PR #43922 完整报告

vllm-project/vllm

docs: clarify ITL acronym in optimization docs

合并时间：2026-05-29 22:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43922

执行摘要

该 PR 仅对优化文档中 ITL 缩写的首次出现进行补全解释，共 2 行修改。无技术影响。

功能与动机

在优化文档中，ITL（inter-token latency）缩写首次出现时未给出全称，可能让不熟悉该缩写的读者产生困惑。PR 作者希望澄清这一术语。

实现拆解

1. 在 docs/configuration/optimization.md 的“Benefits”小节中（第 49 行），将 - It improves ITL and generation decode... 改为 - It improves inter-token latency (ITL) and generation decode...，确保首次出现给出全称。
2. 在“Performance Tuning with Chunked Prefill”小节中（第 55 行），将 - Smaller values (e.g., 2048) achieve better inter-token latency (ITL) because... 改为 - Smaller values (e.g., 2048) achieve better ITL because...，因为此时缩写已被解释，可简化回缩写。

无代码变更。

评论区精华

无实质性讨论。hmellor 直接 approve，Mergify 自动发布文档预览链接。

风险与影响

无风险。仅提升文档可读性。

关联脉络

与近期 PR#43346（修复 KV transfer 影响 ITL 指标）相关，但该 PR 仅澄清术语，不影响任何指标计算逻辑。