

PR #51293 完整报告

vllm-project/vllm

[CI] Re-enable FI autotune in GSM8K config for Qwen3.5-35B-A3B

合并时间: 2026-08-07 10:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51293>

执行摘要

- 一句话: 重新启用 Qwen3.5-35B-A3B CI 的 FlashInfer MoE autotune
- 推荐动作: 不值得精读, 属于低风险配置回归。可留意其关联的 issue #46083 是否被正式关闭, 以及后续 CI 是否稳定; 若想了解 FlashInfer MoE autotune 的机制, 可结合 #46083 和 #45002 一起阅读。

功能与动机

根据 issue #46083, `test_gsm8k_correctness[Qwen3.5-35B-A3B-DEP2]` 在 B200 上因 FlashInfer `trtllm_bf16_moe` autotune 崩溃导致 CI 失败, PR #45002 通过添加 `--no-enable-flashinfer-autotune` 绕开该问题。现在 autotune bug 已修复, PR 作者在 commit `5fba75aefeedcf5b6cc27abf9bc145b6a49873a7` 上验证复现测试通过, 因此移除该参数以重新启用 autotune, 恢复对真实推理路径的 CI 覆盖。

实现拆解

变更入口是 `tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-DEP2.yaml`, 从 `server_args` 中删除一行 `--no-enable-flashinfer-autotune`, 使 FlashInferMoE autotune 回到默认开启状态。

实现步骤如下:

1. 编辑配置文件, 移除禁用 autotune 的参数 (+0/-1 行), 恢复 server 启动参数为原有的 `--max-model-len 4096 --data-parallel-size 2 --enable-expert-parallel`。
2. 执行复现测试: `cd tests && pytest -s -v evals/gsm8k/test_gsm8k_correctness.py --config-list-file=configs/models-qwen35-blackwell.txt -k Qwen3.5-35B-A3B-DEP2`, 确认无失败。
3. 无其他源码、文档或部署配套改动, 属于纯测试配置调整。

关键文件:

- `tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-DEP2.yaml` (模块评测配置; 类别 test; 类型 test-coverage): 唯一变更文件, 通过删除 `--no-enable-flashinfer-autotune` 恢复 FlashInfer MoE autotune 的 CI 覆盖, 直接对应 PR 目的。

关键符号: 未识别

关键源码片段

tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-DEP2.yaml

唯一变更文件，通过删除 `--no-enable-flashinfer-autotune` 恢复 FlashInfer MoE autotune 的 CI 覆盖，直接对应 PR 目的。

```
# Qwen3.5-35B-A3B-DEP2 GSM8K 评测配置
# 变更说明：删除 --no-enable-flashinfer-autotune,
# 恢复 FlashInfer MoE 内核自动调优（默认开启），
# 对应 issue #46083 的 workaround 已被 bug 修复替代。
model_name: "Qwen/Qwen3.5-35B-A3B"
accuracy_threshold: 0.84
tolerance: 0.03
num_questions: 1319
num_fewshot: 5
server_args: >-
  --max-model-len 4096
  --data-parallel-size 2
  --enable-expert-parallel
```

评论区精华

本 PR 的 review 讨论较少：claude[bot] 自动评论指出该 PR 来自 fork，自动 review 被禁用，需维护者手动触发；维护者 ZJY0516 直接批准并回复 'Thanks'。没有关于实现或风险的技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是 FlashInfer MoE autotune 问题复发的可能性：即使作者声称 bug 已修复并跑过复现测试，但该崩溃曾在多个 nightly CI 日期间歇性出现，若修复不彻底会导致 CI 重新变红。另外，重新启用 autotune 会增加 CI 任务中的 autotune 耗时（21 个 profile 的 tuning），但这是恢复默认行为而非新增开销。
- 影响：影响范围仅限 CI 测试覆盖：恢复 LM Eval Qwen3.5 Models (B200) 作业中 Qwen3.5-35B-A3B-DEP2 配置对 FlashInfer autotune 路径的验证。对服务代码、推理行为、用户 API 均无影响；对团队而言，减少了 workaround 配置的遗留债务。
- 风险标记：CI 稳定性依赖外部 bug 修复，测试配置回归

关联脉络

- PR #45002 [Bugfix] fix qwen3.5 ep weight loading: 该 PR 曾添加 `--no-enable-flashinfer-autotune` 作为绕过 autotune 崩溃的 workaround，本次 PR 将其移除，是同一功能的回归恢复。