PR #46365 完整报告

vllm-project/vllm

[Bugfix][CPU] Fix CPU model runner v2

合并时间：2026-06-22 22:52

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46365

执行摘要

- 一句话：修复 CPU model runner v2 的 pin_memory 缺失问题
- 推荐动作：该 PR 简单明确，无需精读。但值得关注 CPU 后端的补丁模式，了解如何通过 monkey-patch 适配无 CUDA 环境。

功能与动机

PR body 指出目的是修复 broken CPU model runner v2 test。由于 CPU 上没有 CUDA 支持，torch.Tensor.pin_memory 会失败，因此需要提供一个 no-op 补丁。

实现拆解

在 vllm/v1/worker/cpu/shm.py 中新增 fake_pin_memory 函数，定义为直接返回 self（即 no-op）。随后在 torch 补丁部分添加 torch.Tensor.pin_memory = fake_pin_memory。该文件是 CPU model runner v2 的 torch API 补丁模块，用 no-op 替代 CUDA 相关操作使代码在 CPU 上正常运行。

关键文件：

- vllm/v1/worker/cpu/shm.py（模块 模型运行器；类别 source；类型 core-logic；符号 fake_pin_memory）：CPU model runner v2 的 torch API 补丁文件，新增 fake_pin_memory 并挂载到 torch.Tensor.pin_memory，修复因缺少该函数导致的测试失败。

关键符号：fake_pin_memory

评论区精华

该 PR 没有 review 评论和 issue 讨论，仅有 jikunshang 的批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅 5 行新增，且为简单的 no-op 函数，不会影响 GPU 或其他后端。但若将来 CPU 后端需要实际 pin_memory 功能，该补丁需被替换。
- 影响：影响范围小，仅修复 CPU model runner v2 的测试失败，使 CPU 后端在调用 pin_memory 时不会崩溃。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR