MARCO HARUNI

AI/ML Engineer | Model Building, LLM Deployment & GPU Systems

Dar es Salaam, Tanzania | Open to remote work and international relocation
mtaniharuni95@gmail.com | Website | GitHub | LinkedIn | Hugging Face | X

TECHNICAL SKILLS

Languages: Python, C++, CUDA C++, Rust, Bash
Model Building: PyTorch, JAX, Flax NNX, Optax, transformers, MoE, tokenization, positional encoding, autoregressive evaluation
Deployment: LLM loading, streaming HTTP/SSE APIs, continuous batching, metrics, Docker, Linux, CI, testing, reproducible GPU execution
Performance: Triton, Pallas, vLLM custom operators, paged attention, KV-cache optimization, benchmarking, numerical correctness

OPEN-SOURCE & INDEPENDENT ENGINEERING

- **Independent specialization:** deep, project-based study of transformers, model training, LLM serving, GPU programming, and performance engineering alongside a BSc in Physiotherapy.
- **Active open-source work:** maintain ongoing public repositories; implemented a real out-of-tree vLLM operator and work with vLLM, SGLang, TensorRT-LLM, and FlashInfer codebases and interfaces.

SELECTED AI/ML PROJECTS

ForgeEngine | 2026-Present

- Build and maintain a Qwen3 4B inference service with explicit model loading, prefill/decode, paged KV caching, continuous batching, sampling, browser/terminal clients, health and metrics endpoints, and streaming chat completion APIs.
- Validated the complete L4 serving path with 93 Python tests, 16 parameter checks, and 3 Rust tests; handled four concurrent requests with zero failures and zero KV-cache leaks.

JAX Addition Transformer / Hugging Face | 2026-Present

- Designed, trained, and exhaustively evaluated an exact 10M-parameter decoder-only transformer using JAX and Flax NNX; reached 100% exact match across all one million ordered input pairs with zero invalid generations.
- Ran a controlled 32-experiment dense-versus-MoE scaling study and built a top-2 ragged-dot MoE model with 17.94M stored parameters.

Forge LLM Stack | 2026-Present

- Build a configurable PyTorch decoder-only pretraining reference with GQA, RoPE, QK-Norm, fused QKVO projection, SDPA, token-budgeted training, dataset preparation, checkpoint/resume, evaluation metrics, and generation.
- Implemented companion byte-level BPE tokenization and positional-encoding laboratories covering sinusoidal embeddings, RoPE variants, ALiBi, NoPE, context scaling, and KV-cache position handling.

LLM Serving Kernels | 2026-Present

- Build eight CUDA/Triton inference projects and a real vLLM plugin; passed 51 L4 GPU correctness tests and measured up to 5.12x speedup over a named PyTorch baseline.

ADDITIONAL EVIDENCE

- **Pallas MoE Kernels** - fused JAX/Pallas routed-MoE forward kernel; 1.84x end-to-end L4 speedup and approximately 81% lower XLA peak memory.
- **Technical Documentation** - architecture notes, benchmark methodology, validation evidence, runbooks, and linked open-source model artifacts.

EDUCATION

Muhimbili University of Health and Allied Sciences (MUHAS) | Dar es Salaam, Tanzania
Bachelor of Science in Physiotherapy, Expected 2027
Independent ML Systems Study, Ongoing - deep, project-based specialization in transformers, model training and inference, GPU programming, profiling, and open-source serving systems.

All selected repositories are active projects with public source code, tests, benchmarks, and documentation.