PR #21804 完整报告

sgl-project/sglang

[Misc] Fix comparator e2e tests: add polars dep + fix dp-attention test

合并时间：2026-04-02 06:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21804

执行摘要

该 PR 修复了比较器端到端测试中的两个关键缺陷：添加缺失的 polars 依赖解决了测试提前崩溃问题，并修正了 dp-attention 测试中的维度标注和步骤限制逻辑。这些修复确保了测试能够实际运行并正确验证功能，提升了测试基础设施的可靠性，对系统运行时无直接影响。

功能与动机

修复比较器端到端测试中的长期缺陷：由于比较器模块 (sglang.srt.debug_utils.comparator) 在顶层导入 polars，但该依赖从未添加到任何依赖组中，导致测试在运行比较前就因 ModuleNotFoundError 崩溃。同时，dp-attention 测试中存在逻辑错误：mlp_output 的维度被错误标注为已全部归约，而实际上在转储点仍为部分状态；且解码步骤的令牌分布在多个 DP rank 上，违反了单 rank 假设。这些问题在之前的 dump 比较器 PR 系列（#19274, #19681）中引入但一直未被发现，因为依赖缺失导致测试从未真正执行比较。

实现拆解

实现分为两个关键文件修改：

1. 依赖修复（python/pyproject.toml）： test=["pandas", "parameterized", "peft>=0.18.0", +"polars", "pytest", "pytest-cov", "diff-cover",] 添加 polars 到 test 依赖组，确保比较器模块能正常导入。
2. 测试逻辑修正（test/registered/debug_utils/test_engine_dumper_comparator_e2e.py）：
   - 将 mlp_output 的 dims 从 't h # tp:replicated' 改为 't h[moe_tp:partial] # tp:replicated'，反映 MLP 输出在转储点仍为部分状态（归约分散在 postprocess_layer() 中发生）。
   - 在 test_dp_attention 中添加 extra_comparator_args 参数： 限制比较到步骤 0（预填充），避免解码步骤中多 DP rank 导致的比较错误；并允许 mlp_output 比较失败，因为 FusedMoE 调度器组合路径可能包含隐式全部归约。

评论区精华

无 review 评论，但 PR body 中提供了详细的修复说明和测试验证链接：

> 这些 bug 在 dump 比较器 PR 系列（#19274, #19681）中引入但从未被发现，因为缺失的 polars 依赖在比较运行前就使比较器子进程崩溃。

作者通过多个 CI 运行链接（如 https://github.com/sgl-project/sglang/actions/runs/23853524625/job/69540028112）验证了修复后测试通过。

风险与影响

风险：

- 依赖添加仅影响测试环境，无运行时风险。
- 测试逻辑修正使比较更准确，但允许 mlp_output 失败可能掩盖 FusedMoE 调度器的潜在问题。
- 限制比较到预填充步骤可能遗漏解码步骤的回归，需后续测试补充。

影响：

- 直接影响测试基础设施，确保比较器端到端测试能正常运行，有助于发现未来代码变更引入的回归。
- 对用户和系统无直接影响，但提升了测试覆盖的可靠性。

关联脉络

该 PR 与历史 PR #19274 和 #19681 关联，这些 PR 引入了 dump 比较器功能但未正确处理依赖和测试逻辑。从近期 PR 分析看，该仓库频繁进行测试基础设施的维护和修复（如 #21873 添加网络超时、#21830 修复 CI 稳定性），表明团队重视测试可靠性。本次修复延续了这一趋势，解决了长期存在的测试缺陷，为后续功能开发提供了更稳定的验证基础。