SANKUR KUNDU

ML Systems Engineer | LLM Training, Inference & GPU Kernels
Bhubaneswar, India (open to relocation) • +91 97761 94201 • sankur.kundu.tw@gmail.com
linkedin.com/in/sankur-kundu • github.com/Shakhtar-Sankur • shakhtar-sankur.github.io • **evidence: the original emails**

17 fixes merged: US Department of Energy, NVIDIA, Apple, Meta, IBM, Massachusetts Institute of Technology

OPEN-SOURCE: 17 FIXES MERGED IN 22 DAYS, BY THE PEOPLE WHO LEAD THE PROJECTS

US DEPARTMENT OF ENERGY NATIONAL LABORATORIES, KOKKOS | Kokkos 5.4 release notes | PR #9738 **Merged 5 Nov 2026**
- Merged by its lead **Damien Lebrun-Grandié** (Los Alamos National Laboratory) after **Oak Ridge National Laboratory** approved it.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY MEEP | Electromagnetics simulator | #3364, #3363 **Merged 5 & 6 Nov 2026**
- **Both fixes merged by its original lead developer Ardavan Oskooi**: "the correct approach."

BARCELONA SUPERCOMPUTING CENTER AUTOSUBMIT | Climate-model workflows | #3405, #3406 **Merged 6 Nov 2026**
- **Two fixes** for Spain's national supercomputing centre, invited by **Bruno P. Kinoshita** for its **4.2.0 release**: "Thank you!!"

NVIDIA ISING-DECODING | Two fixes, merged by maintainer Ivan Basov | #98, #99 **Merged 30 Oct & 2 Nov 2026**

APPLE MLX | Apple's ML framework for Apple silicon, **29k★** | PR #4737 **Merged 22 Oct 2026**
- Fixed a data race in MLX's distributed backend (ThreadSanitizer: 2–4 races per run → **0**); merged by MLX maintainer @zcbenz.

META EXECUTORCH | On-device AI in Instagram & WhatsApp, **5.1k★** | #23677, #23912 **Merged 16 & 28 Oct 2026**
- **Two fixes** to **Meta's** on-device LLM kernels, incl. an out-of-bounds read; **Meta's** Scott Roy: "LGTM! Thanks for the fix!"

TENCENT NCNN | Tencent's on-device AI inference framework, **24k★** | PR #7239 **Merged 5 Nov 2026**
- Fixed Vulkan GPU NaNs in Mish; **approved, extended and merged by its creator nihui**: "Thanks for your contribution!"

RELAY-BP, IBM QUANTUM | Merged by its first author, Tristan Müller | PR #49 **Merged 5 Nov 2026**

CARNEGIE MELLON UNIVERSITY XGRAMMAR | Structured output for vLLM & SGLang | PR #1012 **Merged 29 Oct 2026**

HUGGING FACE OPENENV | RL environments (Meta-PyTorch, NVIDIA, Microsoft), **2.8k★** | #1441, #1442 **Merged 19 Oct 2026**

JAXMARL | NeurIPS 2024; Oxford, University College London & Imperial, **868★** | PR #206 **Merged 26 Oct 2026**

UNIVERSITY OF OXFORD PYZX | Merged by its co-creator, John van de Wetering | PR #563 **Merged 22 Oct 2026**

In review: Google JAX, Google DeepMind MuJoCo, Microsoft ONNX Script, vLLM, SGLang, PyTorch

PUBLIC-SECTOR SCIENCE SOFTWARE: 3 INSTITUTIONS MERGED (ABOVE), 5 IN REVIEW

Weather, climate & oceans: **European Centre for Medium-Range Weather Forecasts**: ecCodes; **UK Met Office & Science and Technology Facilities Council**: **PSyclone**, "otherwise it looks good"; **CSIRO, Australia's national science agency**: emsarray.
Big science: European Organization for Nuclear Research (CERN): Rucio; **Swiss National Supercomputing Centre**: ReFrame.

HONORS & RECOGNITION

- **Harvard's Machine Learning Systems textbook** (Prof. Vijay Janapa Reddi): **built and submitted a full lab** on why temperature-0 LLMs are not deterministic (**PR #2175**), passing the book's test suites (**2,658 tests, 0 failures**); team: "the feature looks great".
- **NASA Conjunction Assessment Risk Analysis (CARA)**: my reproduction of its Monte Carlo collision probabilities (**53/53** cases, **15–19×** faster per core) is **under review by NASA CARA**, referred by its Operations Team Lead (emails).
- **Entrepreneur First** (UK): scouted me from public work, citing my cuBLAS-parity GEMM and a negative result: "which I rarely see".

RESEARCH (SOLE AUTHOR)

- **"Removing the Token Tax from an Existing Model"**, **"The Sampler Was Not the Policy"**, **"Reproducing NASA CARA's Monte Carlo Probability of Collision"** (arXiv, forthcoming).

PROJECTS

ratchet | PyTorch, C++/CUDA: RL post-training (GRPO) from scratch | study page
- Lifted Qwen2.5-0.5B on GSM8K from **48.1% to 52.5%** in 100 steps on 2× T4, scoring all 1,319 test problems.
- Cut weight sync **369×** with in-place GPU updates, and made steps **1.9×** faster by overlapping generation with training.

tessel | Rust: a tile language and compiler for GPU and TPU kernels | study page
- Matched **cuBLAS** on GEMM (**1.00–1.04×**) and beat PyTorch SDPA on causal attention (**1.3–1.9×**) on a T4, checked against fp64.

lipi | PyTorch: cutting the token tax on Indian languages | study page
- Built a tokenizer that needs fewer tokens than **GPT-4o's and Gemini's in 23 of 24** Indian languages (on Hugging Face).

samebit | C, CUDA, PyTorch, MLX: LLM inference in integers, the same bits on every chip
- Ran **Qwen2.5-0.5B (494M params) in integers**: within **0.75%** of float32 perplexity, **bit-identical on NVIDIA T4 GPU and CPU**.

More projects
- **chirp**: GPU gravitational-wave search built from general relativity; finds **62 of 71 LIGO** O4a events blind, all 71 at **95%** of catalog SNR.
- **syndrome**: GPU quantum decoder bit-identical to **IBM's**, **28.7×** faster per shot; **kessler**: matches **NASA CARA** on 53/53.

EXPERIENCE

GIGZEN | Co-founder & Founding Engineer | Bhubaneswar, India **Apr 2026 – Present**
Early-stage startup (name approved by India's Ministry of Corporate Affairs; incorporation in progress) | Sole engineer
- Released **Populace**, an open-source testing tool (**1,600+ npm downloads**) that found **5 bugs** QA missed.

TECHNICAL SKILLS

ML systems: PyTorch, JAX, MLX; GRPO, DDP, ZeRO/FSDP, MoE expert parallelism, paged KV cache, int8 quantization
Languages & GPU: Python, C++20, CUDA (tensor cores), Rust, C, Swift, SQL; Triton-style kernels, Metal, Pallas (TPU)

EDUCATION

INSTITUTE OF ENGINEERING & MANAGEMENT, KOLKATA | B.Tech, Computer Science & Engineering (AI & ML) **2022 – 2026**