PR #40844 完整报告

vllm-project/vllm

[Bugfix] add seq_lens_cpu_upper_bound to CommonAttentionMetadata in mla_runner.py

合并时间：2026-04-25 07:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40844

执行摘要

- 一句话：新增 seq_lens_cpu_upper_bound 参数修复 MLA 注意力测试
- 推荐动作：建议开发者精读以了解 CommonAttentionMetadata 的构造演进方向。同时可借此机会清理已弃用的 _seq_lens_cpu 参数，避免未来断裂。

功能与动机

PR #40654 为 CommonAttentionMetadata 新增了 seq_lens_cpu_upper_bound 字段，但 MLA 注意力基准测试中的构造调用未同步更新，导致运行 benchmark 时在 mla_attention.py 的 build 方法中触发 assert seq_lens_cpu is not None 断言失败。本 PR 修复该回归问题。

实现拆解

1. 在 benchmarks/attention_benchmarks/mla_runner.py 的 _build_attention_metadata 函数中，为 CommonAttentionMetadata 的构造调用添加 seq_lens_cpu_upper_bound=seq_lens_cpu 参数。
2. 该参数位于 seq_lens 之后、_seq_lens_cpu 之前，与生产环境中其他调用保持一致。
3. 仅新增一行，无其他修改。

关键文件：

- benchmarks/attention_benchmarks/mla_runner.py（模块 基准测试；类别 source；类型 core-logic）：修复核心文件：新增 seq_lens_cpu_upper_bound 参数以匹配 CommonAttentionMetadata 接口变更，修复 benchmark 断言失败。

关键符号：_build_attention_metadata

关键源码片段

benchmarks/attention_benchmarks/mla_runner.py

修复核心文件：新增 seq_lens_cpu_upper_bound 参数以匹配 CommonAttentionMetadata 接口变更，修复 benchmark 断言失败。

```
# 在 _build_attention_metadata 函数中构造 CommonAttentionMetadata 对象
common_attn_metadata = CommonAttentionMetadata(
    ...
    seq_lens=seq_lens_gpu,
```

```
    # 新增：#40654 引入的字段，用于避免在 torch.compile 场景下不必要的 GPU-CPU 同步
    seq_lens_cpu_upper_bound=seq_lens_cpu,
    _seq_lens_cpu=seq_lens_cpu, # 已弃用，但为兼容仍保留
    ...
)
```

评论区精华

机器人审查者 gemini-code-assist[bot] 提出：添加 seq_lens_cpu_upper_bound 是正确的修复，但 _seq_lens_cpu 参数已冗余且已弃用，建议直接替换以提高可维护性，避免未来移除时断裂。但该建议未被采纳，PR 保持同时传递两个参数。njhill 批准了 PR。

- 是否应替换弃用参数 _seq_lens_cpu (design): 未采纳：PR 同时保留了 _seq_lens_cpu 和 seq_lens_cpu_upper_bound，保持最小改动。

风险与影响

- 风险：风险较低：仅影响基准测试文件，不涉及生产路径。但保留了已弃用的 _seq_lens_cpu 参数，当该字段最终被移除时 benchmark 可能会再次报错。
- 影响：影响范围仅限于 MLA 注意力基准测试脚本，不涉及生产代码、用户或系统。修复后相关 benchmark 可正常运行为内部开发提供数据。
- 风险标记：已弃用字段保留

关联脉络

- PR #40654 Add seq_lens_cpu_upper_bound to CommonAttentionMetadata: 本 PR 是 #40654 的配套修复，于生产接口变更后更新基准测试的构造调用。