PR #47680 完整报告

vllm-project/vllm

[Bugfix][V1/V2] Fix prompt_logprobs to respect logprobs_mode

合并时间：2026-07-18 04:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47680

执行摘要

- 一句话：修复 prompt_logprobs 忽略 logprobs_mode 配置
- 推荐动作：建议阅读此 PR，特别是函数重命名和配置门控移除的设计决策。review 中关于命名、类型安全和测试覆盖的讨论值得关注。如果使用 logprobs_mode 功能，请验证 processed_* 模式是否符合预期。

功能与动机

用户 issue #35832 报告 prompt_logprobs 始终返回 log_softmax 结果，忽略 logprobs_mode 的 'raw_logits' 等设置，导致 prompt_logprobs 行为与输出 logprobs 不一致。需要在两个引擎中统一行为，使 prompt_logprobs 尊重 logprobs_mode。

实现拆解

1. 共享函数重构：在 vllm/v1/worker/gpu/sample/logprob.py 中将 compute_topk_logprobs 重命名为 compute_topk_scores，增加 logits_mode 参数。当 logits_mode=True 时，直接通过 gather 获取 logits 值并转换为 float32；否则仍执行 compute_token_logprobs（log_softmax）。
2. V1 引擎适配：在 vllm/v1/worker/gpu_model_runner.py 的 _get_prompt_logprobs_dict 中根据 model_config.logprobs_mode 分支：若为 logits 模式，从 logits 直接取 float32；否则调用 sampler.compute_logprobs 计算概率。
3. V2 引擎适配：在 vllm/v1/worker/gpu/sample/prompt_logprob.py 中的 PromptLogprobsWorker 添加 logprobs_mode 参数，并在 compute_prompt_logprobs_with_chunking 中根据模式设置 logits_mode 传递给 compute_topk_scores。
4. 采样器支持：在 vllm/v1/worker/gpu/sample/sampler.py 中移除对 logprobs_mode 的限制检查（NotImplementedError），允许所有模式，并调整 processed_logits 分支条件，使用处理后的 logits。
5. 配置门控移除：在 vllm/config/vllm.py 中从 _get_v2_model_runner_unsupported_features 移除与 logprobs_mode 相关的条目，使 V2 引擎原生支持所有 logprobs_mode。
6. 配套更新：更新 vllm/v1/worker/gpu/spec_decode/rejection_sampler.py、vllm/model_executor/models/diffusion_gemma.py 等调用点，使用新的 compute_topk_scores 并传递 logits_mode。

7. 测试增加：新增 tests/v1/sample/test_logprobs.py 中的 test_prompt_logprobs_mode 单元测试，以及 tests/v1/sample/test_sampling_params_e2e.py 中的 test_prompt_logprobs_respects_logprobs_mode e2e 测试，验证不同模式下返回值差异。

关键文件：

- vllm/v1/worker/gpu/sample/logprob.py（模块 采样层；类别 source；类型 core-logic；符号 compute_topk_logprobs, compute_topk_scores）：核心共享函数的重命名和逻辑扩展：compute_topk_logprobs → compute_topk_scores，新增 logits_mode 条件分支，决定返回 raw logits 还是 log_softmax。所有调用点均依赖于该函数。
- vllm/v1/worker/gpu/sample/prompt_logprob.py（模块 Prompt 采样；类别 source；类型 core-logic；符号 init）：V2 引擎 prompt logprobs 核心实现。注入 logprobs_mode，传递到 compute_prompt_logprobs_with_chunking，决定使用 logits 还是 logprobs 模式。
- vllm/v1/worker/gpu/sample/sampler.py（模块 采样器；类别 source；类型 core-logic）：V2 采样器，移除对 logprobs_mode 的限制检查，允许所有模式；调整 processed_* 分支条件，使用处理后的 logits；调用 compute_topk_scores 并传递 logits_mode。
- vllm/v1/worker/gpu_model_runner.py（模块 模型运行器；类别 source；类型 data-contract）：V1 引擎 prompt logprobs 计算入口，根据 logprobs_mode 分支返回 logits 或 logprobs。
- vllm/config/vllm.py（模块 配置；类别 source；类型 core-logic）：移除 V2 对 logprobs_mode 的配置门控，使 V2 引擎原生支持所有模式。
- vllm/v1/worker/gpu/spec_decode/rejection_sampler.py（模块 推测解码；类别 source；类型 dependency-wiring）：推测解码模块适配新函数，传递 logits_mode；同时处理了 cu_num_logits 传递问题。
- vllm/model_executor/models/diffusion_gemma.py（模块 扩散模型；类别 source；类型 data-contract）：扩散模型自定义采样器调用 compute_topk_scores，需要传递 logits_mode。
- tests/v1/sample/test_logprobs.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_prompt_logprobs_mode）：新增 test_prompt_logprobs_mode 单元测试，验证所有 LogprobsMode 下 prompt_logprobs 正确分支。

关键符号：compute_topk_scores, compute_topk_logprobs, PromptLogprobsWorker.init, compute_prompt_logprobs_with_chunking, Sampler.init, Sampler.call, gpu_model_runner._get_prompt_logprobs_dict

关键源码片段

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

核心共享函数的重命名和逻辑扩展：compute_topk_logprobs → compute_topk_scores，新增 logits_mode 条件分支，决定返回 raw logits 还是 log_softmax。所有调用点均依赖于该函数。

```
def compute_topk_scores(
    logits: torch.Tensor,
    num_logprobs: int,
    sampled_token_ids: torch.Tensor,
```

```
    cu_num_logits: list[int] | None = None,
    logprob_token_ids_state: "LogprobTokenIdsState | None" = None,
    expanded_idx_mapping: torch.Tensor | None = None,
    max_per_req_token_ids: int = 0,
    logits_mode: bool = False, # 新增参数：True 时返回 raw logits，False 时返回 log_softmax
) -> LogprobsTensors:
    assert num_logprobs >= 0
    batch_size, vocab_size = logits.shape

    if max_per_req_token_ids == 0:
        # 快速路径：所有请求未指定自定义 logprob_token_ids
        logprob_token_ids = sampled_token_ids.unsqueeze(-1)
        if num_logprobs > 0:
            topk_indices = torch.topk(logits, num_logprobs, dim=-1).indices
            logprob_token_ids = torch.cat((logprob_token_ids, topk_indices), dim=1)

        if logits_mode:
            # 直接 gather 原始 logit 值并转换为 float32
            scores = logits.gather(-1, logprob_token_ids).to(torch.float32)
        else:
            # 计算 log_softmax: 每个 token 的 logprob
            scores = compute_token_logprobs(logits, logprob_token_ids)
    else:
        # 某些请求指定了自定义 logprob_token_ids
        # 构建 token_ids 矩阵和有效掩码 (kernel 调用省略具体参数 )
        num_cols = max(num_logprobs, max_per_req_token_ids)
        logprob_token_ids = sampled_token_ids.new_zeros((batch_size, 1 + num_cols))
        valid_mask = torch.zeros_like(logprob_token_ids, dtype=torch.bool)
        _fill_logprob_token_ids_kernel[(batch_size,)](
            logprob_token_ids, ..., valid_mask, ...)

        if logits_mode:
            scores = logits.gather(-1, logprob_token_ids).to(torch.float32)
        else:
            scores = compute_token_logprobs(logits, logprob_token_ids)
        scores = scores.masked_fill(~valid_mask, float("-inf"))

    # 计算采样 token 的排名
    token_ranks = torch.empty(batch_size, dtype=torch.int64, device=logits.device)
    _ranks_kernel[(batch_size,)](token_ranks, logits, ...)

    return LogprobsTensors(
        logprob_token_ids=logprob_token_ids,
        logprobs=scores, # 字段仍命名为 logprobs，但实际内容取决于模式
        selected_token_ranks=token_ranks,
        cu_num_generated_tokens=cu_num_logits,
    )
```

评论区精华

- 配置门控移除风险：@fede-kamel 指出“Removing the config gate regresses V2 + raw_logits from 'works via V1 fallback' to 'crash at model load'”。后续更新使 V2 采样器接受所有 logprobs_mode，风险解除。
- 函数命名：@njhill 提出“renaming logprobs to scores... might still be clearer”，@fede-kamel 回应“scores seems right to me”。最终采用 compute_topk_scores。
- 类型安全：@wojciech-wais 指出 PromptLogprobsWorker.__init__ 和 compute_prompt_logprobs_with_chunking 使用 str 类型而非 LogprobsMode，可能导致拼写静默降级（如 "raw_logrobs" 无报错）。此问题未解决。
- 测试覆盖：@wojciech-wais 指出 e2e 测试仅覆盖 raw_logits/raw_logprobs，未测试 processed_* 模式。也未添加相应单元测试。
  - 移除配置门控导致 V2 raw_logits 崩溃 (correctness): 后续更新使 V2 采样器接受所有 logprobs_mode，风险解除。
  - 函数命名：compute_topk_logprobs → compute_topk_scores (design): 最终采用 compute_topk_scores。
  - logprobs_mode 使用 str 而非 LogprobsMode 类型 (correctness): 未修改，仍使用 str。
  - processed_* 模式在 prompt 中未采样测试覆盖 (testing): 未增加对应测试。

风险与影响

- 风险：
  - 回归风险：移除 V2 配置门控后，若 V2 采样器或 prompt 路径未正确处理所有 logprobs_mode，可能导致模型启动失败。已在 review 中验证修复。
  - 行为差异：processed_* 模式在 prompt logprobs 中未应用采样处理器（logit bias、temperature 等），与文档中“processed 包括后处理”的承诺有出入。现有注释提及此限制，但可能造成用户困惑。
  - 类型安全：使用 str 而非 LogprobsMode 导致拼写错误静默采用默认 raw_logprobs 行为，无编译时检查。
  - 内存压力：在 logits 模式中，当前代码先全量将 logits 转换为 float32 再 gather，可能显著增加显存占用，尤其在大型模型上。Codex 评论建议先 gather 再 cast，但未实施。
- 影响：
  - 用户影响：修复了 prompt_logprobs 行为与 logprobs_mode 配置不一致的 bug。使用 raw_logits 或 processed_logits 的用户现在能从 prompt_logprobs 获取原始 logits 值。
  - 系统影响：V2 引擎不再需要因 logits 模式回退到 V1，提升了推理效率和模式一致性。
  - 团队影响：compute_topk_logprobs 更名为 compute_topk_scores，涉及多个模块需同步更新，可能影响其他进行中的分支。
- 风险标记：配置门控移除风险，processed 模式未应用处理器，类型字符串安全风险，大张量全量 cast 可能 OOM

关联脉络

- PR #35885 [Bugfix] Fix prompt_logprobs to respect logprobs_mode: 此 PR 的前身之一，首次尝试修复，覆盖 V1 和 V2 但未合并。该 PR 的 body 和 co-author 提及为此前身。
- PR #36539 [Bugfix] Fix prompt_logprobs to respect logprobs_mode: 另一前身 PR，聚焦 V1 引擎并包含测试。该 PR 的 body 提及合并其最佳部分。