

PR #27688 完整报告

sgl-project/sglang

Fix flaky hicache l3 mmlu nightly test

合并时间: 2026-06-10 10:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27688>

执行摘要

- 一句话: 移除 GLM-5 L3 nightly 中不稳定的 MMLU two-pass 测试
- 推荐动作: 此 PR 为单纯的测试维护变更, 不建议深入精读。但值得注意的决策经验: 对于小样本评估, 精度差异阈值应结合样本量合理设定, 或改用 KL 散度等更稳定的指标。

功能与动机

PR body 明确指出 MMLU two-pass 测试在 100 样本评估下 max accuracy-diff 0.02 的阈值过于严格, 2-3 道题翻转即可超出, 而 GSM8K two-pass 保持稳定。该测试已成为 nightly CI 的 flaky 来源, 需要移除以恢复 CI 稳定性。

实现拆解

1. 移除 MMLU 相关配置: 在 AccuracyTwoPassMixin (test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py) 中删除 mmlu_threshold, num_mmlu_examples, mmlu_num_threads 三个类属性。
2. 移除 `_run_mmlu` 方法: 删除原来用于运行 MMLU 评估的辅助方法, 该方法通过 `sglang.test.run_eval.run_eval` 执行。
3. 移除 `test_mmlu_two_passes` 测试用例: 删除该测试方法, 该方法依赖 `_two_pass` 机制进行精度一致性校验。

关键文件:

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_nightly.py (模块 HiCache 测试; 类别 test; 类型 test-coverage; 符号 `_run_mmlu`, `test_mmlu_two_passes`, `mmlu_threshold`, `num_mmlu_examples`): 唯一变更文件, 删除了 MMLU two-pass 测试相关的属性、方法和测试用例

关键符号: `_run_mmlu`, `test_mmlu_two_passes`

评论区精华

无 review 评论, 仅 PR 作者自己合并。但 PR body 中解释了决策依据: 0.02 精度差异在 100 样本 MMLU 上过于敏感 (2-3 题翻转即超出), 而 GSM8K 使用 200 样本则更稳定。KL 覆盖率测试将在后续 PR 中补充。

- 暂无高价值评论线程

风险与影响

- 风险：无技术风险，仅移除一个不稳定的测试。但需要注意：KL 覆盖率测试尚未到位，该 PR 只移除了 flaky 测试，未替换为更稳定的机制。后续修复（如增大样本量或放宽阈值）未在本 PR 中实现。
- 影响：对用户无影响。对系统：GLM-5 L3 nightly CI 将不再因 MMLU two-pass 测试误报失败，提高 CI 可靠性。对团队：暂时失去对 MMLU 精度一致性的自动化监控，需后续通过 KL 覆盖率或其他测试补充。
- 风险标记：测试覆盖下降（MMLU 精度一致性测试被移除），后续需补充 KL 覆盖测试

关联脉络

- 暂无明显关联 PR