PR #1736 完整报告

THUDM/slime

[docker] Fix IndexCache with mla model

合并时间：2026-03-18 11:32

原文链接：http://prhub.com.cn/THUDM/slime/pull/1736

执行摘要

本次 PR 修复了 Docker 环境中 mla 模型的 IndexCache 问题，通过更新 SGLang 补丁和版本号，确保索引缓存逻辑正确运行。变更涉及核心注意力层，对系统稳定性有积极影响。

功能与动机

由于 PR body 和关联 Issue 为空，动机推断为修复 IndexCache 在 mla 模型中的 bug。IndexCache 是 SGLang 中的索引缓存机制，mla 模型（可能指 Multi-Head Attention 变体）在此处存在兼容性问题，需要调整以避免错误。

实现拆解

主要改动集中在两个文件：

- docker/patch/latest/sglang.patch：修改了 SGLang 库的 DeepseekV2AttentionMLA 类。在 __init__ 方法中添加了 skip_topk 和 next_skip_topk 属性，并调整了 indexer 初始化逻辑，以支持 mla 模型的索引缓存。此外，forward 方法增加了 prev_topk_indices 参数。
- docker/version.txt：版本号从 nightly-dev-20260318a 更新到 nightly-dev-20260318b，标记 Docker 环境的补丁更新。

评论区精华

本次 PR 无 review 评论，因此没有讨论内容可供分析。

风险与影响

风险方面，变更直接修改了核心注意力层的索引逻辑，可能引入回归错误，尤其是在复杂的 mla 模型场景中。由于缺少额外的测试覆盖，变更的正确性依赖现有测试或后续验证。影响上，修复了 Docker 环境中 mla 模型的 IndexCache 问题，提升用户使用体验和系统稳定性，同时版本更新确保了补丁同步。

关联脉络

从历史 PR 分析中，PR #1743 同样更新了 sglang.patch 文件，这表明 Docker 环境中 SGLang 依赖的维护是持续进行的。此外，多个历史 PR 涉及 DeepseekV2 模型的 bugfix，如 PR #1734 和 #1737，显示该模型系列的活跃开发和问题修复。