PR #22408 完整报告

sgl-project/sglang

[CI] Adding Gemma 4 to Nightly CI

合并时间：2026-04-17 10:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22408

执行摘要

- 一句话：在夜间 CI 测试中新增 Gemma 4 系列模型评估项，替换旧版 Gemma 3 测试。
- 推荐动作：该 PR 变更简单直接，适合快速浏览以了解 CI 测试模型的更新情况。值得关注的点是：1) 模型测试套件如何跟进上游模型发布；2) 性能阈值基于实际运行数据调整的实践。但无需深入分析源码逻辑。

功能与动机

根据 PR 正文描述，此变更是为了“Adding Gemma 4 variants to Nightly CI”，并引用了 PR #21952 作为先导。动机是跟进模型生态发展，确保 CI 对新发布的 Gemma 4 模型进行持续集成测试，以验证 SGLang 框架对新模型的支持和性能表现。

实现拆解

1. 更新测试配置字典：修改文件 test/registered/eval/test_vlms_mmmu_eval.py 中的 MODEL_THRESHOLDS 字典。
2. 替换模型条目：将原有的 google/gemma-3-4b-it 和 google/gemma-3n-E4B-it 条目替换为三个新的 Gemma 4 模型条目：google/gemma-4-E4B-it、google/gemma-4-26B-A4B-it（需 2 路张量并行）、google/gemma-4-31B-it（需 2 路张量并行）。
3. 调整性能阈值：基于实际夜间 CI 运行结果（作者在 Issue 评论中提供了运行链接），为新增的 Gemma 4 模型设置了相应的准确率（第一个数值）和延迟阈值（第二个数值）。例如，gemma-4-E4B-it 的阈值从旧版的 (0.360, 10.9) 调整为 (0.26, 15.0)。
4. 无其他配套改动：本次变更仅涉及测试配置文件，没有修改源码、部署脚本或文档。

关键文件：

- test/registered/eval/test_vlms_mmmu_eval.py（模块 VLM 评估；类别 test；类型 test-coverage；符号 MODEL_THRESHOLDS）：这是唯一变更的文件，包含了多模态 VLM 评估测试的模型配置和阈值，直接决定了夜间 CI 测试哪些模型及其通过标准。

关键符号：未识别

关键源码片段

test/registered/eval/test_vlms_mmmu_eval.py

这是唯一变更的文件，包含了多模态 VLM 评估测试的模型配置和阈值，直接决定了夜间 CI 测试哪些模型及其通过标准。

```
MODEL_THRESHOLDS = {
    # ... 其他模型条目保持不变
    # 新增Gemma 4模型测试项，替换原有的Gemma 3
    ModelLaunchSettings("google/gemma-4-E4B-it"): ModelEvalMetrics(0.26, 15.0),
    # 26B版本需要2路张量并行（--tp=2）
    ModelLaunchSettings(
        "google/gemma-4-26B-A4B-it", extra_args=["--tp=2"]
    ): ModelEvalMetrics(0.27, 22.3),
    # 31B版本同样需要2路张量并行
    ModelLaunchSettings(
        "google/gemma-4-31B-it", extra_args=["--tp=2"]
    ): ModelEvalMetrics(0.28, 25.5),
    # ... 后续模型条目保持不变
}
```

评论区精华

本次 PR 没有 Review 评论，仅有的讨论是作者在关联 Issue 中提供了夜间 CI 运行结果的链接（https://github.com/sgl-project/sglang/actions/runs/24543909587），用于佐证阈值调整的依据。这表明变更基于实际测试数据，但缺乏同行对阈值合理性的评审。

- 暂无高价值评论线程

风险与影响

- 风险：1. 测试覆盖风险：替换旧模型测试可能降低对 Gemma 3 的持续监控，但鉴于 Gemma 4 是新一代模型，此风险可控。 2. 阈值准确性风险：新设置的准确率和延迟阈值（如 gemma-4-E4B-it 的 0.26, 15.0）若设置不当，可能导致 CI 测试误报（通过本应失败的测试）或漏报（失败本应通过的测试）。由于缺乏 Review 讨论，阈值的科学性和长期稳定性未经验证。 3. 依赖兼容性风险：PR 正文提到“Pending https://github.com/sgl-project/sglang/pull/21569 upgrade transformer to 5.5.0”，暗示 Gemma 4 模型可能需要更高版本的 Transformer 库支持。若依赖未升级，测试可能失败。
- 影响：1. 对用户影响：无直接影响，这是内部 CI 测试的更新。 2. 对系统影响：夜间 CI 将开始对 Gemma 4 模型进行自动化评估，有助于提前发现与新模型相关的回归问题。 3. 对团队影响：开发团队需要关注 Gemma 4 测试结果，确保框架兼容性；运维团队需确认 CI 环境满足新模型的资源需求（如内存、GPU）。影响范围限于测试流程，程度较低。
- 风险标记：阈值未经验证，依赖待升级

关联脉络

- PR #21952 [CI] Adding Gemma 4 to Nightly CI: PR 正文明确引用此 PR 为先导（"following https://github.com/sgl-project/sglang/pull/21952"），表明这是同一功能线的延续，可能涉及更早的 Gemma 4 CI 集成工作。

- PR #21569 upgrade transformer to 5.5.0: PR 正文提到 "Pending https://github.com/sgl-project/sglang/pull/21569"，暗示 Gemma 4 测试可能依赖 Transformer 库升级，两者在依赖层面关联。