

PR #28617 完整报告

sgl-project/sglang

Fix bench serving base-url-only runs

合并时间: 2026-06-18 15:34

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28617>

执行摘要

- 一句话: 修复 bench_serving 仅传 --base-url 时的崩溃
- 推荐动作: 该 PR 是小型回归修复, 值得关注其体现的设计原则: 将相似逻辑集中到共享函数中以避免重复和条件分支。值得快速合并。

功能与动机

PR #28598 重构了基准测试客户端的 base-URL 解析逻辑, 但未处理 --base-url-only 的使用场景, 导致 bench_serving 在缺少 --port 参数时崩溃。作者在 PR body 中引用了 CI 失败日志表明问题。

实现拆解

1. 在 python/sglang/srt/utils/network.py 中新增 resolve_host_port() 函数, 与 resolve_base_url() 类似但不添加 http:// 前缀, 直接返回 host:port 格式 (通过 NetworkAddress.to_host_port_str() 实现 IPv6 兼容)。
2. 在 python/sglang/bench_serving.py 中, 将 import 语句从导入 NetworkAddress, resolve_base_url 改为导入 resolve_base_url, resolve_host_port。
3. 在 run_benchmark 函数的 gserver 分支中, 将 api_url = args.base_url if args.base_url else _na.to_host_port_str() 替换为 api_url = resolve_host_port(args.base_url, args.host, args.port), 从而复用共享逻辑。
4. 移除不再需要的 _na = NetworkAddress(args.host, args.port) 临时变量。
5. 更新了相关注释以反映 gserver 使用无 scheme 的 host:port 形式。

关键文件:

- python/sglang/srt/utils/network.py (模块 网络层; 类别 source; 类型 core-logic; 符号 resolve_host_port): 新增了 resolve_host_port 函数, 与 resolve_base_url 对称, 用于返回无 scheme 的 host:port 格式。
- python/sglang/bench_serving.py (模块 基准测试; 类别 source; 类型 dependency-wiring): 修改了 import 语句和 gserver 分支的 URL 构建逻辑, 解决了崩溃问题。

关键符号: resolve_host_port

评论区精华

该 PR 没有 review 评论。第二个 commit 的提交信息 (`use resolve_host_port for gserver; drop conditional_na`) 表明 hnyils2002 参与了协作，将条件分支重构为统一的 `resolve_host_port` 调用。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅限于 gserver 后端的 URL 构造逻辑，其他后端不受影响。新增的 `resolve_host_port` 函数与 `resolve_base_url` 结构相同，仅在输出格式上不同，逻辑清晰且易于验证。没有测试覆盖但回归面窄，CI 测试已通过。
- 影响：只影响使用 `--base-url` 且 backend 为 gserver 的场景，修复了此类场景下的崩溃。对其他用户无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #28598 [misc] Share bench HTTP-client base-URL resolution with IPv6-compatible formatting: 该 PR 是 #28598 引入的回归修复，直接关联相同文件和功能逻辑。