Vipul Sharma

vipuls181999@gmail.com — GitHub

EDUCATION

Brown University, RI, US — 2023 — 2025
Master of Science in Computer Science — GPA: 3.625/4

National Institute of Technology (NIT), Bhopal, India — 2017 — 2021
Among the 32 NITs in India. Bachelor of Science in Computer Science & Engineering — GPA: 3.96/4

SKILLS

- **Tools:** PyTorch, SGLang, vLLM, TorchAO, Torch Compiler, NCCL, Triton, CUDA C++, Docker, SLURM, Python, Linux, SQL
- **ML Systems Concepts:** Distributed Training and Inference (FSDP, EP, TP, etc.), Mixed Precision Training and Quantization (MX/NVFP4, MXFP8, etc.), ML Compilers, CUDA Kernels, Multi-GPU Communication Kernels (NVSHMEM)
- **Relevant Coursework:** Systems for Machine Learning (topics), Computational Linguistics (topics), Self-Supervised Learning, Computational Cognitive Science

PROJECTS

Survey of Quantization Formats GitHub — Aug 2025 — Oct 2025

- Presented a **survey on model quantization** techniques for inference optimization at **EleutherAI**. [Recording] [Slides]
- Implemented & **benchmarked** quantized **Hugging Face** models using **TorchAO** & **GemLite**, analyzing **token throughput** across low-precision formats (**MXFP8**, **INT4**, etc.). [Results]
- **Diagnosed** & reported a **performance degradation** in **TorchAO** FP8 weight-only quantization (PR in progress). [Worklog]

NCCL From First Principles GitHub — Aug 2025 — Present

- **Reverse engineering NCCL** and the **communication collectives of NCCL** using only the NCCL communication primitives in CUDA and C++. [Annotated paper]
- Surveyed communication algorithms (**Ring**, **Tree**), & communication paths (**P2P**, **RDMA**, **Shared Memory**, etc.).

RESEARCH EXPERIENCE

Serre Lab & Balestriero Lab, Brown University
Research Assistant — May 2024 — May 2025

- Implemented a **novel pre-training methodology** to address **noise robustness** in **self-supervised** learning for **audio** data. [GitHub] [Project report]
- Developed an **open-source GPU-accelerated** library for differentiable image warping transformations to **evaluate the robustness of vision** models. [PyPI package]
- Implemented a **multi-GPU fine-tuning & inference** pipeline using **FSDP**, **activations checkpointing**, & **gradient accumulation** for evaluating a **1.2B vision** model. [GitHub]

PROFESSIONAL EXPERIENCE

JPMorgan Chase & Co.
SDE, Equities Data & Analytics — Jul 2021 — Jul 2023
One of 4 engineers responsible for JPM's **low-latency trading alerts engine** serving **thousands of clients** and **millions of alerts** a day.

- **Designed** and developed a **distributed disaster recovery system** across 6 data centers, ensuring failover with **zero missed alerts** using **asynchronous messaging queues**.
- Engineered **7 low-latency data pipelines** processing **1M+ events/day**, enabling new real-time alerts.
- **Developed a chatbot** that mined traders' chats & news feed with **Named Entity Recognition** to generate 3000+ trading opportunities alerts every day.
- Redesigned chatbot backend to **reduce cloud infra costs by 70%**.
- One of fewer than 10 recipients of **"SEP Recognition Scroll" award** out of hundreds of new graduates for **infra modernization**.