

PR #44638 完整报告

vllm-project/vllm

[Disagg] return routed_experts on streaming generate responses

合并时间: 2026-06-21 22:37

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44638>

执行摘要

- 一句话: 流式生成响应添加 MoE 路由信息
- 推荐动作: 值得精读以了解 vLLM disagg 服务中 MoE 路由数据的序列化与传输模式。代码简洁, 设计意图清晰, 是良好的功能补充范例。

功能与动机

disagg 服务的非流式路径已在 `GenerateResponseChoice` 中返回 per-token MoE 路由, 但流式路径 `GenerateResponseStreamChoice` 遗漏了该字段。此 PR 旨在补充流式路径的路由信息, 使流式消费者 (如 RL 训练器中的 R3 rollout) 无需额外非流式调用即可获取完整路由数据。

实现拆解

1. 扩展流式响应协议模型 (`vllm/entrypoints/serve/disagg/protocol.py`): 在 `GenerateResponseStreamChoice` 类中添加 `routed_experts: str | None = None` 字段, 类型与语义与非流式的 `GenerateResponseChoice.routed_experts` 一致。
2. 编码并注入路由数据 (`vllm/entrypoints/serve/disagg/serving.py`): 在 `serve_tokens_stream_generator` 函数中, 当 `output.routed_experts` 不为 `None` 时, 使用 `io.BytesIO + np.save + base64` 编码为 `base64` 字符串, 并传递给 `GenerateResponseStreamChoice` 的 `routed_experts` 字段; 否则保持 `None`。该逻辑完全复用了非流式路径的序列化方式。

关键文件:

- `vllm/entrypoints/serve/disagg/protocol.py` (模块 协议层; 类别 source; 类型 core-logic; 符号 `GenerateResponseStreamChoice`): 定义流式响应数据模型, 新增 `routed_experts` 字段。
- `vllm/entrypoints/serve/disagg/serving.py` (模块 服务层; 类别 source; 类型 core-logic; 符号 `serve_tokens_stream_generator`): 实现流式生成中的 `routed_experts` 编码与注入逻辑。

关键符号: `serve_tokens_stream_generator`

关键源码片段

vllm/entrypoints/serve/disagg/protocol.py

定义流式响应数据模型，新增 routed_experts 字段。

```
class GenerateResponseStreamChoice(BaseModel):
    index: int
    logprobs: ChatCompletionLogProbs | None = None
    finish_reason: str | None = None
    token_ids: list[int] | None = None
    # Per-token expert routing decisions, base64-encoded ``.numpy`` bytes.
    # ``None`` if not available or feature disabled.
    routed_experts: str | None = None
```

vllm/entrypoints/serve/disagg/serving.py

实现流式生成中的 routed_experts 编码与注入逻辑。

```
# within serve_tokens_stream_generator, for each output:
routed_experts_b64 = None
if output.routed_experts is not None:
    buf = io.BytesIO()
    np.save(buf, output.routed_experts)
    routed_experts_b64 = base64.b64encode(buf.getvalue()).decode("ascii")

chunk = GenerateStreamResponse(
    request_id=request_id,
    choices=[
        GenerateResponseStreamChoice(
            index=i,
            logprobs=logprobs,
            finish_reason=finish_reason,
            token_ids=as_list(delta_token_ids),
            routed_experts=routed_experts_b64, # <-- new field
        )
    ],
)
```

评论区精华

该 PR 无 review 评论，由 ywang96 和 njhill 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：纯新增字段，默认值为 None，当 enable_return_routed_experts 关闭时不影响现有行为。序列化方式与非流式路径一致，已通过现有测试覆盖。
- 影响：影响范围限定在 disagg 服务流式响应路径，对已有消费者无破坏性变更。流式消费者可据此获取 per-token MoE 路由信息，优化 RL 训练中的 R3 rollout 流程。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR