SANKUR KUNDU

Curriculum Vitae | ML Systems Engineer & Independent Researcher
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, Massachusetts Institute of Technology, NVIDIA, Apple, Meta, IBM Quantum

SUMMARY

ML systems engineer and independent researcher (B.Tech, Computer Science and Engineering, 2026), sole author of **three manuscripts** and co-founder of Gigzen. Seventeen of my fixes were merged in 22 days **by the people who lead the projects**, into the **US Department of Energy's Kokkos** (merged by its lead at **Los Alamos National Laboratory** after approval from **Oak Ridge National Laboratory**), the **Massachusetts Institute of Technology's (MIT) MEEP** (two, both merged by its original lead developer), the **Barcelona Supercomputing Center's (BSC) Autosubmit** (two, invited for its 4.2.0 release and credited in its changelog), **NVIDIA's Ising-Decoding** (two), **Apple MLX**, **Meta ExecuTorch** (two), **Tencent ncnn** (merged by its creator), **IBM Quantum's Relay-BP** (merged by its first author), **Carnegie Mellon University's XGrammar**, **Hugging Face OpenEnv** (two), the **University of Oxford's PyZX** and **JaxMARL**, and I built and submitted a full lab to **Harvard's Machine Learning Systems textbook** (Prof. Vijay Janapa Reddi). Fixes to public-sector science software are in review at the **European Organization for Nuclear Research (CERN)**, the **European Centre for Medium-Range Weather Forecasts (ECMWF)**, the **UK Met Office** and **Science and Technology Facilities Council (STFC)** (PSyclone), the **Swiss National Supercomputing Centre (CSCS)** and **CSIRO**, Australia's national science agency. I build systems from scratch and benchmark them against the leaders: GPU kernels at **cuBLAS** parity, an LLM engine matching **vLLM** token for token, a collision-risk engine matching **NASA's Conjunction Assessment Risk Analysis (CARA)** team on all 53 of its test cases, a 494M-parameter LLM that gives **identical bits on an NVIDIA GPU and a CPU**, a gravitational-wave search that finds **62 of 71 Laser Interferometer Gravitational-Wave Observatory (LIGO)** O4a detections blind, and a test showing the **Vesuvius Challenge's** official ink model finds **79% of real ink** in an unread Herculaneum scroll.

HONORS & RECOGNITION

- **Harvard's Machine Learning Systems textbook** (Prof. Vijay Janapa Reddi): **built and submitted a full lab**, The Same-Bits Contract, on why temperature-0 LLMs are not deterministic (**PR #2175**, Nov 2026). It passes the book's test suites (**2,658 tests, 0 failures**); its team: "the feature looks great" (discussion).
- **NASA Conjunction Assessment Risk Analysis (CARA)**: my independent reproduction of CARA's Monte Carlo results on all 53 of its test cases is **under review by NASA CARA**: its Operations Team Lead referred it to CARA's analysts (20 Oct 2026; emails).
- **Entrepreneur First** (UK): its Talent Investor scouted me from public work alone (15 Oct 2026), citing my cuBLAS-parity compiler and a negative result I chose to publish, "which I rarely see" (email).

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

US DEPARTMENT OF ENERGY NATIONAL LABORATORIES, KOKKOS | GPU performance portability | PR #9738 **Merged 5 Nov 2026**

- Kokkos::sort did not compile on **NVIDIA CUDA** or **AMD HIP** for value types without a default constructor; it now sorts an index permutation and moves the values (**20/20** GPU sort tests on an NVIDIA T4).
- **Merged by Damien Lebrun-Grandié**, who leads the Kokkos project at **Los Alamos National Laboratory** ("Generally the approach looks good"), after approval by **Daniel Arndt** of **Oak Ridge National Laboratory**; reviewed by **Andrey Prokopenko** (Oak Ridge), who proposed my next fix. Listed under Bug Fixes in the **Kokkos 5.4 release notes**.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY (MIT) MEEP | Electromagnetics simulator | PR #3364, #3363 **Merged 5 & 6 Nov 2026**

- Remaining-time estimates fell to **0.25×** the real time; now **0.97–1.02×** (issue #2694, where **Prof. Steven G. Johnson** asked for a pull request). Merged by **Ardavan Oskooi**, MEEP's original lead developer and first author of its reference paper, 5 Nov.
- **Second fix,** #3363 (issue #3170, filed by Prof. Johnson), **merged by Oskooi, 6 Nov**: raising an error "is the correct approach because in Python, failing fast and explicitly is strongly preferred over silently mutating inputs."

BARCELONA SUPERCOMPUTING CENTER (BSC), AUTOSUBMIT | Spain's national supercomputing centre; climate-model workflows | #3405, #3406 **Merged 6 Nov 2026**

- On PostgreSQL, job timestamps such as 20261027164231 overflowed 32-bit columns and `autosubmit run` crashed (NumericValueOutOfRange); they are now 64-bit, checked on PostgreSQL 16. Maintainer **Bruno P. Kinoshita** invited the fix, "as we need it for the 4.2.0 release", and merged it: "Looks great! ... Thank you!!"; **credited in the Autosubmit 4.2.0 changelog**.
- **Second fix,** #3406, at his request ("That'd be helpful!!!"): a CI test that failed on busy runners now waits for the process to exit (**5 of 5** passes with every CPU busy). Merged: "Thanks (again) @Shakhtar-Sankur !!!"

RELAY-BP, IBM QUANTUM | Quantum error-correction decoder (arXiv 2506.01779), © IBM, **113★** | PR #49 **Merged 5 Nov 2026**

- Serial and parallel batches decoded **1,970 of 2,000** shots differently; now **0**. Relay-BP's first author **Tristan Müller (IBM Quantum)** chose this fix ("thank you for opening this! I think 1. is fine!") and merged it, 5 Nov: "Thank you! Happy with this, will merge it."

NVIDIA ISING-DECODING | NVIDIA's AI decoders for quantum error correction | #98, #99 **Merged 30 Oct & 2 Nov 2026**

- The tests and offline docs looked only for .pt checkpoints, so users of the SafeTensors models **NVIDIA** publishes on **Hugging Face** failed them; they now load the newest published file, with a test that it loads bit for bit. Maintainer Ivan Basov approved and merged it after review rounds, 30 Oct: "Merged, thank you for the contribution and for working through the review rounds."
- **Second fix,** #99, split out at his request: the documented checkpoint setting is now honoured for .pt files; merged 2 Nov: "Merged, thank you for the follow-up and for adding the precedence test."

APPLE MLX | Apple's ML framework for Apple silicon, **29k★** | PR #4737 **Merged 22 Oct 2026**

- Fixed a data race in the distributed ring backend that links Macs for multi-machine training (ThreadSanitizer: 2–4 races per run → **0**; ring tests pass on 8 ranks). Approved and merged by MLX maintainer @zcbenz, 22 Oct.

META EXECUTORCH | On-device AI in Instagram & WhatsApp, **5.1k★** | #23677, #23912 **Merged 16 & 28 Oct 2026**

- The 2- and 4-bit quantized embedding kernels of on-device LLMs read memory outside the table on an invalid token id; added the missing bounds check. Approved by **Meta** maintainers Scott Roy ("LGTM! Thanks for the fix!") and Jacob Stevens; merged 16 Oct.
- **Second fix,** #23912: the .out variants read past their scales; revised after **Meta** review, merged 28 Oct (86/86 tests pass).

TENCENT NCNN | Tencent's on-device AI inference framework, **24k★** | PR #7239 **Merged 5 Nov 2026**

- Mish returned NaN on Vulkan GPUs for large inputs: exp(x) overflowed; clamped it in the Mish layer and the fused activation (**30 NaN outputs → 0**). **Approved, extended with tests and merged by ncnn's creator, nihui**: "Thanks for your contribution!"

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

- Jump-forward decoding crashed (UnicodeDecodeError) whenever a grammar forces part of a UTF-8 character, the bug behind issue #132; it now returns only whole characters. Merged by maintainer Linzhang Li, 29 Oct.

UNIVERSITY OF OXFORD PYZX | ZX-calculus quantum-circuit optimiser, **553★** | PR #563 **Merged 22 Oct 2026**

- Three optimisers put an X rotation on the wrong qubit, changing the unitary of **25 of 100** random Clifford circuits (0 after the fix). Merged by co-creator John van de Wetering, 22 Oct: "Thanks for the fix! This was a regression from an earlier PR."

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

- The rollout manager behind 8 Q-learning baselines zeroed the first observation of every new episode after an auto-reset, so agents chose their first action blind; two of three new tests fail without the fix. Approved and merged by lead author Alex Rutherford, 26 Oct: "Good find!!"

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

- Fixed reward rubrics for RL training that ran pre-hooks after scoring and scored async judges as unawaited coroutines; both merged 19 Oct: "177 rubric tests pass".

In review

- **Google JAX** #41859 (**36.5k★**; **Google** assigned it to JAX co-author Matthew Johnson, 21 Oct): jax.grad through lax.reduce_window with a custom reduction has failed since 2021; its derivative is now a fold over the window offsets; the **242** new test cases fail on main.
- **Google DeepMind MuJoCo** #3756: MJX, MuJoCo's JAX backend, silently dropped plugin forces (a PID actuator: **0.441 in MuJoCo, 0.0 in MJX**); it now rejects them. Found by comparing MJX with MuJoCo on every model in the repository and **MuJoCo Menagerie**.
- **Microsoft ONNX Script** #3157 (the library behind **PyTorch's** ONNX exporter): its attention fusion made **GPT-2 unloadable in ONNX Runtime**; the fusion now keeps the original output shape (max difference **1.2e-7**; the new test fails on main).
- **Allen Institute for AI (Ai2) OLMo-core** #945: interleaved data loading trained on only the **first 5%** of a larger source, the same examples every epoch; each epoch now draws a fresh random sample (data seen over 10 epochs **5% → 40%**).
- **vLLM** #60955 (**95.7k★**): **SGLang** #47305 (**38.4k★**): **PyTorch** #199875; **MIT Han Lab** #957; **ETH Zurich** DaCe #2729 (maintainer: "once green can be merged").

GOVERNMENT & PUBLIC RESEARCH: 5 FIXES MERGED AT 3 INSTITUTIONS, 5 IN REVIEW

- **Merged (details above): US Department of Energy (DOE) national laboratories**, Kokkos #9738, by its lead at Los Alamos National Laboratory; **Barcelona Supercomputing Center (BSC)**, Spain's national supercomputing centre, Autosubmit #3405 and #3406, credited in its 4.2.0 changelog; **Massachusetts Institute of Technology (MIT)**, MEEP #3364 and #3363, by its original lead developer.
- **Weather & climate, European Centre for Medium-Range Weather Forecasts (ECMWF): ecCodes** #684, its GRIB library for global weather data (intergovernmental, **35 states**, HQ Reading, UK): reversed-scan Gaussian grids got wrong coordinates (**~6,000 points**); now none, all **112** tests pass.
- **Weather & climate, UK Met Office and Science and Technology Facilities Council (STFC) Hartree Centre: PSyclone** #3668, code generator of the Met Office's next-generation **LFRic** model: Fortran BIND(C) is no longer dropped, and BIND(C) functions no longer crash it (**6,502** tests pass); its maintainer: "otherwise it looks good".
- **Particle physics, European Organization for Nuclear Research (CERN): Rucio** #8862, the data management system of the **Large Hadron Collider** experiments: failed uploads and downloads reported success (exit code 0); **11 commands** now return the real error code.
- **National supercomputing, Swiss National Supercomputing Centre (CSCS), ETH Zurich: ReFrame** #3750, its supercomputer testing framework: a regression test for an issue on its 4.10 milestone (**1,451** tests pass).
- **Oceans, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia's national science agency: emsarray** #233: forecast datasets got the wrong time coordinate; candidates are now ranked as its maintainer suggested (**440** tests pass).

MANUSCRIPTS AND RELEASES

- S. Kundu. "Removing the Token Tax from an Existing Model: A Vocabulary Extension for 24 Indian Languages That Keeps Every Other Token." Manuscript, 2026; arXiv (cs.CL) submission in progress (paper).
- S. Kundu. "The Sampler Was Not the Policy: Finding and Removing a Silent Training–Inference Mismatch in a From-Scratch GRPO Stack." Technical report, 2026; arXiv (cs.LG) submission in progress (paper).
- S. Kundu. "Reproducing NASA CARA's Monte Carlo Probability of Collision: An Open Two-Body Implementation, Checked on All 53 of CARA's Test Conjunctions." Manuscript, 2026; arXiv (physics.space-ph) submission in progress (paper).
- Released lipi-Indic 96K (on Hugging Face) and samebit, both Apache-2.0.

ML SYSTEMS: TRAINING, INFERENCE AND RL

ratchet | PyTorch, C++/CUDA: reinforcement-learning 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.
- Found fp16 weight sync silently erasing most updates, so the sampler ran a policy the trainer never had (gap **5.2 → 0.04** nats).

relay | C++, CUDA, Swift: a disaggregated, fault-tolerant LLM inference cluster | study page

- Wrote flash-attention and flash-decoding kernels that cut 2,048-token prefill **8.7×** (3.7 → 0.43 s) and batch-32 decode **5.4×** on a T4.

- Kept output identical through failures: watchdogs reroute requests from killed, frozen or stalled workers (**384/384 through 60 kills**; none lost in 300 fault-injection seeds).

samebit | C, CUDA, PyTorch, MLX: LLM inference in integers, the same bits on every chip

- Ran **Qwen2.5-0.5B** (494M parameters, 24 layers, grouped-query attention) **entirely in integers** (int8 weights, int16 activations, int32 residual stream, int64 sums): WikiText-2 perplexity **23.16 vs 22.99** for float32 (**+0.75%**); 93.25% of top-1 tokens unchanged.
- Showed **bit-identity on an NVIDIA Tesla T4** (public Kaggle run): it reproduces the x86 CPU's logit digests on real text, with each dot product split over 1–64 blocks and summed by atomicAdd in a new order every run; float16 shifts logits by up to **0.25**.

switchyard | CUDA, C++: mixture-of-experts serving with expert parallelism | study page

- Wrote routing, sorting and grouped expert GEMM/GEMV kernels on tensor cores and peer-to-peer dispatch: OLMoE-1B-7B decodes **2.2× faster than Hugging Face**, bit-identical to one GPU, with dispatch **1.25× faster than NCCL**.
- Showed dropping tokens at full experts raises perplexity 12.9 → 20.2 even at capacity factor 2, so the kernels never drop one.

crucible | Python on Linux: sandboxed environments for RL on coding agents | study page

- Trained with GRPO on ratchet, every reward from a sandbox: MBPP pass@1 **34.9%** → **38.6%** (3 seeds; paired test, p = 0.013).

tandem | C++, CUDA: distributed training without torch.distributed or NCCL | study page

- Built DDP, ZeRO-1/2/3, GPipe and 1F1B on my own ring all-reduce, training **bit-identically to PyTorch DDP/FSDP on NCCL**.

GPU KERNELS AND COMPILERS

tessel | Rust, zero dependencies: 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×**), torch.compile on RMSNorm (1.27×) on a T4, checked against fp64.
- Built an LLM engine on them (paged KV, batching, CUDA graphs) matching **vLLM token for token at 1.05× its throughput**.

kiln | Rust, zero dependencies: an ML compiler from ONNX to SIMD C and CUDA | study page

- Gave the lowest latency on all three benchmark models on 4 CPU cores: BERT in **117 ms** vs 149 for torch.compile and 189 for JAX/XLA; on a T4, SmolLM2 **1.65×** faster than torch.compile.

SCIENCE: GRAVITATIONAL WAVES, ANCIENT SCROLLS, SPACE AND QUANTUM COMPUTING

chirp | Python, PyTorch (CUDA): an independent search for gravitational waves in public LIGO data

- Searched every two-detector event of the **LIGO**–Virgo–KAGRA O4a catalog (GWTC-4.0) with a 3,886-template bank on one T4 in **17 minutes**: **all 71** with complete data re-found at a median **95%** of the catalog's network SNR, 67 within 50 ms (public run).
- Added gating and the Allen chi-squared veto with reweighted SNR, as **LIGO's** PyCBC searches use: searched blind over 188 s of data per event, the loudest trigger is the real event for **62 of 71** (48 without the veto), at a cost of 2% of the events' own SNR (public run).

ink-sensitivity | Python, PyTorch: calibrating ink detection for the unread Herculaneum scrolls (Vesuvius Challenge)

- The official 9 µm model finds **79%** of the real-ink windows at a 5% false-positive rate (**AUC 0.94**; 168 ink vs 168 controls, no untouched window flagged), and transplanted ink scores like the same text left in place.

kessler | C++/CUDA, Python: an independent auditor of space traffic | live globe

- Matched **NASA CARA's** Monte Carlo collision probability on **all 53** of its test conjunctions (chi-squared p = 0.67), including the 28 where the standard 2D method is wrong, **15–19× faster** per core than **NASA's** library.
- Screens all **531 million pairs** of 32,593 tracked objects a day ahead in **28 s** on a T4, matching CelesTrak SOCRATES to 0.24 m.

syndrome | C++, CUDA: an open GPU decoder for IBM's quantum LDPC codes | study page

- Matched **IBM's** Relay-BP reference **bit for bit on 256/256 shots** of the [[144,12,12]] gross code (391,320 edges).
- Decoded a single shot **28.7× faster than IBM's decoder** (median 37.5 ms vs 1.08 s) and a batch **4.0×** faster on all CPU threads.

LANGUAGE AI

lipi | PyTorch: cutting the token tax on Indian languages | study page

- Measured 24 Indian languages on 15 tokenizers (Odia: 4.99× English on GPT-4o) and built lipi-Indic (on Hugging Face), which needs fewer tokens than **GPT-4o's and Gemini's in 23 of 24** (1.30× vs 1.92×, 1.58×).
- Extended Qwen2.5-0.5B: Odia takes **8.6×** fewer tokens, predicts **20% better** in equal GPU time; Odia→English chrF++ 16.0 → 24.8.

DISTRIBUTED SYSTEMS AND VERIFICATION

quorumdb | Rust, zero dependencies: a distributed SQL database, PostgreSQL wire-compatible | study page

- Built LSM storage, self-rebalancing multi-Raft, serializable transactions (Percolator) and a SQL planner: **1,500 tpmC** (TPC-C-like).
- Found **6 real bugs** with deterministic simulation (**10,000 Raft clusters**) and TLA+, which catch **all 15 planted bugs**.

EXPERIENCE

GIGZEN | Co-founder & Founding Engineer | Bhubaneswar, India **Apr 2026 – Present**

Early-stage startup (company name approved by India's Ministry of Corporate Affairs; incorporation in progress) | Sole engineer

- Built **Waggle**, a three-app Android delivery platform (TypeScript/React, PostgreSQL): **53 tables, 154 SQL functions**.
- Moved dispatch, UPI reconciliation and GST billing into the database; tested its access rules with **150 automated attacks**: all blocked.
- Released **Populace**, an open-source simulated-user tester (npm, **1,600+ downloads**) whose first run found **5 missed defects**.

TECHNICAL SKILLS

GPU and compilers: CUDA (tensor cores, mma.sync, ldmatrix, graphs, IPC), NVRTC, Triton-style tile languages, Metal, Pallas, SIMD
ML & distributed systems: GRPO, DDP, ZeRO, FSDP, MoE expert parallelism, paged KV, int8; Raft, MVCC, Percolator, LSM, TLA+
Languages and platforms: C++20, CUDA, C, Rust, Python, Swift, TypeScript, SQL; PyTorch, JAX, MLX, PostgreSQL, Docker, Kubernetes

EDUCATION

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