Aditya Kamat

NY,USA | adityakamat007@gmail.com | +1 934-949-7012 | linkedin.com/in/adityakamat24 | adityakamat.vercel.app

SUMMARY

Machine Learning & Research Engineer with experience scaling transformer models and optimizing LLM inference using PyTorch, CUDA, FlashAttention, and ONNX. Skilled in distributed training (FSDP), quantization, and speculative decoding. Passionate about open-source AI and building performant, reproducible ML systems.

EDUCATION

Masters of Science in Data Science
State University of New York at Stony Brook • May 2026 • 3.82/4.0

EXPERIENCE

Research Assistant – Applied AI in Robotics Jan 2025 – May 2025
Stony Brook University — Interacting Robotic Systems Lab

- Developed a high-fidelity **Unreal Engine** simulator to evaluate LLM-driven robotic agents on real-world tasks like pouring and grasping in closed-loop environments.
- Modeled contact dynamics with **neural Signed Distance Fields (SDFs)**, reducing failure rates by **45%** and improving controller robustness across object types.
- Integrated **Smoothed Particle Hydrodynamics (SPH)** for fluid simulation; enabled robotic manipulation under noisy sensor input.
- Automated multi-step physical workflows by translating **natural language into control sequences** using LLM APIs and task planning pipelines.

Data Scientist Intern April 2023 – April 2024
Rivach, India

- Fine-tuned **GPT-style transformer models** with **PyTorch & HuggingFace**, improving task-specific summarization and sentiment outputs by **35%**.
- Built RAG pipelines with **FAISS** + **OpenAI embeddings**, using optimized chunking and **async I/O** to minimize inference latency.
- Designed reproducible ML workflows using **W&B, Docker**, and version-controlled pipelines on **GCS**; benchmarked outputs using **BLEU** and token-level metrics.
- Engineered scalable **ETL systems** for **10M+ rows**, exposing materialized views and REST dashboards via **Power BI** for real-time monitoring.

PROJECTS

Verifier-Guided Speculative Decoding for LLM Inference July 2025 – Present

- Implemented **verifier-guided speculative decoding (VGSD)** with parallel proposal-verification loops for transformer LLMs, enabling fast token generation with correctness guarantees.
- Achieved up to **1.6× latency reduction** over greedy decoding; evaluated **KV cache reuse**, token acceptance rates, and failure cases across varied sequence lengths.

Lightweight LLM Inference for GPUs with Limited VRAM May 2025 – June 2025

- Built **INT8-quantized ONNX Runtime deployment** with optimized **CUDA execution**, reducing **VRAM usage by 60%** while maintaining real-time performance.
- Deployed as a **FastAPI microservice** with streaming, async batching, and REST APIs; tested on **RTX 3060** with sustained inference throughput.

Scaling GPT-2 with FlashAttention and PyTorch FSDP Mar 2025 – Apr 2025

- Scaled **124M GPT-2** using **FlashAttention + FSDP (mixed precision)**, achieving **3.8× training speedup** and **42% memory savings** on A100/3090 clusters.
- Built distributed training pipelines with GPU monitoring dashboards to track utilization, memory trends, and convergence metrics.

Accelerating Transformer Inference with FlashAttention and Triton Kernels Feb 2025 – Mar 2025

- Developed **custom fused attention kernels** in **Triton**, integrated into PyTorch, and reduced transformer inference latency by **1.7×**.
- Packaged kernels as a reusable PyTorch extension; used `nvprof` and `nvdisasm` to optimize **warp-level scheduling** and memory access patterns.

SKILLS

Languages: Python, C++, CUDA, Java, Bash, SQL
ML & Research: Transformers, LLMs, RAG, FlashAttention, Speculative Decoding, Quantization, Attention Mechanisms, SDFs, Mixed Precision, Interpretability, Reinforcement Learning (basic)
Systems & Infra: PyTorch, HuggingFace, ONNX Runtime, Triton Kernels, Distributed Training (FSDP, DDP), CUDA Profiling (Nsight, nvprof), MLflow, W&B, Docker, GCS, AWS, Linux
APIs & Deployment: FastAPI, REST, Async Batching, Token Streaming, JAX, TensorFlow
Data & Analysis: Power BI, Ray, Dask, Pyspark, statsmodels, Matplotlib, Seaborn
Certifications: AWS Academy Cloud Foundations, NPTEL – Big Data Computing, ORACLE – Database Programming with SQL