PR #22292 完整报告

sgl-project/sglang

[CI] Fix stage-b-test-1-gpu-large (0) timeout by reordering LoRA tests and using tokenizer from cache

合并时间：2026-04-08 11:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22292

执行摘要

本 PR 通过重命名测试方法以调整执行顺序，并优化 tokenizer 加载逻辑，修复了 CI 中 stage-b-test-1-gpu-large (0) 测试频繁超时的问题。变更集中于测试文件 test_bench_serving_1gpu_part1.py 和工具函数 run_bench_serving，预计可节省约 5 分钟测试时间，提升 CI 稳定性和效率，对生产代码无直接影响。

功能与动机

stage-b-test-1-gpu-large (0) 测试在 CI 中频繁超时，根本原因在于：

- unittest 按字母顺序执行测试方法，导致 test_lora_* 在 test_offline_* 之前运行。
- LoRA 测试跳过 CI_OFFLINE 模式，首次服务器启动时 tokenizer 加载会调用 Hugging Face Hub API（如 list_repo_files），在冷启动时阻塞数分钟。
- 即使模型已本地缓存，基准测试客户端仍使用仓库 ID 加载 tokenizer，可能触发额外 API 调用。 PR body 引用 CI 日志数据，显示 test_lora_online_latency 服务器启动耗时 6 分 49 秒，而 test_offline_throughput_default 仅 46 秒，重排顺序后可将冷启动时间从约 7 分钟降至 2 分钟。

实现拆解

1. 测试重命名：在 test/registered/perf/test_bench_serving_1gpu_part1.py 中，将 test_lora_online_latency 和 test_lora_online_latency_with_concurrent_adapter_updates 重命名为 test_online_lora_latency 和 test_online_lora_latency_with_concurrent_adapter_updates，确保按字母顺序时 test_offline_* 优先执行，预热 HF 缓存。
2. Tokenizer 路径优化：在 python/sglang/test/test_utils.py 的 run_bench_serving() 函数中添加以下逻辑：python bench_tokenizer = tokenizer if bench_tokenizer is None: try: from sglang.srt.utils import find_local_repo_dir local_dir = find_local_repo_dir(model, revision=None) if local_dir and os.path.isdir(local_dir): bench_tokenizer = local_dir except Exception: pass 当本地缓存可用时，将 tokenizer 解析为本地路径，避免基准测试客户端调用 Hub API。

评论区精华

Review 中仅 Fridge003 批准，无评论内容，表明变更直接且无争议。提交历史显示作者从最初延长超时时间（提交 52485f7）转向根本原因修复（提交 75ac42b 和 60cd126），体现了

从临时缓解到系统性解决的决策过程。

风险与影响

- 风险：测试重命名可能影响依赖名称的脚本或文档；本地路径解析依赖 find_local_repo_dir 函数，需确保其可用性和异常处理稳健。
- 影响：显著减少 CI 超时风险，提升测试效率，对用户无直接影响，但团队需注意测试重命名可能带来的轻微混淆。

关联脉络

与近期 CI 优化 PR 如 #22301（优化 coredump 上传）、#22297（撤销测试模型更新）和 #22284（添加快速失败机制）同属提升 CI 稳定性的努力。本 PR 专注于解决外部依赖延迟问题，通过调整测试顺序和资源加载策略，为类似场景提供了参考模式。