

PR #36570 完整报告

sgl-project/sglang

[CI] Lower the AWQ Marlin MMLU threshold to 0.80

合并时间: 2026-08-27 05:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36570>

执行摘要

- 一句话: 降低 AWQ Marlin MMLU 测试阈值至 0.80
- 推荐动作: 该 PR 价值较低, 但值得关注其背后的测试稳定性问题。建议在后续工作中考虑对评测结果进行多次采样取均值, 或使用置信区间来判断是否真正回归, 而非单纯依赖单次阈值。

功能与动机

PR 描述指出: 该评测在多次运行中分数波动范围为 0.816-0.840, 而阈值设为 0.83, 导致只要少数 MMLU 题目翻转测试就会失败。因此需要降低阈值来减少 CI 的不稳定性。

实现拆解

变更仅涉及测试文件 `test/registered/quant/test_awq.py` 中的一行断言: 将 `test_mmlu` 方法中的阈值从 0.83 调整为 0.80。具体实现如下:

1. 定位到 `TestAWQMarlinBfloat16` 类的 `test_mmlu` 方法, 该方法通过 `run_eval` 运行 MMLU 评测并取得 `metrics["score"]`。
2. 将 `self.assertGreater(metrics["score"], 0.83)` 改为 `self.assertGreater(metrics["score"], 0.80)`。
3. 该调整仅针对测试验证逻辑, 不涉及任何生产代码、配置或依赖变化。

关键文件:

- `test/registered/quant/test_awq.py` (模块 量化测试; 类别 test; 类型 test-coverage) : 调整了 AWQ Marlin MMLU 测试的通过阈值, 是本次唯一变更文件。

关键符号: `test_mmlu`

关键源码片段

`test/registered/quant/test_awq.py`

调整了 AWQ Marlin MMLU 测试的通过阈值, 是本次唯一变更文件。

```
# test/registered/quant/test_awq.py
# 调整 MMLU 评测通过阈值, 从 0.83 降至 0.80, 以容忍评测结果波动。
class TestAWQMarlinBfloat16(CustomTestCase):
    # ... setUpClass 和 tearDownClass 略
```

```
def test_mmlu(self):
    args = SimpleNamespace(
        base_url=self.base_url,
        model=self.model,
        eval_name="mmlu",
        num_examples=256,
        num_threads=32,
    )

    metrics = run_eval(args)
    # 原阈值为 0.83, 实测波动 0.816-0.840, 降低至 0.80 以减少 CI 不稳定。
    self.assertGreater(metrics["score"], 0.80)
```

评论区精华

该 PR 没有 review 评论或讨论。作者通过 `/rerun-test test_awq.py` 触发了一次 CI 重跑，结果显示通过，但阈值调整本身未引发进一步技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：降低阈值可能掩盖实际的精度回归。当前设定的 0.80 与观测到的波动下限 0.816 之间留有较大余量，但若未来模型或量化实现出现性能下降，测试可能无法及时捕捉。建议配套一个更严格的相对阈值或监控机制，但该变更本身风险较低。
- 影响：影响范围仅限于 AWQ Marlin 量化模型的 MMLU 测试项。对用户、系统无直接影响，主要改善 CI 稳定性，减少因评测随机性导致的失败。对团队而言，测试更稳定，但可能降低对精度回归的敏感性。
- 风险标记：测试阈值降低可能掩盖回归

关联脉络

- PR #36275 fix(moe): guard FP8 delegate activation params: 同为量化相关测试调整，涉及量化路径的稳定性。