PR #27101 完整报告

sgl-project/sglang

[Gemma4] Use hard GSM8K accuracy floor for 31B MTP test

合并时间：2026-06-03 13:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27101

执行摘要

- 一句话：修复 Gemma4 31B MTP 测试的 GSM8K 阈值
- 推荐动作：无需精读。该 PR 是 CI 测试的小幅稳定性改进。值得注意的设计决策是基于 40 次实际运行校准阈值，而非使用占位符值——这是测试可靠性工程的最佳实践。

功能与动机

Gemma4 31B MTP 测试的 GSM8K 阈值从占位观测分数推导得出，这些分数“来自 cookbook，未从该测试的实际 MTP 首次 200 样本运行中测量”。两个 top-k 条目都是相同的 0.805，表明是占位符。因此推导出的阈值 0.775 没有实际校准数据支撑。

实现拆解

1. 在 test/registered/spec/test_gemma4_mtp_31b_extra.py 中，删除了 GSM8K_SCORE_MARGIN 常量、OBSERVED_GSM8K_SCORES 字典以及基于 min(OBSERVED_GSM8K_SCORES.values()) - GSM8K_SCORE_MARGIN 的动态阈值计算。
2. 将 GSM8K_SCORE_THRESHOLD 直接替换为硬编码值 0.75，并添加注释说明其为硬精度下限。
3. 保留了 ACCEPT_LENGTH_THRESHOLD = 1.5 和底层的 assertGreaterEqual(mtp_score, GSM8K_SCORE_THRESHOLD) 断言逻辑，均未变动。

关键文件：

- test/registered/spec/test_gemma4_mtp_31b_extra.py（模块 测试配置；类别 test；类型 test-coverage）：唯一变更文件；用基于 40 次 MTP 运行校准的硬编码常量替换了占位符驱动的 GSM8K 阈值计算。

关键符号：未识别

关键源码片段

test/registered/spec/test_gemma4_mtp_31b_extra.py

唯一变更文件；用基于 40 次 MTP 运行校准的硬编码常量替换了占位符驱动的 GSM8K 阈值计算。

```
# 之前：基于占位符推导（0.805 - 0.03 = 0.775）
# GSM8K_SCORE_MARGIN = 0.03
```

```
# OBSERVED_GSM8K_SCORES = {1: 0.805, 3: 0.805}
# GSM8K_SCORE_THRESHOLD = min(OBSERVED_GSM8K_SCORES.values()) - GSM8K_SCORE_
MARGIN

# 之后：基于 40 次运行校准的硬精度下限（最低 0.765，均值 ~0.780）
# 0.75 在方差余量与回归捕获之间取得平衡
GSM8K_SCORE_THRESHOLD = 0.75
```

评论区精华

PR 中没有 review 讨论。唯一一条评论来自 *gemini-code-assist[bot]*，提示已达到每日配额限制，与变更内容无关。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。阈值从 0.775 降低到 0.75，降低了假阳性概率（即测试因随机方差失败的可能性）。校准数据（40 次运行，最低 0.765）表明 0.75 提供了约 0.015 的余量，足以覆盖正常波动。主要风险是如果实际性能显著下降（例如低于 0.75），测试可能仍能通过，但考虑到阈值仅降低了 0.025，且校准数据扎实，此风险可控。
- 影响：影响仅限于 Gemma4 31B MTP 测试（*test/registered/spec/test_gemma4_mtp_31b_extra.py*）。降低的阈值将使 CI 测试更稳定，减少因随机方差导致的误报失败，同时仍能捕获真实回归。对用户或系统性能无其他影响。
- 风险标记：测试稳定性改进，低风险，仅限 CI

关联脉络

- 暂无明显关联 PR