PR #21745 完整报告

sgl-project/sglang

Fix disaggregation hybrid attention ci

合并时间：2026-03-31 16:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21745

执行摘要

- 一句话：降低 disaggregation hybrid attention CI 测试的 accuracy 阈值以解决 flaky 问题。
- 推荐动作：该 PR 简单，值得快速浏览以了解 CI 调整；关注点在于 TODO 注释和 issue #21744 的后续处理。

功能与动机

根据 PR body，动机是 'Adjust the threshold for flaky CI test.'，并记录 issue #21744 用于跟踪精度修复，以解决 CI 失败问题。

实现拆解

仅修改了 test/registered/distributed/test_disaggregation_hybrid_attention.py 文件：在两个 test_gsm8k 方法中，将 assertGreater 阈值从 0.93 降低到 0.90，并添加了包含 issue 链接的 TODO 注释。

关键文件：

- test/registered/distributed/test_disaggregation_hybrid_attention.py（模块 test）：修改了 disaggregation hybrid attention 的 GSM8K 测试阈值，直接影响 CI 通过性

关键符号：TestDisaggregationHybridAttentionGSM8K.test_gsm8k, TestDisaggregationHybridAttentionMambaDPDecode.test_gsm8k

评论区精华

review 中，gemini-code-assist[bot] 建议在 TODO 注释中添加 issue 链接以提高可维护性，建议被采纳，无其他争议。

- TODO 注释添加 issue 链接 (documentation): 建议被采纳，作者添加了 issue 链接

风险与影响

- 风险：风险在于降低了测试标准，可能掩盖 disaggregation hybrid attention 的精度问题；TODO 注释指向的 issue #21744 需及时解决，否则可能导致回归未被检测到。
- 影响：对用户无直接影响，纯 CI 内部调整；系统 CI 更稳定，但测试覆盖短期减弱；团队需跟踪 issue #21744 以确保精度问题修复。

- 风险标记：降低测试阈值，TODO 待修复

关联脉络

- PR #21733 [CI]Remove msgm-en and mmlu tests which cause timeout: 同样涉及 CI 测试调整，解决超时问题
- PR #21714 Fix human-eval CI install on 5090 runners: 修复 CI 安装问题，与本 PR 同为 CI 稳定性改进