BHARGAV KUMAR NATH

Machine Learning Engineer

Leeds, United Kingdom · +44 7563 131573 · bhargavkumarnathh@gmail.com
linkedin.com/in/bhargavkumarnath · github.com/BhargavKumarNath · precisionmechanism.dev

ML engineer for LLM inference: CUDA kernels, KV-cache memory, quantization and RAG.

TECHNICAL SKILLS

Languages & Data:	Python, C++/CUDA, Rust, SQL, Bash, pandas, Polars, NumPy, DuckDB, PySpark
ML & Deep Learning:	PyTorch, Transformers, quantization (QAT, PTQ, FP8), scikit-learn, LightGBM, XGBoost
LLM Systems:	LLM inference, paged KV-cache, GPU profiling, RAG, LangGraph, FAISS, Qdrant, RAGAS
MLOps:	MLflow, Docker, GitHub Actions (CI/CD), FastAPI, pytest, Hydra, drift monitoring, Linux, GCP

SOFTWARE & SYSTEMS EXPERIENCE

Applied Machine Learning Engineer (Contract) | Mercor, Remote — Oct 2026 - Present

- Implemented decoder-only Transformer architectures from first principles in PyTorch, spanning 4-8 layers and **10-40M parameters** on TinyStories.
- Developed a document-grounded RAG pipeline with structure-aware chunking, hybrid lexical/semantic retrieval and cross-encoder reranking; evaluated retrieval relevance independently of answer faithfulness.
- Produced recorded technical coding demonstrations for a frontier AI lab, implementing ML tasks while explaining architectural decisions, mathematical reasoning and implementation validation.

Open Source Systems Developer | Core AI & Developer Tooling — Aug 2026 - Present

- **apple/coreai-optimization** (merged PR #71): fixed a silent QAT bug that gave shared weights mismatched precision and schedule; reviewed by Apple maintainers.
- **bloomberg/memray** (PR #1006, landed in main with my authorship): fixed multi-thread allocations collapsing into one mislabelled CSV row, at 98.5% patch coverage.

Data Analyst Intern | Airports Authority of India, Regional HQ — Jul 2023 - Aug 2023

- Deduplicated 1,053+ SAP ERP asset records and flagged **~22%** for correction or exclusion.
- Showed backfilled ticket dates would leak a failure model's target; advised fixing data first.

Data Engineering Intern | Indian Institute of Technology, Guwahati — Jul 2022 - Aug 2022

- Normalised a MySQL schema to 3NF across 15+ modules with enforced foreign keys.
- Added covering indexes after EXPLAIN profiling, cutting a lookup from **1.8 s to 14 ms**.

Data Analyst | M/S Sanjog Trading, India — Jul 2020 - Nov 2021

- Vectorised Python ETL (pandas, Polars) over 500K+ records: full refresh **35 min to 42 s**.
- Shipped XGBoost forecasts at **12-18% MAPE**, retrained on a schedule with MAPE/MAE monitoring.

SELECTED PROJECTS

Custom CUDA Kernel Library for LLM Inference | CUDA C++, zero-copy Rust/PyO3, PyTorch [GitHub] [Live]

- Three of 12 kernels cut Llama-3-8B block latency **65.4 to 50.8 ms** (1.29x) and peak memory **23.3%**.

PageForge: Paged KV-Cache for LLM Serving | Rust, CUDA, PyTorch [GitHub] [Live]

- Built a Rust allocator and CUDA kernels: **8x less VRAM** (75.5 vs 604 MB) at +33% latency.

EDUCATION

MSc Data Science and Analytics (Merit) | University of Leeds, UK — Sep 2024 - Nov 2025

BTech Computer Science and Engineering (First Class) | Assam Don Bosco University — Aug 2020 - Jun 2024

ADDITIONAL EXPERIENCE

Crew Member (part-time) | McDonald's, UK — Sep 2025 - Present