PR #45867 完整报告

vllm-project/vllm

[Bugfix][Gemma4] Render reasoning on assistant turns without tool_calls

合并时间：2026-06-18 04:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45867

执行摘要

- 一句话：修复 Gemma4 模板中 reasoning 在无工具调用时被丢弃
- 推荐动作：建议阅读以了解聊天模板中 thinking channel 的渲染逻辑，以及如何通过微小条件修复纠正模型行为。该 PR 展示了回归修复的典型协作流程。

功能与动机

修复 #45553 引入的模板回归：thinking channel 守卫 and message.get('tool_calls') 导致 assistant 消息没有 tool_calls 时 reasoning 内容被静默丢弃，例如模型在使用工具推理后给出最终回答的场景。

实现拆解

1. 定位回归源：#45553 在聊天模板中为 thinking channel 加上了 message.get('tool_calls') 条件，意图是只在工具调用轮次渲染 reasoning，却误过滤了没有 tool_calls 的 assistant 消息。
2. 修改条件判断：将 {%- if thinking_text and thinking_gate and message.get('tool_calls') -%} 改为 {%- if thinking_text and thinking_gate -%}，使 reasoning 渲染不再依赖 tool_calls 存在。
3. 同步注释：将 (tool-call turns only) 注释更新为通用描述。
4. 测试验证：运行 124 个现有单元测试全部通过，且作者提供了需要展示推理的多轮对话样例。

关键文件：

- examples/tool_chat_template_gemma4.jinja（模块 聊天模板；类别 other；类型 core-logic）：核心变更文件：移除 thinking channel 渲染中的 tool_calls 条件，使得不含 tool_calls 的 assistant 消息也能正确显示 reasoning 内容。

关键符号：未识别

评论区精华

• m4r1k 指出 #45832 已覆盖 adjust_request 部分，本 PR 只需专注于模板修改，作者同意在 #45832 合并后 rebase。 • bbrowning 审查时指出注释 (tool-call turns only) 不再准确，lucianommartins 已修正。

- 覆盖 adjust_request 修复的范围 (design): 作者同意在 #45832 合并后 rebase，最终 PR 只保留模板修改。
- 同步注释反映新行为 (documentation): 注释更新为通用描述。

风险与影响

- 风险：变更仅 2 行（条件移除 + 注释更新），逻辑简单，但可能引发极低风险的 reasoning 过渲染（当 thinking_text 和 thinking_gate 都满足但本不该渲染时）。然而 thinking_gate 本身有控制逻辑（仅用户之后或 preserve_thinking），且现有测试覆盖正常场景，风险可控。
- 影响：影响范围局限于使用 Gemma4 模型并启用 reasoning 和工具调用的用户。修复后，工具链末尾的 assistant 回答能正确展示推理内容，提升多轮交互体验。由于改动极小且经测试验证，对系统稳定性和性能无影响。
- 风险标记：低风险

关联脉络

- PR #45553 [Bugfix][Gemma4] Fix offline parser truncation, adjust_request token leak, and chat template sync: 引入 thinking channel 守卫 message.get('tool_calls')，导致本 PR 修复的 regression。
- PR #45832 [Bugfix][Gemma4] Fix parsing when thinking is disabled: 修复同一 regression 中的 adjust_request 部分，本 PR 基于其上只包含模板修改。