PR #22210 完整报告

sgl-project/sglang

[CI] Relax transformers MMLU threshold from 0.65 to 0.64

合并时间：2026-04-07 06:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22210

执行摘要

本 PR 将 Transformers 模型测试中的 MMLU 评估阈值从 0.65 降低到 0.64，以解决 CI 环境中测试不稳定性导致的误报失败。这是一个纯粹的测试配置调整，不涉及任何功能代码变更，旨在提高 CI 的可靠性。

功能与动机

PR body 和 commit 消息明确指出，test_transformers_models.py 中的 MMLU 评估在 CI 环境中存在不稳定性（flaky），在 0.65 阈值边界上会失败，例如观察到得分 0.640625。调整阈值是为了减少 CI 的误报失败，确保测试的可靠性。

实现拆解

仅修改了 test/registered/models/test_transformers_models.py 文件中的两个类属性：

- TestTransformersFallbackEndpoint.mmlu_lower_bound：0.65 → 0.64
- TestTransformersFallbackTorchAO.mmlu_lower_bound：0.65 → 0.64

这两个变更都只涉及数值调整，不改变测试逻辑或功能代码。

评论区精华

review 中只有 gemini-code-assist[bot] 的自动评论，确认了变更内容并表示没有反馈。没有人工 review 讨论，因此没有技术争议或设计权衡的讨论。

风险与影响

风险分析：

- 仅修改测试阈值，不涉及生产代码，无回归风险。
- 降低阈值可能掩盖模型性能的实际下降，但 PR body 指出这是针对 CI 不稳定性的调整，且从 0.65 到 0.64 的变化很小（1.5% 相对变化）。
- 需要确保 0.64 阈值仍能有效捕获模型性能问题，但考虑到这是针对 flaky 测试的调整，风险可控。

影响分析：

- 对用户：无直接影响，这是内部测试配置调整。
- 对系统：无功能影响，仅影响测试通过标准。

- 对团队：减少 CI 失败噪音，提高开发效率，但需要监控后续测试结果以确保调整合理。

关联脉络

从近期历史 PR 分析可见，本 PR 与以下 PR 有相似之处：

- PR #22194 "[Qwen3-Specv2]: Fix flaky ci"：同样通过调整测试阈值（KL 散度阈值）来修复 CI 不稳定性。
- PR #22190 "Update test coverage report" 和 PR #22189 "Update test skills and guide"：同属测试相关 PR，反映了团队对测试质量和稳定性的持续关注。

这表明团队在积极维护 CI 稳定性，通过阈值调整、测试规范更新等方式减少误报失败，提高开发效率。