PR #38562 完整报告

vllm-project/vllm

[Bugfix][MLA] Change default SM100 MLA prefill backend back to TRT-LLM

合并时间：2026-03-31 00:51

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38562

执行摘要

本 PR 修复了在 SM100 GPU 上 MLA prefill 默认后端错误导致 Kimi-K2.5 模型输出不可用的问题，通过将配置默认值改回TRT-LLM后端实现临时修复，同时暴露了配置命名的维护性风险。

功能与动机

PR 旨在解决 issue #36763，即在 SM100 硬件上，FA4 MLA prefill 后端对 Kimi-K2.5 模型产生不可用输出。因此，更改默认后端以恢复模型功能，等待 FA4 问题解决。

实现拆解

仅修改了 vllm/config/attention.py 文件中的 AttentionConfig 类。关键变更如下：

```
use_trtllm_ragged_deepseek_prefill: bool = True # 从 False 改为 True
```

这使 TRT-LLM 成为 SM100 上 MLA prefill 的默认后端。

评论区精华

- 命名问题：gemini-code-assist[bot] 评论："The name of this configuration flag... is very specific to 'deepseek'... please consider a more generic name." 这指出了设计缺陷，但未在本次修复中解决。
- 控制机制：mgoin 问："Where do we control FA4 MLA prefill?" 作者回复解释回退逻辑，并链接到 PR #32623 进行未来清理。

风险与影响

- 风险：默认后端切换可能引入新问题；标志命名混淆增加维护成本；依赖 TRT-LLM 后端的稳定性。
- 影响：用户端修复了特定模型问题，系统端可能影响其他 MLA prefill 模型的性能或正确性。

关联脉络

本 PR 是临时修复，关联到未来的 PR #32623，后者计划清理 MLA prefill 接口。这表明团队正在处理注意力后端选择的架构问题，可能涉及更大的重构。