

PR #54788 完整报告

vllm-project/vllm

[Bugfix][Spec Decode] Honour the draft's moe_backend on Model Runner V2

合并时间: 2026-09-09 07:50

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54788>

执行摘要

- 一句话: 修复 MRV2 草稿模型忽略用户指定 MoE 后端导致启动失败的问题。
- 推荐动作:

功能与动机

当使用 MRV2 (v0.28.1rc1+) 时, 若目标模型 (如 Qwen3.6-35B-A3B-NVFP4) 被量化而草稿模型 (MTP 头) 未量化, 用户通过 `--speculative-config '{`

实现拆解

1. 核心源码路径 - `vllm/v1/worker/gpu/spec_decode/eagle/utils.py` (core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +10/-0
2. 测试与验证 - `tests/v1/spec_decode/test_draft_moe_backend_override.py` (test-coverage) : 测试配套; 涉及符号 `_KernelConfig`, `_CacheConfig`, `_SpeculativeConfig`, `_VllmConfig`; 包含 测试覆盖调整、导入关系调整、控制流调整; +87/-0; 关键符号: `_KernelConfig`, `_CacheConfig`, `_SpeculativeConfig`, `_VllmConfig`

关键文件:

- `tests/v1/spec_decode/test_draft_moe_backend_override.py` (模块 `draft/moe/backend/override`; 类别 `test`; 类型 `test-coverage`; 符号 `_KernelConfig`, `_CacheConfig`, `_SpeculativeConfig`, `_VllmConfig`) : 测试配套; 涉及符号 `_KernelConfig`, `_CacheConfig`, `_SpeculativeConfig`, `_VllmConfig`; 包含 测试覆盖调整、导入关系调整、控制流调整; +87/-0
- `vllm/v1/worker/gpu/spec_decode/eagle/utils.py` (模块 `vllm/v1`; 类别 `source`; 类型 `core-logic`) : 源码主路径; 包含 控制流调整、配置键调整; +10/-0

关键符号: `_KernelConfig`, `_CacheConfig`, `_SpeculativeConfig`, `_VllmConfig`, `_config`, `_Captured`, `init`, `_capture_draft_config`

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR