PR #40316 完整报告

vllm-project/vllm

[Docs] Fix thinking_token_budget docs

合并时间：2026-04-20 16:09

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40316

执行摘要

该 PR 修复了推理输出文档中一个 curl 示例的错误，将 thinking_token_budget 参数从 extra_body 对象移出到请求体的顶层，确保文档与实际 API 使用方式一致。这是一个纯文档修复，风险极低，影响仅限于提升开发者体验。

功能与动机

根据 PR body 描述，extra_body 是用于 OAI SDK 的，而不是 curl 命令。因此，文档中的 curl 示例错误地将 thinking_token_budget 参数放在了 extra_body 对象内，这可能导致用户在使用 curl 时产生混淆。本次修正旨在提供准确的 API 使用指导，避免误导开发者。

实现拆解

1. 变更入口：修改 docs/features/reasoning_outputs.md 文件，这是推理输出功能的文档页面。
2. 核心修正：调整 curl 示例中的 JSON 结构，将 thinking_token_budget 参数从 extra_body 对象移出，直接作为请求体的顶层字段。
3. 清理冗余：删除 extra_body 包装，简化示例，使其更清晰。
4. 无配套改动：这是一个纯文档修复，不涉及源码、测试、配置或部署文件的任何变更。

评论区精华

review 讨论非常简短，主要确认了变更的正确性：

- gemini-code-assist[bot] 指出："This pull request updates the documentation ... to move the thinking_token_budget parameter from the extra_body object to the top-level request body in a curl example."
- DarkLight1337 批准并致谢。 没有出现争议或未解决的疑虑。

风险与影响

- 风险分析：风险极低。仅修改文档，不影响代码逻辑、性能、安全或兼容性。唯一潜在风险是文档中可能还存在其他类似错误，但本 PR 范围有限，未涉及。
- 影响分析：正面影响。对用户而言，文档更准确，避免使用 curl 时的困惑；对系统和团队无负面影响，反而减少了后续支持成本。

关联脉络

从近期历史 PR 分析看，本 PR 是一个独立的文档修复，与其他 PR 无直接功能关联。类似的小幅文档修正可见于 PR #40266（修复 token_embed 文档拼写错误），体现了团队对文档细节的持续维护。