PR #37622 完整报告

vllm-project/vllm

[Bugfix] Fix Step3 pipeline parallel KeyError for residual tensor

合并时间：2026-05-29 18:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37622

执行摘要

- 一句话：修复 Step3 模型流水线并行中 residual 的 KeyError
- 推荐动作：值得合并，修复明确且低风险。建议团队后续为流水线并行的 IntermediateTensors 初始化编写单元测试，防止类似问题在新模型中复现。

功能与动机

用户在使用 Step3 模型（Qwen3.5）进行流水线并行推理时遇到 KeyError: 'residual'，根本原因是 intermediate_tensors 初始化时缺少 "residual" 键，而非首节点的 forward 方法依赖该键。Issue #37543 明确报告了此问题并给出了修复提议。

实现拆解

1. 定位问题：在 vllm/model_executor/models/step3_text.py 的 __init__ 方法中，第 347-349 行调用 make_empty_intermediate_tensors_factory 时只传入了 ["hidden_states"]。
2. 修复变更：将键列表修改为 ["hidden_states", "residual"]，与非首节点 forward 方法中 intermediate_tensors["residual"] 的访问保持一致。
3. 一致性验证：同类模型 step1.py 已包含 ["hidden_states", "residual"]，step3p5.py 因不使用 residual 而无需修改，本次修复使 step3_text.py 与 step1.py 对齐。变更仅涉及一行代码，无测试、配置或部署配套改动。

关键文件：

- vllm/model_executor/models/step3_text.py（模块 模型层；类别 source；类型 data-contract；符号 Step3TextModel.init, make_empty_intermediate_tensors_factory）：修复核心文件：在 make_empty_intermediate_tensors_factory 的键列表中增加 'residual'，使 intermediate_tensors 在非首流水线节点上正确包含 residual 张量。

关键符号：Step3TextModel.init

关键源码片段

vllm/model_executor/models/step3_text.py

修复核心文件：在 make_empty_intermediate_tensors_factory 的键列表中增加 'residual'，使 intermediate_tensors 在非首流水线节点上正确包含 residual 张量。

```
# vllm/model_executor/models/step3_text.py

# 修复前：只初始化了 hidden_states，非首节点 forward 访问 residual 时 KeyError
# self.make_empty_intermediate_tensors = make_empty_intermediate_tensors_factory(
# ["hidden_states"], config.hidden_size
# )

# 修复后：同时初始化 hidden_states 和 residual，与 forward 方法中的解包保持一致
self.make_empty_intermediate_tensors = make_empty_intermediate_tensors_factory(
    ["hidden_states", "residual"], config.hidden_size
)

# forward 方法中非首节点的使用（保持不变）：
# else:
# assert intermediate_tensors is not None
# hidden_states = intermediate_tensors["hidden_states"]
# residual = intermediate_tensors["residual"] # 此处依赖 residual 键
```

评论区精华

审阅者 jeejeelee 要求提供 GSM8K 测试结果，作者 JMonde 指出这只是初始化 bug 修复，模型输出完全一致，但表示愿意运行评估。最终审阅者回复忽略该要求并批准 PR。此外，审阅者提到 CI 失败，作者进行了合并主分支操作以修复。

- GSM8K 测试必要性 (other): 审阅者后续未再要求，直接批准了 PR。
- CI 失败修复 (other): PR 最终通过 CI。

风险与影响

- 风险：风险极低。改动仅在一行调用中增加一个字符串键值，且与 step1.py 的一致模式吻合。但缺少针对流水线并行场景的单元测试，回归测试仅依赖现有 CI，可能遗漏边界情况（如混合张量并行与流水线并行的配置）。
- 影响：正面影响：修复了 Step3 模型（Qwen3.5 等）在流水线并行部署下的推理崩溃，使该功能可正常使用。影响范围仅限于使用流水线并行且模型基于 Step3TextModel 的用户。无性能或兼容性影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #37543 [Bug]: 推理 vllm，出现如下报错，KeyError：residual: 本 PR 直接修复该 Issue 报告的问题。