowflake**, **MongoDB**) to create **scalable, production-ready AI systems**.

## TECHNICAL SKILLS

- **Programming Languages & Systems:** C++, Python, C#, Java, JavaScript
- **AI/ML Frameworks & Libraries:** PyTorch, TensorFlow, TRL, OpenCV
- **GPU & High-Performance Computing:** CUDA Kernels, OpenCL, Numba
- **Simulation, 3D, & Visualization Tools:** Unity3D, Blender
- **Data Engineering & Storage:** Snowflake, SQL, MongoDB
- **Web & Front-end Technologies:** React.js, Angular
- **Cloud & MLOps:** AWS (S3, EC2, SageMaker), Docker, GitHub Actions, Google Colab

## WORK EXPERIENCE

Jul’21
–
Present

### PowerSchool

Growth Path:

Associate Software Engineer I (Jul'21 - Mar'22) → Associate Software Engineer II (Apr'22 - Mar'23) → Software Engineer I (Apr'23 - Present)

As **Software Engineer I:**

***Projects Undertaken:***

**DeepSeek-R1 Replica & LLaMA Fine-Tuning**

**Technologies Used**: LLaMA, PPO, TRL, YAML

**Responsibilities:**

- Recreated core capabilities of DeepSeek-R1 prior to its official release, showcasing forward-leaning research and engineering initiative.
- Fine-tuned **LLaMA** using **PPO** with **TRL**, optimizing model behavior through reinforcement learning.
- Enforced structured **YAML** outputs and applied domain-specific reward shaping to improve factual accuracy and reasoning consistency.

## EDUCATION

2021
**Master of Computer Applications (MCA)**
*Birla Institute of Technology, Mesra*
*CGPA: 8.4*

2018
**Bachelor of Computer Applications (BCA)**
*Birla Institute of Technology, Mesra*
*CGPA: 7.74*

## PERSONAL PROJECTS

### Cortana++ (In Progress)
- Building a lightweight ML engine in C++/CUDA supporting tensor ops, broadcasted matmul, reduction, and Dense layers.
- Designing for high performance with minimal dependencies and tight GPU integration.

### Darwin's Silicate Organism (Design Complete)
- Architected a modular AI framework inspired by survival-driven intelligence, focusing on memory, reward, and curiosity.
- Implementation pending; built to explore open-ended learning behaviors.

### TwinSqueeze – Siamese Network for Embedding Compression
- Developed a Siamese network using MiniLM-L6-v2, fine-tuned on STS-B, to compress 384-dimensional embeddings while retaining domain-specific semantic accuracy.
- Incorporated NEFTune-style loss regularization (α = 0.75) to enhance training stability and generalization.

### GreedyContext
- Developed a context-selection system reducing LLM token usage and latency via semantic trails of relevant messages.
- Used sentence embeddings, cosine similarity, and greedy graph traversal for optimal message selection.

### Custom DeepFake Pipeline
- Built a DeepFake inference engine on AMD GPU using OpenCL, without external ML libraries.
- Manually implemented ANN inference, showcasing low-level GPU compute control.

### Unity3D Self-Driving Simulation
- Simulated self-driving in Unity3D using a Blender-designed car and track with continuous-action PPO agents.
- Integrated GPS-based routing to navigate new, dynamic paths.

### Conway's Game of Life - Genetic Algorithm
- Created a genetic algorithm to evolve minimal seed patterns for full-grid coverage.
- Focused on efficiency and emergent complexity in cellular automata.

### Vector-Free Retrieval-Augmented Generation (RAG) System
**Technologies Used**: Pandas, Cosine Similarity
**Responsibilities:**
- Engineered a highly efficient RAG system with **<150ms** latency, enabling near real-time performance.
- Eliminated reliance on vector databases by using **pandas** and **cosine similarity**, simplifying infrastructure while preserving retrieval quality.

### Siamese Network for Embedding Compression
**Technologies Used**: PyTorch, Contrastive Loss
**Responsibilities:**
- Designed and trained a custom Siamese neural network to compress high-dimensional embeddings.
- Improved similarity scoring accuracy while reducing memory and computational overhead.

### PowerBuddy – Core Algorithm Design
**Technologies Used**: NLP, SQL, LLMs, Code Execution Frameworks
**Responsibilities:**
- Developed an advanced **NLP-to-SQL** translation engine to convert natural language into executable database queries.
- Built a custom code interpreter inspired by **OpenAI's** architecture, enabling PowerBuddy to process and execute user logic dynamically.
- Enabled multi-modal interaction, positioning PowerBuddy as a central intelligent assistant in the platform.

### Partial JSON Streaming for LLMs
**Technologies Used**: JSON, LLM APIs
**Responsibilities:**
- Created a resilient streaming method for partial **JSON** outputs from **LLMs**, addressing a common bottleneck in real-time AI applications.
- Resolved failures in traditional parsers, enabling consistent downstream consumption of incomplete or partial responses.

### LLM Fine-Tuning with SFT & DPO
**Technologies Used**: Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), custom reward models
**Responsibilities:**
- Trained large language models using both **SFT** and **DPO** techniques, improving performance on complex reasoning and instruction-following tasks.
- Incorporated structured reward signals and enforced output formatting to ensure alignment with task-specific constraints.

### AI Tutor & Career Counselor – PowerBuddy Tools
**Technologies Used**: GPT, LLM APIs
**Responsibilities:**
- Developed an AI-powered tutoring assistant to provide on-demand academic support across subjects.
- Implemented a personalized career guidance system to help students explore academic paths and future career options, improving student engagement and planning.

As **Associate Software Engineer II**:
**Highlights:**
- Designed and implemented ERP module APIs using C# (ASP.NET) and Angular 4, supporting business logic and user workflow.
- Resolved production-level defects, refined UI elements, and supported ongoing improvements to CI/CD pipeline processes.
- Handpicked by the ERP Director mid-year to join the AI team, in recognition of strong technical contributions and adaptability.
- Worked on advanced AI features, including student knowledge modeling, debugging adaptive learning systems, dynamic test difficulty adjustment, and personalized learning content delivery.

As **Associate Software Engineer I:**
**Highlights:**
- Contributed to UI enhancements by adding custom fields to ERP screens and resolving reported production bugs.
- Redesigned the application header bar to meet updated UI/UX specifications and design requirements.
- Provided ongoing support for minor UI revisions and participated in code maintenance tasks.