PR #21156 完整报告

sgl-project/sglang

[Fix][Eval] Keep `--dataset-path` scoped to `longbench_v2`

合并时间：2026-03-24 17:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21156

执行摘要

本 PR 修复了 run_eval.py 中 gpqa 评估因错误使用 --dataset-path 参数导致的 FileNotFoundError。通过恢复硬编码数据集 URL 并明确参数仅作用于 longbench_v2，解决了运行错误，但移除了 gpqa 的自定义路径能力，建议关注后续是否需恢复此功能以避免功能回归。

功能与动机

PR 旨在修复由 #20469 引入的 bug。--dataset-path 参数原设计仅用于 longbench_v2 评估（默认值为 THUDM/LongBench-v2），但变更后 gpqa 意外读取该参数，导致尝试打开默认数据集作为本地 CSV 路径而失败。如 PR body 所述："--dataset-path appears to be meant for longbench_v2 only"，修复后恢复 gpqa 的先前行为，保持参数作用域正确。

实现拆解

变更集中在 python/sglang/test/run_eval.py 文件的 run_eval 函数中：

变更点	原代码	新代码	影响
gpqa 评估	filename = getattr(args, "dataset_path", None) or ("https://...")	filename = ("https://...")	移除对 dataset_path 的依赖，硬编码数据集 URL
longbench_v2 评估	data_source = getattr(args, "dataset_path", None)	data_source = args.dataset_path	确保参数被直接使用，作用域限定

此实现简单直接，但牺牲了 gpqa 的配置灵活性。

评论区精华

review 中，gemini-code-assist[bot] 指出关键问题：

> "This change fixes a bug ... However, it also removes the ability to specify a custom dataset path for gpqa ... This is a functional regression."

并建议长期解决方案（如添加 --gpqa-dataset-path 参数）或临时修复（检查默认值）。Fridge003 批准了 PR，表明可能接受当前方案，但讨论未解决回归问题，揭示了参数作用域设计的重要性。

风险与影响

- 风险：功能回归——gpqa 评估无法指定自定义数据集路径，影响测试灵活性；参数作用域混淆可能引发未来类似 bug。
- 影响：用户层面，默认配置下 gpqa 评估能正常运行，但自定义路径用户需寻找替代方案；系统层面，仅测试脚本变更，核心功能不受影响；团队层面，提示需优化参数设计以避免交叉污染。

关联脉络

从同仓库历史 PR 看，本 PR 与 #21276（回滚测试脚本变更）和 #21195（修复测试 bug）类似，均涉及测试评估模块的 bugfix 主题。虽然 PR body 提及 #20469，但未在提供的近期列表中出现，推测为早期变更。整体上，这反映了测试脚本维护中的常见问题：参数作用域管理需谨慎，以避免意外行为。建议未来 PR 考虑更模块化的参数设计，如为不同评估类型添加专用参数。