Vansh Nawander

vansh.nawander@research.iiit.ac.in | LinkedIn | GitHub | Website | +91-8125544202

Education

International Institute of Information Technology, Hyderabad (IIIT-H) Hyderabad, India
M.S. by Research, Computer Science 2025 – Present
Advisor: Dr. Manish Shrivastava (LTRC) | Focus: Diffusion Models, Quantization, LLM Inference Systems

Vardhaman College of Engineering Hyderabad, India
B.Tech, Computer Science and Engineering 2020 – 2024

Research Interests

GPU architectures & kernel optimization • LLM inference & training systems • Model quantization • Deep learning • Diffusion models • World models

Experience

AI/ML Consultant – Alonzo AI April 2025 – March 2026

- Fine-tuned and deployed production ASR and TTS systems for speech applications.
- Built an end-to-end agentic speech system integrating speech recognition, language understanding, and speech synthesis into a unified pipeline.
- Handled backend engineering and NLP-based search infrastructure.

AI/ML Engineer – Datazoic Oct 2024 – Jun 2025

- Fine-tuned large language models on A100 GPU servers and deployed them to production.
- Built RAG-based question answering systems for retrieval over thousands of documents.
- Engineered AI agents with LangGraph for a CRM platform, enabling automated actions and improved UX.

Projects

FlashInfer Kernel Contest @ MLSys 2026 `[GitHub]`

- Competing in the FlashInfer AI Kernel Generation Contest at MLSys 2026, writing high-performance GPU kernels (fused MoE) for NVIDIA Blackwell (B200) GPUs.
- Implemented kernels in Triton and CUDA; benchmarked on B200 hardware via Modal cloud infrastructure. Produced a technical report analysing kernel performance against the FlashInfer-Bench evaluation suite.

CUDA & GPU Kernel Sessions `[GitHub]`

- Implemented GPU kernels from scratch: tiled matrix multiplication, memory-optimised CUDA kernels, and Triton kernels; includes Mojo GPU puzzle solutions and GPUMode practice exercises.
- Stack: C++, CUDA, Triton, Mojo, Python.

Language Model from Scratch

- Built a transformer-based language model from scratch in PyTorch, implementing multi-head attention, positional encoding, and the full training pipeline.
- Covered tokenisation, pretraining objectives, and scaling considerations; used as a deep-dive into LLM internals and architecture design.

Publications

- V. Nawander et al., "Span Identification and Summarization," *Workshop Proceedings, NAACL 2025.* [Link]

Skills

Systems & GPU	CUDA, Triton, C, C++, GPU kernel optimisation
ML & DL	PyTorch, JAX, vLLM, HuggingFace, Scikit-learn, NumPy, Pandas
LLM / AI	LangChain, LangGraph, DSPy, ADK, RAG, fine-tuning (A100)
Languages	Python, C++, C, JavaScript
Web / Backend	FastAPI, Flask, React.js, Node.js, Express.js
Infra / DB	Linux, Redis, Celery, AWS, MySQL, PostgreSQL, MongoDB

Achievements & Leadership

- **Smart India Hackathon 2023 – National Winner:** Won the national-level competition for problem statement #1384 by the Ministry of Commerce and Industry.
- **IndoML Datathon (NIQ Sponsored):** Won 1st and 5th prizes at the IndoML conference datathon challenge.
- **Microsoft ImagineCup 2024:** Qualified for the first round.