

PR #47333 完整报告

vllm-project/vllm

[Misc] Update request-extras parity for batch chat completion

合并时间: 2026-07-04 22:19

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47333>

执行摘要

- 一句话: 补全批量聊天请求缺少的请求扩展字段
- 推荐动作: 变更简洁、安全, 可以直接合并, 无需深度精读。但对理解 vLLM 批量请求与单请求的字段对齐策略有一定参考价值。

功能与动机

PR 描述指出: 批量聊天补全会将每次对话转换为普通 `ChatCompletionRequest`, 但 `BatchChatCompletionRequest` 没有显式声明普通补全已支持的几个请求扩展字段, 导致功能缺失。具体可通过模型字段查询验证。

实现拆解

1. 协议层新增字段 (`protocol.py`): 在 `BatchChatCompletionRequest` 类的字段定义区 (`# vLLM extensions` 段落) 新增四个字段 —— `media_io_kwargs`、`mm_processor_kwargs`、`priority`、`cache_salt`, 默认值分别取 `None`、`None`、`0`、`None`, 与 `ChatCompletionRequest` 中的定义对齐。
2. 业务层简化调用 (`batch_serving.py`): 将 `create_batch_chat_completion` 方法中传递给 `engine_client.generate` 的 `priority=request.priority if hasattr(request, "priority") else 0` 简化为 `priority=request.priority`, 因为现在字段已稳定存在。

关键文件:

- `vllm/entrypoints/openai/chat_completion/protocol.py` (模块 协议定义; 类别 `source`; 类型 `core-logic`; 符号 `BatchChatCompletionRequest`): 核心变更文件: 在 `BatchChatCompletionRequest` 模型中新增四个请求扩展字段, 与 `ChatCompletionRequest` 对齐。
- `vllm/entrypoints/openai/chat_completion/batch_serving.py` (模块 服务逻辑; 类别 `source`; 类型 `core-logic`; 符号 `create_batch_chat_completion`): 配套简化: 移除 `hasattr` 回退, 直接使用 `request.priority`, 因为字段已稳定存在。

关键符号: `create_batch_chat_completion`

关键源码片段

`vllm/entrypoints/openai/chat_completion/protocol.py`

核心变更文件：在 `BatchChatCompletionRequest` 模型中新增四个请求扩展字段，与 `ChatCompletionRequest` 对齐。

```
# file: vllm/entrypoints/openai/chat_completion/protocol.py
# 在 BatchChatCompletionRequest 类中新增以下字段，置于 # vLLM extensions 下方

media_io_kwargs: dict[str, dict[str, Any]] | None = None
mm_processor_kwargs: dict[str, Any] | None = None
priority: int = Field(default=0, ge=_INT64_MIN, le=_INT64_MAX)
cache_salt: str | None = None
```

评论区精华

无 reviewer 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅新增可选字段，默认值均为 `None` 或 `0`，完全向后兼容；业务层改动仅为移除 `hasattr` 回退逻辑，功能行为不变。
- 影响：影响范围小，仅涉及批量聊天补全 API 的入参模型。用户现可直接在批量请求中设置 `media_io_kwargs`、`mm_processor_kwargs`、`priority`、`cache_salt`，无需手动转为单请求再拼装。
- 风险标记：暂无

关联脉络

- PR #46966 [Misc] Validate Pooling `cache_salt` Values: 涉及相同字段 `cache_salt` 的验证逻辑，可参考其处理方式。