Kashiful Haque

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

Engineer building production-grade AI systems; I have deployed fine-tuned LLMs (and vLLM for inference) into live infra, designed multi-service discord bots with redis/postgres state and delivered reverse image search on the fly during interviews. Experienced in end-to-end ownerships; model training → scalable infra (docker, k8s) → API delivery. My work has been used by hundreds and integrated into customer facing apps.

Work Experience

American Express (via IntraEdge) *Bengaluru, India*
Engineer III 02/2025 – Present

- Built CC Listener module for Bridge Intelligence, capturing real-time CCs captions for thousands of Webex meetings, streaming them into Redis for downstream summarization using GPT-4o.
- Deployed CC Listener to Amex's Hydra platform (RedHat OpenShift) across dev, QA, and pre-prod using Helm, Jenkins, and Docker; navigated complex internal infra and CI/CD pipelines.
- Built a hybrid semantic + fuzzy (trigram) retrieval system over 2M+ 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 to incorporate internal documentation and company workflows.
- Deployed the fine-tuned model via vLLM on an NVIDIA T4 cluster; leveraged dynamic batching and quantized inference for sub-second 95th-percentile latency (TTFT).
- Led development of "Beacon", a print-request automation tool powered by a fine-tuned ModernBERT NER model; parsed email and chat inputs to trigger document workflows.
- Built AskDB, an AI agent used by 300+ internal users; used llama 3.1 to translate business queries into SQL, fetch data and auto-generate charts and summaries.

Fiery (an Epson company, formerly EFI) *Bengaluru, India*
Internship 01/2023 – 07/2023

- Built product mockup pipeline using ImageMagick, achieving cost-effective design automation.

Corteva Agriscience *Hyderabad, India*
Internship 07/2022 – 12/2022

- Migrated Flask monolith to containerized microservices, reducing AutoML job-launch latency by 40%.

Projects

Mini-Numpy in Rust • github

- Implemented a lightweight NumPy clone in Rust with Python bindings to deeply understand tensor internals and performance optimizations in ML frameworks.

Boo: AI-powered Discord Bot • github • deepwiki

- Architected a multi-service AI bot running as a Docker Compose stack with multiple services.
- 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

- python, typescript, golang, rust, c++, sql
- fastapi, docker, k8s, vllm, ollama, huggingface, peft, qlora
- pytorch, numpy, pandas, sklearn, spacy, cuda, redis, vector db