

# PR #45432 完整报告

vllm-project/vllm

[Docs] Expand llm-d integration page

合并时间: 2026-07-28 23:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45432>

## 执行摘要

此 PR 扩展了 vLLM 文档中的 llm-d 集成页面，从两句话的简短提示变为包含功能概述、性能数据和入门指引的完整文档。变更仅涉及一个文件，无代码路径影响。

## 功能与动机

原 llm-d 集成页面只有一句话介绍并链接到外部项目，没有说明 llm-d 在 vLLM 之上提供了哪些额外价值。PR 旨在让读者无需点击外部链接即可判断 llm-d 是否适合其使用场景，同时保持文档简洁且链接导向。

## 实现拆解

1. 文件变更: 修改 docs/deployment/integrations/llm-d.md, 将原有的两行内容替换为结构化文档。
2. 新增「What llm-d adds to vLLM」小节: 列举五项核心能力——前缀感知路由、分布式 KV 缓存管理、预填充 / 解码分离、宽专家并行和 SLO 感知自动扩缩容, 每项均附有指向 llm-d 官方文档的链接。
3. 新增「Performance」小节: 展示跨加速器的代表性基准测试结果, 并提供完整列表和可复现测试的链接。
4. 新增「Get started」小节: 提供快速入门、指南和架构参考链接。
5. 保留原有内容: KServe 部署路径 (LLMInferenceService) 说明保持不变。

无可用源码片段 (纯文档变更)。

## 评论区精华

无 review 评论; PR 获得单一审核者 (robertgshaw2-redhat) 批准。

## 风险与影响

- 风险: 低。纯文档变更, 无代码路径风险。但引用的外部链接和性能数据可能随时间过时, 需定期维护。
- 影响: 对用户更友好, 便于快速了解 llm-d 功能。对系统无影响。

## 关联脉络

无关联 PR 或 Issue。该 PR 独立改进文档质量。