PR #28894 完整报告

sgl-project/sglang

fix(runner): prevent eager token buffer under-allocation

合并时间：2026-06-25 14:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28894

执行摘要

- 一句话：修复 eager 模式 token 缓冲不足的问题
- 推荐动作：建议精读，这是一个典型的最小化修复案例：通过取 max 替代条件分支，安全地修复潜在的 under-allocation 问题。值得关注的是 reviewer 对测试覆盖的提问，后续可考虑补充单元测试。

功能与动机

PR body 明确指出：启用 chunked prefill 后，eager 缓冲容量可能对真实 extend batch 来说太小，导致运行时 tensor copy 形状不匹配。本次变更旨在使用安全的上界来调整 eager token 容量，确保其至少达到全局最大 token 预算。

实现拆解

1. 定位问题逻辑：在 python/sglang/srt/model_executor/runner/eager_runner.py 的 __init__ 方法中，原始代码通过条件判断设置 prefill_ceiling：当 sa.chunked_prefill_size 存在且大于 0 时使用 sa.max_prefill_buffer_tokens()，否则回退到 mr.max_total_num_tokens。这种二选一的逻辑在 chunked prefill 场景下可能低估实际需要的 token 数。
2. 简化并强化上界：将条件表达式改为 max(mr.max_total_num_tokens, sa.max_prefill_buffer_tokens())，无论 chunked prefill 是否启用，都取两者的最大值作为安全上界。这样确保了扩展 batch 场景下 eager 缓冲容量至少与全局最大 token 预算一致，避免 under-allocation。
3. 删除冗余条件：移除了 if sa.chunked_prefill_size and sa.chunked_prefill_size > 0 的分支判断，代码量减少 5 行，逻辑更清晰。
4. 测试配套：本次改动未包含新增测试用例，review 过程中 reviewer mingfeima 询问了测试覆盖情况。

关键文件：

- python/sglang/srt/model_executor/runner/eager_runner.py（模块 运行时；类别 source；类型 data-contract）：包含核心修复逻辑，修改了 prefill_ceiling 的计算方式。

关键符号：init

评论区精华

reviewer mingfeima 在合并前询问：“@jiayisunx is this change covered in current test cases?”这表明团队关注该修复是否已有测试覆盖。目前无进一步回复记录，但 PR 最终被批准合并，推测可能已有隐式测试覆盖或风险较低。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险低：修改为单向强化上界（取 max），只会增加缓冲容量，不会减少，因此不会因容量不足而引发新的 shape 不匹配问题。
  - 性能影响轻微：增加缓冲容量可能略微增加内存占用，但由于取 max 的目标值本身已存在，实际增加量有限。
  - 测试覆盖不足：没有新增针对性测试用例，若后续逻辑变更可能未被及时发现。
- 影响：
  - 用户 / 系统：修复了 chunked prefill 启用时可能出现的 tensor copy 运行时错误，提升稳定性。
  - 团队：代码简化，消除隐含的条件分支，降低后续维护成本。
  - 影响范围：仅影响 eager runner 初始化阶段的缓冲容量计算，不涉及 decode 或其他路径。
  - 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR