

PR #45813 完整报告

vllm-project/vllm

[Doc] Clarify fastokens availability

合并时间: 2026-07-06 21:33

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45813>

执行摘要

该 PR 在文档 [docs/configuration/optimization.md](#) 中补充说明 `VLLM_USE_FASTOKENS` 环境变量从 vLLM v0.23.0 开始可用, 并指导用户升级 vLLM 以使用该功能。变更仅 3 行新增、2 行删除, 纯文档修正。

功能与动机

Issue #43446 报告, 用户按照文档设置 `VLLM_USE_FASTOKENS=1` 但遇到变量未识别的问题, 因为该变量在较早版本中不存在。文档未注明版本要求。PR 旨在消除歧义, 防止用户因版本不匹配而困惑。

实现拆解

- 文件: `docs/configuration/optimization.md`
- 变更: 在描述 fastokens 后端的段落中, 将原本的 "Enable it by setting `VLLM_USE_FASTOKENS=1`:" 改写为 "`VLLM_USE_FASTOKENS` is available in vLLM v0.23.0 and later. If your installed vLLM version does not recognize the environment variable, upgrade vLLM before enabling the override:"
- 结果: 清晰地说明了版本依赖, 并给出升级建议, 代码示例不变。

(无源码变更, 仅文档文字修改。)

评论区精华

无 review 评论, njhill 直接批准合并。

风险与影响

风险: 无技术风险。 影响: 帮助用户快速定位并解决版本问题, 降低 support 成本。

关联脉络

关联 Issue #43446, 该 issue 由用户提出文档缺失版本标注。此 PR 为该 issue 的修复。无其他关联 PR。