

Kashiful Haque

+918240868544 • haque.kashiful7@gmail.com • [github](#) • [linkedin](#) • [ifkash.dev](#)

Applied ML Systems Engineer; bridging models with production infra. Experienced in fine-tuning LLMs, optimizing inference (vLLM, quantization, batching) and building scalable AI-powered applications (Docker, K8s, Redis, Postgres). Proven track record of deploying research into production systems, adopted across teams organization wide.

work experience

Wand AI (via Nityo)

Palo Alto, USA

Senior ML Engineer

11/2025 – Present

American Express (via IntraEdge)

Bengaluru, India

Engineer III

02/2025 – 11/2025

- Architected a low-latency streaming infra (Redis + GPT-4o summarization) for thousands of meetings; deployed to RedHat OpenShift (Amex Hydra) with Helm/Jenkins.
- Built a hybrid semantic + trigram retrieval system over 200K+ Confluence pages (runbooks); reduced latency from 30s (native search) to under 2s.
- Designed a multimodal runbook generation pipeline using GPT-4o, converting raw incident data (text + screenshots) into standardized Confluence pages.

Fiery (an Epson company, formerly EFI)

Bengaluru, India

Associate Software Engineer

07/2023 — 02/2025

- Fine-tuned Mistral-7B using QLoRA on 4070 Ti Super for enterprise knowledge injection, incorporated into a RAG pipeline.
- Built efficient inference at scale via vLLM on NVIDIA T4 cluster; leveraged dynamic batching and quantized inference for sub-second 95th-percentile latency (TTFT).
- [Led development of “Fiery Scribe”](#), a print-request automation tool powered by a fine-tuned ModernBERT NER model; parsed email and chat inputs to trigger document workflows. [Demoed at Printing United 2024, Las Vegas.](#)
- Built AskDB, an agentic AI system for natural language → SQL, adopted by 300+ active users across departments.

Intern

01/2023 — 07/2023

- Built a product mockup generation system using ImageMagick. [read blog here](#)

Corteva Agriscience

Hyderabad, India

Internship

07/2022 – 12/2022

- Migrated old Flask monolith API to microservices architecture to speed up deployments.

projects

NoPokeDB: a Lightweight Vector Database • [github](#) • [pypi](#) • [blog](#)

- Designed a disk-backed vector db in Python using hnswlib for ANN search and SQLite for metadata.
- Added durability via a write-ahead oplog with crash recovery; supported batch inserts, auto-resize, CRUD.
- 2K+ PyPI downloads!

Mini-Numpy in Rust • [github](#)

- Implemented a lightweight NumPy clone in Rust with Python bindings, exploring numerical backend design for ML frameworks.

Boo: AI-powered Discord Bot • [github](#) • [deepwiki](#) • [blog](#)

- Architected a multi-service infra with LLM orchestration for an AI bot.
- Features: conversational AI (LLMs), image analysis & generation, weather, GIF & HN search via function/tool calling.

education

IIT Madras BS, Data Science and Applications

2020 – 2024

skills

- pytorch, huggingface, peft, qlora, vllm, ollama, vector db, embeddings, inference optimization, cuda
- python, go, rust, fastapi, docker, k8s, redis, postgres, sql, helm, jenkins