

PR #39570 完整报告

vllm-project/vllm

[Fix] Sync gemma4 chat template from hf

合并时间: 2026-05-02 11:06

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/39570>

执行摘要

本 PR 将 vLLM 内置的 Gemma4 聊天模板与 HuggingFace 官方模板同步，修复了参数格式化逗号处理、添加 reasoning_content 支持、修正思维令牌换行、移除媒体令牌多余换行、改进 tool_response 注入逻辑等多项问题。变更仅限于示例模板文件 [examples/tool_chat_template_gemma4.jinja](#)，不涉及运行时代码，风险可控但对多轮工具调用的正确性至关重要。

功能与动机

PR 描述指出：HuggingFace 在 24h 前更新了 Google Gemma4 的聊天模板，导致 vLLM 中的示例模板过时，某些场景下工具解析出现异常。原模板存在多项技术债务，包括硬编码逗号分隔、缺乏 reasoning_content 字段（与 OpenAI 兼容 API 不匹配）、思维令牌缺少换行、媒体令牌含多余空白符、以及 tool_response 处理边界情况错误。本次同步旨在恢复与模型原生行为的一致。

实现拆解

- 参数格式化重构：在 format_parameters 宏中引入 filter_keys 参数和 add_comma 跟踪机制，替代原 hardcoded 逗号分隔，修复 OBJECT 和 nullable 字段排版问题。
- 新增 reasoning_content 支持：支持 reasoning 和 reasoning_content 消息字段用于思维通道渲染，并确保 <lthinkl> 令牌后附带换行符以符合 OpenAI 兼容 API。
- 媒体令牌格式修正：移除 <limagel> 和 <lvideol> 令牌前后的多余 ` ` 填充，匹配模型预期的格式。
- tool_response 注入边界处理：当 tool_call 无法匹配 tool response 时，注入 <ltool_response> 令牌而非 <turnl>，防止 turn 错误打开；在生成提示中跳过 <lturn>model 打开标签以避免重复。
- 同步上游最新代码：最终 commit 拉取了 2026-04-28 的 HuggingFace commit 145dc25，包含之前所有调整并增加了递归调用时的过滤参数。

[examples/tool_chat_template_gemma4.jinja](#)

唯一被修改的文件，也是 PR 核心变更所在，直接同步上游模板并修复多个格式问题。

```
{%- 重写的 format_parameters 宏，使用 add_comma 跟踪解决 hardcoded 逗号问题 -%}
{%- macro format_parameters(properties, required, filter_keys=false) -%}
{%- set standard_keys = ['description', 'type', 'properties', 'required', 'nullable'] -%}
```

```

{% set ns = namespace(found_first=false) -%}
{% for key, value in properties | dictsort -%}
  {% set add_comma = false -%}
  {%# 如果 filter_keys 为 true, 则跳过 standard_keys, 避免递归时重复输出 #}
  {% if not filter_keys or key not in standard_keys -%}
    {% if ns.found_first %},{% endif -%}
    {% set ns.found_first = true -%}
    {{ key }}:{
    {%# 先处理 description, 设置 add_comma 标志 #}
    {% if value['description'] -%}
      description:<\"|>{{ value['description'] }}<\"|>
      {% set add_comma = true -%}
    {% endif -%}
    {%# STRING 和 ARRAY 类型提前处理, 确保逗号正确 #}
    {% if value['type'] | upper == 'STRING' -%}
      ...
    {% elif value['type'] | upper == 'ARRAY' -%}
      ...
    {% endif -%}
    {%# nullable 和 OBJECT 放在后面, 利用 add_comma 动态拼接 #}
    {% if value['nullable'] %}
      {% if add_comma %},{% else -%}{% set add_comma = true -%}{% endif -%}
      nullable:true
    {% endif -%}
    {% if value['type'] | upper == 'OBJECT' -%}
      {% if add_comma %},{% else -%}{% set add_comma = true -%}{% endif -%}
      properties:{
        {{- format_parameters(value['properties'], value['required'] | default([]), filter_keys=true)
        -}}
      }
    {% if value['required'] -%}
      {% if add_comma %},{% else -%}{% set add_comma = true -%}{% endif -%}
      required:[
        {% for item in value['required'] | default([]) -%}
          <\"|>{{- item -}}<\"|>
          {% if not loop.last %},{% endif -%}
        {% endfor -%}
      ]
    {% endif -%}
    {% endif -%}
    {% if add_comma %},{% else -%}{% set add_comma = true -%}{% endif -%}
    type:<\"|>{{value['type'] | lower }}<\"|>
  }
{% endif -%}
{% endfor -%}
{% endmacro -%}

```

评论区精华

Gemini Code Assist (bot): "The logic for suppressing the `<|turn>model` opening tag at the generation prompt relies on `ns.prev_message_type` being `tool_response` or `tool_call` to indicate an open turn. However, if the assistant message contained content, the turn was explicitly closed but `ns.prev_message_type` still holds the previous value. This results in the model generating its next response outside of any turn tags."

bbrowning: "I compared before/after results with BFCL multi_turn as well as inspecting manual prompts and this looks good. There would be an issue if messages were sent in that contained both content and tool calls, but Gemma 4 doesn't seem to generate that in practice so it ends up being a non-issue for the majority of use-cases."

最终 reviewer 确认缺陷存在但对主流场景影响有限，核准同步。

风险与影响

- 技术风险：turn 标签重置不完善可能在某些边缘场景下产生格式错误，但 Gemma 4 模型自身行为使触发概率低。模板输出与旧版本不兼容，需确保上层依赖兼容性。
- 影响范围：所有使用 Gemma 4 模型并启用工具调用的 vLLM 用户。变更为正向修复，无回退风险。建议用户升级后对多轮工具调用进行回归测试。

关联脉络

- 关联 Issue #39027 为本次同步的原始需求来源。
- 同一仓库近期还有多个 Gemma 4 相关 PR（如 #40796 修复软令牌限幅），表明 vLLM 对 Gemma 4 模型系列持续完善。
- 本次模板同步也是工具调用工具链（tool-calling）稳定性提升的一部分。