Kashiful Haque

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

Software Engineer with 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 (client for IntraEdge) *Bengaluru, India*

- Designed & shipped a real-time NLP pipeline (JS, Webex APIs, OpenAI) that ingests and summarizes 1,000+ live WebEx transcripts summary, cutting down manual meeting note taking.
- Building a search engine for company confluence pages, can be queried using natural language
- Built an NLP-powered Confluence search service (FastAPI, PGVector) indexing 2M+ pages, supporting sub-500ms 95th-percentile latency and 90% conversational-query accuracy.

Associate Software Engineer 07/2023 – 02/2025
Fiery (an Epson company, formerly known as EFI) *Bengaluru, India*

- Deployed LLaMa 3.1-8B via vLLM on a 4070 Ti Super cluster for 500 users; engineered dynamic batching and employed mixed-precision inference to maintain sub-300ms 95th-percentile TTFT
- Fine-tuned Mistral-7B with LoRA adapters on 4×1080 Ti GPUs, embedding domain knowledge and boosting task accuracy by 12%.
- 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 and audio fingerprinting solutions; integrated RESTful APIs with Redis caching

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.

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%.

Education

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

Projects

Boo • git repo
Python, Discord.py, Go, PostgreSQL, Cloudflare Workers, Linode

Skills

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