

Kashiful Haque

+91 8240868544 • me@ifkash.dev • [github](#) • [huggingface](#) • [linkedin](#) • [ifkash.dev](#)

ML Engineer with 4 YOE shipping production LLM systems: agent runtimes in Go and Python, hybrid retrieval, and fine-tuned LLMs on vLLM. Pre-trained a 360M LLaMA from scratch, aligned Llama-3.2 3B with ORPO, and built an LLM inference engine in C++. Targeting pre-training, RL post-training, and GPU inference roles: CUDA, HPC, and low-level runtimes.

Work Experience

wand.ai

Bangalore

Backend AI / ML Engineer

11/2025 – Present

- Built Temporal-backed agent scheduling in Go: schedule APIs, overlap policies, Kafka events, and an agent scheduling tool.
- Hardened the agent loop: context compaction, tool-argument validation, prompt-injection-safe attachments, and citations.
- Built agent observability: conversation-scoped trace IDs, span hierarchy, and Kafka monitoring events with error context.
- Implemented deferred OAuth/OIDC credentials: a spec transpiler emits auth gateways, OAuth triggers, and token refresh.
- Added dynamic tools that agents define in the workflow spec and execute against workflow state variables at runtime.

American Express

Bangalore

Engineer III

02/2025 – 11/2025

- Built live Webex captioning: Webex webhooks launch headless Chromium running the browser-only Webex SDK, which streams transcripts through Node.js into Redis in real time.
- Contributed to an incident assistant that merged captions, Slack, and an internal portal into 5-minute LLM summaries.
- Replaced Confluence search with hybrid retrieval over 200k+ incident runbooks: a chunk, keyword, and embedding ingestion pipeline feeds Elasticsearch BM25 + vector search; cut p95 query latency from 30s+ to under 2s.

Fiery

Bangalore

Associate Software Engineer

01/2023 – 02/2025

- Built Fiery Scribe (flagship): turns customer print requests into printer XML, work that needed trained operators. Moved from LLM tool calling to fine-tuning, then to an in-browser ModernBERT classifier with the LLM only verifying.
- Built AskDB: a local LLM agent that turns management's questions into SQL, then writes Python to build business reports, so managers no longer wait on engineers.
- Built Fiery Chat, a hackathon winner that became internal and customer-facing products; fixed acronym hallucinations with QLoRA SFT on a hand-curated dataset, merged into a 7B model.
- Served models via vLLM on T4 clusters (customers, sub-second p95 TTFT) and Ray Serve on 4x 1080 Ti (internal); trained on an RTX 4070 Ti.

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