IEEE/ACM SC EESP WORKSHOP 2026

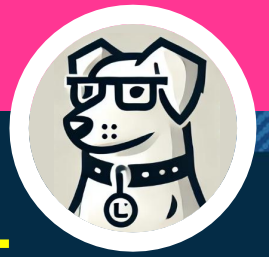
LASSI-EE: LLM-Based Automated Refactoring of Parallel Scientific Codes for Energy Efficiency

Matthew T. Dearing, University of Illinois Chicago; **Yiheng Tao**, University of Illinois Chicago;
Xingfu Wu, Argonne National Laboratory; **Zhiling Lan**, University of Illinois Chicago;
Valerie Taylor, Argonne National Laboratory



SC26 hpc unites | Chicago

Can LLMs generate energy-efficient parallel scientific code?



The energy challenge for HPC

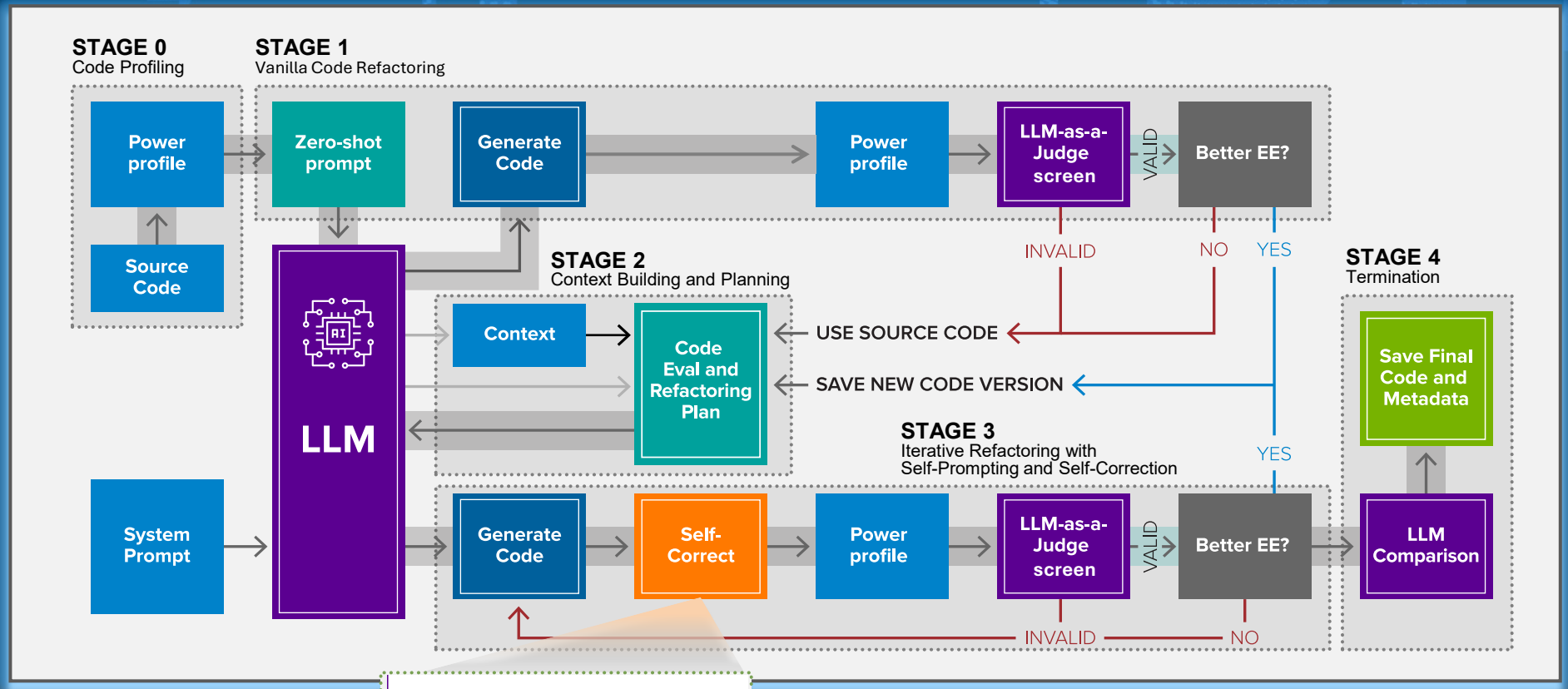
- Exascale systems require 24 to 39 MW of power.
- HPC efficiency efforts (hardware, networking, scheduling) often overlook application code.
- Energy optimization takes deep expertise and manual tuning.
- Recent LLM code generation targets correctness or runtime, not energy.

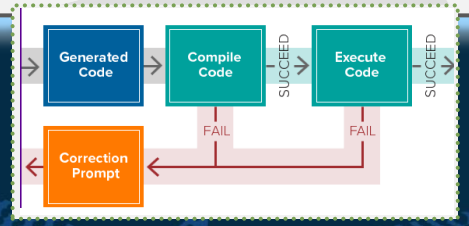
Our approach: LASSI-EE

- **L**LM-based **A**utomated **S**elf-correcting pipeline for generating parallel **Sc**ientific codes for **E**nergy **E**fficiency.
- Automated refactoring guided by power measured on the target GPU.
- Addresses the software dimension of HPC energy efficiency.

Builds on LASSI (CLUSTER LLMxHPC Workshop 2024)

LASSI-EE Design: Augment LLM capability for energy-efficient scientific code





Power is measured on the target GPU at every iteration.

22 applications (20 HeCBench, XSBench, miniMDock),
NVIDIA A100 and AMD MI100,
30 trials each, 1,320 trials, all compiled at -O3.

Results: energy-reduction@*k*

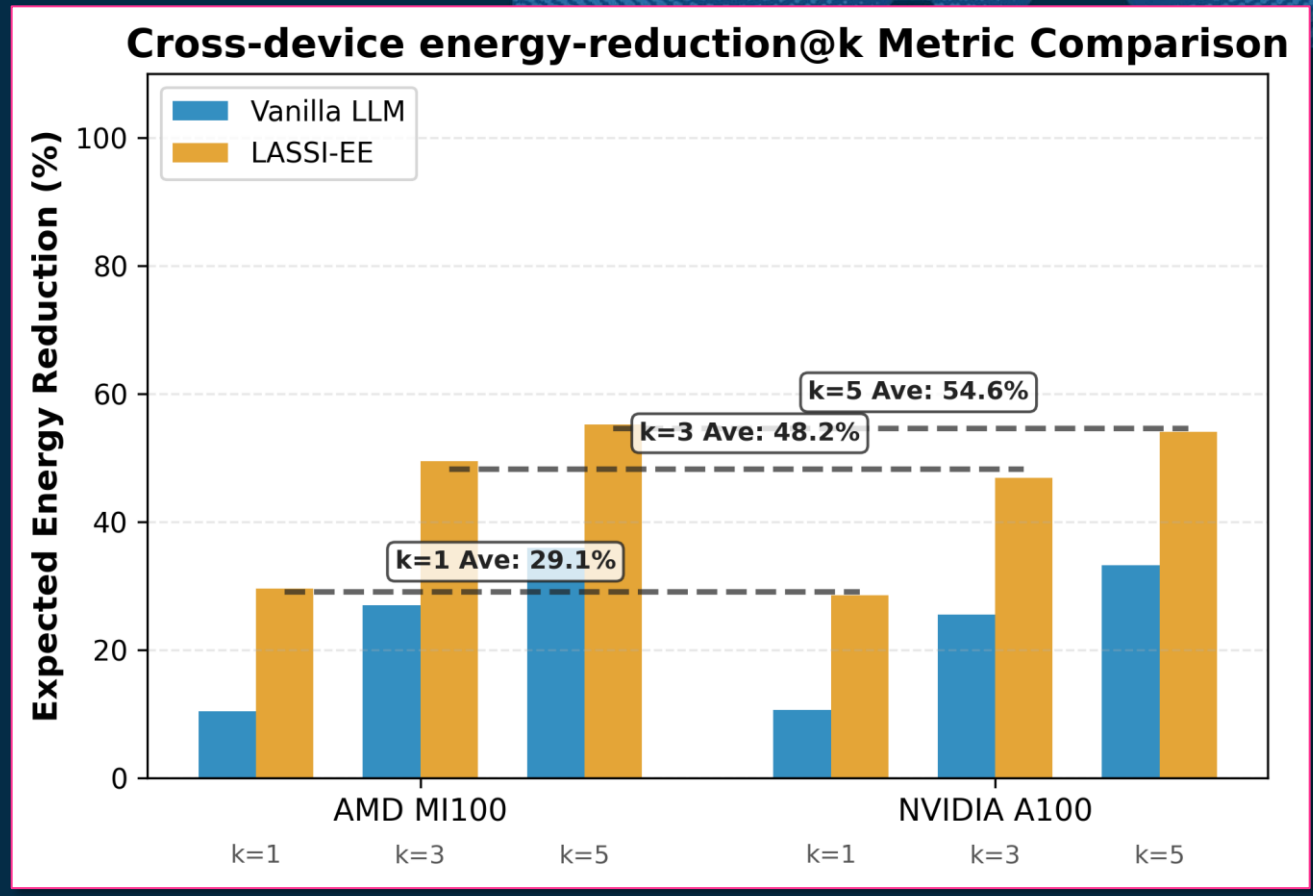
- Generate *k* candidates, keep the most efficient. *What savings can I expect?*

Expected energy savings:

29% at *k* = 1 48% at *k* = 3

- 2.8× the vanilla LLM at *k* = 1
- *k* = 3 is a favorable operating point

Counts all attempts, including failures. Returned passing refactorings average 34% to 36%.



Energy savings with sustained performance



Runtime & precision preserved

- **93.2%** of application-device pairs reduce both energy and runtime.
- **6.8%** reduce energy through power alone, which a runtime-only objective would not produce.
- No precision-reducing transformations among about 3,500 optimizations.

Adapts to application & hardware

- Different code, same outcomes: only **3.5%** of optimizations are identical across AMD and NVIDIA, yet average savings differ by 2.3 pp.
- Runtime reductions are similar on both platforms: **26.2%** (A100) and **29.6%** (MI100).

Code open-source at **`github.com/SPEAR-UIC/LASSI-EE`**
Next up: LASSI-Verify, empirical verification of LLM-refactored code, see it at the Correctness 2026 workshop (Nov 16).

Q&A

