PR #24754 完整报告

sgl-project/sglang

Reduce gemma4 moe deterministic test runtime

合并时间：2026-05-09 11:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24754

执行摘要

- 一句话：缩短 gemma4 MoE 确定性测试运行时间
- 推荐动作：建议合并。该 PR 在保持测试有效性的前提下显著缩短了 CI 时间，属于高效的运维改进。

功能与动机

减少回归测试在 CI 中的运行时间，从 420 秒缩短到 107 秒，以加速 CI 流水线。

实现拆解

1. 在文件 test/registered/core/test_gemma4_moe_deterministic.py 中，将 register_cuda_ci 的 est_time 参数从 420 改为 107。
2. 将 NUM_REQUESTS 从 200 减少为 180。
3. 删除了 docstring 中描述原始 bug 复现行为和修复效果的 3 行内容。

关键文件：

- test/registered/core/test_gemma4_moe_deterministic.py（模块 测试；类别 test；类型 test-coverage）：唯一变更文件：调整了期望运行时间和请求数量，并精简了文档字符串。

关键符号：未识别

关键源码片段

test/registered/core/test_gemma4_moe_deterministic.py

唯一变更文件：调整了期望运行时间和请求数量，并精简了文档字符串。

```
# gemma4/moe/deterministic: 回归测试配置调整
# 变更前：est_time=420, NUM_REQUESTS=200
# 变更后：est_time=107, NUM_REQUESTS=180

register_cuda_ci(est_time=107, suite="stage-b-test-2-gpu-large")
# ...
NUM_REQUESTS = 180
CONCURRENCY = 128
MAX_TOKENS = 256
```

评论区精华

该 PR 没有 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅调整测试参数和文档，未修改任何核心逻辑。减少请求数可能略微降低测试覆盖率，但 180 个请求仍足以触发并验证原 issue #24394 中的 OOB 问题。
- 影响：影响范围仅限于单个测试文件。该 PR 使 CI 中 Gemma4 MoE 确定性测试的运行时间缩短约 75%，加快 CI 反馈速度。
- 风险标记：测试微调不影响核心逻辑

关联脉络

- PR #24724 [Spec] Disambiguate verified_id into bonus_token(s) / accept_tokens: 同属近期 PR，但与本 PR 无直接关联。