PR #38664 完整报告

vllm-project/vllm

[CI][ROCm] Add Qwen3.5-35B-A3B-MXFP4 model eval into CI

合并时间：2026-04-03 12:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38664

执行摘要

本 PR 在 vLLM 的 ROCm CI 流水线中新增了 Qwen3.5-35B-A3B-MXFP4 量化模型的 GSM8K 评估配置，通过创建 YAML 配置文件和更新模型列表实现。这是一个低风险的 CI 基础设施变更，旨在验证该模型在 AMD 硬件上的推理准确性，不影响核心功能。变更已通过简单 review 并合并。

功能与动机

根据 PR 标题和作者说明，动机是将已在本地通过 TP2 验证的 Qwen3.5-35B-A3B-MXFP4 模型（来自 Hugging Face 仓库 amd/Qwen3.5-35B-A3B-MXFP4）纳入持续集成测试。PR body 中提到“一旦 mxfp4 模型公开，就将此 PR 标记为就绪”，表明这是一个预先准备的配置，待模型公开后即可启用 CI 验证，确保该量化模型在 vLLM 框架下的兼容性和性能。这反映了 vLLM 团队对扩展量化模型（尤其是 MXFP4 格式）和 ROCm 平台支持的趋势。

实现拆解

实现仅涉及两个配置文件的改动，无代码逻辑变更：

文件路径	变更	关键内容
tests/evals/gsm8k/configs/Qwen3.5-35B-A3B-MXFP4-TP2.yaml	新增	定义模型评估参数： - model_name: "amd/Qwen3.5-35B-A3B-MXFP4" - accuracy_threshold: 0.82 - tolerance: 0.03 - num_questions: 1319 - num_fewshot: 5 - server_args: "--max-model-len 4096 --tensor-parallel-size 2"
tests/evals/gsm8k/configs/models-qwen35-mi355.txt	修改	添加一行 Qwen3.5-35B-A3B-MXFP4-TP2.yaml，将新配置注册到 CI 模型列表中

评论区精华

Review 讨论非常简短，核心交锋仅一点：

> AndreasKaratzas在初始提交的 diff 中评论："Don't forget to change this part too once the model is out. Thank you for this contribution :)"
>
> BowenBao回复："nice catch"

这指向初始提交中模型路径为本地路径 /shareddata/amd/Qwen3.5-35B-A3B-MXFP4，作者在后续提交中将其更新为远程 Hugging Face URL，确保 CI 可访问。tjtanaa 最终批准（LGTM），无其他争议。

风险与影响

- 技术风险：极低。仅添加配置，未修改核心代码，无回归、性能或安全风险。主要风险是 CI 依赖外部模型仓库（Hugging Face），若模型不可访问或变更，可能导致 CI 失败。
- 影响分析：
  - 对用户无直接影响，不改变 vLLM 运行时功能。
  - 对系统扩展了 CI 测试覆盖，新增一个量化模型在 ROCm 平台上的评估，有助于提前发现部署问题。
  - 对团队为 AMD 提供了针对特定量化模型的持续验证能力，加强了 vLLM 对 ROCm 和 MXFP4 量化格式的支持生态。影响程度低，仅涉及 CI 配置。

关联脉络

从近期历史 PR 可见相关脉络：

- PR #38292：同样在 ROCm CI 中添加量化模型（gpt-oss w4a8）评估配置，属于同一类 CI 扩展活动。
- PR #38832：修复 Qwen3.5 模型在 NVFP4 量化下的崩溃问题，与本 PR 关注的 Qwen3.5-35B-A3B-MXFP4 模型同属 Qwen3.5 家族，反映团队对 Qwen3.5 模型量化支持的持续投入。
- PR #33657：为 Qwen3.5 模型在 XPU 上启用 GDN 注意力支持，与本 PR 在 ROCm 上支持 Qwen3.5 量化模型类似，均属扩展 vLLM 对 Qwen3.5 模型在不同平台和配置下的覆盖。

整体上，这些 PR 共同体现了 vLLm 在 v1 版本中积极扩展对多样化模型（尤其是 Qwen 系列）、量化格式（如 MXFP4、NVFP4）和硬件平台（ROCm、XPU）的支持，以提升框架的兼容性和生态系统。