

PR #46939 完整报告

vllm-project/vllm

[Misc] Forward request-level prompt extras for cross-encoder scoring

合并时间: 2026-07-04 22:18

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46939>

执行摘要

- 一句话: 转发跨编码器打分请求级 prompt extras
- 推荐动作: PR 变更简洁清晰, 值得精读以理解跨编码器打分的请求流。设计上清晰分离了模板级 extras (chat_template_kwargs) 与引擎级 extras (mm_processor_kwargs, cache_salt), 是一个好范例。

功能与动机

PR body 指出跨编码器打分请求的 `cache_salt` 和 `mm_processor_kwargs` 在引擎处理阶段被丢弃, 导致缓存隔离失败、多模态处理器覆盖未生效。例如 `cache_salt: "user-1"` 和 `mm_processor_kwargs: {"max_pixels": 3048}` 在 main 分支上未传递至引擎输出。

实现拆解

1. 新增导入 `extract_target_prompt`: 在 `vllm/entrypoints/pooling/scoring/io_processor.py` 中导入 `vllm.renderers.inputs.preprocess.extract_target_prompt`, 用于定位引擎 prompt 中的可变目标字典。
2. 提取引擎级 extras: 在 `_pre_process` 方法开头, 从 `prompt_extras` 中构建 `engine_prompt_extras` 字典, 仅保留 `mm_processor_kwargs` 和 `cache_salt` 两个键, 并过滤掉值为 `None` 的项。
3. 注入引擎 prompt: 在 `tok_params.apply_post_tokenization` 之后, 如果 `engine_prompt_extras` 非空, 则调用 `extract_target_prompt` 获取引擎 prompt 中的目标子字典 (`target_prompt`), 然后将 `engine_prompt_extras` 合并进去。此设计确保了 `chat_template_kwargs` 仍保持在模板处理阶段, 不混入引擎 prompt。

关键文件:

- `vllm/entrypoints/pooling/scoring/io_processor.py` (模块入口层; 类别 source; 类型 dependency-wiring): 唯一变更文件, 修复了跨编码器打分中请求级 prompt extras 未传递至引擎的核心问题。新增导入 `extract_target_prompt`, 并在 `_pre_process` 中提取 `mm_processor_kwargs` 和 `cache_salt` 注入引擎 prompt。

关键符号: `ScoringIOProcessor._pre_process`

关键源码片段

vllm/entrypoints/pooling/scoring/io_processor.py

唯一变更文件，修复了跨编码器打分中请求级 prompt extras 未传递至引擎的核心问题。新增导入 `extract_target_prompt`，并在 `_pre_process` 中提取 `mm_processor_kwargs` 和 `cache_salt` 注入引擎 prompt。

```
# vllm/entrypoints/pooling/scoring/io_processor.py
# 新增导入 extract_target_prompt
from vllm.renderers.inputs.preprocess import extract_target_prompt

class ScoringIOProcessor(PoolingIOProcessor):
    # ...

    def _pre_process(
        self,
        scoring_data: ScoringData,
        tok_params: TokenizeParams,
        pooling_params: PoolingParams | None,
        chat_template: str | None = None,
        max_tokens_per_query: int = 0,
        max_tokens_per_doc: int = 0,
        prompt_extras: dict[str, Any] | None = None,
    ) -> tuple[Sequence[EngineInput], list[PoolingParams]]:
        # 移除旧的 todo 注释
        arrival_time = time.time()
        # 从 prompt_extras 中提取引擎级 extras: cache_salt 和 mm_processor_kwargs
        engine_prompt_extras = (
            {
                k: v
                for k in ("mm_processor_kwargs", "cache_salt")
                if (v := prompt_extras.get(k)) is not None
            }
            if prompt_extras
            else None
        )

        # ... 原始 data_1/data_2 处理逻辑 ...

        for q, d in zip(data_1, data_2):
            _, engine_prompt = self.get_score_prompt(
                data_1=q,
                data_2=d,
                encode_kwargs=tok_params.get_encode_kwargs(),
                chat_template=chat_template,
                max_tokens_per_query=max_tokens_per_query,
                max_tokens_per_doc=max_tokens_per_doc,
                # chat_template_kwargs 仅用于模板，不混入引擎 extras
                chat_template_kwargs=prompt_extras.get("chat_template_kwargs")
                if prompt_extras
```

```
        else None,
    )

    # ... token_type_ids 处理 ...

    tok_params.apply_post_tokenization(self.tokenizer, engine_prompt)
    # 注入引擎级 extras 到目标 prompt 中
    if engine_prompt_extras:
        target_prompt = extract_target_prompt(self.model_config, engine_prompt)
        target_prompt.update(engine_prompt_extras)
    engine_inputs.append(
        self.renderer.process_for_engine(engine_prompt, arrival_time)
    )
    return engine_inputs, pooling_params_list
```

评论区精华

无 review 讨论。仅有一条自动 review 提示无法对 fork 进行审查，一条 approver 的 LGTM 评论和一条 CI 重试提示。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。
- 变更仅涉及 ScoringIOProcessor._pre_process 中的一行注释移除（# todo: support prompt_extras）和约 12 行新增逻辑。
- 新增代码仅在 prompt_extras 非空且包含相关键时才激活，不影响现有行为。
- 使用 extract_target_prompt 获取目标字典，该方法已存在于主线，不会引发意外的键错误。
- 影响：影响范围仅限于跨编码器打分请求（/score 端点）。修复后，请求级 cache_salt 和 mm_processor_kwargs 将被正确传递到引擎，为缓存隔离和多模态处理提供一致性。
- 风险标记：变更范围小，无测试文件变更

关联脉络

- PR #47082 [Misc] Preserve cross-encoder pooling extra kwargs: 同一模块（交叉编码器打分）的同类修复，修复了 extra_kwargs 被覆盖的问题。本 PR 在此基础上进一步补全引擎级 extras 的传递。
- PR #46966 [Misc] Validate Pooling cache_salt Values: 同为 pooling 模块的 cache_salt 相关改进，本 PR 确保 cache_salt 被正确传递。