PR #38155 完整报告

vllm-project/vllm

[ROCm][CI] Add LM Eval Qwen3.5 Models test for MI355

合并时间：2026-03-27 00:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38155

执行摘要

此 PR 在 vllm 仓库的 CI 配置中添加了一个新的测试条目，用于在 AMD MI355 GPU 上运行 Qwen3.5 模型的 GSM8K 评估测试。变更涉及更新 Buildkite YAML 文件和新增模型配置文件，以增强 ROCm 平台的测试覆盖。

功能与动机

PR 动机源于扩展 ROCm 平台测试范围的需求。作者在 PR body 中明确表示："Adds a new CI entry for running Qwen3.5 model evaluation on MI355 GPUs"，旨在验证 Qwen3.5 模型在 AMD 硬件上的正确性。

实现拆解

主要改动点如下：

- 在 .buildkite/test-amd.yaml 中新增一个 Buildkite 步骤： ```yaml
- label: LM Eval Qwen3.5 Models (MI355) timeout_in_minutes: 120 mirror_hardwares: [amdexperimental, amdproduction, amdgfx950nightly, amdmi355] commands: - pytest -s -v evals/gsm8k/test_gsm8k_correctness.py --config-list-file=configs/models-qwen35-mi355.txt ```
- 创建配置文件 tests/evals/gsm8k/configs/models-qwen35-mi355.txt，内容为 Qwen3.5-35B-A3B-DEP2.yaml。

评论区精华

Review 讨论聚焦于两个关键点：

- 标签准确性：gemini-code-assist[bot] 指出初始标签 "LM Eval Qwen3.5 Models (B200-MI355)" 混淆了 NVIDIA B200 和 AMD MI355 硬件，建议修改为 "MI355" 以避免误解。
- 依赖项管理：tjtanaa 建议添加 Qwen 系列模型的父类依赖，而 AndreasKaratzas 回应已添加 qwen、qwen2、qwen3 文件以覆盖所有情况，同时移除不相关的 qwen3_next 文件，优化 CI 资源使用。

风险与影响

风险较低，但需注意：

- 依赖项不精确可能导致 CI 运行不必要文件，增加资源浪费。
- 配置错误可能使测试失败，影响 ROCm 平台测试可靠性。 影响方面，此 PR 对用户无直接感知，但能提升系统在 AMD 硬件上的模型验证能力，团队需维护好依赖项以确保 CI 效率。

关联脉络

从近期历史 PR 看，此 PR 与 #38014（添加 b200 测试）和 #38161（修复 ROCm CI 测试）类似，都属于 CI 测试增强。这反映了团队在扩展多硬件平台测试覆盖上的持续努力，尤其是针对 Qwen 模型和 ROCm 生态。