PR #25033 完整报告

sgl-project/sglang

Fix kimi k2.5 mla eagle + dp attention

合并时间：2026-05-12 11:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25033

执行摘要

- 一句话：修复 Kimi K2.5 MLA EAGLE 在 DP 注意力下的 embedding 越界
- 推荐动作：建议合并，该修复解决了明确的 runtime 错误。但应跟踪后续是否添加对应测试。

功能与动机

修复 Kimi K2.5 模型在 MLA + EAGLE3 推理时，由于 input_ids 中包含多模态 sentinel（MM_PAD_SHIFT_VALUE+hash），直接调用 embed_tokens(input_ids) 会导致索引越界（out-of-bounds）错误。PR body 引用相似修复 #21391。

实现拆解

1. 问题定位：在 kimi_k25_eagle3.py 的 Eagle3MLADecoder 的 forward 方法中，当 input_embeds 为 None 时，原始代码直接调用 self.embed_tokens(input_ids)。但多模态模式下，input_ids 可能包含 MM_PAD_SHIFT_VALUE+hash 这样的 sentinel 值，远超词表大小，导致 embedding 层索引越界。
2. 修复方案：改用 forward_batch.mm_input_embeds（由 target 模型产生的多模态嵌入）作为默认 embedding，仅在非多模态扩展阶段回退到 self.embed_tokens。具体逻辑：
   - 首先尝试使用 forward_batch.mm_input_embeds。
   - 如果是扩展模式（extend）且包含多模态输入，且不是 draft-extend 阶段，则保留 mm_input_embeds[:-1]，并对最后一个 token（即 appended next-token）单独执行 embed_tokens(input_ids[-1].unsqueeze(0))，然后拼接。
   - 如果 mm_input_embeds 为 None，则回退到 self.embed_tokens(input_ids)。
3. 改动范围：仅修改 python/sglang/srt/models/kimi_k25_eagle3.py 文件，共 +15/-1 行，逻辑集中在新条件的嵌入选择部分。无测试文件变更。

关键文件：

- python/sglang/srt/models/kimi_k25_eagle3.py（模块 模型层；类别 source；类型 data-contract；符号 forward）：该文件是 Kimi K2.5 EAGLE3 模型实现，本次修复的核心改动所在。修改了 forward 方法中的 embedding 选取逻辑，避免多模态 sentinel 值导致 OOB 错误。

关键符号：forward

关键源码片段

python/sglang/srt/models/kimi_k25_eagle3.py

该文件是 Kimi K2.5 EAGLE3 模型实现，本次修复的核心改动所在。修改了 forward 方法中的 embedding 选取逻辑，避免多模态 sentinel 值导致 OOB 错误。

```
# 修改后：通过 target 产生的 mm_input_embeds 避免 sentinel OOB
if input_embeds is None:
    # MM 位置在 input_ids 中保存了 MM_PAD_SHIFT_VALUE+hash 哨兵值（远大于 vocabulary_
    size）。
    # 对这些位置使用 target 产生的 mm_input_embeds，仅对附加的下一个 token 调用 embed_
    tokens，
    # 以避免 embed 索引越界（OOB）。
    embeds = forward_batch.mm_input_embeds
    if (
        forward_batch.forward_mode.is_extend()
        and forward_batch.contains_mm_inputs()
        and not forward_batch.forward_mode.is_draft_extend(include_v2=True)
    ):
        assert embeds is not None
        # 保留 mm_input_embeds[:-1]（已有嵌入），仅对最后一个 token 执行 embed
        embeds = torch.cat(
            [embeds[:-1], self.embed_tokens(input_ids[-1].unsqueeze(0))]
        )
    if embeds is None:
        embeds = self.embed_tokens(input_ids)
else:
    embeds = input_embeds
```

评论区精华

该 PR review 评论数为 0，无讨论线程。作者在合并后在 issue 评论中说明“已在本地测试，暂时没有测试覆盖此场景，后续可能添加 CI 测试”。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：改动集中在 input_embeds 为 None 的分支，改变了默认 embedding 的来源。对于非多模态场景，forward_batch.mm_input_embeds 可能为 None，此时会回退到原始逻辑，行为不变。对于多模态场景，新逻辑依赖 mm_input_embeds 的正确性，若 target 模型未正确传递该字段，可能产生错误 embedding。
  2. 缺少测试覆盖：作者明确说明无测试，存在潜在质量问题。
  3. 影响范围：仅影响 Kimi K2.5 MLA + EAGLE3 的 DP 注意力路径，其他模型或配置不受影响。 - 影响：用户影响：修复了 Kimi K2.5 模型在特定配置（MLA + EAGLE3 + DP 注意力）下的崩溃问题，提升了稳定性和可用性。系统影响：改动极小，不引入新依赖或

配置项。团队影响：无。

- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #21391 Fix MLA EAGLE3 with MM inputs: PR body 中提及此 PR 与 #21391 类似，修复相同类别的多模态嵌入索引问题。