Kashiful Haque

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

3 YOE building ML platforms and developer tooling (FastAPI, Docker), experienced in fine-tuning and deploying scalable GenAI models (QLoRA, vLLM) across distributed infra.

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.
- *Python, FastAPI, Langchain, PGVector, Streamlit, Node.js, Puppeteer, WebEx JS SDK, OpenShift*

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.
- *vLLM, Python, PyTorch, QLoRA, peft, SFT, Mistral, Llama, SQL, FastAPI*

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

- Created a product mockup pipeline using ImageMagick and Node.js, achieving cost-effective design automation.
- *Node.js, ImageMagick, Angular*

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

- Migrated a Flask monolith to containerized microservices (Docker, Kubernetes), reducing AutoML job-launch latency by 40%.
- *Python, Flask, Docker, Kubernetes*

Education

Indian Institute of Technology Madras 2020 – 2024
Bachelor of Science, *Data Science and Applications*

Skills

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