# PR #25363 完整报告

sgl-project/sglang

benchmark/lora: make number of LoRA adapters configurable

合并时间：2026-05-21 09:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25363

## 执行摘要

- 一句话：LoRA 基准测试适配器数量可配置
- 推荐动作：值得合并，作为 LoRA 路线图的基础设施改进。开发者可参考此模式将其他硬编码基准测试参数化。

## 功能与动机

当前 LoRA 基准测试辅助脚本硬编码 NUM_LORAS = 4，使得测试不同适配器数量的多 LoRA 服务变得困难。此 PR 是 Q2 LoRA 路线图的一部分，旨在支持多 LoRA 推理的基准测试和优化。

## 实现拆解

1. benchmark/lora/launch_server.py：将硬编码的 NUM_LORAS 和 LORA_PATH 替换为命令行参数 --base-model-path、--lora-path 和 --num-loras，并相应调整 launch_server 函数。
2. benchmark/lora/lora_bench.py：移除对 launch_server 中常量的导入，添加相同的默认常量，并使用 args.num_loras 随机选择 LoRA 适配器。
3. 数据记录增强：在两个脚本的 JSON 输出中添加 base_model_path、base_only 和 num_loras 字段。

关键文件：

- benchmark/lora/launch_server.py（模块 基准测试；类别 source；类型 core-logic；符号 launch_server）：核心启动脚本，将硬编码的 LoRA 数量和路径改为可配置参数。
- benchmark/lora/lora_bench.py（模块 基准测试；类别 source；类型 dependency-wiring；符号 async_request_openai_completions, benchmark, run_benchmark）：基准测试客户端，移除对启动器常量的导入，改用命令行参数。

关键符号：launch_server, async_request_openai_completions, benchmark, run_benchmark

## 关键源码片段

### benchmark/lora/launch_server.py

核心启动脚本，将硬编码的 LoRA 数量和路径改为可配置参数。

```
import argparse
import os

# 从硬编码模块常量改为默认值常量
DEFAULT_BASE_MODEL_PATH = "meta-llama/Llama-2-7b-hf"
DEFAULT_LORA_PATH = "winddude/wizardLM-LlaMA-LoRA-7B"
DEFAULT_NUM_LORAS = 4

def launch_server(args):
    # 从命令行参数读取路径
    base_path = args.base_model_path
    lora_path = args.lora_path

    if args.base_only:
        cmd = f"python3 -m sglang.launch_server --model-path {base_path} "
    else:
        # 非 base-only 模式时校验 num_loras > 0
        if args.num_loras <= 0:
            raise ValueError(
                "--num-loras must be greater than 0 unless --base-only is set"
            )
        cmd = f"python3 -m sglang.launch_server --model-path {base_path} --lora-paths "
        for i in range(args.num_loras):
            lora_name = f"lora{i}"
            cmd += f"{lora_name}={lora_path} "
    cmd += f"--disable-radix "
    # ... 其余参数拼接逻辑保持不变
    print(cmd)
    os.system(cmd)

# 在 argparse 中添加三个新参数
if __name__ == "__main__":
    parser = argparse.ArgumentParser()
    parser.add_argument("--base-model-path", type=str, default=DEFAULT_BASE_MODEL_PATH,
                        help="Base model path or Hugging Face model ID.")
    parser.add_argument("--lora-path", type=str, default=DEFAULT_LORA_PATH,
                        help="LoRA adapter path or Hugging Face model ID used for all registered "
                             "LoRA adapters.")
    parser.add_argument("--num-loras", type=int, default=DEFAULT_NUM_LORAS,
                        help="Number of LoRA adapters to register. For example, 4 registers "
                             "lora0, lora1, lora2, and lora3.")
    # ... 原有参数保持不变
```

## benchmark/lora/lora_bench.py

基准测试客户端，移除对启动器常量的导入，改用命令行参数。

```
# 移除 from launch_server import LORA_PATH, NUM_LORAS 导入
# 改为本地定义默认值
DEFAULT_BASE_MODEL_PATH = "meta-llama/Llama-2-7b-hf"
```

```
DEFAULT_NUM_LORAS = 4

async def async_request_openai_completions(request_func_input, pbar=None):
    # ... 使用 args.num_loras 替代全局常量
    if args.base_only:
        payload = {"text": prompt, "sampling_params": {"max_new_tokens": request_func_input.
        output_len}}
    else:
        payload = {
            "text": prompt,
            "sampling_params": {"max_new_tokens": request_func_input.output_len},
            "lora_path": f"lora{random.randint(0, args.num_loras - 1)}",
        }
    # ... 其余保持不变

def run_benchmark(args_):
    # ... 在解析参数后添加校验
    if not args.base_only and args.num_loras <= 0:
        raise ValueError("--num-loras must be greater than 0 unless --base-only is set")
    # 使用 args.base_model_path 替代原先的 LORA_PATH["base"]
    model_id = args.base_model_path
    tokenizer_id = args.base_model_path

if __name__ == "__main__":
    parser = ArgumentParser(description="Benchmark the online lora serving throughput.")
    parser.add_argument("--base-model-path", type=str, default=DEFAULT_BASE_MODEL_PATH,
                        help="Base model path or Hugging Face model ID.")
    parser.add_argument("--num-loras", type=int, default=DEFAULT_NUM_LORAS,
                        help="Number of LoRA adapters used by the benchmark. Must match the "
                             "server launcher.")
    # ... 原参数
```

# 评论区精华

PR 获得 jybsuper 的快速批准，无实质性讨论。

- 暂无高价值评论线程

# 风险与影响

- 风险：低风险。仅修改基准测试辅助脚本，不涉及核心推理逻辑。可能出现的问题包括：用户使用旧参数脚本会导致错误，但通过保留默认值和清晰的帮助信息可缓解。
- 影响：影响范围小，仅限于 LoRA 基准测试工作流。用户现在可以通过命令行灵活指定基础模型路径、LoRA 适配器路径和适配器数量，便于多 LoRA 推理性能研究。
- 风险标记：基准测试工具变更

# 关联脉络

- PR #25095 [Roadmap] Lora (2026 Q2): 此 PR 是 Q2 LoRA 路线图的一部分，旨在实现多 LoRA 推理基准测试。