PR #23879 完整报告

sgl-project/sglang

test: relax TestMLADeepseekV3.test_gsm8k threshold 0.62 -> 0.60

合并时间：2026-04-28 06:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23879

执行摘要

- 一句话：降低 DeepSeek-V3 测试精度阈值至 0.60
- 推荐动作：建议合入，属于合理的 CI 维护。后续可观察模型精度趋势，必要时引入更鲁棒的评估指标。

功能与动机

CI 中 TestMLADeepseekV3.test_gsm8k 的得分已持续回落至 0.61 附近，频繁触发阈值接近告警。PR 作者指出同一测试文件中的其他三个类已使用 0.60，此次调整使基线保持一致。

实现拆解

1. 修改一行断言：在 test/registered/mla/test_mla_deepseek_v3.py 中，将 test_gsm8k 方法的 self.assertGreater(metrics["score"], 0.62) 改为 self.assertGreater(metrics["score"], 0.60)。
2. 未修改其他代码或配置，仅调整阈值。

关键文件：

- test/registered/mla/test_mla_deepseek_v3.py（模块 测试；类别 test；类型 test-coverage）：唯一修改文件，调整了 TestMLADeepseekV3.test_gsm8k 的断言阈值，降低至 0.60，与同类测试对齐。

关键符号：test_gsm8k

关键源码片段

test/registered/mla/test_mla_deepseek_v3.py

唯一修改文件，调整了 TestMLADeepseekV3.test_gsm8k 的断言阈值，降低至 0.60，与同类测试对齐。

```
# test/registered/mla/test_mla_deepseek_v3.py
# 降低 gsm8k 评估断言阈值，避免 CI 误报

class TestMLADeepseekV3(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = "lmsys/sglang-ci-dsv3-test"
```

```
        cls.base_url = DEFAULT_URL_FOR_TEST
        other_args = ["--trust-remote-code", "--chunked-prefill-size", "256"]
        if is_cuda():
            other_args.extend(["--enable-torch-compile", "--cuda-graph-max-bs", "2"])
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=other_args,
        )

    def test_gsm8k(self):
        args = SimpleNamespace(
            base_url=self.base_url,
            model=self.model,
            eval_name="gsm8k",
            api="completion",
            max_tokens=512,
            num_examples=200,
            num_threads=128,
        )
        metrics = run_eval(args)
        print(metrics)
        # 阈值从 0.62 放宽至 0.60，与同文件其他测试类保持一致
        self.assertGreater(metrics["score"], 0.60)
```

评论区精华

无实质性讨论，审核者 Kangyan-Zhou 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。阈值放宽可能掩盖模型性能退化，但 0.60 仍是合理底线，且与其他同类测试一致。后续可通过监控长期趋势补偿。
- 影响：影响范围限于 CI 稳定性：减少因阈值波动导致的误报，降低人工介入频率。用户、系统无影响。
- 风险标记：阈值降低可能掩盖精度退化

关联脉络

- 暂无明显关联 PR