PR #51906 完整报告

vllm-project/vllm

[Frontend] Add routed-experts prompt offset

合并时间：2026-08-13 23:09

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51906

执行摘要

- 一句话：R3 新增 prompt 偏移参数，统一 numpy 序列化入口
- 推荐动作：值得精读。重点关注三点：① 协议字段 → SamplingParams → 引擎消费的参数链路如何保持单一来源；② 两轮评审如何把 helper 抽象、docstring 精简、额外测试三处过度设计拉回（分别内联、保留、删除），是很好的开源协作示范；③ " 流式 R3 先实现后回退 " 的决策过程对同类 API 扩展有参考意义。若团队在 R3 / Mask Replay 功能线继续迭代，建议以本 PR 的 numpy2base64 为公共序列化基础。

功能与动机

PR body 明确定位：新增 routed_experts_prompt_start "allowing clients to omit an already-known prompt prefix from returned R3"。在 SamplingParams 的详细 docstring 中给出了使用场景：多轮 agent 场景下，上一轮已返回的路由数据无需重复传输，客户端把偏移设为已返回前缀长度即可避免重复数据。PR body 还专门说明与 #49555（sparse-attention indexer top-k）不是重复工作：本 PR 只做 R3 面向前端的共享管道（参数透传 + 序列化收敛），不改变路由计算与模型输出。

实现拆解

1. 参数链路扩展：vllm/sampling_params.py 的 SamplingParams 新增 routed_experts_prompt_start: int = 0；vllm/entrypoints/openai/completion/protocol.py 与 chat_completion/protocol.py 的 CompletionRequest / ChatCompletionRequest 新增同名 Field(default=0, ge=0)，并在 to_sampling_params 中透传给 SamplingParams，保持与 repetition_detection 等字段同一条转发链路；同时把响应字段 routed_experts 的 dtype 注释从 uint8/uint16 扩展为包含 int32。
2. 序列化收敛：vllm/utils/serial_utils.py 新增 numpy2base64（BytesIO + np.save + pybase64，显式 allow_pickle=False）；三个 serving 模块（vllm/entrypoints/scale_out/token_in_token_out/serving.py、openai/completion/serving.py、openai/chat_completion/serving.py）删除各自内联的 io / numpy / pybase64 编码代码并移除对应导入，统一改调 numpy2base64。编码结果与旧实现一致，但关闭了 pickle 能力并集中管理格式。
3. 配置兼容性校验：vllm/config/vllm.py 的 VllmConfig.__post_init__在 enable_return_routed_experts 分支新增 DCP > 1 或 PCP > 1 时抛出 ValueError，理由是 context parallelism 改变 KV 分块与 slot_mapping 映射，slot 索引的 routed_experts

缓冲无法对齐；早期 commit 曾抽象 _check_capture_feature_compat helper，评审后改为内联。

4. 流式行为定稿：提交历史显示最早实现过流式返回 routed_experts（736e3e6），随后 070324d 明确回退，最终保持 OpenAI 流式不返回 R3 的既有行为；141d683 又把偏移校验限定在 HTTP 层。
5. 测试与收尾：tests/utils_/test_serial_utils.py 新增 round-trip；tests/entrypoints/openai/completion/test_completion.py 与 tests/entrypoints/openai/test_stop_token_ids.py 新增字段转发断言；test_return_routed_experts.py 提取 assert_valid_routed_experts 供复用；按评审删除 tests/v1/engine/test_output_processor.py 中不必要的测试（af1ff28）。PR body 注明 49 个测试通过，未跑模型评估（仅传输层改动）。

关键文件：

- vllm/utils/serial_utils.py（模块 序列化；类别 source；类型 core-logic；符号 numpy2base64）：新增 numpy2base64 工具函数，成为 R3 载荷编码的唯一入口；显式 allow_pickle=False 消除 pickle 安全隐患，三个 serving 入口都依赖它。
- vllm/entrypoints/openai/completion/protocol.py（模块 请求协议；类别 source；类型 core-logic；符号 CompletionRequest, CompletionResponseChoice）：CompletionRequest 新增 routed_experts_prompt_start 请求字段并在 to_sampling_params 中透传，是功能的对外协议入口；同时修正 routed_experts 响应注释的 dtype 文档（加入 int32）。
- vllm/entrypoints/openai/chat_completion/protocol.py（模块 请求协议；类别 source；类型 core-logic；符号 ChatCompletionRequest, ChatCompletionResponseChoice）：ChatCompletionRequest 同名改动，保持 chat 与 completion 两条 OpenAI API 面行为一致。
- vllm/config/vllm.py（模块 配置校验；类别 source；类型 core-logic；符号 VllmConfig）：新增 DCP/PCP 与 R3 的不兼容校验，防止 context parallelism 下 slot 索引的路由缓冲错位；评审后由独立 helper 简化为内联实现。
- vllm/sampling_params.py（模块 采样参数；类别 source；类型 core-logic；符号 SamplingParams）：SamplingParams 新增 routed_experts_prompt_start 字段，是引擎参数侧的落点；最终 diff 仅 +2 行，说明评审建议的详细 docstring 被保留。
- vllm/entrypoints/scale_out/token_in_token_out/serving.py（模块 令牌服务；类别 source；类型 dependency-wiring；符号 ServingTokens）：tokens 服务替换为 numpy2base64，移除 io / numpy / pybase64 依赖，与历史 PR #49577 同文件。
- tests/entrypoints/openai/test_return_routed_experts.py（模块 集成测试；类别 test；类型 test-coverage；符号 assert_valid_routed_experts, test_routed_experts）：提取 assert_valid_routed_experts 断言供多测试复用，保持 R3 端到端覆盖的同时减少重复。
- tests/utils_/test_serial_utils.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_numpy2base64_round_trip）：numpy2base64 round-trip 测试，验证编码函数与 np.load 解码的往返一致性。

关键符号：numpy2base64, to_sampling_params, assert_valid_routed_experts, test_completion_request_forwards_routed_experts_prompt_start, test_numpy2base64_round_trip, test_routed_experts_prompt_start_is_forwarded

关键源码片段

vllm/utils/serial_utils.py

新增 numpy2base64 工具函数，成为 R3 载荷编码的唯一入口；显式 allow_pickle=False 消除 pickle 安全隐患，三个 serving 入口都依赖它。

```
# vllm/utils/serial_utils.py —— R3 载荷编码的唯一入口
# routed_experts 是 shape 为 (num_tokens - 1, num_layers, num_experts_per_tok) 的
# 整数 ndarray（专家 ID，dtype 可为 uint8 / uint16 / int32），JSON 无法直接承载
# 二进制，因此统一编码为 base64 字符串随响应返回。


def numpy2base64(array: np.ndarray) -> str:
    """Encode a NumPy array using its `.npy` representation."""
    with io.BytesIO() as buffer:
        # allow_pickle=False：拒绝 pickle 反序列化，避免解码端 np.load 碰到
        # 不可信 base64 载荷时触发任意代码执行；专家 ID 为纯数值，序列化结果
        # 与旧实现字节级一致，仅显式关闭了不必要的 pickle 能力。
        np.save(buffer, array, allow_pickle=False)
        # pybase64 基于 SIMD，大数组下比标准库 base64 快约 3 倍；
        # 返回 .npy 字节流的 ascii 编码，解码端用 np.load 即可还原 shape 与 dtype。
        return pybase64.b64encode(buffer.getbuffer()).decode("ascii")
```

vllm/entrypoints/openai/completion/protocol.py

CompletionRequest 新增 routed_experts_prompt_start 请求字段并在 to_sampling_params 中透传，是功能的对外协议入口；同时修正 routed_experts 响应注释的 dtype 文档（加入 int32）。

```
# vllm/entrypoints/openai/completion/protocol.py —— CompletionRequest 新增可选字段
# 该字段只影响 R3 返回载荷的裁剪，不参与任何采样计算；默认 0 表示返回全部
# prompt token 的路由数据，兼容所有存量请求。

    routed_experts_prompt_start: int = Field(
        default=0,
        ge=0,
        description="Skip the first N prompt tokens from returned routed-expert data.",
    )

    # ... 其余字段 ...

    # to_sampling_params 把协议字段透传给 SamplingParams，与 repetition_detection、
    # thinking_token_budget 等字段保持同一条转发链路；chat 请求的
    # ChatCompletionRequest 做了完全一致的改动。
    return SamplingParams(
```

```
            max_tokens=max_tokens,
            # ... 其余参数 ...
            repetition_detection=self.repetition_detection,
            thinking_token_budget=self.thinking_token_budget,
            routed_experts_prompt_start=self.routed_experts_prompt_start,
        )
```

vllm/config/vllm.py

新增 DCP/PCP 与 R3 的不兼容校验，防止 context parallelism 下 slot 索引的路由缓冲错位；评审后由独立 helper 简化为内联实现。

```
# vllm/config/vllm.py —— VllmConfig.__post_init__ 中 R3 兼容性校验
# 评审前曾抽象出 _check_capture_feature_compat 辅助函数，Isotr0py 认为不必要，
# 最终合并版本把 DCP/PCP 检查内联进既有分支，保持校验逻辑集中可读。

        if (
            self.model_config is not None
            and self.model_config.enable_return_routed_experts
        ):
            # PP > 1 时路由捕获跨 stage 不完整，早先已有拒绝逻辑。
            if self.parallel_config.pipeline_parallel_size > 1:
                raise ValueError(
                    "--enable-return-routed-experts is incompatible with "
                    "pipeline parallelism (PP > 1)."
                )

            # 新增：context parallelism 会改变 KV 分块与 slot_mapping 映射，
            # 使 slot 索引的 routed_experts 缓冲无法对齐，显式报错优于返回错位数据。
            if (
                self.parallel_config.decode_context_parallel_size > 1
                or self.parallel_config.prefill_context_parallel_size > 1
            ):
                raise ValueError(
                    "--enable-return-routed-experts is incompatible with context "
                    "parallelism (DCP > 1 or PCP > 1)."
                )
```

评论区精华

评审意见集中在三处过度设计回退与一处未采纳的简化建议：

- Isotr0py 质疑为 DCP 检查单独抽象 _check_capture_feature_compat 函数不必要，合并代码采纳后改为在既有分支内联判断。
- Isotr0py 认为 SamplingParams 的详细 docstring（说明多轮场景语义）比精简版更好，最终 sampling_params.py 的 diff 仅 +2 行，说明详细描述被保留。
- Isotr0py 建议删除 output processor 的新增测试，最终提交 af1ff28 "Remove redundant output processor test" 落实。

- njhill 建议 numpy2base64 直接对 None 返回 None 以省去调用处三元表达式，但合并后的 head 版本仍保留 numpy2base64(...) if output.routed_experts is not None else None 写法，该建议未采纳。
- R3 兼容性检查是否需要独立 helper 函数 (design): 采纳建议，合并版本改为在 enable_return_routed_experts 分支内联 DCP/PCP 判断，最终 config 改动为 +8 行。
- SamplingParams 字段 docstring 详细程度 (documentation): 最终 sampling_params.py 的 diff 仅为 +2 行新增字段，说明详细 docstring 被保留，评审意见生效。
- output processor 新增测试是否必要 (testing): 已删除，最终提交 af1ff28 "Remove redundant output processor test"，测试文件不在合并变更列表中。
- numpy2base64 对 None 的简化处理 (design): 未采纳，合并后的 head 版本仍保留 numpy2base64(...) if output.routed_experts is not None else None 写法，类型上保持 numpy2base64 只接受 ndarray。

风险与影响

- 风险：
  1. 偏移无上界校验：routed_experts_prompt_start 仅声明 ge=0，且 141d683 刻意把校验限定在 HTTP 层。若客户端传入超过实际 prompt 长度的偏移，或与 tokenizer 计数口径不一致（多模态占位符场景），返回的 routed_experts 数组可能为空或与 token_ids 错位，属数据契约风险，依赖客户端自行保证。
  2. DCP/PCP 组合新增拒绝语义：存量启用 --enable-return-routed-experts 且开启 context parallelism 的用户会直接启动失败。这是刻意的正确性保护（避免错位数据），但对既有配置是行为变更，需要文档同步提示。
  3. 序列化路径统一：numpy2base64 与原实现字节级一致，差异仅在 allow_pickle=False，不会影响纯数值的专家 ID 数组；round-trip 测试覆盖了该行为。
  4. 流式路径无 R3 断言：回退流式 R3 后相关测试被删除，OpenAI 流式不返回 R3 的行为靠约定维持，后续若重开流式 R3 需重新补覆盖。 - 影响：对客户端：新字段可选、向后兼容，多轮 agent 与 RL 数据回流场景可显著降低 routed_experts payload 体积。对系统：纯前端传输与校验改动，不改变模型输出与采样结果；新增配置组合校验影响面为启用 R3 的 context parallelism 用户。对团队：serial_utils.py 成为 R3 载荷编码的唯一入口，后续流式 R3、Mask Replay 等回放功能可直接复用，降低三处重复编码的维护成本。 - 风险标记：偏移无上界校验，DCP/PCP 组合新增拒绝语义，流式 R3 先实现后回退，序列化统一为 allow_pickle=False

关联脉络

- PR #49577 [Feature] Mask Replay: 同改 vllm/entrypoints/scale_out/token_in_token_out/serving.py，且 sampling_mask 与 routed_experts 同属 v1 采样回放能力族；本 PR 的 numpy2base64 为该类回放载荷提供统一编码入口。
- PR #50874 [Bugfix][R3] Size monolithic routing replay buffer for DP: R3 链路引擎 / 内核侧配套修复（DP/EP 下路由回放缓冲），与本 PR 的前端参数透传互补，共同完善 routed_experts 全链路。

- PR #49555 Return sparse-attention indexer top-k (referenced in PR body): PR body 明确对比说明与其不是重复工作：该 PR 返回 sparse-attention indexer top-k，本 PR 只做 R3 共享管道（偏移透传 + 序列化收敛）。