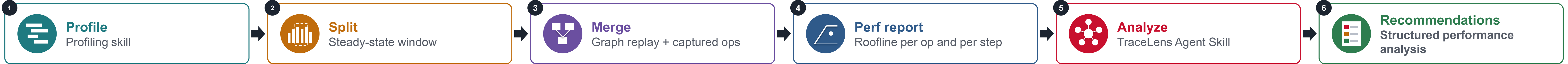
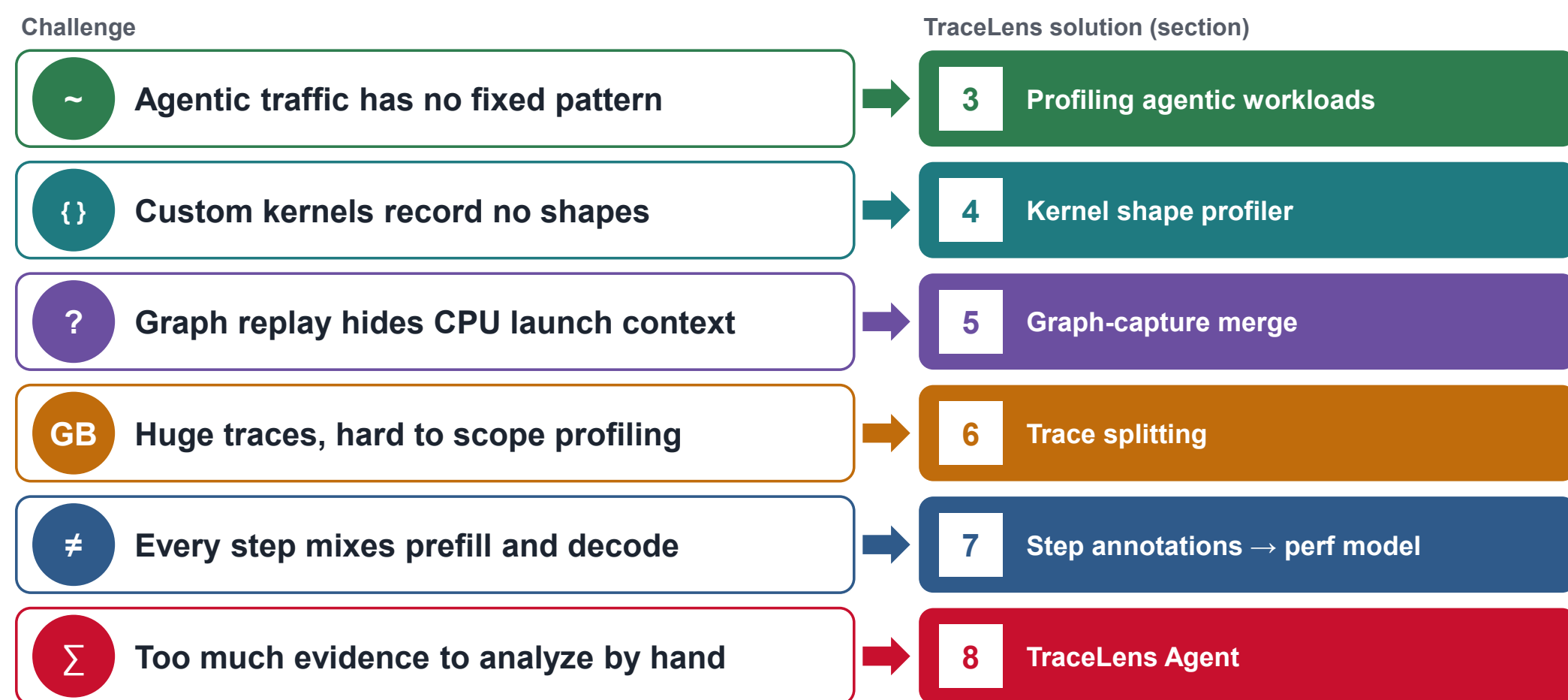


AUTOMATED ANALYSIS

From a live agentic serving run to performance analysis and optimization recommendations

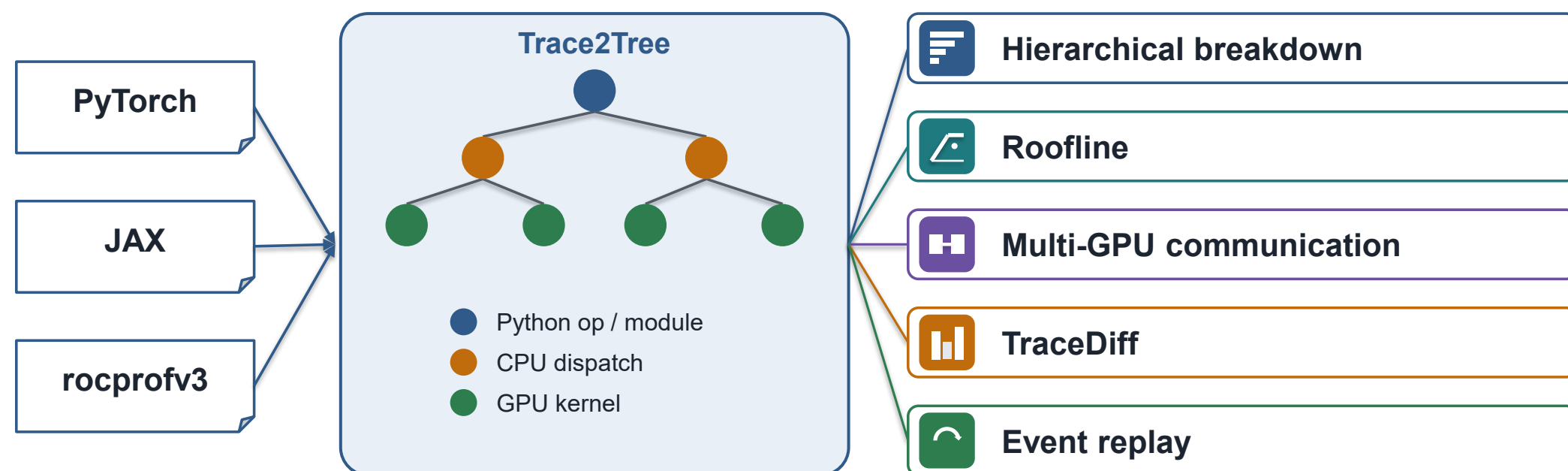


1 Challenges of Analyzing Inference Workloads



2 What is TraceLens?

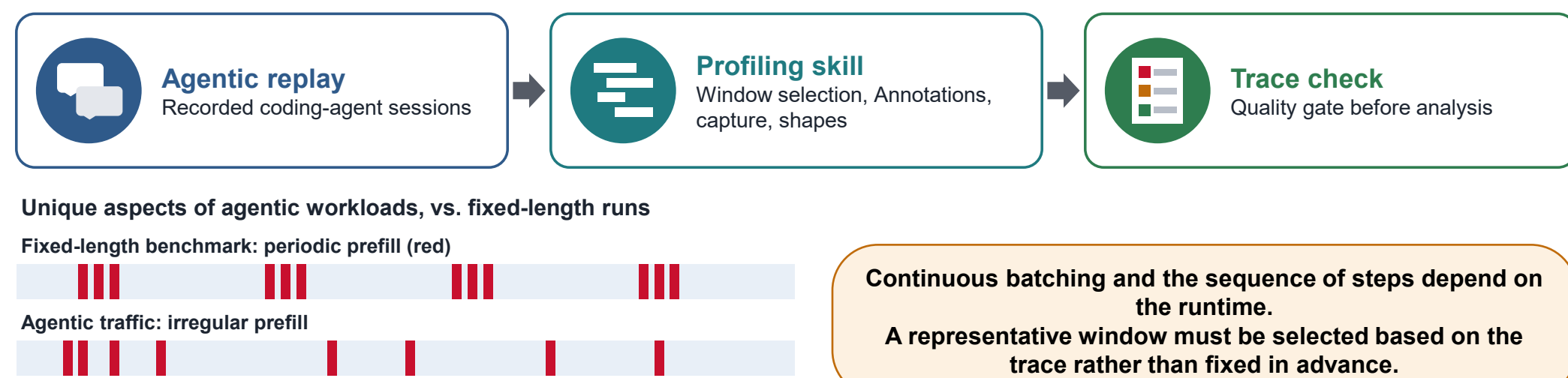
Open-source library that turns PyTorch, JAX and rocprofv3 traces into actionable reports.



Toy example: unified perf report, one row per op × input shape

name	category	Input Dims	time (ms)	TFLOP/s	TB/s	bound	% of roofline	% of time
gemm_a16w16	GEMM	(128,4096) × (12288,4096)	493	387	3.1	memory	43	24
paged_attention	SDPA	Q (128,32,128), KV 2048	157	41	4.0	memory	61	8

3 Profiling agentic workloads

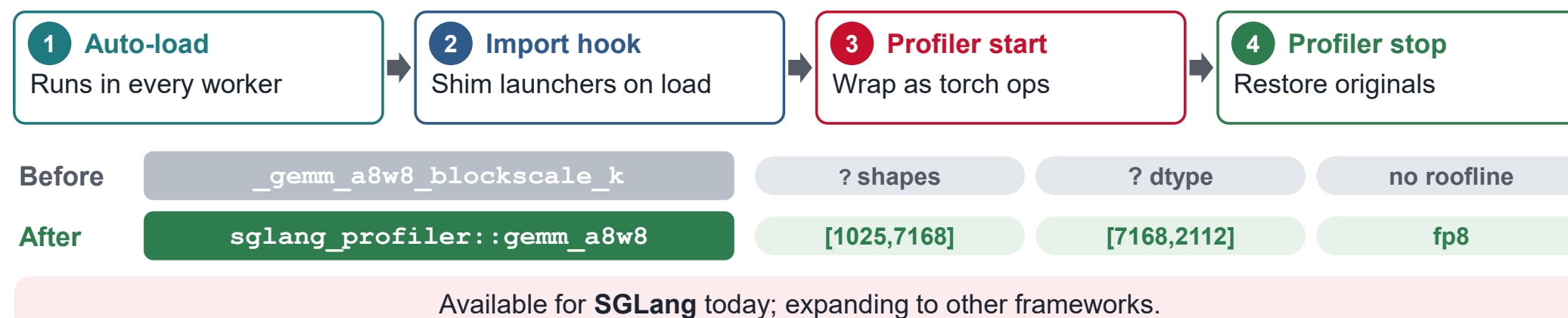


Upstreamed into serving frameworks

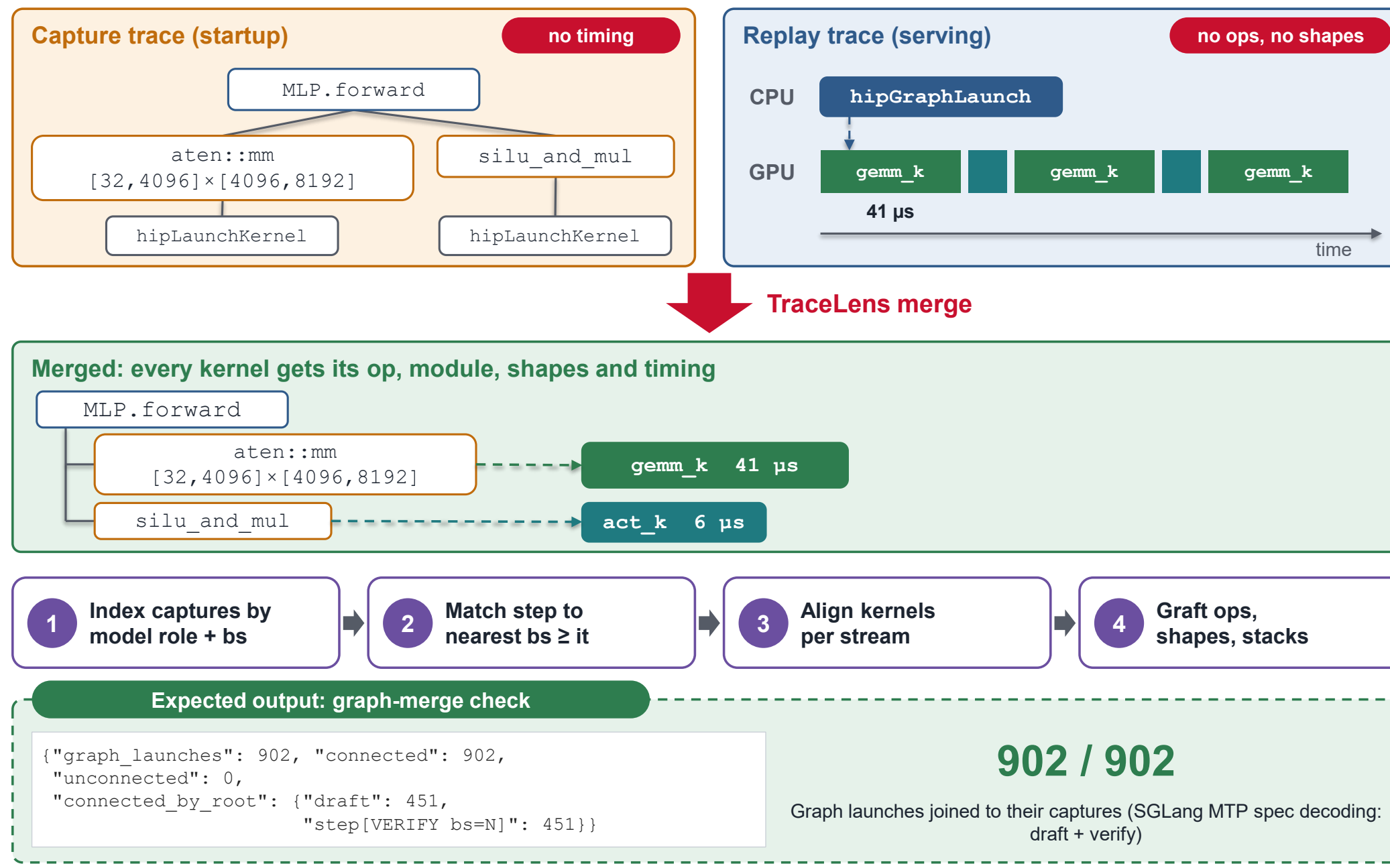
Annotations + graph-capture profiling: vLLM v0.26+, SGLang v0.5.18+, ATOM

4 Kernel shape profiler

Custom kernels bypass the dispatcher and show no shapes; the profiler hooks the Python runtime instead.

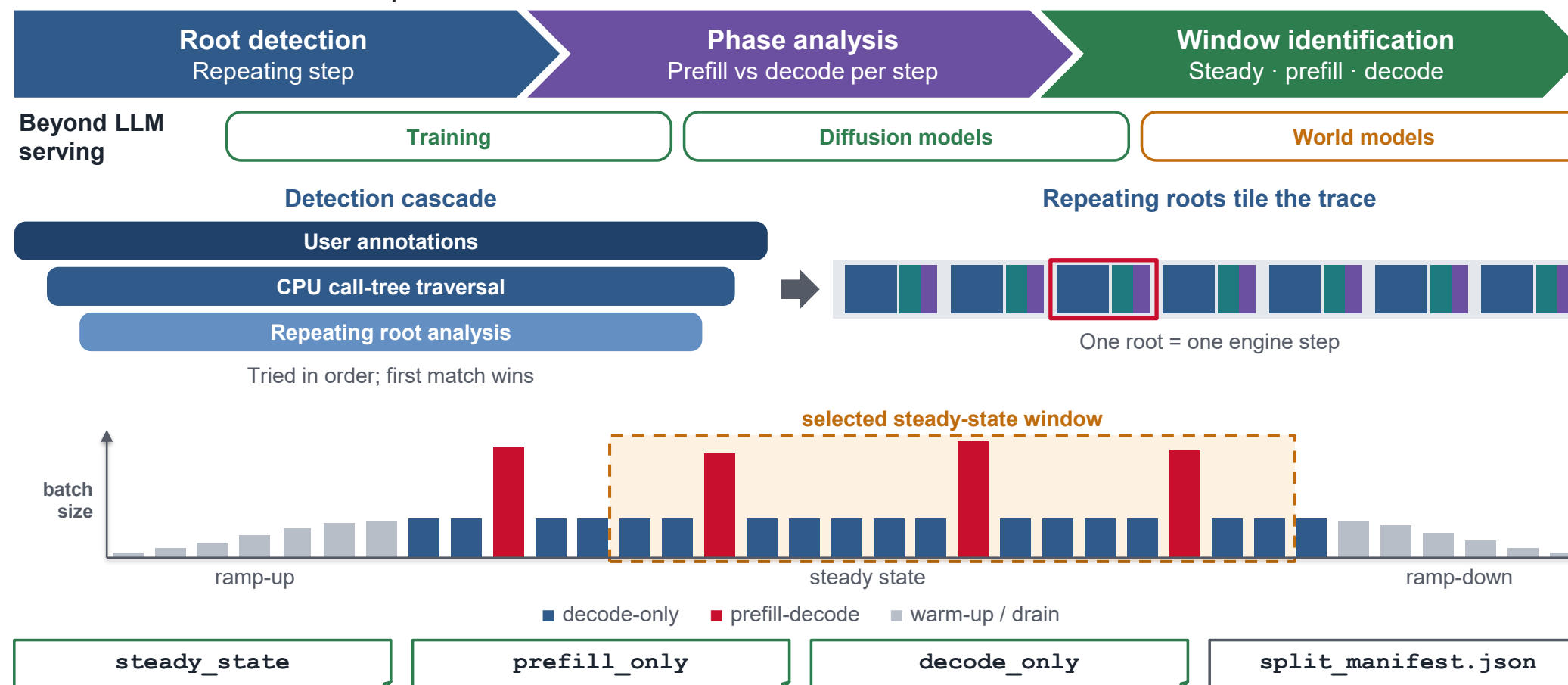


5 Graph-capture merge



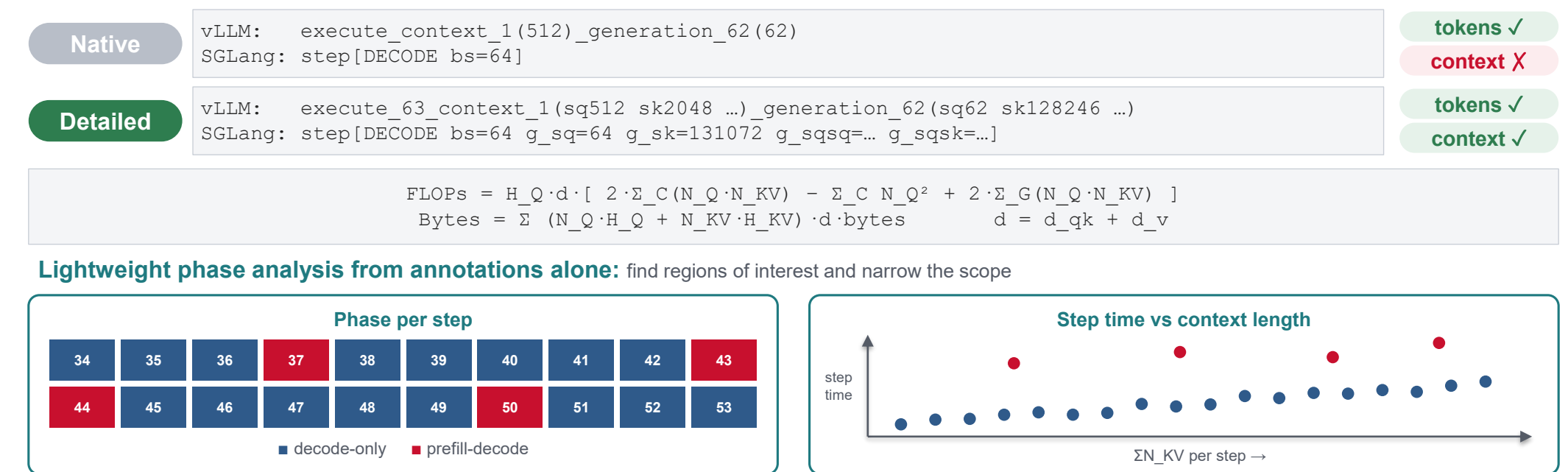
6 Trace splitting: a focused window

Multi-GB trace in, small representative traces out, with or without annotations.

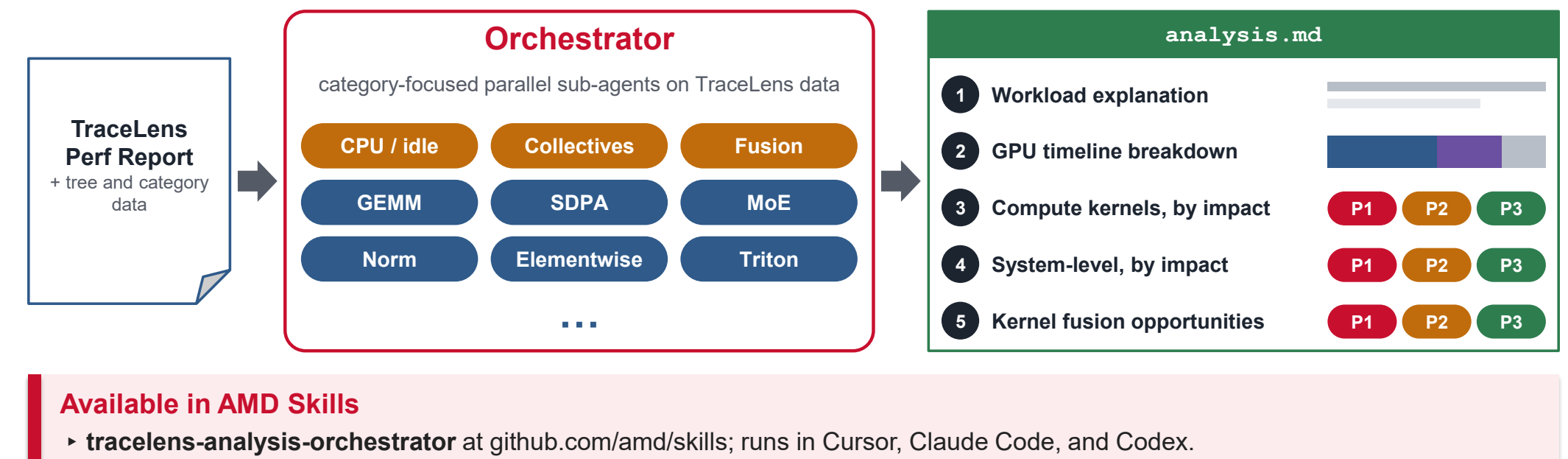


7 Step annotations → analytical perf model

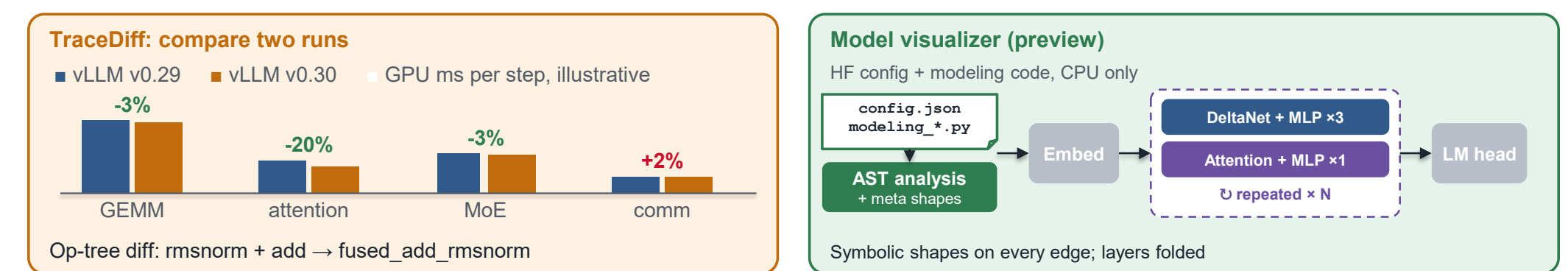
Attention roofline needs per-request query and KV lengths, which kernel traces do not record.



8 TraceLens Agent



9 More features



10 Get started



11 TraceLens powers Hyperloom

ROCm Hyperloom autonomously optimizes inference on AMD Instinct GPUs; TraceLens is its analysis engine.

