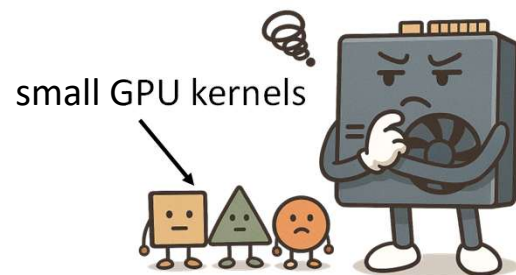




"All Core Computing"

Fully utilize GPU performance on small GPU kernels

- Small GPU kernels cannot fill whole GPU.
Batch calculation is not always suitable.
- Using multi-processes/multi-threads on one GPU with MPS, MIG, and Green Context is promising.



Using both GPU and CPU

- Performance of latest CPUs are not negligible.
- GPU is suitable for low-precision, CPU is suitable for high-precision.

Question: How to implement? How to optimize?

- Satoshi Ohshima, et al., "A Study on the Performance and Usability of Managed Memory and Unified Memory for Accelerating Numerical Calculation Program", IEEE, MCSoc, 2025, doi.org/10.1109/MCSoc67473.2025.00017

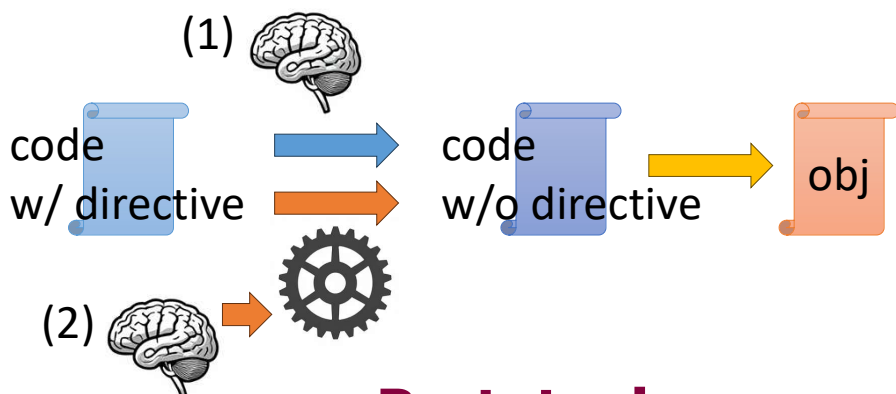
Directive-based Language with LLM

Directive-based languages are widely used in HPC, but developing a custom one is difficult.

LLMs may be useful

- (1) LLM as translator
- (2) LLM generated translator

- How to build?
- What are the concerns?



Prototyping now...

Please check my website for the latest information.