PR #45279 完整报告

vllm-project/vllm

docs, kv_offloading: add docs for selective offload

合并时间：2026-06-17 19:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45279

执行摘要

- 一句话：新增 per-request selective offload 文档
- 推荐动作：值得阅读，尤其是使用 KV offloading 的用户，可了解如何精细控制 per-request offload 行为。

功能与动机

之前引入了 selective_offload 配置（commit 864990e8d），但缺少使用文档。本 PR 旨在补充说明和指导，帮助用户理解和使用该功能。

实现拆解

在 docs/features/kv_offloading_usage.md 文件中新增一个二级标题 "## Per-Request Selective Offload"，包含以下内容：

1. 功能解释：说明 max_offload_tokens 的作用——限制请求中可被 offload 的 token 数量，前 max_offload_tokens 个 token 被 offload，之后的 token 跳过 store 路径。
2. 参数表格：列出键名、类型（非负 int）和说明，包括 0 表示禁用 offload，省略或设为 None 表示不设上限，非 int、负数或 bool 值会告警并视为无上限。
3. 注意事项：使用 !!! note 提醒该功能为实验性，可能变更。
4. 示例：提供一个 OpenAI-compatible completions 请求的 JSON 示例，展示如何设置 kv_transfer_params 中的 max_offload_tokens。

关键文件：

- docs/features/kv_offloading_usage.md（模块 文档；类别 docs；类型 documentation）：唯一变更文件，新增 selective offload 配置说明、示例和注意事项。

关键符号：未识别

评论区精华

无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：纯文档变更，无技术风险。但需确保文档内容与代码实现一致，避免误导用户。
- 影响：影响范围小，仅新增文档段落，帮助用户理解和使用 selective offload 功能，不涉及代码逻辑变更。
- 风险标记：暂无

关联脉络

- PR #45595 [KV Connector][Offloading] Avoid blocking the engine to flush offloads on idle: 同为 KV offloading 相关 PR，涉及 offloading 调度优化。