PR #24464 完整报告

sgl-project/sglang

Add --random-input-len to send_one.py

合并时间：2026-05-06 08:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24464

执行摘要

- 一句话：send_one.py 新增随机输入 len 参数
- 推荐动作：该 PR 代码清晰，改动集中，可直接合入。建议其他开发者在 profiling prefill 性能时使用 --random-input-len 参数。

功能与动机

开发者在 profiling 时需要确保 prefill 阶段不被 radix cache 命中，从而获得真实的性能数据。PR body 明确指出："Useful for profiling to ensure the full prefill is captured"。

实现拆解

1. 新增数据字段：在 BenchArgs dataclass 中添加 random_input_len: Optional[int] = None 和 random_input_vocab_size: int = 32768 两个字段。
2. 新增 CLI 参数：在 add_cli_args 中注册 --random-input-len 和 --random-input-vocab-size，提供详细的帮助文本说明其用途。
3. 重构输入构造逻辑：在 send_one_prompt 函数中，当 random_input_len 非 None 时，使用 random.choices 在 [0, vocab_size) 范围内生成随机 token ID：单请求直接生成列表，批量请求时根据 different_prompts 决定是每个请求独立随机还是共享同一个随机序列。
4. 调整请求体构建：将原来始终使用 text 字段的请求改为根据是否有随机输入动态选择 input_ids 或 text 字段。移除了原有的批量 prompt 构造代码，将其整合到非随机输入的分支中。
5. 添加断言：在使用随机输入时，禁止同时使用 image 或 json 模式，避免逻辑冲突。

关键文件：

- python/sglang/test/send_one.py（模块 测试工具；类别 test；类型 test-coverage；符号 BenchArgs, send_one_prompt）：唯一的变更文件，实现了所有新增逻辑：参数定义、随机输入生成、请求体构造。

关键符号：BenchArgs, send_one_prompt

关键源码片段

python/sglang/test/send_one.py

唯一的变更文件，实现了所有新增逻辑：参数定义、随机输入生成、请求体构造。

```
# python/sglang/test/send_one.py（关键新增部分）

@dataclasses.dataclass
class BenchArgs:
    # ... 原有字段 ...
    random_input_len: Optional[int] = None # 随机输入长度，None 表示使用文本 prompt
    random_input_vocab_size: int = 32768 # 随机 token 的 vocab 范围，默认 DeepSeek 常见
    vocab 大小

    @staticmethod
    def add_cli_args(parser: argparse.ArgumentParser):
        # ... 原有参数 ...
        parser.add_argument(
            "--random-input-len",
            type=int,
            default=BenchArgs.random_input_len,
            help="Generate a random prompt of exactly this many tokens (random token IDs). "
            "Each request in the batch gets unique random IDs, avoiding radix cache hits. "
            "Useful for profiling to ensure the full prefill is captured.",
        )
        parser.add_argument(
            "--random-input-vocab-size",
            type=int,
            default=BenchArgs.random_input_vocab_size,
            help="Vocab size for --random-input-len. Token IDs are sampled from "
            "[0, vocab_size). Default: 32768.",
        )

def send_one_prompt(args: BenchArgs):
    base_url = f"http://{args.host}:{args.port}"

    # 构建输入：随机 token 或文本 prompt
    if args.random_input_len is not None:
        n = args.random_input_len
        v = args.random_input_vocab_size
        if args.batch_size == 1:
            # 单请求：直接生成一个随机序列
            input_ids = random.choices(range(v), k=n)
        else:
            if args.different_prompts:
                # 批量且不同 prompt：每个请求独立随机，避免重复命中 cache
                input_ids = [
                    random.choices(range(v), k=n) for _ in range(args.batch_size)
                ]
            else:
                # 批量且相同 prompt：所有请求共享同一个随机序列
                input_ids = [random.choices(range(v), k=n)] * args.batch_size
    else:
        input_ids = None
```

```
        # ... 原有文本 prompt 构建逻辑 ...

    # 构造请求 JSON
    json_data = {
        **({"input_ids": input_ids} if input_ids is not None else {"text": prompt}),
        "image_data": image_data,
        "sampling_params": {
            "temperature": args.temperature,
            "max_new_tokens": args.max_new_tokens,
            "frequency_penalty": args.frequency_penalty,
            "presence_penalty": args.presence_penalty,
        },
        # ... 其他字段 ...
    }
```

评论区精华

该 PR 没有 review 评论或设计讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅影响测试工具 send_one.py，不涉及核心推理路径。新增参数默认值为 None，不会改变现有行为。主要风险在于：如果服务端不支持 input_ids 字段，随机输入可能失败；但该工具本身用于测试，问题容易发现。
- 影响：仅影响使用 send_one.py 进行性能测试的开发者。不会影响生产服务或其他组件。影响程度低。
- 风险标记：仅测试工具变更，低风险

关联脉络

- PR #24296 [Fix] Handle nixlRemoteDisconnectError in NixlKVSender: 同为 send_one.py 所在仓库的测试工具改动，但无直接关联。
- PR #24439 fix(req_pool): bump pool.size to match actual tensor row count after #24243: 同为性能相关修复，但无直接关联。