PR #45467 完整报告

vllm-project/vllm

[Model Runner V2] Fix `openai.InternalServerError: Error code: 500 - 'list index out of range'`

合并时间：2026-06-13 16:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45467

执行摘要

- 一句话：修复 Model Runner V2 中 num_logprobs=-1 导致 500 错误
- 推荐动作：值得精读。虽然改动微小，但反映了 Model Runner V2 中参数传递路径的边界处理完整性，可作为模型运行器参数校验的参考模式。

功能与动机

用户使用 VLLM_USE_V2_MODEL_RUNNER=1 并设置 top_logprobs=-1 时，收到 openai.InternalServerError: Error code: 500 - 'list index out of range'。根据 OpenAI 协议，top_logprobs=-1 表示返回所有 logprobs，但 Model Runner V2 的采样状态未处理该值，导致后续数组访问越界。

实现拆解

1. 定位问题根因：在文件 vllm/v1/worker/gpu/sample/states.py 的 add_request 方法中，原先只处理了 num_logprobs is None（代表无 logprobs 请求）的情况，而 num_logprobs == -1（代表请求所有 logprobs）未被处理，直接赋值为 -1，后续索引 vocab_size 数组时越界。
2. 修复方案：在原有 if 分支后增加 elif num_logprobs == -1: 分支，将 num_logprobs 设置为 self.vocab_size，即请求所有 token 的 logprobs。
3. 影响范围：仅涉及一行源码变更（+2/-0），无需修改测试文件，新增的 elif 分支逻辑清晰，不会影响其他采样参数的处理。

关键文件：

- vllm/v1/worker/gpu/sample/states.py（模块 采样器；类别 source；类型 core-logic；符号 add_request）：修复的核心文件，在 add_request 方法中增加对 num_logprobs==-1 的边界处理，将 -1 转换为 vocab_size。

关键符号：add_request

关键源码片段

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

修复的核心文件，在 add_request 方法中增加对 num_logprobs==-1 的边界处理，将 -1 转换为 vocab_size。

```
# 文件：vllm/v1/worker/gpu/sample/states.py
# 修复前：num_logprobs==-1 未被处理，直接赋值为 -1，导致后续索引 vocab_size 数组时越界
# 修复后：将 -1 转换为实际的 vocab_size，请求所有 token 的 logprobs

def add_request(self, req_idx: int, sampling_params: SamplingParams) -> None:
    self.temperature.np[req_idx] = sampling_params.temperature
    self.top_p.np[req_idx] = sampling_params.top_p
    top_k = sampling_params.top_k
    if top_k <= 0 or top_k > self.vocab_size:
        top_k = self.vocab_size
    self.top_k.np[req_idx] = top_k
    self.min_p.np[req_idx] = sampling_params.min_p

    seed = sampling_params.seed
    self.seeds_set[req_idx] = seed is not None
    if seed is None:
        seed = np.random.randint(_NP_INT64_MIN, _NP_INT64_MAX)
    self.seeds.np[req_idx] = seed

    num_logprobs = sampling_params.logprobs
    if num_logprobs is None:
        # None 表示不请求 logprobs，使用特殊值 NO_LOGPROBS
        num_logprobs = NO_LOGPROBS
    elif num_logprobs == -1:
        # -1 表示请求所有 logprobs，转换为实际的 vocab_size
        num_logprobs = self.vocab_size
    self.num_logprobs[req_idx] = num_logprobs
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：该变更仅增加一个边界条件处理，风险极低。但需注意：
  - 若 self.vocab_size 非常大（如 DeepSeek V3 的 4096 或更大），请求所有 logprobs 可能带来显著的内存和带宽压力。
  - 该修复仅在 Model Runner V2 路径生效，V1 路径是否已有类似处理需确认（PR body 未提及 V1 状态）。
  - 影响：影响范围：仅影响使用 VLLM_USE_V2_MODEL_RUNNER=1 且设置 top_logprobs=-1 或 logprobs=-1 的用户。影响程度：低。修复了一个明确的 bug，且改动极小。但需要用户知晓 -1 的特殊语义，未来可能需要在文档中注明。
- 风险标记：缺少测试覆盖

关联脉络

- PR #42667 [Model Runner v2] Migration from v1 to v2, with Qwen and DSv2 MOE models [3/N]: 本 PR 修复了 Model Runner V2 中未被处理的边界情况，属于同一功能线的后续修复。