

PR #33467 完整报告

sgl-project/sglang

[CI] Fix scheduler max new tokens test fixture

合并时间: 2026-08-04 11:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33467>

执行摘要

- 一句话: 修复调度器 max-new-tokens 测试 fixture
- 推荐动作: 值得快速浏览, 了解测试 fixture 如何与生产代码演进保持同步; 可作为小规模回归修复的范例。

功能与动机

PR body 说明: 'Fix the scheduler max-new-tokens test fixture now that #33448 reads `dcp_size` from `server_args`.' 关联 Issue #33448 指出 DCP 下请求长度上限计算使用聚合 KV 池, 'dcp_size scaling matches existing usage a few lines above the Scheduler clamp', 因此 Scheduler 的 max_new_tokens clamp 开始读取 `server_args.dcp_size`, 测试 fixture 未同步导致回归。

实现拆解

1. 背景: #33448 修改了 `python/sglang/srt/managers/scheduler.py` 中 `init_req_max_new_tokens` 的逻辑, 使其从 `server_args.dcp_size` 获取 DCP 并行度, 用于将请求输入长度折算到聚合 KV 池的上限判断中。
2. 修复: 在 `test/registered/unit/managers/test_scheduler_init_req_max_new_tokens.py` 的 `_new_scheduler` 方法中, 为 `Scheduler.__new__` 构造的实例补充 `server_args` 属性, 并设置 `dcp_size=1`, 与默认非 DCP 场景一致。
3. 验证: 作者通过 `/rerun-test` 触发该测试文件的重新运行, github-actions bot 报告 1 个测试通过, 确认 fixture 生效。
4. 影响面: 这是一次纯测试配套改动, 不涉及生产代码、配置文件或部署逻辑, 风险面极小。

关键文件:

- `test/registered/unit/managers/test_scheduler_init_req_max_new_tokens.py` (模块 调度器测试; 类别 test; 类型 test-coverage; 符号 `_new_scheduler`): 修复 Scheduler max-new-tokens 测试 fixture, 补上 `server_args.dcp_size`, 使测试适配 #33448 的改动。

关键符号: `_new_scheduler`

关键源码片段

`test/registered/unit/managers/test_scheduler_init_req_max_new_tokens.py`

修复 Scheduler max-new-tokens 测试 fixture，补上 server_args.dcp_size，使测试适配 #33448 的改动。

```
# test/registered/unit/managers/test_scheduler_init_req_max_new_tokens.py
# 构造最小可用的 Scheduler 实例，用于被测的 init_req_max_new_tokens 逻辑。
# 注意：由于 #33448 之后，Scheduler 会从 server_args.dcp_size 读取 DCP 并行度，
# 这里必须补齐该属性，否则 fixture 会因 AttributeError 失败。
def _new_scheduler(
    self,
    max_req_len: int = 128,
    max_total_num_tokens: int = 1024,
    page_size: int = 1,
) -> Scheduler:
    scheduler = Scheduler.__new__(Scheduler) # 绕过 __init__，手工注入所需字段
    scheduler.max_req_len = max_req_len
    scheduler.max_total_num_tokens = max_total_num_tokens
    scheduler.page_size = page_size
    # 用 SimpleNamespace 模拟 server_args，其中 dcp_size 固定为 1，即默认非 DCP 场景
    scheduler.server_args = SimpleNamespace(dcp_size=1)
    scheduler.max_new_tokens_limit = envs.SGLANG_MAX_NEW_TOKENS_LIMIT.get()
    return scheduler
```

评论区精华

kpham-sgl 在评论中承认 'Oh nice catch. I should have let CI ran full 🙄'，说明 #33448 的 CI 未跑完整导致回归遗漏。随后作者通过 /rerun-test 触发针对该文件的 rerun，github-actions bot 报告 1 个测试通过。

- kpham-sgl 承认 CI 未全跑导致遗漏 (other): PR 作者发现并提交 fixture 修复，评审者批准合并。
- /rerun-test 验证修复 (testing): 修复后的测试通过，确认 fixture 有效。

风险与影响

- 风险：风险极低：仅修改测试 fixture。唯一潜在问题是 dcp_size=1 掩盖了 DCP 场景的覆盖缺口——该测试无法验证 dcp_size > 1 时的 clamp 逻辑，但这也符合该测试只关注非 DCP 调度语义的定位。
- 影响：影响 CI 稳定性和测试可靠性：修复了 #33448 引入的测试失败，使 test_scheduler_init_req_max_new_tokens 恢复通过。对用户无直接影响。
- 风险标记：仅测试文件变动，测试未覆盖 DCP 参数组合

关联脉络

- PR #33448 [DCP] Bound a request by the aggregate KV pool, not one rank's share: 本 PR 修复了 #33448 引入的测试 fixture 回归：Scheduler 从 server_args 读取 dcp_size 后，测试未同步导致失败。