Kashiful Haque

+91 8240868544 • haque.kashiful7@gmail.com • github • linkedin • ifkash.dev

Systems-oriented ML Engineer with 3.5+ years of experience building and optimizing large-scale LLM inference and training pipelines. Strong focus on model internals, acceleration techniques (quantization, QLoRA, batching), and structured evaluation systems. Experienced in designing low-latency, high-throughput NLP systems and collaborating closely with research teams to translate ideas from papers into production-grade multilingual and domain-adapted models.

Work Experience

wand.ai — *Palo Alto (Remote)*
Backend AI / ML Engineer — 11/2025 – Present

- Designing execution-layer infrastructure for LLM-based agents, focusing on pipelined inference, tool orchestration, and structured task environments.

American Express — *Bangalore*
Engineer III — 02/2025 – 11/2025

- Architected distributed, low-latency streaming pipelines with backpressure control and token-level streaming semantics, directly applicable to real-time LLM inference workloads.
- Deployed and optimized production systems on RedHat OpenShift (Kubernetes, Helm, Jenkins), improving system throughput and reliability under high concurrency.
- Rebuilt large-scale search infrastructure over 200k+ documents using hybrid semantic + lexical retrieval, reducing p95 latency from 30s to under 2s.

Fiery — *Bangalore*
Associate Software Engineer — 01/2023 – 02/2025

- Fine-tuned LLMs (Mistral-7B, Llama-8B) using QLoRA for domain adaptation; instrumented GPU memory usage and profiled kernel-level bottlenecks during training and inference.
- Delivered high-throughput inference pipelines using vLLM with dynamic batching and quantized kernels, achieving sub-second p95 time-to-first-token on T4 GPU clusters.
- Developed Fiery Scribe, an NER-driven automation engine using a fine-tuned ModernBERT model, focused on domain-specific text understanding and robustness.
- Implemented LLM-to-SQL agent systems with partial-plan evaluation, structured failure recovery, and programmatic validation of model outputs.

Corteva Agriscience — *Hyderabad*
Internship — 07/2022 – 12/2022

- Worked on applied ML pipelines and migration of Flask monolith to microservices.

Projects

smoltorch: Minimal Autograd Engine • github • pypi • blog

- Implemented a reverse-mode autograd engine with tape-based computation graphs and topological scheduling.
- Designed NumPy-backed tensor operations with broadcasting and a minimal training loop inspired by PyTorch internals, deepening understanding of model architecture and optimization mechanics.

tinyndarray: Mini NumPy in Rust • github

- Built a stride-aware ndarray implementation with slicing and broadcasting, mirroring tensor layouts used in modern ML frameworks.
- Exposed Rust numerical kernels to Python via PyO3, enabling experimentation with efficient tensor operations and future JIT/acceleration work.

nopokedb: Lightweight Vector Database • github • pypi • blog

- Designed a disk-backed HNSW vector database with crash recovery, oplog durability, and low memory footprint.
- Implemented fast ANN search and metadata filtering, relevant for retrieval-augmented multilingual and domain-adapted NLP systems.

Boo: AI-powered Discord Bot • github • deepwiki • blog

- Built a containerized, agentic LLM system with isolated execution, deterministic traces, and structured tool APIs.
- Added unit-test–based evaluation hooks and feedback signals, aligning with preference-based learning and RL-style evaluation paradigms.

Education

IIT Madras

BS, *Data Science and Applications* 2020 – 2024

Skills

- Languages: Python, Go, Rust, C++
- ML / NLP: PyTorch, LLM fine-tuning, quantization (QLoRA), vLLM, ModernBERT, RAG systems
- Systems & Acceleration: CUDA-aware profiling, dynamic batching, low-latency inference, distributed pipelines
- Infra: Docker, Kubernetes, OpenShift, Redis, PostgreSQL
- Research & Evaluation: Paper implementation, structured evals, error analysis, test-based feedback, agentic environments