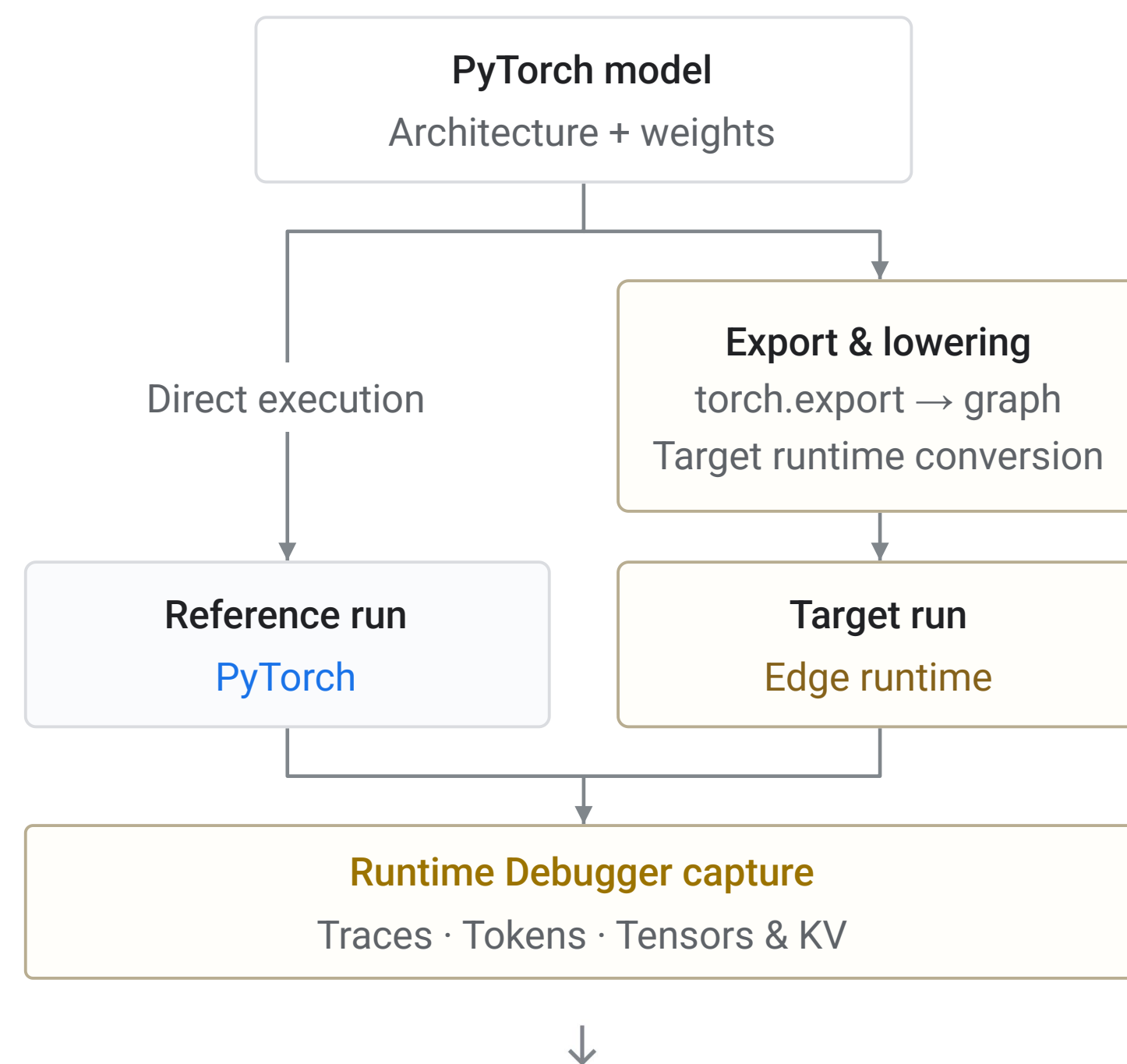


Debug Workflow · Weeks → Hours

A workflow for debugging edge models by tracing inference discrepancies to their root causes.

01 Runtime Export · Run · Capture



02 Debugger Align · Compare · Inspect

- Trace Diff**

Reference

Target
- Token Diff**

Reference

Target
- Analysis Tools**

Line Chart

Heatmap

Graph Diff

Logit-Lens

03 Manual / Agentic Analysis Explain · Validate

Analyze evidence with agents

Form a hypothesis about the observed failure.

Verify with a controlled rerun

Debugger Overview · Compare → Inspect

Gemma 4 E2B

Reference

Linux Desktop · PyTorch · CPU FP32

Prefill

... Logged mood of 2 / 1 0 ... R aining ...

Decode

... log mood ... Great time playing pickle ...

Turn 3 → Turn 4

Prefill

... Logged mood ... Great time playing ...

Decode

... get history ... show dashboard ": true ...

Target · Debugging

Samsung S24 · Edge Runtime · GPU FP16

Prefill ✓

... Logged mood of 2 / 1 0 ... R aining ...

Decode ⚠

... log mood ... great time playing pickle ...

Turn 3 → Turn 4

Prefill ⚠

... Logged mood ... great time playing ...

Decode ⚠

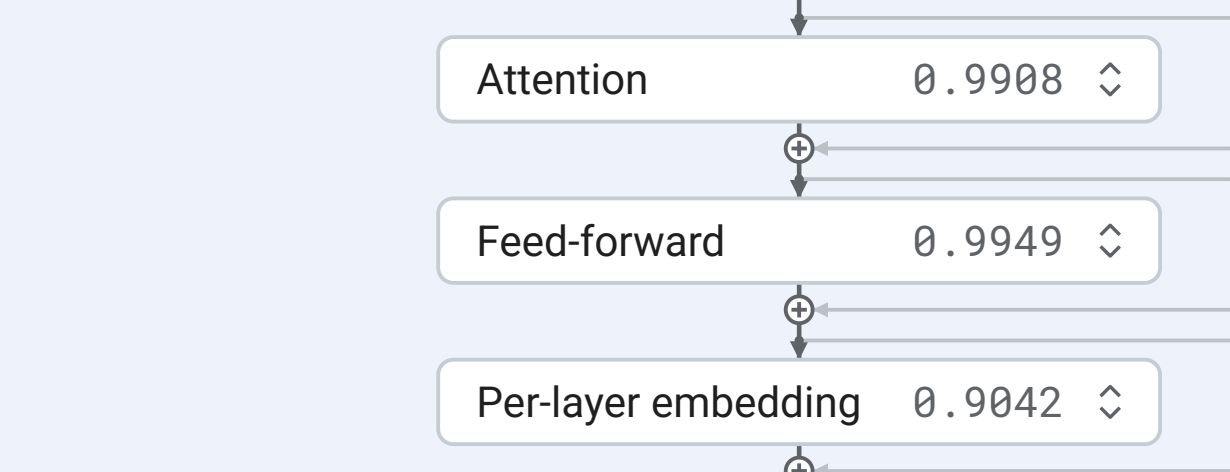
... get history ... show dashboard ": true ...

Analysis Tools

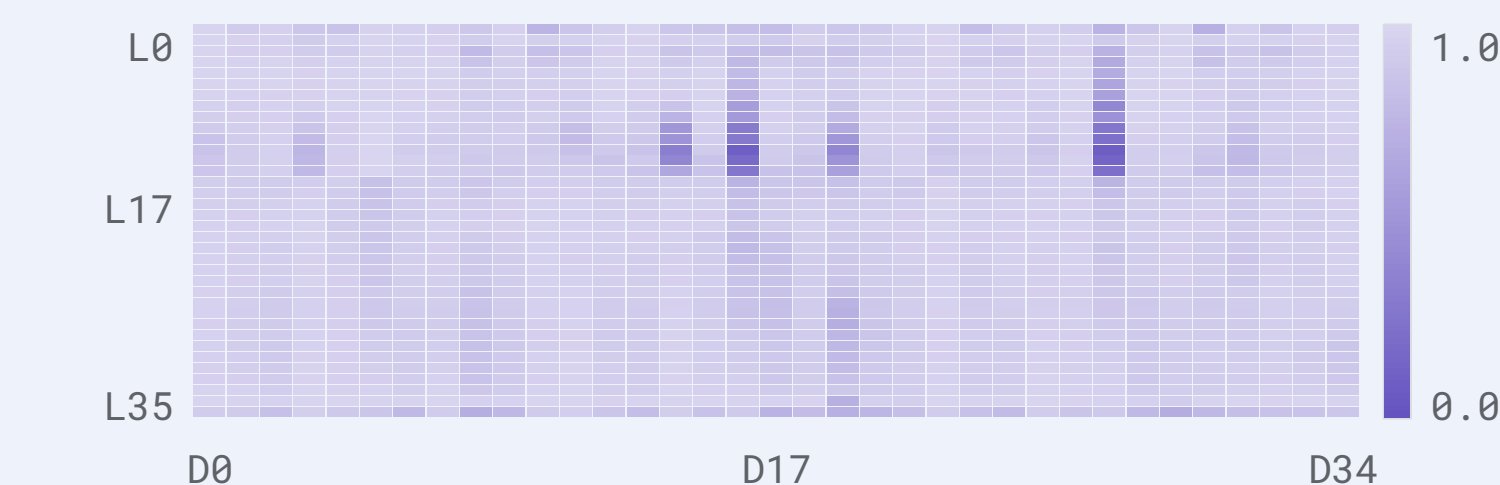
Line Chart KV Cache Diff Cosine similarity



Graph Diff Intermediate Tensor Diff Cosine similarity



Heatmap Intermediate Tensor Diff Cosine similarity



Logit-Lens Layer-wise Token Predictions Vocabulary Rank



Turn 4 → Turn 5

Prefill

... R aining heavily ... Great time playing ...

Decode

... . I have also shown you ... mood dashboard . <turn> ...

Turn 4 → Turn 5

Prefill ⚠

... R aining heavily ... great time playing ...

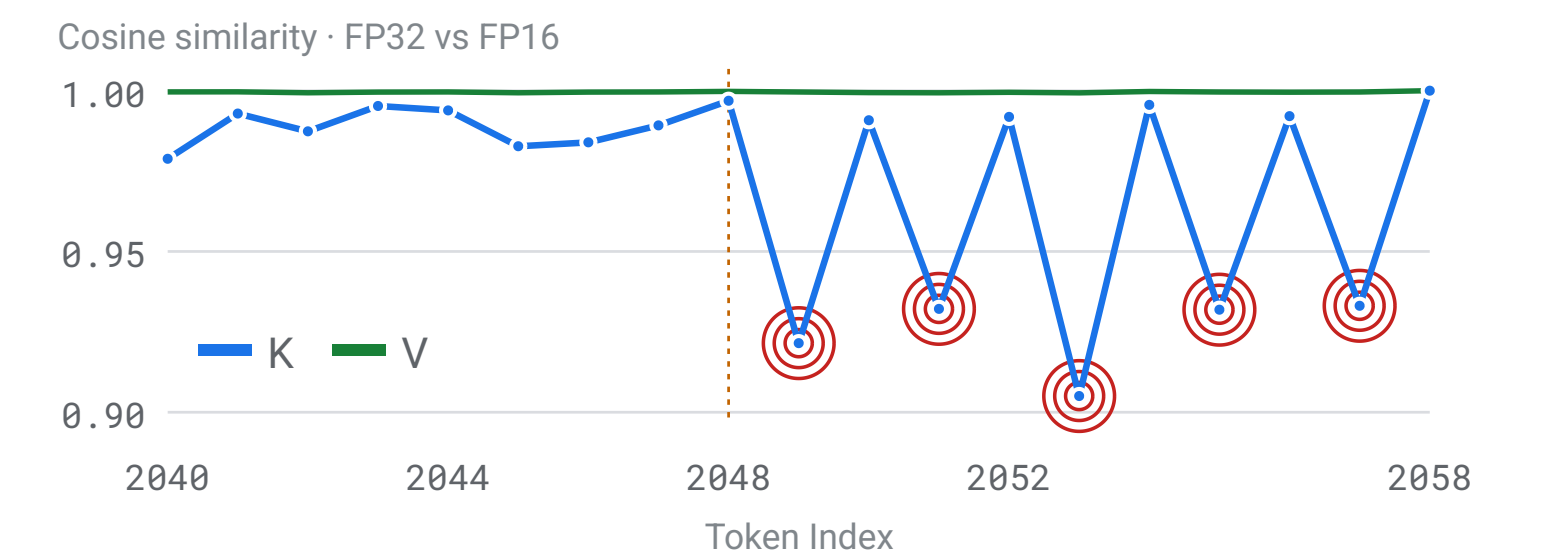
Decode ✖

... Here is your mood dashboard : ... 2 0 2 6 - 7 - 7 : 2 / 1 0 (R ... 0 2 6 - 7 - 8 : 9 / 1 0 (great time ...

Real-World Debugging · Root Causes

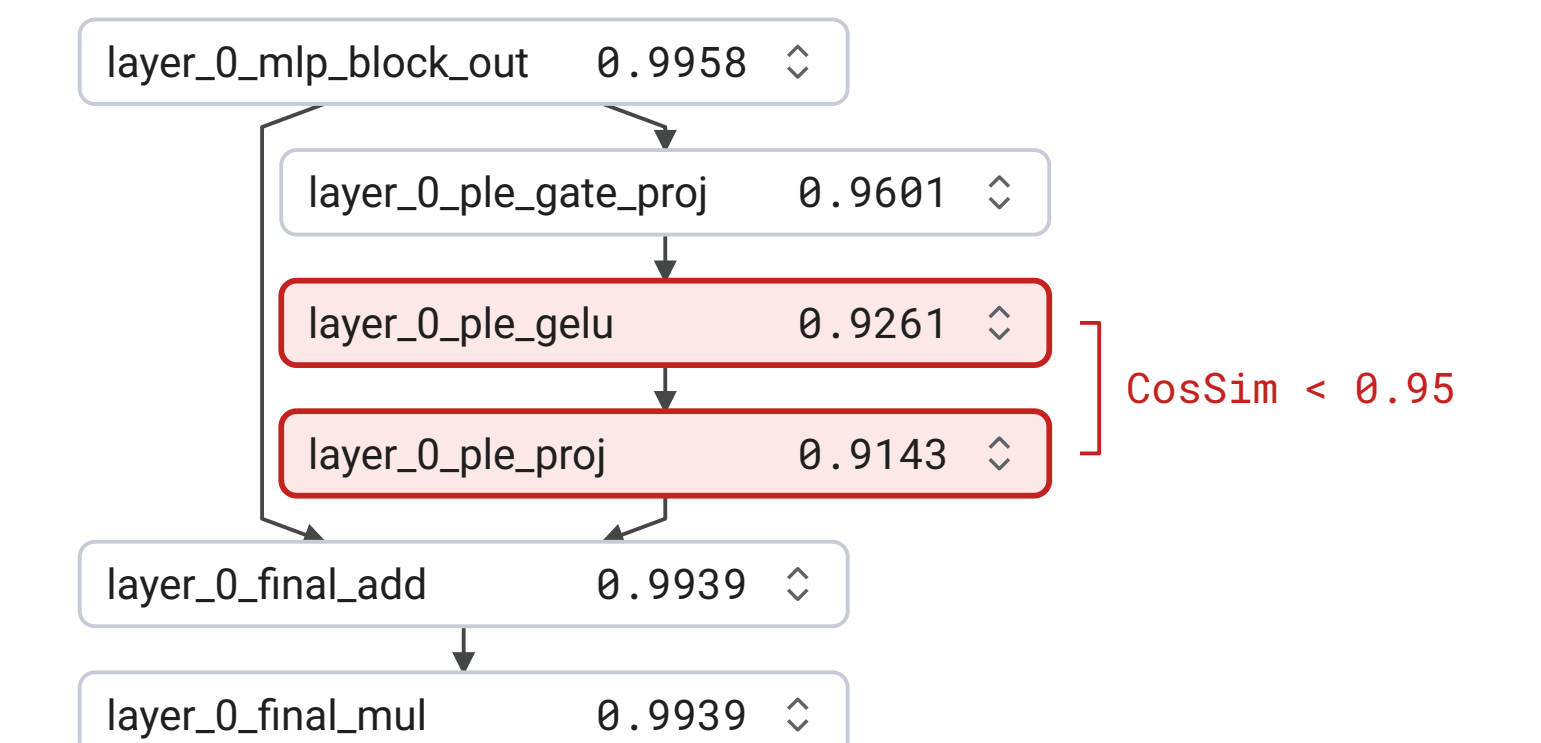
Gemma 4 E2B · Edge Runtime · GPU

Bug 1 — FP16 generates “026” instead of “2026”



Root Cause · FP16 rounds RoPE positions above 2048.

Bug 2 — FP16 output differs on Samsung S24 vs S25 GPUs

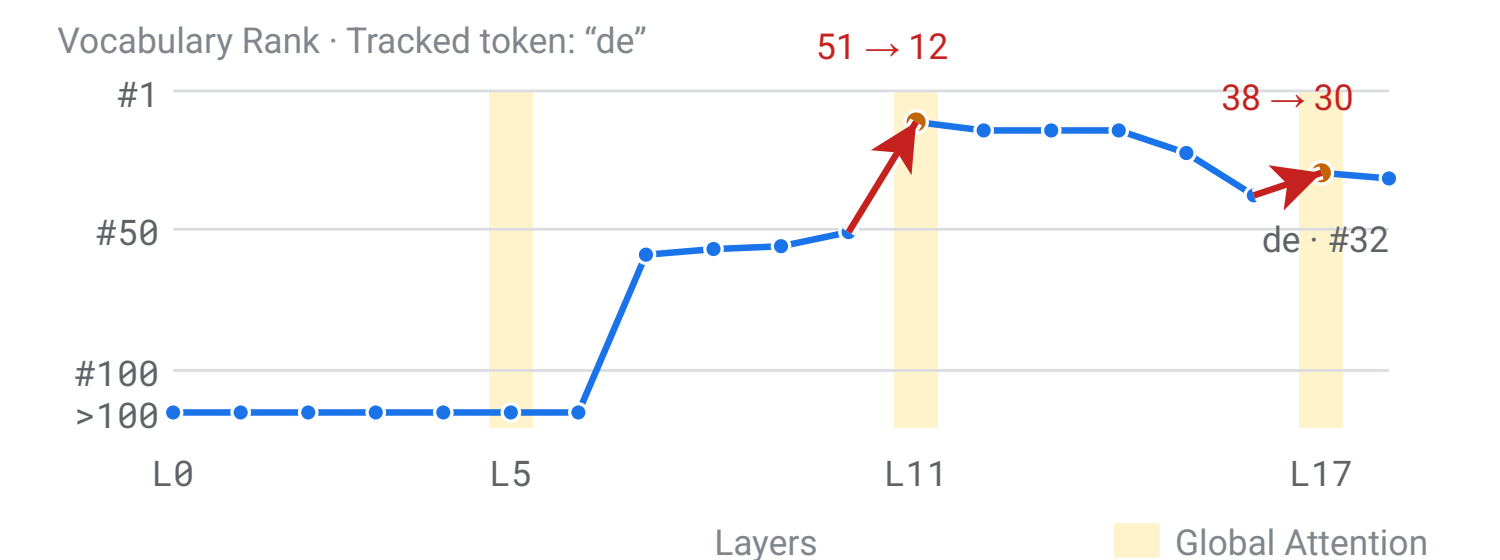


Root Cause · S24 GPUs flush FP16 subnormal PLE weights to zero. S25 GPUs do not.

Gemma 3 270M · All runtimes

Bug 3 — Proofreading repeats text over 512 context tokens

Mid-generation, a full passage repeats until tokens run out.



Root Cause · Relevant context falls outside the 512-token local attention window.

