PR #7211 完整报告

PaddlePaddle/FastDeploy

[benchmark] update tools

合并时间：2026-04-07 16:25

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7211

执行摘要

- 一句话：新增随机 token_ids 基准测试数据集，支持纯 token 输入的性能评估。
- 推荐动作：该 PR 值得快速浏览，特别是关注 RandomTokenDataset 的实现和 random_flag 的处理逻辑。设计决策简单直接，但需要注意 review 中提到的随机数种子问题是否已修复。对于负责基准测试的工程师，建议检查随机数生成逻辑以确保数据多样性。

功能与动机

PR body 中的 Motivation 部分为空，但从代码变更可以推断，该 PR 旨在丰富基准测试场景，提供纯 token 输入的性能评估能力，避免文本编码对基准测试结果的影响。fastdeploy-bot 在 review 中提及“新增随机 token_ids 生成可以丰富基准测试场景”，验证了这一推断。

实现拆解

实现分为三个关键部分：1) 在 benchmark_dataset.py 中新增 RandomTokenDataset 类，其 sample 方法生成随机 token_ids 列表；2) 在 backend_request_func.py 的 async_request_eb_openai_chat_completions 函数中添加 random_flag 处理逻辑，当启用时直接使用 prompt_token_ids 作为输入；3) 在 benchmark_serving.py 中注册 random_token_ids 数据集选项，并更新命令行帮助信息。

关键文件：

- benchmarks/benchmark_dataset.py（模块 benchmarks）：新增 RandomTokenDataset 类，是核心功能实现所在，但包含随机数种子设置错误的 bug。
- benchmarks/backend_request_func.py（模块 benchmarks）：添加 random_flag 处理逻辑，支持直接使用 prompt_token_ids 作为输入，是功能集成的关键。
- benchmarks/benchmark_serving.py（模块 benchmarks）：注册 random_token_ids 数据集选项，扩展了命令行接口，影响用户使用方式。

关键符号：RandomTokenDataset.sample, async_request_eb_openai_chat_completions

评论区精华

review 中主要讨论集中在代码正确性和文档准确性上。fastdeploy-bot 指出两个问题：1) P0 级别 bug：random.seed(21) 在循环内重复设置，导致所有请求生成相同 token_ids，失去了随机性；2)文档字符串不准确：类描述为“生成随机英文单词”，实际功能是生成随机token_ids。

PR 作者在提交时未回应这些问题，但 reviewer EmmonsCurse 直接批准并跳过 CI，可能意味着问题在后续被接受或忽略。

- 随机数种子设置错误 (correctness): 未在 review 中看到明确修复结论，但 reviewer EmmonsCurse 批准了 PR，可能问题被接受或忽略。
- 文档字符串不准确 (documentation): 建议修复文档字符串，但未在 review 中看到修改确认。

风险与影响

- 风险：主要风险包括：1) 随机数生成逻辑错误（random.seed 在循环内）导致基准测试数据缺乏多样性，可能影响测试结果的代表性；2) 缺少单元测试，新功能未经过自动化测试验证；3) 文档字符串不准确可能误导其他开发者。这些风险集中在 benchmark_dataset.py 文件中，但影响范围仅限于基准测试工具本身，不涉及核心推理路径。
- 影响：对用户的影响：为基准测试提供了新的随机 token_ids 数据集选项，方便用户进行纯 token 输入的性能评估。对系统的影响：仅影响基准测试工具，不改变推理引擎或模型执行逻辑。对团队的影响：丰富了测试场景，但需要关注随机数生成逻辑的正确性以确保测试有效性。
- 风险标记：随机数生成逻辑错误，缺少测试覆盖，文档不准确

关联脉络

- 暂无明显关联 PR