

Kashiful Haque

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

Almost 4 YOE in Machine Learning, building high-throughput services, scalable hybrid search and production ML inference. Pretrained a 360M llama-style LLM on NVIDIA H100 (6B tokens) and building an LLM inference engine in C++; Learning internals of PyTorch/NumPy by implementing my own autograd engine and ndarray in Rust.

Work Experience

wand.ai

Backend AI / ML Engineer

Palo Alto (Remote)

11/2025 – Present

- Built a config-driven agent workflow runtime with custom DSL.

- Implemented workflow validation + schema enforcement + versioning to prevent production incidents from config drift.

American Express

Bangalore

Engineer III

02/2025 – 11/2025

- Built hybrid search over 200k+ documents using dense embeddings + keyword retrieval, powering internal knowledgebase.
- Reduced p95 latency from 30s to 2s by refactoring pipelines, caching, and optimizing I/O + query execution.

Fiery

Bangalore

Associate Software Engineer

01/2023 – 02/2025

- Fine-tuned and productionized domain LLMs using SFT/QLoRA, serving via vLLM on NVIDIA T4 and 4070Ti clusters with sub-second p95 latency for internal AI products.
- Built Fiery Scribe, an NLP system translating natural-language print requests into complex printer XML configurations, automating operator workflows and reducing GPU inference costs by >80% using a fine-tuned ModernBERT model.
- Developed AskDB, an LLM-powered analytics agent that maps natural-language queries to SQL over production databases and auto-generates structured reports, eliminating manual data requests for business KPIs.

Projects

smol-llama: 360M LLaMA Pre-training • [github](#) • [huggingface](#)

- Implemented a 360M parameter LLaMA model from scratch in PyTorch, featuring GQA, RoPE, RMSNorm, and SwiGLU.
- Pre-trained on 6B tokens of FineWeb on 1x H100, achieving 75k tok/s throughput via FlashAttention.
- Engineered a cost-effective training pipeline (<\$60 total) with gradient accumulation, automatic checkpoint versioning to Hugging Face, and W&B experiment tracking.

smoltorch: Minimal Autograd Engine • [github](#) • [pypi](#) • [blog](#)

- Implemented a reverse-mode autograd engine with tape-based computation graphs and topological scheduling.
- Designed NumPy-backed tensor ops with a minimal training loop inspired by PyTorch internals.

tinyndarray: Mini NumPy in Rust • [github](#)

- Built stride-aware ndarray with slicing and broadcasting, mirroring tensor layouts used in ML frameworks.
- Exposed Rust numerical kernels to Python via PyO3 for efficient tensor operations and future JIT/acceleration work.

nopokedb: Lightweight Vector Database • [github](#) • [pypi](#) • [blog](#)

- Designed a disk-backed HNSW vector database with crash recovery, oplog durability, and low memory footprint.
- Implemented fast ANN search and metadata filtering, relevant for retrieval-augmented and domain-adapted NLP systems.

banana.cpp • [github](#)

- Building an LLM inference engine with KV-cache, speculative decoding and continuous batching.
- Achieved 10x speedup through CPU parallelization + fused kernel optimizations.

Education

IIT Madras

BS, Data Science and Applications

2020 – 2024

Skills

- Python, Go, Rust, C++
- PyTorch, LLM fine-tuning, quantization (QLoRA), vLLM, RAG
- Docker, Kubernetes, OpenShift, Redis, PostgreSQL