PR #42010 完整报告

vllm-project/vllm

[CI][XPU]Ignore some lora tests from LoRA Intel CI pipeline

合并时间：2026-05-08 17:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42010

执行摘要

- 一句话：移除 Intel CI 中 Qwen 测试并设置 spawn 模式
- 推荐动作：该 PR 属于常规 CI 维护，值得精读程度低。可关注其与 PR #41895 等 spawn 兼容性修复的关联。

功能与动机

根据 PR body 和提交信息，目的是忽略 LoRA Intel CI pipeline 中某些测试（尤其是 Qwen 相关）以减少 CI 耗时，并设置 VLLM_WORKER_MULTIPROC_METHOD=spawn 以解决 XPU 环境下的多进程兼容性问题。背景可能源于之前 PR #41895 等中对 XPU/ROCm spawn 兼容性的修复。

实现拆解

1. 移除特定测试项：在 LoRA Models 步骤中，删除了 test_qwen35_densemodel_lora.py、test_llama_tp.py、test_olmoe_tp.py 的调用，仅保留 test_transformers_model.py、test_chatglm3_tp.py、test_minicpmv_tp.py。
2. 添加 spawn 环境变量：在所有 5 个命令块（Runtime + Utils、Fused/MoE Kernels、Punica Kernels、Punica FP8/XPU Ops、LoRA Models）中，在 pytest 命令前均增加了 export VLLM_WORKER_MULTIPROC_METHOD=spawn &&。
3. 移除 Mixtral 测试的宽松退出：在 LoRA Models 步骤中，原本 test_mixtral.py 使用 (pytest ... || true) 允许失败，变更后保留该模式。
4. 精简 Multimodal 步骤：原本完整的 LoRA Multimodal 步骤被截断，具体内容未完全展示，但根据差异推测仅保留了头部。

关键文件：

- .buildkite/intel_jobs/lora_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：仅有的变更文件，包含所有配置修改。

关键符号：未识别

评论区精华

仅有自动化 bot 评论，无实质性讨论。维护者 jikunshang 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅修改 CI 配置文件，不涉及任何源码逻辑变更。主要风险是被移除的测试不再在 Intel CI 中运行，可能导致回归未被及时发现。但移除的测试可能本身不稳定或与 XPU 不兼容，属于有意的缩减。
- 影响：影响范围：仅影响 Intel GPU 上的 LoRA CI pipeline。影响程度：低。CI 执行时间预计缩短，移除的测试项将不再覆盖，但核心 LoRA 功能测试仍保留。
- 风险标记：测试覆盖缩减，仅 CI 配置变更

关联脉络

- PR #41895 [Bugfix] Fix XPU/ROCm compatibility in spawn_new_process_for_each_test: 同样涉及 XPU/ROCm 的 spawn 兼容性修复，本 PR 通过设置 VLLM_WORKER_MULTIPROC_METHOD=spawn 来类似地解决多进程问题。