

ML kernel optimization

Performance work on the operations, memory movement, and execution patterns that limit a model on its target hardware. Every change is checked against a baseline for speed, resource use, and numerical behavior.

WHAT WE CAN DO

- 01

Profiling & attribution

Find which kernels, operators, transfers, or synchronization points dominate the workload.
- 02

Kernel implementation

Develop or revise target-specific compute kernels around data layout and execution characteristics.
- 03

Memory & precision

Investigate access patterns, intermediate buffers, fusion opportunities, and precision tradeoffs.
- 04

Correctness & benchmarks

Compare output tolerances and repeatable performance across representative shapes and workloads.

TECHNOLOGIES & WORKING AREAS

CUDA / Triton / Kernel profiling / Operator fusion / Memory layout / Precision / benchmarking

A CLEAR PATH TO HANDOVER.

WHAT YOU RECEIVE

- Profiling report and prioritized bottlenecks
- Optimized kernels or operator changes in scope
- Correctness checks and numerical tolerances
- Reproducible benchmarks with hardware details
- Integration notes and measured tradeoffs

HOW THE WORK PROGRESSES

- 01 **Reproduce the workload**
Target hardware, shapes, precision, and baseline.
- 02 **Profile the bottleneck**
Compute, memory, transfers, and synchronization.
- 03 **Optimize & compare**
Change one limiting operation and remeasure.
- 04 **Validate integration**
Numerical behavior and workload-wide impact.

ILLUSTRATIVE EVALUATION PLAN

Kernel benchmark

Compare the baseline and revised operation across agreed tensor shapes.

End-to-end impact

Check that a local improvement translates to the actual inference or training workload.

TO SCOPE YOUR PROJECT

- / Model, operators, and representative inputs
- / Target GPU or compute architecture
- / Current profiler traces or benchmark code
- / Accuracy tolerances and performance priorities

LET'S DISCUSS YOUR PROJECT

hello@embedly.dev

