

PR #32096 完整报告

sgl-project/sglang

Fix get_server_args import lint error

合并时间: 2026-07-23 00:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32096>

执行摘要

- 一句话: 修复 qwen3.5 未导入 get_server_args
- 推荐动作: 该 PR 是紧急导入修复, 技术价值较低。但建议关注: ①相似问题是否存在在其他模型文件中存在; ②#24651 的 PR 审查是否漏检了导入。可以精读 #24651 的合并情况。

功能与动机

PR body 明确指出 `qwen3_5.py:181` 调用了 `get_server_args()` 但未导入 (由 #24651 引入), 导致 CI 中 `ruff` 报错 `F821 Undefined name 'get_server_args'`, 且 AMD AITER all-reduce fusion 路径会在运行时抛出 `NameError`。

实现拆解

1. 定位缺少导入的文件: `python/sglang/srt/models/qwen3_5.py`。
2. 在现有 `from sglang.srt.runtime_context import (...)` 的多行导入块中增加一行 `get_server_args`。该块原本包含 `get_exec`、`get_forward`、`get_parallel`、`get_stream`。
3. 提交调整后的代码, 确保 `get_server_args` 函数在模块作用域内可被调用。
4. 无其它配套变更, 因为是纯导入修复。

关键文件:

- `python/sglang/srt/models/qwen3_5.py` (模块 模型; 类别 source; 类型 bugfix): 修复缺少的 `get_server_args` 导入的唯一文件。

关键符号: 未识别

关键源码片段

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

修复缺少的 `get_server_args` 导入的唯一文件。

```
# python/sglang/srt/models/qwen3_5.py # 修复前的导入块缺少 get_server_args
from sglang.srt.runtime_context import ( get_exec, get_forward, get_parallel,
# 缺少 get_server_args ← 导致 F821 lint 错误和运行时 NameError get_stream, ) # 修
复后显式导入 get_server_args from sglang.srt.runtime_context import ( get_exec,
get_forward, get_parallel, get_server_args, # ← 新增导入 get_stream, ) 上述片
```

段展示了修复前后的对比。关键改动是在 `sglang.srt.runtime_context` 的多行导入中增加了 `get_server_args`，使得模块末尾调用的 `get_server_args()` 可以正常解析。

评论区精华

该 PR 无 review 评论。仅有 bot (gemini-code-assist) 的过期通知，无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅为添加一行导入语句，不影响已有逻辑；`get_server_args` 函数存在于同一模块 (`sglang.srt.runtime_context`) 且已由 #24651 正确导出。但该 fix 可能遗漏其他文件中的相同问题（检查是否有其他模型文件也缺少该导入）。
- 影响：影响范围小，仅限于 `python/sglang/srt/models/qwen3_5.py`。修复后 CI lint 检查通过，且 AMD AITER 路径 (`_use_aiter` 条件) 不会因 `NameError` 中断。不影响非 AMD 用户或未启用 AITER 的场景。
- 风险标记：低风险，缺少测试覆盖

关联脉络

- PR #24651 [AMD] Add fused all-reduce RMSNorm per-group quant for Qwen3.5 FP8: 引入 `get_server_args` 使用的 PR，但遗漏了导入。