PR #42778 完整报告

vllm-project/vllm

[Model Runner V2] Fix prompt logprobs calculation `Sizes of tensors must match` error

合并时间：2026-05-18 23:27

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42778

执行摘要

- 一句话：修复 V2 模型运行器中 prompt logprobs 张量形状不匹配错误
- 推荐动作：值得精读用于理解 Model Runner V2 中 prompt logprobs 的处理流程，特别是跨请求变长张量切片的处理模式。该 PR 本身逻辑清晰简单，可作为参考学习。

功能与动机

该 PR 是 Model Runner V1 迁移到 V2（Issue #41286）的一部分。当启用 VLLM_USE_V2_MODEL_RUNNER=1 并运行 tests/v1/sample/test_logprobs.py 中的特定测试时，由于批量请求的 num_prompt_logprobs 不同，原始代码试图直接拼接不同形状的张量，导致 RuntimeError: Sizes of tensors must match。

实现拆解

此 PR 仅修改 vllm/v1/worker/gpu/sample/prompt_logprob.py 中的一处核心逻辑，共 +8/-2 行。

1. 提取每个请求的 num_prompt_logprobs：在遍历 input_batch.req_ids 的循环中，新增 req_num_prompt_logprobs = int(num_prompt_logprobs[i])，获取当前请求实际需要的 logprobs 个数。
2. 计算当前请求的宽度 width：新增 width 变量，若 req_num_prompt_logprobs == -1（表示请求所有 logprobs），则宽度取 prompt_logprobs.shape[1]（即全列数）；否则宽度为 req_num_prompt_logprobs + 1（+1 是因为 logprobs 包含当前 token 的预测概率，后续逻辑会去掉最后一项）。这是修复的核心：将原先固定的全列切片改为动态宽度。
3. 使用 :width 进行切片：在构造 LogprobsTensors 对象时，对 logprob_token_ids 和 logprobs 两个张量均使用 [start_idx:end_idx, :width] 切片，而非原来的 [start_idx:end_idx]。selected_token_ranks 的切片保持不变（仍为 [start_idx:end_idx]），因为其维度含义不同。

该改动仅修改源码，未涉及测试配置或部署文件，但 PR body 中提及已通过原有测试验证。

关键文件：

- vllm/v1/worker/gpu/sample/prompt_logprob.py（模块 采样器；类别 source；类型 core-logic）：修复逻辑所在的核心源码文件，修改了 compute_prompt_logprobs 函数中的张量切片逻辑。

关键符号：compute_prompt_logprobs

关键源码片段

vllm/v1/worker/gpu/sample/prompt_logprob.py

修复逻辑所在的核心源码文件，修改了 compute_prompt_logprobs 函数中的张量切片逻辑。

```
# vllm/v1/worker/gpu/sample/prompt_logprob.py
# 在 compute_prompt_logprobs 方法中，关键的修复如下：

req_is_prompt_chunked = is_prompt_chunked[i]
req_num_prompt_logprobs = int(num_prompt_logprobs[i]) # 获取当前请求的 num_prompt_
logprobs
start_idx = query_start_loc_np[i]
end_idx = query_start_loc_np[i + 1]
# 省略断言和 end_idx 调整 ...

# 计算当前请求的切片宽度：-1 表示请求所有列，否则为 num_prompt_logprobs + 1
width = (
    prompt_logprobs.shape[1]
    if req_num_prompt_logprobs == -1
    else req_num_prompt_logprobs + 1
)

logprobs = (
    None
    if start_idx >= end_idx
    else LogprobsTensors(
        # 关键修复：使用 :width 切片，而非整列切片
        logprob_token_ids=prompt_token_ids[start_idx:end_idx, :width],
        logprobs=prompt_logprobs[start_idx:end_idx, :width],
        selected_token_ranks=prompt_ranks[start_idx:end_idx],
    )
)
```

评论区精华

该 PR 无人工 review 评论讨论，gemini-code-assist[bot] 仅提供了代码变更摘要而未提出问题。njhill 直接给予了批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险非常低。变更只涉及单个文件中的 8 行逻辑，且是典型的张量切片边界修复。selected_token_ranks 保持一维切片不变，预期行为正确。若 num_prompt_logprobs 值异常（如为 0 或小于 -1），可能产生空切片，但已有 start_idx >= end_idx 的保护检查会返回 None。该路径仅在启用 VLLM_USE_V2_MODEL_RUNNER 时生效，不影响默认的 Model Runner V1。

- 影响：影响范围局限于 Model Runner V2 下的 prompt logprobs 计算，仅影响启用该功能的用户。影响程度为修复性，无破坏性变更。对系统性能无影响（仅修复正确性）。
- 风险标记：暂无

关联脉络

- PR #41286 [Feature]: Migration from Model Runner v1 to Model Runner v2: 该 PR 是 Model Runner V2 迁移任务清单中的一部分（已标记完成）。