

PR #32091 完整报告

sgl-project/sglang

[CI] Fix failures on main

合并时间: 2026-07-23 00:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32091>

执行摘要

- 一句话: 修复 CI main 分支导入和测试断言问题
- 推荐动作: 建议快速合并以恢复 CI 健康状态。该 PR 可作为示例: 展示如何在引入新 API 后及时更新相关模块的导入, 以及如何因生产代码变更而修正测试断言。

功能与动机

PR #24651 在 `qwen3_5.py` 中引入了 `get_server_args` 的调用却未添加对应导入, 导致 Ruff F821 错误; PR #31312 中对 `batch_size` 的切片返回新的张量对象, 导致使用 `assertIs` 的 n-gram 测试失败。这些失败阻碍了 CI 流水线的正常运行。

实现拆解

1. 修复导入问题: 在 `python/sglang/srt/models/qwen3_5.py` 的 `from sglang.srt.runtime_context import` 列表中新增 `get_server_args`。
2. 修复测试断言: 在 `test/registered/unit/model_executor/model_runner_components/test_ngram_embedding_manager.py` 中, 将三处 `self.assertIs(...)` 替换为 `self.assertTrue(torch.equal(...))`, 以检查张量值而非对象身份。

关键文件:

- `python/sglang/srt/models/qwen3_5.py` (模块 模型层; 类别 source; 类型 core-logic) : 添加缺失的 `get_server_args` 导入, 修复 Ruff F821 lint 错误。
- `test/registered/unit/model_executor/model_runner_components/test_ngram_embedding_manager.py` (模块 测试; 类别 test; 类型 test-coverage) : 修复 n-gram 测试断言, 将对象恒等检查替换为张量值相等检查, 以适应生产代码切片行为。

关键符号: 未识别

关键源码片段

`python/sglang/srt/models/qwen3_5.py`

添加缺失的 `get_server_args` 导入, 修复 Ruff F821 lint 错误。

```
# python/sglang/srt/models/qwen3_5.py (partial)
from sglang.srt.runtime_context import (
    get_exec,
```

```
get_forward,  
get_parallel,  
get_server_args, # 添加此行以修复 Ruff F821: 未定义名称  
get_stream,  
)
```

test/registered/unit/model_executor/model_runner_components/test_ngram_embedding_manager.py

修复 n-gram 测试断言，将对象恒等检查替换为张量值相等检查，以适应生产代码切片行为。

```
# test/registered/unit/model_executor/model_runner_components/test_ngram_embedding_  
manager.py (partial)  
# 修正前：使用对象恒等检查，但由于切片返回新对象而失败  
# self.assertIs(kwargs["row_indices"], req_pool_indices)  
# 修正后：使用值相等检查  
self.assertTrue(torch.equal(kwargs["row_indices"], req_pool_indices))  
self.assertTrue(torch.equal(kwargs["column_starts"], info.out_column_starts))  
self.assertTrue(torch.equal(kwargs["req_lens"], info.out_req_lens))
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：导入修复遵循已有模式，仅添加一行 import；测试断言从对象恒等改为值相等，与生产代码语义一致，不会遗漏回归。
- 影响：直接影响 CI 流水线，恢复 main 分支的 lint 检查和 n-gram 单元测试的通过状态。不影响用户功能或生产行为。
- 风险标记：暂无

关联脉络

- PR #24651 [AMD] Add fused all-reduce RMSNorm per-group quant for Qwen3.5 FP8: 引入了 get_server_args 的调用但未添加导入，导致本 PR 的修复需求。
- PR #31312 Fix LongCat n-gram token-table crashes on padded batches: 修改了 n-gram 更新逻辑中的切片行为，导致现有测试断言失败。
- PR #32083 [CI] Fix n-gram tensor equality test: 与本 PR 部分重复，本 PR 包含了该修复。