ANSH VERMA

ansh0verma@gmail.com | +91-8398822550 | github.com/crankyastrologer | Ansh.site | linkedin.com/in/ansh-vermaai/

WORK EXPERIENCE

AI/ML Engineer (joined as intern), Wasserstoff **June 2025**

- Architected "Eval-Agent," a high-fidelity "Judge-as-a-Service" framework for automated LLM execution trace analysis, enabling deep-dive "archaeology" into complex multi-agent workflows.
- Engineered a Dual-Storage Trace Ecosystem using OpenTelemetry (OTel), balancing sub-second LLM retrieval (optimized truncated JSONs) with a full-content archive for human-in-the-loop forensic analysis.
- Developed an Adversarial "Legal Suite" of agents (Distinguisher, Proceduralist, Shepardizer) to stress-test agentic decision-making, increasing logic divergence detection between model versions.
- Pioneered Advanced Browser Automation on complex government portals by integrating Vision and Grounding Models with Playwright and PyTorch, automating high-friction manual web workflows.
- Dramatically reduced recommendation chat latency, cutting first-token response time from 5 minutes to 20 seconds via optimized streaming of structured LLM responses.

Intern, IIIT Delhi **June 2023 - Aug 2023**

- Deployed local servers acting as interface between the supercomputer and frontend to make available trajectory predictor functions
- Created new data processing pipelines to speed website load time by up to 500%
- Developed the front end of the website to incorporate new features as well as refining the old
- ones helping push user engagement
- Setting a Local MongoDB server to feed data to the supercomputer along with automating the
- data mining process

PROJECTS

Dineleap

- Developed a QR-based ordering system using the Hono web framework, and deployed dynamic QR code menus for the initial client restaurant.
- Established robust API contracts using Zod and OpenAPI documentation, while integrating Redis caching for ultra-low latency order and menu delivery.
- Managed all customer data and generated real-time analytics (sales, popular items) using MongoDB and integrated CDN services for dynamic image optimization.

GPT-2 from Scratch: High-Performance CUDA LLM Implementation

- Engineered a custom GPT-2 inference/training engine in C++ and CUDA, implementing Transformer internals as raw kernels to bypass high-level framework overhead.
- Optimized memory throughput via coalesced access patterns and parallel reduction kernels for LayerNorm and Embedding layers, significantly reducing global memory latency.
- Developed a specialized Attention transpose kernel to restructure token data into head-contiguous layouts, minimizing bus traffic for high-speed linear algebra.
- Scaled execution across Streaming Multiprocessors (SMs) by implementing grid-stride loops for GeLU and vector addition, ensuring hardware-agnostic performance.
- Documented a low-level "Kernel Journal," mapping theoretical Transformer mathematics to thread-level execution logic for AI systems engineering.

EDUCATION

2021-2025

B.Tech, Computer Science and Engineering

Northcap University

- Specialization Artificial Intelligence and Machine learning
- 9 CGPA

ADDITIONAL INFORMATION

- **ML libraries:** PyTorch, TensorFlow, Scikit-learn, Stablebaseline, Langchain, Pandas, Seaborn, Qdrant, langchain, cuda
- **Fullstack skills:** MongoDB, Express.js, Flask, Svelte, Hono, Next.js, Fastapi
- **Languages:** Python, C/C++, Java, JavaScript, TypeScript
- **Certifications:** AWS Cloud Practitioner, NPTEL Business Analytics and Data Mining (Silver Medal | Top 5%), NPTEL Programming in Modern C++ (Silver Medal | Top 2%)