

PR #28280 完整报告

sgl-project/sglang

Allow overriding tokenizer path in benchmark harness

合并时间: 2026-06-16 04:07

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28280>

执行摘要

- 一句话: 支持 benchmark 命令行指定本地 tokenizer 路径
- 推荐动作: 本 PR 为实验性小功能增强, 适合快速合并。建议后续考虑对 vllm 后端也支持该选项, 以保持一致性。

功能与动机

在对外部 SGLang 服务器进行基准测试时, tokenizer 路径默认从 /server_info 接口获取。但某些场景下 (如 server_info 不可信、使用本地定制 tokenizer、或离线测试), 需要手动指定 tokenizer 路径。该 PR 通过新增 CLI 选项满足此需求, 确保 benchmark 的准确性和可复现性。

实现拆解

1. 新增字段: 在 python/sglang/test/bench_one_batch_server_internal.py 的 BenchArgs dataclass 中添加 local_tokenizer_path: str = "" 字段。
2. 新增 CLI 参数: 在 add_cli_args 方法中注册 --local-tokenizer-path 参数, 类型为字符串, 默认值为空。
3. 修改 tokenizer 加载逻辑: 在 run_benchmark_internal 函数中, 获取 tokenizer 路径的代码块之前新增判断: 若 bench_args.local_tokenizer_path 非空, 则直接使用; 否则回退到原逻辑 (优先使用 /server_info 中的 tokenizer_path, 其次解析 prefill 信息)。

关键文件:

- python/sglang/test/bench_one_batch_server_internal.py (模块 基准测试; 类别 test; 类型 test-coverage; 符号 BenchArgs.local_tokenizer_path, BenchArgs.add_cli_args, run_benchmark_internal): 单文件变更, 新增 --local-tokenizer-path 选项并修改 tokenizer 加载逻辑。

关键符号: BenchArgs.add_cli_args, run_benchmark_internal

关键源码片段

`python/sglang/test/bench_one_batch_server_internal.py`

单文件变更, 新增 --local-tokenizer-path 选项并修改 tokenizer 加载逻辑。

位于 BenchArgs dataclass 中新增字段

```

@dataclasses.dataclass
class BenchArgs:
    base_url: str = ""
    local_tokenizer_path: str = "" # 新增字段，默认为空字符串

# 在 add_cli_args 中注册参数（简化展示）
@staticmethod
def add_cli_args(parser: argparse.ArgumentParser):
    parser.add_argument("--base-url", type=str, default=BenchArgs.base_url)
    parser.add_argument(
        "--local-tokenizer-path",
        type=str,
        default=BenchArgs.local_tokenizer_path,
        help=(
            "Local tokenizer path to use when benchmarking an external "
            "SGLang server via --base-url. Defaults to the tokenizer path "
            "reported by /server_info."
        ),
    )

# 在 run_benchmark_internal 中修改 tokenizer 路径获取逻辑
# 原代码：if "tokenizer_path" in server_info:
# 新代码：优先使用用户指定的路径
if bench_args.local_tokenizer_path:
    tokenizer_path = bench_args.local_tokenizer_path
elif "tokenizer_path" in server_info:
    tokenizer_path = server_info["tokenizer_path"]
elif "prefill" in server_info:
    tokenizer_path = server_info["prefill"][0]["tokenizer_path"]

```

评论区精华

Review 中 `gemini-code-assist[bot]` 评论指出：当前 `--local-tokenizer-path` 仅在 `sglang` 后端生效，当 `backend` 为 `vllm` 时该选项被静默忽略，建议扩展支持 `vllm` 后端。该建议未被采纳或进一步回应，可能因为该 PR 主要针对 `SGLang` 场景，但确实是一个值得跟踪的遗留问题。

- `vllm` 后端也应支持 `--local-tokenizer-path (design)`：未被采纳或回应，可能因 PR 针对 `SGLang` 场景；但理论上可追踪作为后续改进。

风险与影响

- 风险：改动仅限 `benchmark` 测试脚本，不涉及生产代码或核心推理路径，风险极低。唯一潜在风险是：若用户误传无效路径，代码会继续尝试从 `/server_info` 回退，行为与未传入时一致，不会崩溃或产生误导性结果。
- 影响：影响范围仅限于单 `batch benchmark` 脚本，对外部 `SGLang` 服务器的测试用户有利。不会影响正常推理服务、其他测试或现有 `CI` 流程。改动小，无兼容性问题。
- 风险标记：测试专用变更

关联脉络

- 暂无明显关联 PR