PR #21492 完整报告

sgl-project/sglang

Fix benchmark generating empty prompts when random_input_len is small

合并时间：2026-03-27 07:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21492

执行摘要

- 一句话：修复 benchmark 在 random_input_len=1 时生成空提示的 bug，防止服务器错误。
- 推荐动作：这个 PR 是一个简单的 bugfix，变更逻辑直白。对于工程师，如果关心 benchmark 正确性或 CI 稳定性，可以快速浏览以了解问题根源。但整体上，无需深入分析设计决策。

功能与动机

根据 PR body 描述，当 random_input_len=1 时，减去特殊 token 后输入长度可能为 0，导致 tokenizer.decode([]) 返回空字符串，服务器抛出 'ValueError: texts cannot be empty and tokenizer must be initialized' 错误。修复旨在保证输入长度至少为 1，避免此类错误。

实现拆解

实现集中在 python/sglang/benchmark/datasets/random.py 文件的 sample_random_requests 函数中。关键改动是将计算输入长度的行从 input_lens[i] = max(0, input_lens[i] - num_special_tokens) 修改为 input_lens[i] = max(1, input_lens[i] - num_special_tokens)。这确保了在调整特殊 token 后，输入长度至少为 1，从而防止生成空提示。

关键文件：

- python/sglang/benchmark/datasets/random.py（模块 benchmark/datasets）：修复了 sample_random_requests 函数中计算输入长度时的 bug，防止生成空提示。

关键符号：sample_random_requests

评论区精华

review 中仅有一个 bot 评论，总结了变更内容并表明无反馈提供。没有实质性的技术讨论或争议点。结论是变更被接受，无未解决疑虑。

- Bot review of the fix (other)：变更被接受，无争议。

风险与影响

- 风险：风险较低。变更仅影响 benchmark 数据集的输入长度计算，可能导致生成的提示长度分布轻微变化（从可能为 0 变为至少 1）。但考虑到 random_input_len 通常用于测试，

这种变化对实际 benchmark 结果影响有限。此外，修复解决了服务器错误，减少了系统不稳定性。

- 影响：对用户：benchmark 测试将更稳定，不会因空提示而失败；对系统：防止了服务器在处理空输入时抛出异常，提高了健壮性；对团队：解决了 CI 中报告的问题，减少了调试时间。影响范围限于使用 random 数据集进行 benchmark 的场景。
- 风险标记：benchmark 输入长度变化

关联脉络

- PR #21413 Api add flush cache timeout: PR body 中提到该问题是在 PR #21413 的 CI 运行中发现的，因此关联。