Kashiful Haque

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

Work Experience

wand.ai — Palo Alto (Remote)
Backend AI / ML Engineer — 11/2025 – Present

- Building AI agents to automate workforce.

American Express — Bangalore
Engineer III — 02/2025 – 11/2025

- Engineered a real-time system to capture Webex transcripts and generate automated GPT-4o summaries.
- Built large-scale search over 200k+ docs using hybrid semantic + lexical retrieval, reducing p95 latency from 30s to under 2s.

Fiery — Bangalore
Associate Software Engineer — 01/2023 – 02/2025

- Fine-tuned LLMs using SFT/QLoRA; profiled kernel-level bottlenecks during training and inference.
- High-throughput inference using vLLM, achieving sub-second p95 time-to-first-token on T4 GPU clusters.
- Developed Fiery Scribe, an automation engine that translates natural language into printer instructions via fine-tuned ModernBERT model, slashing costs by replacing GPU-dependent LLMs.
- Developed "AskDB," an AI agent that converts queries into SQL and Python-generated reports, eliminating manual data requests and providing instant access to business KPIs.

Corteva Agriscience — Hyderabad
Internship — 07/2022 – 12/2022

- Worked on applied ML pipelines and migration of Flask monolith to microservices.

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 the FineWeb dataset on a single H100, achieving 75k tokens/sec throughput via Flash Attention 2, `torch.compile`, and bfloat16 mixed precision.
- 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.

Boo: AI-powered Discord Bot • github • deepwiki • blog

- Built a distributed, multi-service AI agent with realtime chat, multimodal vision analysis and tool-calling.
- Engineered a secure execution environment, a coding gym, using a sandboxed Python runtime and handed it over to Boo.

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