PR #43697 完整报告

vllm-project/vllm

[Docs] Fix MLA prefill backend default docs

合并时间：2026-05-27 18:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43697

执行摘要

- 一句话：修正 MLA prefill 后端默认选择文档
- 推荐动作：建议审阅者确认文档预览内容正确后直接合并。这是一次纯粹的文档修正，没有逻辑和行为变更。

功能与动机

文档与实际自动选择行为不一致。PR 描述指出 'MLA prefill backend docs were not aligned with the current auto-selection behavior'，需要更新描述以反映 FlashAttention 优先，以及 Blackwell 上的完整回退顺序。

实现拆解

1. 修改源码注释：在 tools/pre_commit/generate_attention_backend_docs.py 中，将 parse_mla_prefill_backends 函数内注释从 'default Blackwell backend' 改为 'the highest-priority automatic backend'，以更准确地表示标记含义。
2. 更新文档生成逻辑：在同一文件中，修改 generate_mla_section 函数内的文档字符串，将原先的 'TRT-LLM Ragged is the default on Blackwell' 和 'FlashAttention is used as the default' 替换为描述实际自动选择流程的文本：优先尝试 FlashAttention，Blackwell 上回退顺序为 TRT-LLM Ragged、FlashInfer、TokenSpeed MLA，其他 GPU 仅考虑 FlashAttention。
3. 同步更新生成的文档：docs/design/attention_backends.md 根据上述变动自动重新生成，包含与源码一致的最新描述。

关键文件：

- tools/pre_commit/generate_attention_backend_docs.py（模块 文档生成；类别 source；类型 core-logic；符号 parse_mla_prefill_backends, generate_mla_section）：控制自动生成文档的代码，修改了注释和文档内容以反映实际的自动选择逻辑。
- docs/design/attention_backends.md（模块 文档；类别 docs；类型 documentation）：最终生成的文档文件，是用户实际阅读的内容，反映文档修正的最终效果。

关键符号：parse_mla_prefill_backends, generate_mla_section

关键源码片段

tools/pre_commit/generate_attention_backend_docs.py

控制自动生成文档的代码，修改了注释和文档内容以反映实际的自动选择逻辑。

```
# tools/pre_commit/generate_attention_backend_docs.py
# 关键代码段：修改注释和文档字符串

def parse_mla_prefill_backends() -> list[dict[str, Any]]:
    # ...
    # Add marker for the highest-priority automatic backend.
    marker = ""
    if backend_name == priority_order[0] and priorities.get("blackwell"):
        marker = " ‡ "

def generate_mla_section(...) -> str:
    lines.extend([
        "",
        # 更新后的文档：描述实际自动选择逻辑
        "> ** ‡ ** Automatic selection tries FlashAttention first. On Blackwell",
        "> (SM100), the fallback order is TRT-LLM Ragged, FlashInfer, then",
        "> TokenSpeed MLA. On other GPUs, only FlashAttention is considered.",
        "",
        # ...
    ])
```

评论区精华

无 review 讨论，PR 被直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅涉及注释和自动生成文档的文本更新，未变更任何运行时行为或 API。文档预览链接已由 mergify 自动生成，可验证内容正确性。
- 影响：用户影响：阅读 MLA prefill 后端文档的用户将获得准确的默认选择描述，避免误解。系统影响：无。团队影响：维护了文档与代码的一致性和可维护性。
- 风险标记：暂无

关联脉络

- PR #43325 [MLA][Attention] Add OOT MLA prefill backend registration mechanism: 实现新的 MLA prefill 后端注册机制，本 PR 修正其相关文档。