# PR #38632 完整报告

vllm-project/vllm

[CI] fix LM Eval Qwen3.5 Models (B200)

合并时间：2026-03-31 21:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38632

## 执行摘要

本 PR 修复了 Qwen3.5-397B 模型在 GSM8K 评估中的 CI 测试失败，通过将 max_num_seqs 从 1024 降低到 512 以避免 Mamba cache blocks 不足错误，变更仅限于配置文件调整，风险低且影响面小。

## 功能与动机

CI 测试失败源于错误：RuntimeError: 'max_num_seqs (1024) exceeds available Mamba cache blocks (600)...'，根据 PR body 描述，此问题由 PR #38270 引入。动机是快速修复评估测试以恢复 CI 稳定性，确保模型评估能正常运行。

## 实现拆解

仅修改一个 YAML 配置文件：tests/evals/gsm8k/configs/Qwen3.5-397B-A17B-NVFP4-DEP2.yaml。变更点是在 server_args 中添加 --max-num-seqs 512 参数，完整配置示例如下：server_args:>- --max-model-len 4096 --data-parallel-size 2 --enable-expert-parallel --max-num-seqs 512 此调整直接降低并发序列数，匹配 Mamba 缓存的硬件限制。

## 评论区精华

review 过程简单：gemini-code-assist[bot] 评论无反馈，ProExpertProg 直接批准，表明变更被接受且无技术争议或深入讨论。

## 风险与影响

- 风险：极低，仅配置文件参数变更，但需确保新值 512 在长期测试中不引发性能问题或与其他模型配置冲突；目前缺少自动化测试验证此调整。
- 影响：限于特定模型的评估测试，修复 CI 失败以提升测试可靠性，对生产环境或用户功能无直接影响。

## 关联脉络

- 与 PR #38270 直接相关，后者引入了导致 CI 失败的问题，本 PR 作为后续修复。
- 近期类似 PR 如 #38612 也涉及 CI 测试修复，显示团队在维护测试稳定性方面的持续努力。

- 从仓库历史看，此类小范围配置调整常见于 bugfix 标签，反映了对测试基础设施的精细化维护。