Kashiful Haque

+91 8240868544 • me@ifkash.dev • github • huggingface • linkedin • ifkash.dev

ML Engineer with 4 YOE across LLM pre-training, post-training with RL, and high-performance inference in C++ and Rust. Pre-trained a 360M LLaMA from scratch, fine-tuned production LLMs, and applied GRPO/ORPO for behavioral alignment. Targeting roles in pre-training, RL-based post-training, and GPU inference — CUDA, HPC, and low-level runtime work.

Work Experience

wand.ai — *Bangalore*

Backend AI / ML Engineer — 11/2025 – Present

- Built observability for agent workflows: span hierarchy, trace IDs, and Kafka-based monitoring events for e2e run visibility.
- Implemented OAuth/OIDC credential flows with deferred auth, token refresh, and variable propagation across agentic tasks.
- Built workflow specs: generates auth gateways, OAuth triggers, and injects credential references into tool environments.
- Added dynamic tool support: agents can define, reference, and execute tools that resolve workflow state variables at runtime.
- Maintain a Go service and contribute to a gRPC-based internal API powering agent execution.

American Express — *Bangalore*

Engineer III — 02/2025 – 11/2025

- Built a hybrid retrieval system over 200k+ internal docs combining dense embeddings and keyword retrieval; cut p95 query latency from 30s to 2s via caching and query optimizations.
- Shipped a live Webex meeting-summary bot (headless Chrome + Webex SDK as virtual participant); streamed real-time transcripts, generated rolling LLM summaries, and posted updates to Slack every 5 minutes.
- Profiled large-scale transformer training and inference (FlashAttention, KV-cache, batching) to improve embedding throughput and serving efficiency.

Fiery — *Bangalore*

Associate Software Engineer — 01/2023 – 02/2025

- Built Fiery Scribe: translates natural-language print requests to printer XML using ModernBERT + LLM, replacing brittle rule-based parsing.
- Built AskDB: an LLM agent that maps natural language to SQL over production DBs and generates business reports for non-technical users.
- Built Fiery Chat: internal RAG assistant on the company knowledge base; fine-tuned the base model with SFT to eliminate acronym hallucinations across new-hire onboarding and query routing.
- Fine-tuned domain LLMs using SFT/QLoRA, served via vLLM on T4 clusters with sub-second p95 TTFT; evaluated SGLang but chose vLLM for tool-calling support at the time.
- Deployed multi-GPU inference (4x 1080 Ti) using Ray Serve for request routing and scaling.

ML Engineering Work

banana.cpp: LLM Inference Engine in C++ • github

- Inference engine with KV-cache, speculative decoding, and continuous batching; 10x speedup via CPU parallelization and fused kernel optimizations.

Llama-3.2 3B Politeness Alignment using ORPO • hf • blog

- Applied ORPO with TRL + Unsloth to align Llama-3.2 3B to a behavioral policy (respond politely or refuse).
- Built a 32k-row behavior preference dataset; validated convergence via reward margins and log-prob dynamics in W&B.

smol-llama: 360M LLM pre-training • github • hf • wandb • blog

- Trained a 360M LLaMA from scratch in PyTorch with GQA, RoPE, RMSNorm, and SwiGLU on 6B tokens (H100, 75K tok/s).

Micro-Llama: bare-metal inference on RP2040 • github • hf • wandb

- Trained a 213K-param LLaMA on TinyStories (JAX/Flax); bare-metal greedy inference in C on RP2040 (264KB SRAM).

Education

IIT Madras

BS, *Data Science and Applications* — 2020 – 2024

Skills

- Python, C++, Rust, Go
- PyTorch, FlashAttention, CUDA, vLLM, SGLang, Ray Serve, gRPC, TRL, QLoRA, Docker, Kubernetes, Redis, PostgreSQL