PR #42180 完整报告

vllm-project/vllm

docs: clarify Gemma 4 assistant speculative decoding

合并时间：2026-05-10 11:08

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42180

执行摘要

- 一句话：澄清 Gemma 4 辅助模型须用 MTP 路径
- 推荐动作：建议精读。该 PR 是处理“文档与实现不一致”的标准范例，值得其他特性维护者参考。

功能与动机

用户尝试将 Gemma 4 辅助模型作为通用草稿模型使用，导致初始化失败（Issue#42005）。文档未明确说明 Gemma 4 辅助模型实际走 MTP 路径，造成用户困惑。

实现拆解

1. 在 docs/features/speculative_decoding/mtp.md 新增小节：标题为“Gemma 4 Assistant Models”，解释 Gemma 4 辅助模型通过 method: mtp 使用，并给出完整命令行示例，明确说明 E2B、E4B、26B-A4B、31B 等模型变体均通过 model_type: gemma4_assistant 映射到内部的 Gemma4MTPModel。
2. 在 docs/features/speculative_decoding/README.md 添加醒目的注意事项：使用 !!! note 提醒读者，Gemma 4 辅助模型不是通用草稿模型，必须使用 method: mtp，并提示看到 method=draft_model 日志时需升级 vLLM 版本。
3. 在 docs/models/supported_models.md 的 Gemma 4 条目添加交叉引用：在已有说明后增加一句话，指向 mtp.md 中的具体示例。

关键文件：

- docs/features/speculative_decoding/mtp.md（模块 文档；类别 docs；类型 documentation）：新增 Gemma 4 辅助模型的用法小节和命令行示例，是核心变更。
- docs/features/speculative_decoding/README.md（模块 文档；类别 docs；类型 documentation）：添加醒目的注意事项，提醒读者避免错误用法。
- docs/models/supported_models.md（模块 文档；类别 docs；类型 documentation）：为 Gemma 4 模型条目增加交叉引用，引导用户阅读 mtp.md。

关键符号：未识别

评论区精华

无实质审核评论。PR 由维护者 DarkLight1337 批准，自动化工具（Claude、Gemini）无技术反馈。

- 暂无高价值评论线程

风险与影响

- 风险：纯文档变更，无代码修改，无回归风险。
- 影响：对用户：消除了 Gemma 4 辅助模型配置时的歧义，减少错误尝试。对系统：无影响。
- 风险标记：暂无

关联脉络

- PR #42005 [Doc]: Gemma 4 assistant speculative decoding docs do not match actual behavior on vLLM 0.20.1: 本 PR 直接修复该 Issue 指出的文档误导问题。