PR #23060 完整报告

sgl-project/sglang

[fix] Fix dynamic chunking profiling crash on GLM-5 models

合并时间：2026-04-23 19:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23060

执行摘要

- 一句话：修复动态分块 profiling 在 GLM-5 模型上的崩溃
- 推荐动作：建议合并，该修复精准定位了 profiling 路径中缺失的标志初始化问题，改动极小且正确性明确。

功能与动机

Issue #23057 报告：在 GLM-5 等使用 DeepEP 的模型上，--enable-dynamic-chunking 导致 profiling 崩溃，原因为 profiling 路径绕过 prepare_mlp_sync_batch，未设置 _is_extend_in_batch。该错误还引发 KV cache 泄漏和 NCCL 超时。

实现拆解

1. 添加导入：在 scheduler_pp_mixin.py 中导入 set_is_extend_in_batch（来自 sglang.srt.layers.dp_attention）。
2. 显式设置标志：在 profile_and_init_predictor 方法的 profiling 循环中，在调用 model_runner.forward() 之前，添加 set_is_extend_in_batch(batch.forward_mode.is_extend())，确保在 forward 时该标志已正确初始化。
3. 无其他文件变更：仅修改一个文件，新增 3 行代码。

关键文件：

- python/sglang/srt/managers/scheduler_pp_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 profile_and_init_predictor）：核心调度器，包含动态分块 profiling 逻辑。修复了在此处 profiling 循环中缺失的 _is_extend_in_batch 设置。

关键符号：profile_and_init_predictor

关键源码片段

python/sglang/srt/managers/scheduler_pp_mixin.py

核心调度器，包含动态分块 profiling 逻辑。修复了在此处 profiling 循环中缺失的 _is_extend_in_batch 设置。

```
# python/sglang/srt/managers/scheduler_pp_mixin.py (modified)
# 在 profile_and_init_predictor 方法的 profiling 循环中，添加标志设置
# 导入 set_is_extend_in_batch（已添加）
```

```
from sglang.srt.layers.dp_attention import (
    get_attention_dp_rank,
    get_attention_dp_size,
    is_dp_attention_enabled,
    set_is_extend_in_batch, # 新增导入
)

# ... 在循环内部，forward 之前
forward_batch = ForwardBatch.init_new(model_worker_batch, model_runner)
set_is_extend_in_batch(batch.forward_mode.is_extend()) # 新增：确保标志正确设置
_ = model_runner.forward(
    forward_batch=forward_batch, pp_proxy_tensors=pp_proxy
)
```

评论区精华

- gemini-code-assist[bot]建议进一步设置 forward_batch 上的 is_extend_in_batch 和 all_extend_in_batch 属性，以提高兼容性。但最终未采纳，PR 已被批准。
- 建议额外设置 forward_batch 属性 (design): 未被采纳，当前修复已解决崩溃问题，且转发批处理属性可能在后续其他路径中设置。PR 被批准。

风险与影响

- 风险：低风险。仅添加一行函数调用，不改变现有逻辑。但需确认是否还有其他类似路径（如非 PP 模式）存在相同问题。
- 影响：修复了特定模型（使用 DeepEP 的 MoE 模型如 GLM-5）在动态分块下的功能性崩溃，确保动态分块可用。影响范围有限，但修复了关键路径上的静默错误。
- 风险标记：核心路径修复

关联脉络

- 暂无明显关联 PR