PR #52007 完整报告

vllm-project/vllm

[CI Bug] Fix ci qwen3.5

合并时间：2026-08-13 01:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52007

执行摘要

- 一句话：下调 Qwen3.5 eval 的 max-num-seqs 修复 CI 回归
- 推荐动作：此 PR 属于低价值配置修复，无需精读；值得留意的点：修 CI 时优先选择机制最简单、影响面最小的参数（降 --max-num-seqs 而非加 batched token 上限），以及将其他配置的修复拆分到后续 PR 以缩小 review 范围。

功能与动机

PR body 明确说明这是 #51908 的 forward fix，目标修复 Buildkite CI #83443 失败。推测 #51908 新增的 Qwen3.5-397B-A17B-NVFP4 gsm8k eval 配置在 CI 环境上因批大小设置过大而失败，因此通过下调 --max-num-seqs 在不动模型逻辑的前提下稳定评测流程。

实现拆解

1. 定位回归源：#51908 引入 Qwen3.5-397B-A17B-NVFP4-DEP2 两个 gsm8k eval 配置后，Buildkite CI #83443 失败；本 PR 作为 forward fix 在测试配置层调整参数，避免改动模型代码。
2. 初步尝试：在 Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml 的 server_args 中追加 --max-num-batched-tokens 8192 限制单次前向的 token 总量（来自 review diff 记录）。
3. review 转向：ZJY0516 建议直接降低 --max-num-seqs 更直接；作者采纳，最终提交将 MTP 配置的 --max-num-seqs 由 384 降为 256，非 MTP 配置由 512 降为 256。
4. 改动范围收敛：两个 YAML 各仅 1 行变化，无源码、无断言、无依赖变更；jeejeelee 提出的其他测试检查由 #52009 承接。
5. 验证与合入：提交 a434d69 触发 Buildkite CI #83575，review 通过后合入 main；配置改动只影响 gsm8k eval 的并发上限。

关键文件：

- tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml（模块 评测配置；类别 test；类型 test-coverage）：MTP 版本的 gsm8k eval 配置，是 CI 回归修复的主要对象；--max-num-seqs 由 384 降为 256，并曾在 review 中尝试追加 --max-num-batched-tokens 8192 后被否决。
- tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2.yaml（模块 评测配置；类别 test；类型 test-coverage）：非 MTP 版本的 gsm8k eval 配置，--max-num-seqs 由 512 降为 256，与 MTP 配置保持一致的收敛策略。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2-MTP.yaml

MTP 版本的 gsm8k eval 配置，是 CI 回归修复的主要对象；--max-num-seqs 由 384 降为 256，并曾在 review 中尝试追加 --max-num-batched-tokens 8192 后被否决。

```
# Qwen3.5-397B-A17B-NVFP4-DEP2-MTP 的 gsm8k 评测配置
# 修复 CI 回归：将 --max-num-seqs 从 384 下调到 256，
# 避免 batch 过大时显存压力导致 Buildkite CI 失败
#  （forward fix for PR#51908，详见 PR#52007）
model_name: "nvidia/Qwen3.5-397B-A17B-NVFP4"
accuracy_threshold: 0.88
tolerance: 0.03
num_questions: 1319
num_fewshot: 5
max_tokens: 12000
server_args: >-
  --max-model-len 16384
  --data-parallel-size 2
  --enable-expert-parallel
  --max-num-seqs 256
  --spec-method mtp
  --spec-tokens 3
```

tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2.yaml

非 MTP 版本的 gsm8k eval 配置，--max-num-seqs 由 512 降为 256，与 MTP 配置保持一致的收敛策略。

```
# Qwen3.5-397B-A17B-NVFP4-DEP2（非 MTP）的 gsm8k 评测配置
# CI 回归修复：--max-num-seqs 由 512 下调至 256
model_name: "nvidia/Qwen3.5-397B-A17B-NVFP4"
accuracy_threshold: 0.88
tolerance: 0.03
num_questions: 1319
num_fewshot: 5
server_args: >-
  --max-model-len 4096
  --data-parallel-size 2
  --enable-expert-parallel
  --max-num-seqs 256
```

评论区精华

核心交锋是修复手段的选择：作者初版尝试在 MTP 配置追加 --max-num-batched-tokens 8192，ZJY0516 建议直接降低 --max-num-seqs（"I think decrease max-num-seqs will be better"），作者采纳并落地；jeejeelee 提醒检查其他测试是否同样受影响（"Could you

check other testing as well?"），作者回应其余测试由 #52009 覆盖。mgoin 最终 approve。

- 检查其他 gsm8k eval 测试是否同样受影响 (testing): 作者回复其他测试由 #52009 覆盖，未在本 PR 内扩散改动范围。
- 修复手段：降 max-num-seqs 还是限制 batched tokens (design): 作者采纳该建议，最终提交将两个配置的 --max-num-seqs 分别从 384/512 降到 256。

风险与影响

- 风险：风险极低。改动局限于两个 gsm8k eval 配置的 server_args，不涉及运行时代码。潜在影响：降低 --max-num-seqs 会压缩评测时的活跃序列并发，可能轻微降低 eval 吞吐，但 accuracy_threshold/tolerance 不变，正确性判定不受影响。需要确认 256 是否在所有 CI 硬件环境稳定；其他同类配置的收敛验证依赖 #52009。
- 影响：对用户与运行时无影响。对团队的意义主要是恢复 Qwen3.5 eval 的 CI 门禁，并为大模型 eval 配置提供了通过降低并发序列数规避资源压力的先例；后续 #52009 继续扩大覆盖范围。
- 风险标记：低风险配置调整，依赖 #52009 验证其他测试，eval 并发度下调

关联脉络

- PR #51908 （被引用的引入 PR）：PR body 明确说明本 PR 是 #51908 的 forward fix；#51908 引入 Qwen3.5 gsm8k eval 配置并导致 Buildkite CI #83443 失败。
- PR #52009 （作者提到的后续测试覆盖 PR）：review 中作者回应其他受影响测试由 #52009 覆盖，本 PR 只收敛修复两处配置。