

PR #53379 完整报告

vllm-project/vllm

[Bugfix] Fix Kimi K3 loading with interleaved weight streams

合并时间: 2026-09-08 22:08

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53379>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `vllm/models/kimi_k3/nvidia/model.py` (data-contract) : 源码主路径; 涉及符号 `process_weights_after_loading`; 包含 配置键调整、符号定义调整; +8/-2; 关键符号: `process_weights_after_loading`
2. 测试与验证 - `tests/models/kimi_k3/test_weight_loading.py` (test-coverage) : 测试配套; 涉及符号 `_FakeKimiLinearModel`, `init`, `load_weights`, `finalize_mega_moe_weights`; 包含 测试覆盖调整、导入关系调整、配置键调整; +72/-0; 关键符号: `_FakeKimiLinearModel`, `init`, `load_weights`, `finalize_mega_moe_weights`

关键文件:

- `tests/models/kimi_k3/test_weight_loading.py` (模块 `weight/loading`; 类别 `test`; 类型 `test-coverage`; 符号 `_FakeKimiLinearModel`, `init`, `load_weights`, `finalize_mega_moe_weights`) : 测试配套; 涉及符号 `_FakeKimiLinearModel`, `init`, `load_weights`, `finalize_mega_moe_weights`; 包含 测试覆盖调整、导入关系调整、配置键调整; +72/-0
- `vllm/models/kimi_k3/nvidia/model.py` (模块 `vllm/models`; 类别 `source`; 类型 `data-contract`; 符号 `process_weights_after_loading`) : 源码主路径; 涉及符号 `process_weights_after_loading`; 包含 配置键调整、符号定义调整; +8/-2

关键符号: `_FakeKimiLinearModel`, `init`, `load_weights`, `finalize_mega_moe_weights`, `test_interleaved_composite_weights_finalize_kimi_once_after_loading`, `process_weights_after_loading`

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