Kashiful Haque

+918240868544 • haque.kashiful7@gmail.com • github • linkedin • ifkash.dev

Systems-focused ML engineer with 3.5 YOE building numerical computing components, inference-optimized pipelines and agentic coding environments. I work at the intersection of ML systems, low-level frameworks and RL for code, designing execution sandboxes, structured task environments and deep-dive into unfamiliar codebases to build testable evals. Previously fine-tuned and deployed models and optimized inference systems org-wide.

work experience

Wand AI (via Nityo) — *Palo Alto (Remote)*

Backend AI/ML Engineer — 11/2025 – Present

- Building internal ML systems and execution-layer infra; focusing on agent tooling and structured task pipelines.

American Express (via IntraEdge) — *Bangalore*

Engineer III — 02/2025 – 11/2025

- Architected a distributed low-latency streaming pipeline with backpressure control and token-stream handling; deployed on RedHat OpenShift using Helm/Jenkins.
- Rebuilt search infra for 200k+ Confluence documents using hybrid semantic + trigram retrieval; improved p95 latency from 30s ⟶ <2s.
- Reverse-engineered internal systems to build testable interfaces and mocks for multi-step automation pipelines.

Fiery — *Bangalore*

Associate Software Engineer — 01/2023 — 02/2025

- Fine-tuned Mistral-7B/Llama-8B with QLoRA; instrumented memory usage, evaluated kernel bottlenecks, and integrated into a scalable RAG pipeline.
- Delivered high-throughput inference using vLLM with dynamic batching and quantized kernels, hitting sub-second p95 TTFT on T4 clusters.
- Built Fiery Scribe, an NER-driven automation engine using a fine-tuned ModernBERT model; designed structured evaluation pipelines and test harnesses for model outputs.
- Developed AskDB, an LLM⟶SQL agent system with programmatic query analysis, partial-plan evaluation, and structured error recovery.

Corteva Agriscience — *Hyderabad*

Internship — 07/2022 – 12/2022

projects

smoltorch: minimal autograd engine • github • pypi • blog

- Reverse-mode autograd engine with tape-based graphs and topological scheduling.
- NumPy-backed tensor ops, broadcasting, and a minimal training loop inspired by PyTorch internals.

tinyndarray: mini numpy in rust • github

- Stride-aware ndarray implementation in Rust with slicing + broadcasting, mirroring ML framework tensor layouts.
- Python bindings via PyO3 enabling fast numerical kernels and early graph/JIT experimentation.

nopokedb: lightweight vector db • github • pypi • blog

- Disk-backed HNSW vector DB with oplog durability, crash recovery, and minimal RAM usage.
- Fast metadata lookups + efficient ANN search via SQLite + hnswlib.

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

- Agentic, containerized code-execution sandbox with filesystem isolation, resource limits, deterministic traces, and multi-step tool APIs.
- Added unit-test–based evaluation hooks and structured feedback signals, forming basis for RL coding environments.

education

IIT Madras BS, *Data Science and Applications* — 2020 – 2024

skills

- python, go, rust, c++, pytorch, vllm, cuda, docker, k8s, redis, postgres, agentic systems (code sandboxes, test-based evals)