Institute of SCIENCE TOKYO

Supercomputing Research Center, IIR

AI-Assisted GPU Application Modernization

HAIRDESC: Advanced HPC-AI R&D Support Center

HAIRDESC Advanced HPC-AI R&D Support Center

Overview

Science Tokyo is a core institution of HAIRDESC, the Advanced HPC-AI R&D Support Center that helps application teams move scientific codes to GPUs with AI assistance. TSUBAME4.0 provides 960 NVIDIA H100 GPUs for training, evaluation and real-machine validation. Our principle is simple: generated code is only a candidate until it compiles, preserves numerical results and improves measured performance on the target system. This principle links three activities. We build an auditable workflow for GPU migration, we develop the Swallow family of open Japanese language models, and we turn the methods into portable GPU education. Fig. 1 shows the evidence loop for a single kernel: on an H100, 36 generate-verify-measure-repair attempts produced 25 correct FP32 GEMM kernels and a best of 574 TFLOPS, 78% of the optimized cuBLAS reference. Failures were not hidden; they became evidence for the next repair, and the same trail from prompt to verified result is what application teams can reproduce.

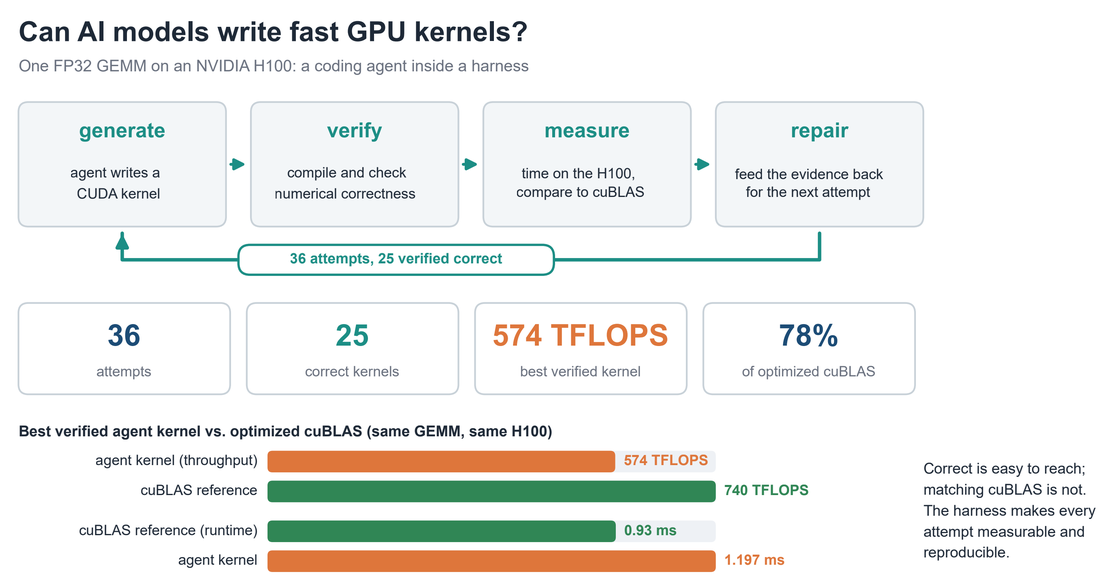


Fig. 1 Can AI models write fast GPU kernels? Hopper H100, FP32 GEMM: 36 attempts, 25 correct, best 574 TFLOPS (78% of cuBLAS) through repeated generate, verify, measure and repair cycles.

Research

Science Tokyo develops an AI-coding-first workflow for modernizing scientific applications. A user supplies the target code and a performance goal. A project manager decomposes the work, coding agents propose bounded changes, and the target supercomputer performs compilation, execution, correctness checks and timing. Failures and measurements are returned to the agents, so the next proposal responds to evidence rather than model confidence, and a human can intervene at every stage. Each accepted change keeps its source revision, compiler and runtime configuration, correctness criterion, timing method and baseline, so performance claims stay reproducible for the next team.

The second track is Swallow (Fig. 2): open English-weight models adapted by continual pre-training, supervised fine-tuning and reinforcement learning with verifiable rewards.

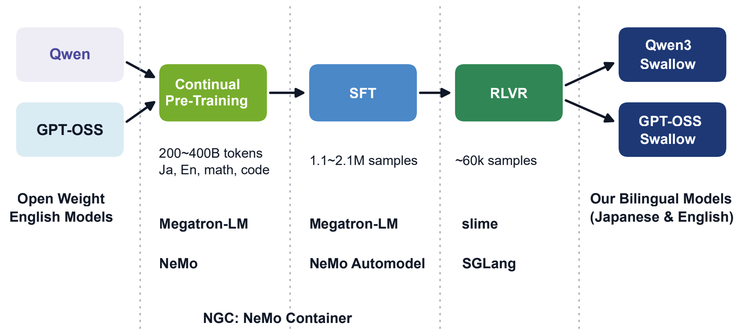


Fig. 2 Swallow open-model pipeline, 2026: 132 models, 19 datasets, 2.4M+ downloads.

What HAIRDESC Enables

HAIRDESC turns individual demonstrations into a reusable national capability. First, application teams gain a measurable route from legacy CPU code to GPU execution: one bounded change, one correctness check and one performance result at a time. Second, shared evidence makes both success and failure transferable; compiler settings, numerical criteria, runtime conditions and performance traces are reused by later teams instead of being rediscovered. Third, open models and controlled execution support scientific environments where code cannot be sent to external services. Finally, the same workflow becomes an education program: researchers learn portable programming, profiling, libraries, multi-GPU execution and complete-application modernization on real machines. Science Tokyo therefore contributes more than hardware or a model: a connected path from machines, through verified methods and open models, to people who can modernize scientific applications with evidence.

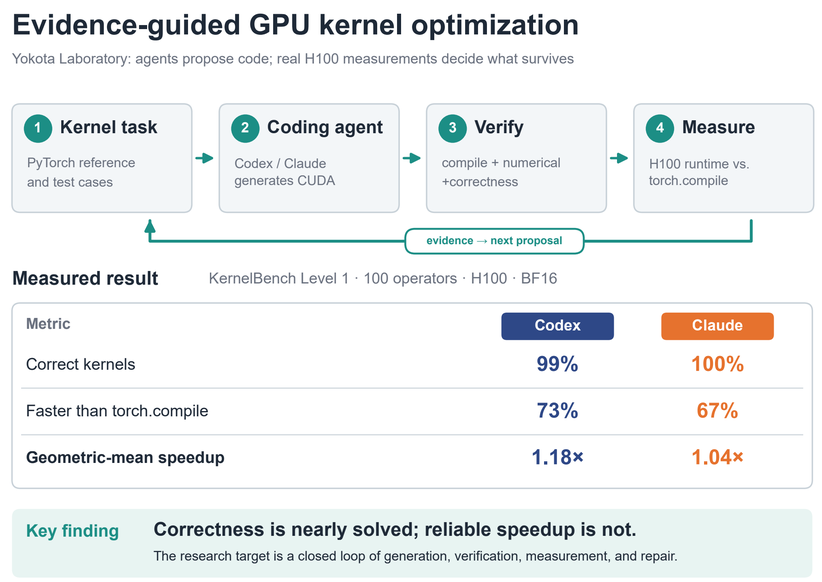


Fig. 3 Yokota Laboratory's evidence-guided GPU kernel optimization study. KernelBench Level 1 (100 single operators), H100, BF16: two agent environments reach 99% and 100% correct kernels; 73% and 67% are faster than torch.compile; geometric-mean speedup 1.18x and 1.04x. Source: W. Martin, Agent Harnesses for GPU Kernel Optimization (2026)

Highlights in 2026

- SALMON 2.2.2 (TD-DFT): an agent harness with optimization skills cut the end-to-end run from 538.8 to 487.6 min (9.5%); 8,000 sampled outputs, four science tables and six checkpoints are identical.
- GENESIS: a GPU-resident redesign reaches 24-30x over stock 2.1.6 on one 8xB200 node for 1.07M- and 1.69M-atom systems; all 56 tests pass (preliminary, Sept. 2026).
- 2nd HAIRDESC Symposium, Tokyo, 7 Oct. 2026; Nagoya University joined as the fourth core institution.

Contact

Advanced HPC-AI R&D Support Center (HAIRDESC):
hairdesc-qa-general@hairdesc.jp https://hairdesc.jp/
Science Tokyo site:
Supercomputing Research Center, Institute of Innovative Research, Institute of Science Tokyo

https://www.isct.ac.jp/