PR #41846 完整报告

vllm-project/vllm

Fix: Nemotron 3 rescue whitespace-only final_content, not just None

合并时间：2026-05-10 10:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41846

执行摘要

- 一句话：修复 NemotronV3 解析器 whitespace-only 内容
- 推荐动作：该 PR 变更简单，可直接合并。建议后续添加对应 edge case 的单元测试，防止回归。

功能与动机

当 Nemotron 模型在 enable_thinking=False 模式下输出带空白符的内容时，final_content 不为 None 但为空白字符串，原条件 final_content is None 无法触发交换，导致实际回答留在 reasoning 字段，客户端收到空白响应。参考上游修复：https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4/commit/4f0cf9daaeb7a4d5e23f80a00e7ed15f0e03caf6

实现拆解

1. 修改条件判断：在 vllm/reasoning/nemotron_v3_reasoning_parser.py 的 extract_reasoning 方法中，将第 29 行条件 and final_content is None 改为 and (final_content is None or not final_content.strip())。
2. 逻辑说明：strip() 方法会移除字符串首尾空白，若结果为 '' 则返回 True，因此 not final_content.strip() 为真时表示 final_content 全为空白字符，此时应执行交换动作。
3. 影响范围：仅修改一处逻辑表达式，无新增依赖或配置变动。

关键文件：

- vllm/reasoning/nemotron_v3_reasoning_parser.py（模块 推理解析；类别 source；类型 core-logic；符号 extract_reasoning）：核心修改文件，修复了 whitespace-only final_content 的判断条件，与上游模型修复对齐。

关键符号：extract_reasoning

关键源码片段

vllm/reasoning/nemotron_v3_reasoning_parser.py

核心修改文件，修复了 whitespace-only final_content 的判断条件，与上游模型修复对齐。

```
# vllm/reasoning/nemotron_v3_reasoning_parser.py
from vllm.reasoning.deepseek_r1_reasoning_parser import DeepSeekR1ReasoningParser
```

```
class NemotronV3ReasoningParser(DeepSeekR1ReasoningParser):
    """Reasoning parser for Nemotron V3 models."""

    def extract_reasoning(
        self, model_output: str, request: ChatCompletionRequest | ResponsesRequest
    ) -> tuple[str | None, str | None]:
        reasoning, final_content = super().extract_reasoning(model_output, request)
        chat_template_kwargs = getattr(request, "chat_template_kwargs", None)

        # 当 enable_thinking=False 或 force_nonempty_content=True 时，
        # 将 reasoning 与 final_content 交换，确保用户看到内容而非推理。
        # 原条件仅检查 final_content is None，但模型可能输出空白字符串（如 "\n"），
        # 导致交换被跳过。现增加空白检测：not final_content.strip()。
        if (
            chat_template_kwargs
            and (
                chat_template_kwargs.get("enable_thinking") is False
                or chat_template_kwargs.get("force_nonempty_content") is True
            )
            and (final_content is None or not final_content.strip()) # 关键修改
        ):
            reasoning, final_content = final_content, reasoning

        return reasoning, final_content
```

评论区精华

gemini-code-assist[bot] 建议在 reasoning 也为空时避免交换，但 reviewer bbrowning 认为当前修改与上游模型修复一致，且改动极小，批准合并。未添加测试，但 reviewer 鼓励后续补充单元测试。

- 增加空白字符串检测 (correctness): 维持当前修改，未来可补充测试覆盖更多 edge case。

风险与影响

- 风险：低风险。改动仅一行，逻辑与上游模型 HuggingFace 仓库中的修复完全一致。潜在风险：若 reasoning 也是空白，交换结果可能仍为空白，但此种情况实际罕见，且不影响正确性。
- 影响：影响范围极小，仅影响 NemotronV3 模型在 enable_thinking=False 或 force_nonempty_content=True 模式下的输出。用户将正确收到内容而非空白响应。
- 风险标记：缺少测试覆盖

关联脉络

- PR #42026 [Bugfix] Preserve leading/trailing whitespace in GLM non-streaming tool parser: 同为模型 tool parser 的空白处理修复，涉及类似 whitespace 敏感逻辑。