PR #26653 完整报告

sgl-project/sglang

test: stabilize Gemma4 26B-A4B MTP GSM8K test with deterministic inference + tuned threshold

合并时间：2026-05-29 11:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26653

执行摘要

- 一句话：使 Gemma4 MTP 测试确定性运行并调整阈值
- 推荐动作：值得精读，可作为如何使用确定性推理和系统数据校准来稳定 CI 测试的范例。PR body 中的统计过程是一种值得团队推广的严谨方法。

功能与动机

test/registered/spec/test_gemma4_mtp_26b_a4b_extra.py 在 main 分支上持续失败。原始阈值 0.42 基于不可达的占位分数（0.45）。在非确定性模式下，topk=3 的分数在 0.33-0.50 间波动（std≈0.06），导致断言无法通过。

实现拆解

1. 服务器参数中启用确定性推理：在 _common_server_args 中添加 --enable-deterministic-inference，使每次运行的 GSM8K 分数完全一致（N=20 中 std=0），消除了由投机验证路径放大的批量组成非确定性。
2. 校准观察分数和阈值：根据确定性运行的实际测量值更新 OBSERVED_GSM8K_SCORES（topk=1 从 0.450 改为 0.445，topk=3 从 0.450 改为 0.440）。GSM8K_SCORE_THRESHOLD 自动重算为 0.41（min(0.445,0.440) - 0.03）。
3. 更新注释：在数据和服务器参数旁增加详细注释，说明校准过程和确定性推理对可重现性的重要性。

关键文件：

- test/registered/spec/test_gemma4_mtp_26b_a4b_extra.py（模块 SpecTest；类别 test；类型 test-coverage；符号 OBSERVED_GSM8K_SCORES, GSM8K_SCORE_THRESHOLD, TestGemma4MTP26BA4B._common_server_args）：本 PR 唯一修改的文件，核心变更包括：服务器参数中添加 --enable-deterministic-inference、校准 GSM8K 分数阈值、更新注释。

关键符号：_common_server_args, _server_args

关键源码片段

test/registered/spec/test_gemma4_mtp_26b_a4b_extra.py

本PR唯一修改的文件，核心变更包括：服务器参数中添加--enable-deterministic-inference、校准 GSM8K 分数阈值、更新注释。

```
# 校准：确定性模式下观察到的分数（200 例，5-shot，greedy，triton，TP=2）
# 使用 --enable-deterministic-inference 后每个 topk 分数可重现（N=20 中 std=0）
# topk=1 -> 0.445, topk=3 -> 0.440
OBSERVED_GSM8K_SCORES = {1: 0.445, 3: 0.440}
GSM8K_SCORE_THRESHOLD = min(OBSERVED_GSM8K_SCORES.values()) - GSM8K_SCORE_
MARGIN # 0.41

@classmethod
def _common_server_args(cls) -> list[str]:
    args = [
        "--attention-backend", "triton",
        "--dtype", "bfloat16",
        # ... 其他参数 ...
        "--skip-server-warmup",
        # 启用批量不变内核使 GSM8K 分数每次运行可重现；
        # 不开启时 topk=3 分数会在 ~0.33-0.50 间波动。
        "--enable-deterministic-inference",
    ]
    if TENSOR_PARALLEL_SIZE > 1:
        args += ["--tp-size", str(TENSOR_PARALLEL_SIZE)]
    return args
```

评论区精华

无 review 讨论。PR body 中大量数据对比（N=20 统计）清晰证实了非确定性是方差原因，以及确定性推理后 std=0 的效果。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响测试文件 test_gemma4_mtp_26b_a4b_extra.py，不修改任何生产代码。确定性模式会小幅度减慢每次运行（~40s vs ~35s），并禁用了重叠调度器和自定义 all-reduce，但这是确保 CI 稳定性的有意权衡。
- 影响：直接影响：Gemma4 26B-A4B MTP GSM8K 测试从不稳定、持续失败变为稳定、可重现。间接影响：为同类测试建立了使用确定性推理和基于测量校准阈值的模式。
- 风险标记：仅变更测试文件（低风险），确定性推理会略微降低速度

关联脉络

- 暂无明显关联 PR