Aditya Kamat

• New York, United States • adityakamat007@gmail.com • +19349497012
• in/adityakamat24 • adityakamat.carrd.co

SUMMARY

Machine Learning engineer with experience building and deploying LLMs, scalable inference systems, & retrieval-augmented pipelines. Skilled in model optimization, distributed training, & end-to-end delivery of AI products powered by PyTorch, HuggingFace, & CUDA.

EDUCATION

Masters of Science in Data Science
State University of New York at Stony Brook • December 2025 • 3.82 • New York

EXPERIENCE

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

- Built a simulation environment in **Unreal Engine** to train and evaluate **AI-driven robotic assistants** for physical tasks like object manipulation and pouring, integrating **LLMs and policy networks** for decision-making.
- Developed **grasping and interaction logic** using **neural network-generated signed distance fields (SDFs)**, improving physical collision modeling and reducing interaction failures by **45%**.
- Implemented **real-time fluid simulations** using **Smoothed Particle Hydrodynamics (SPH)** and connected control modules to simulated robotic arms, enabling end-to-end AI feedback for fine-grained motion planning.
- Integrated **vision and language models (LLMs)** via custom APIs to interpret natural language tasks into robotic control sequences, enabling multi-step task automation in simulated environments.

Data Scientist Intern April 2023 – April 2024
Rivach, India

- Fine-tuned custom **GPT-style transformer models** using **PyTorch** and **HuggingFace** for **document summarization** and **sentiment analysis**, improving automation efficiency by **35%**.
- Built scalable **RAG pipelines** using **FAISS** and **OpenAI embeddings** for knowledge retrieval; optimized **chunking** and **async data loaders** to improve context relevance.
- Designed modular training and evaluation pipelines with **W&B tracking**, **Dockerized jobs**, and version control via **Git** and **GCS**; implemented model metrics like **perplexity**, **BLEU**, and **token-level accuracy**.
- Optimized **SQL-based preprocessing** for **10M+ row datasets** using **ETL batching**, **materialized views**, and **indexes**; deployed monitoring dashboards via **Power BI** and **REST APIs**.

PROJECTS

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

- Designed **low-memory inference infrastructure** using **ONNX Runtime** and **INT8 quantized transformers**; optimized **CUDA execution** to reduce **VRAM usage** by **60%**.
- Deployed model as a **containerized microservice** with **tokenized streaming**, **async batching**, and **REST endpoints**, supporting real-time inference on **RTX 3060 GPUs**.

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

- Scaled a **124M GPT-2** model using **FlashAttention** and **PyTorch FSDP** with **mixed precision**, achieving **3.8× speedup** and **42% memory savings**.
- Built **distributed training pipelines** and **visual dashboards** to analyze GPU utilization, memory usage, and convergence trends on **A100/3090 clusters**.

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

- Built **custom fused attention kernels** in **Triton** & integrated them into **PyTorch**, reducing inference latency by **1.7×**.
- Packaged kernels as a reusable **PyTorch extension** and profiled them with `NVProf` and `nvdisasm` to optimize **warp-level parallelism** and **memory coalescing**.

SKILLS

Programming Languages: Python, C++, C, Java, JavaScript, SQL, R, HTML/CSS, Bash
AI & ML: Transformers, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Agentic Models, Generative Models, Attention Mechanisms, Quantization, FlashAttention, Sequence Modeling, Tokenization Algorithms, Model Checkpointing, Hyperparameter Optimization, Interpretability, Safety-aware ML,Tokenization Pipelines
DevOps & Infra: Distributed Training (FSDP, DDP), Async Data Loading, Experiment Tracking, Reproducible Pipelines, Profiling (torch.profiler, Nsight Systems, nvprof), Cloud Buckets (GCS/S3), Git, Docker, Versioning, Linux, Google Cloud, AWS, CUDA
Data Analysis & Visualization: Power BI, RStudio, Matplotlib, Seaborn, Excel, Ray, Dask, Pyspark, Qdrant, statsmodels
Web & Frameworks: ReactJS, Firebase, MySQL, Web APIs, PyTorch, TensorFlow, Keras, HuggingFace Transformers, ONNX Runtime, FastAPI, Triton Kernels, statsmodels, MLflow, Weights & Biases (W&B), JAX (basic)

ACHIEVEMENTS & INVOLVEMENT

Certifications: AWS Academy Cloud Foundations, NPTEL - Big Data Computing, ORACLE - Database Programming With SQL
Hackathons: Placed **2nd** among **100+ participants** at both the **Epitome (24h)** and **Zignasa-2k23 (36h)** hackathons for building real-time, impactful tech solutions. Contributed to social-good innovations at the **VIVITSU (48h)** Hackathon.