PR #26800 完整报告

sgl-project/sglang

Fix the EAGLE chunked-prefill next-token chain (#26329)

合并时间：2026-05-31 09:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26800

执行摘要

- 一句话：修复 EAGLE chunked prefill draft 链发散 bug
- 推荐动作：建议精读本次变更，理解 chunked prefill 与推测解码交互的细节。值得关注的设计决策是：通过在 ScheduleBatch 中存储 chunked_req_next_prompt_token，将 chunked 状态的查询与 draft worker 解耦。此外，建议尽快将多层 EAGLE worker 中的 TODO 落实为实际修复，并补充端到端测试。

功能与动机

修复 EAGLE chunked prefill 场景下 draft 链与目标模型发散的问题。当 EAGLE 使用 chunked prefill 时，非最终 chunk 的 tail token 应使用下一个 prompt token 而不是已验证的 next token，否则 draft chain 会偏离目标，导致生成质量下降。

实现拆解

1. 在 ScheduleBatch 中添加 chunked_req_next_prompt_token 字段（python/sglang/srt/managers/schedule_batch.py）：新增 _compute_chunked_req_next_prompt_token 函数，通过检查 chunked_req.fill_ids 长度与 origin_input_ids 长度，确定下一个 prompt token。在 init_new 中调用该函数初始化新字段。
2. 新增 _eagle_prefill_tail_tokens 工具函数（python/sglang/srt/speculative/eagle_utils.py）：该函数接收 batch 和 next_token_ids，对于 chunked request，用 chunked_req_next_prompt_token 替换对应的 tail token，否则保持原值。同时保留了原 apply_eagle_prefill_input_rotation 中的 TODO 以便后续统一。
3. 在单层 EAGLE draft worker 中应用（python/sglang/srt/speculative/eagle_worker_v2.py）：修改 _draft_extend_for_prefill，调用 _eagle_prefill_tail_tokens 获取正确的 tail tokens，替换原来的 next_token_ids[i]。
4. 在多层 EAGLE draft worker 中添加 TODO（python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py）：由于多层 worker 使用 rotate_input_ids_triton 且存在同样的 chain divergence 问题，添加 TODO 注释引用 PR#26329，待后续修复。
5. 测试配套：本次变更没有包含直接对应的测试文件，但修复本身为 bugfix，涉及核心推测解码路径。

关键文件：

- python/sglang/srt/speculative/eagle_utils.py（模块 推测解码；类别 source；类型 core-logic；符号 _eagle_prefill_tail_tokens）：新增 _eagle_prefill_tail_tokens 核心函数，实现 chunked-aware 的 tail token 替换逻辑；同时保留了原始函数中的 TODO。
- python/sglang/srt/managers/schedule_batch.py（模块 调度器；类别 source；类型 data-contract；符号 _compute_chunked_req_next_prompt_token）：添加 chunked_req_next_prompt_token 字段和 _compute_chunked_req_next_prompt_token 函数，用于计算并存储 chunked request 的下一个 prompt token。
- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic）：修改 _draft_extend_for_prefill 方法，使用 _eagle_prefill_tail_tokens 获取正确的 tail tokens，修复 chunked prefill 下的 draft 链发散。
- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic）：添加 TODO 注释，指出存在相同的 chunked prefill chain divergence 问题，待后续修复。

关键符号：_eagle_prefill_tail_tokens, _compute_chunked_req_next_prompt_token

关键源码片段

python/sglang/srt/speculative/eagle_utils.py

新增 _eagle_prefill_tail_tokens 核心函数，实现 chunked-aware 的 tail token 替换逻辑；同时保留了原始函数中的 TODO。

```
def _eagle_prefill_tail_tokens(
    batch: ScheduleBatch, next_token_ids: torch.Tensor
) -> torch.Tensor:
    """Per-seq tail token for EAGLE prefill rotation; uses next prompt token for
    non-final chunks (chunked-prefill chain consistency, see PR #26329)."""
    # 默认使用 verified next token 作为 tail token
    tail_tokens = next_token_ids.to(batch.input_ids.dtype)
    # 如果 batch 有 chunked request 的 next prompt token
    next_prompt_token = batch.chunked_req_next_prompt_token
    if next_prompt_token is not None:
        # 遍历找到 chunked request，替换其 tail token 为 next prompt token
        for i, r in enumerate(batch.reqs):
            if r is batch.chunked_req:
                tail_tokens = tail_tokens.clone()
                tail_tokens[i] = next_prompt_token
                break
    return tail_tokens
```

python/sglang/srt/managers/schedule_batch.py

添加 chunked_req_next_prompt_token 字段和 _compute_chunked_req_next_prompt_token 函数，用于计算并存储 chunked request 的下一个 prompt token。

```
def _compute_chunked_req_next_prompt_token(
    chunked_req: Optional[Req],
) -> Optional[int]:
```

```
    """根据 chunked request 的 fill_ids 计算下一个 prompt token.
    如果 chunked_req 为 None 或已经完成预填充，返回 None."""
    if chunked_req is None:
        return None
    fill_len = len(chunked_req.fill_ids)
    if fill_len >= len(chunked_req.origin_input_ids):
        return None
    return int(chunked_req.origin_input_ids[fill_len])

# 在 ScheduleBatch 类中新增字段：
@dataclasses.dataclass
class ScheduleBatch(ScheduleBatchDisaggregationDecodeMixin):
    # ...
    chunked_req_next_prompt_token: Optional[int] = None
    # ...

    @classmethod
    def init_new(cls, ...):
        # ...
        batch = cls(
            # ...
            chunked_req=chunked_req,
            chunked_req_next_prompt_token=_compute_chunked_req_next_prompt_token(chunked_
            req),
            # ...
        )
```

python/sglang/srt/speculative/eagle_worker_v2.py

修改 _draft_extend_for_prefill 方法，使用 _eagle_prefill_tail_tokens 获取正确的 tail tokens，修复 chunked prefill 下的 draft 链发散。

```
# 导入新函数
from sglang.srt.speculative.eagle_utils import (
    TreeMaskMode,
    _eagle_prefill_tail_tokens, # 新增
    build_tree_kernel_efficient,
    per_step_draft_out_cache_loc,
)

# 在 _draft_extend_for_prefill 中使用：
def _draft_extend_for_prefill(self, batch, target_hidden_states, next_token_ids, ...):
    if not batch.forward_mode.is_idle():
        # Chunked-prefill-aware tail tokens (see PR #26329).
        tail_tokens = _eagle_prefill_tail_tokens(batch, next_token_ids)
        pt = 0
        for i, extend_len in enumerate(batch.extend_lens):
            input_ids = batch.input_ids[pt : pt + extend_len]
            batch.input_ids[pt : pt + extend_len] = torch.cat(
                (input_ids[1:], tail_tokens[i].reshape(1))
```

```
        )
        pt += extend_len
```

评论区精华

Code review bot（gemini-code-assist[bot]）提出了两个改进建议：

1. 在 eagle_utils.py 中直接使用 _eagle_prefill_tail_tokens：建议移除 apply_eagle_prefill_input_rotation 中的 TODO，直接调用新函数以确保所有 EAGLE prefill 旋转路径的一致性。
2. 在多层 EAGLE worker 中应用相同修复：建议在 multi_layer_eagle_worker_v2.py 中导入并使用 _eagle_prefill_tail_tokens，避免循环依赖可考虑局部导入，并传递正确的 tail tokens 给 rotate_input_ids_triton。 目前这些建议未被采纳，仍以 TODO 形式留待后续。

- 在 eagle_utils.py 中直接使用 _eagle_prefill_tail_tokens (design): 未被采纳，保留了 TODO。
- 在多层 EAGLE worker 中应用相同修复 (design): 未被采纳，仅添加了 TODO。

风险与影响

- 风险：
  1. 回归风险（低）：仅修改了 chunked prefill 下的 draft 路径，非 chunked 场景行为不变。但缺少针对 chunked prefill + EAGLE 的专项测试，可能遗漏边界情况。
  2. 多层 EAGLE worker 未同步修复：multi_layer_eagle_worker_v2.py 中仅添加了 TODO，实际 bug 仍然存在，如果用户使用多层 EAGLE + chunked prefill，draft 链仍然会发散。
  3. 性能影响（极低）：新函数 _eagle_prefill_tail_tokens 引入了额外的循环和克隆操作，但仅在 chunked request 时执行，开销可忽略。 - 影响：影响范围：仅限于使用 EAGLE 推测解码且启用 chunked prefill 的场景。修复确保 draft 链与目标模型一致，提升生成质量。影响程度：对于受影响用户，这是一个正确的 bug 修复；对于不使用 chunked prefill 或 EAGLE 的用户，无影响。团队影响：需要在多层 EAGLE worker 中跟进修复。 - 风险标记：缺少测试覆盖，多层 EAGLE worker 未同步修复

关联脉络

- PR #26329 Original issue/PR for the chain divergence fix: 本 PR 直接关联 issue/PR #26329，多个 TODO 和注释中明确引用。