Kashiful Haque

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

3 YOE building ML platforms & developer tooling (FastAPI, Docker), fine-tuning and deploying scalable GenAI models (LoRA, vLLM) across distributed systems.

Work Experience

Engineer III 02/2025 – Present

American Express (IntraEdge) *Bengaluru, India*

- Shipped a real-time NLP pipeline that live summarizes 1,000+ WebEx meetings weekly, cutting down manual meeting note taking
- Built an NLP-powered Confluence search service (FastAPI, PGVector) indexing 2M+ pages, supporting sub-1500ms 95th-percentile search results latency
- Tech stack: *Python, FastAPI, Langchain, PGVector, Streamlit, Node.js, Puppeteer, WebEx JS SDK, OpenShift*

Associate Software Engineer 07/2023 – 02/2025

Fiery (an Epson company, formerly known as EFI) *Bengaluru, India*

- Fine-tuned Mistral-7B using LoRA adapters on 4070 Ti Super, baking company knowledge into model
- Deployed the fine-tuned model on a 4x1080 Ti GPU cluster via vLLM; engineered dynamic batching and employed mixed-precision inference for sub-750ms 95th-percentile TTFT
- Led development of "Beacon", an NER-based print-request automation service integrating with emails and natural language interfaces, leveraging a custom fine-tuned ModernBERT model
- Built AskDB, a natural language to insights platform, used by 300+ internal users to query company DBs and auto-generate visual insights
- Tech stack: *vLLM, Python, PyTorch, LoRA, SFT, Mistral, Llama, Ollama, SQL, FastAPI*

Data Scientist, Intern 01/2023 – 07/2023

Fiery (an Epson company, formerly known as EFI) *Bengaluru, India*

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

Fullstack Developer, Intern 07/2022 – 12/2022

Corteva Agriscience *Hyderabad, India*

- Migrated Flask monolith to microservices (Python, Docker, Kubernetes), optimizing AutoML job triggering and reducing job-launch latency by 40%
- Tech stack: *Python, Flask, Docker, Kubernetes*

Education

Indian Institute of Technology Madras 2020 – 2024

Bachelor of Science, *Data Science and Applications*

Skills

- Python, TypeScript, C++, Go
- FastAPI, Flask, SQLAlchemy, Docker, Kubernetes, vLLM, Hugging Face, LoRA
- PyTorch, NumPy, Pandas, scikit-learn, spaCy, NLTK, CUDA, Langchain
- MySQL, PostgreSQL, SQLite, Redis, Chroma, Qdrant, PGVector