

PR #43745 完整报告

vllm-project/vllm

[misc] Bump cutedsl version to 4.5.2

合并时间: 2026-05-27 23:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43745>

执行摘要

此 PR 将 `requirements/cuda.txt` 中 `nvidia-cutlass-dsl[cu13]` 的版本从 4.5.0 升级到 4.5.2，属于常规依赖补丁更新，风险极低。

功能与动机

升级 `cutedsl` 依赖至最新补丁版本，以获取可能的 bug 修复或性能改进。PR body 未提供具体说明，但从版本号变化推断为常规维护。

实现拆解

- 单一文件变更: `requirements/cuda.txt` 第 24 行版本号从 `==4.5.0` 改为 `==4.5.2`。
- 无其他文件修改，无测试或文档配套。

无，变更仅涉及版本号字符串。

评论区精华

无 review 讨论。WoosukKwon 直接批准。

风险与影响

- 风险: 极低。补丁版本升级通常向后兼容。但若 `cutedsl` 4.5.1/4.5.2 引入 `cu13` 特定回归，可能影响 `FlashAttention 4` 功能。缺少自动化测试覆盖。
- 影响: 仅影响 `CUDA` 环境下依赖解析，对用户功能无直接可见变化。

关联脉络

作者 `zyongye` 同时负责 `DeepSeek V4` 相关 PR (如 #43710)，该模型可能依赖 `cutedsl`；另有近期 `AMD ROCm` 兼容性工作 (#43710)，但本 PR 仅限 `NVIDIA CUDA`。