

PR #48008 完整报告

vllm-project/vllm

[docs] Fix the docs build

合并时间: 2026-07-09 06:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48008>

执行摘要

- 一句话: 修复文档构建的 Docstring 格式问题
- 推荐动作: 该 PR 值得合并, 修复了文档构建的警告, 属于基础的维护性变更。

功能与动机

作者发现 `mkdocs build --strict` 产生 5 个警告, 导致最新本文档未发布到 <https://docs.vllm.ai/en/latest/>。PR body 列出了这些警告, 分别来自 `vllm/model_executor/layers/fused_moe/layer.py` 和 `vllm/v1/kv_offload/tiering/p2p/manager.py`。

实现拆解

1. 修复 `FusedMoE` Docstring 格式: 在 `vllm/model_executor/layers/fused_moe/layer.py` 中, 将 `reduce_results` 参数描述的换行符后移, 解决了 `griffe` 解析警告。
2. 替换 `expert_mapping` 为 `ckpt_names`: 同一文件中, 将已废弃的 `expert_mapping` 参数替换为新增的 `ckpt_names` 参数描述, 解决了 `griffe` 解析警告。

关键文件:

- `vllm/model_executor/layers/fused_moe/layer.py` (模块 模型执行器; 类别 `source`; 类型 `data-contract`; 符号 `FusedMoE`) : 修复了两个 Docstring 警告: `reduce_results` 参数换行格式和 `expert_mapping` 替换为 `ckpt_names`
- `vllm/v1/kv_offload/tiering/p2p/manager.py` (模块 KV 卸载; 类别 `source`; 类型 `dependency-wiring`; 符号 `P2PSecondaryTierManager`) : 修复了两个警告: 添加 `Any` 导入和 `**kwargs` 类型标注

关键符号: `FusedMoE`, `P2PSecondaryTierManager.init`

关键源码片段

`vllm/v1/kv_offload/tiering/p2p/manager.py`

修复了两个警告: 添加 `Any` 导入和 `**kwargs` 类型标注

```
# 在 manager.py 中, 为 __init__ 的 **kwargs 添加类型标注
from typing import TYPE_CHECKING, Any
```

```
class P2PSecondaryTierManager(SecondaryTierManager):
```

```
def __init__(
    self,
    offloading_spec: OffloadingSpec,
    primary_kv_view: memoryview,
    tier_type: str = "p2p",
    host: str = "0.0.0.0",
    port: int = 7777,
    backends: list[str] | None = None,
    num_threads: int = 4,
    **kwargs: Any, # 添加类型标注, 修复 griffe 警告
) -> None:
```

评论区精华

该 PR 没有 review 评论。reviewer [simon-mo](#) 直接批准了变更。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改了 Docstring 格式和类型标注，不影响运行时行为。
- 影响：影响范围有限。仅修复文档构建问题，确保最新文档能正确生成并发布。对用户和系统功能无任何影响。
- 风险标记：仅文档变更

关联脉络

- 暂无明显关联 PR