PR #26726 完整报告

sgl-project/sglang

fix(spec-dec): treat `num_nextn_predict_layers=0` the same as absent for EAGLE3 drafts

合并时间：2026-06-06 07:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26726

执行摘要

- 一句话：修复 EAGLE3 draft num_nextn_predict_layers=0 时层数计算错误
- 推荐动作：建议尽快合入并发布补丁，该修复解决了 EAGLE3 的一个显式崩溃问题，且风险极低。同时建议在相关测试中增加 num_nextn_predict_layers=0 的边界测试用例。

功能与动机

用户报告使用 nvidia/Kimi-K2.5-Thinking-Eagle3 作为 EAGLE3 draft 时，SGLang 在首次 draft forward 中崩溃，错误为 IndexError: list index out of range。根因是该 draft 模型的 config.json 中 num_nextn_predict_layers=0，而代码中的 MTP 检测仅检查 is not None，导致 draft worker 将层数设为 0，KV 缓存池大小为 0，后续 set_mla_kv_buffer 越界。

实现拆解

在 python/sglang/srt/model_executor/model_runner.py 中修改 model_has_mtp_layers 的判断条件：从原有的 self.model_config.num_nextn_predict_layers is not None 改为 _nnpl is not None and _nnpl > 0。该一行语义修正确保了当 num_nextn_predict_layers 被显式设为 0 时（例如 nvidia/Kimi-K2.5-Thinking-Eagle3），不被视为 MTP 模型，从而正确使用 num_hidden_layers 计算层数。其他逻辑完全不变。

关键文件：

- python/sglang/srt/model_executor/model_runner.py（模块 模型运行器；类别 source；类型 data-contract）：包含核心层数计算逻辑，修改了 MTP 检测条件的语义，从 is not None 改为 is not None and > 0，这是修复 bug 的唯一变更文件。

关键符号：未识别

关键源码片段

python/sglang/srt/model_executor/model_runner.py

包含核心层数计算逻辑，修改了 MTP 检测条件的语义，从 is not None 改为 is not None and > 0，这是修复 bug 的唯一变更文件。

```
# python/sglang/srt/model_executor/model_runner.py
# 原逻辑：model_has_mtp_layers = self.model_config.num_nextn_predict_layers is not None
# 新逻辑：显式排除值为 0 的情况（EAGLE3 draft 可能携带完整 DSv3 配置且设为 0）
```

```
_nnpl = self.model_config.num_nextn_predict_layers
model_has_mtp_layers = _nnpl is not None and _nnpl > 0

model_num_layers = (
    self.model_config.num_nextn_predict_layers
    if self.is_draft_worker and model_has_mtp_layers
    else max(
        self.model_config.num_hidden_layers,
        self.model_config.num_attention_layers,
    )
)
# 后续代码不变
```

评论区精华

无讨论内容。两个 reviewer 直接批准，没有提出疑问或争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。新条件 is not None and > 0 比旧条件更保守：对于所有已有 MTP draft（num_nextn_predict_layers > 0）行为不变；对于省略该字段的模型（None）行为不变；仅对于显式设为 0 的情况从之前的错误行为纠正为正确行为。回归测试覆盖了 MTP 正值和字段缺失场景，0 值场景为新增边界条件。
- 影响：直接影响使用 EAGLE3 draft 且 draft 模型配置中 num_nextn_predict_layers=0 的用户（当前已知为 nvidia/Kimi-K2.5-Thinking-Eagle3）。修复后该组合可以正常启动和推理，并在 GSM8K 上达到 94% 准确率。对其他用户完全透明，无行为变更。
- 风险标记：边界条件

关联脉络

- PR #27338 [Bug] Fix EAGLE draft CUDA-graph kv_indices under-allocation for topk > 1: 共同属于 speculative-decoding 的 bugfix，且都涉及 EAGLE 推测解码的 KV 分配问题，与本 PR 在错误路径上可能有交互。