

PR #30372 完整报告

sgl-project/sglang

[fix] Fix two trunk test regressions due to flexkv change (#29701)

合并时间: 2026-07-07 16:58

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30372>

执行摘要

- 一句话: 修复 FlexKV 引入的两项 trunk 测试回归
- 推荐动作: 建议快速合入以恢复 trunk CI。该 PR 体现了对代码规范（使用 `runtime_context` 而非遗留全局访问器）和测试隔离的重视，值得所有贡献者参考。

功能与动机

PR #29701 合入后，`main` 分支上 `base-a-test-cpu` 的两个单元测试出现回归。

`test_legacy_global_ratchet` 发现 `get_global_server_args()` 调用次数达到 281 次，超过基线 280，而新的代码应使用 `sglang.srt.runtime_context` 访问器；

`test_fallback_to_radix_cache` 由于 Mock 的 `server_args` 具有真值 `enable_flexkv / flexkv_config_file`，进入了 FlexKV 分支并因 `os.environ["FLEXKV_CONFIG_PATH"] = <MagicMock>` 抛出 `TypeError`。此 PR 旨在修复这两个问题，使 trunk CI 变绿。

实现拆解

1. 替换 FlexKV 中已废弃的 `get_global_server_args()` 调用 - 文件：
`python/sglang/srt/mem_cache/storage/flexkv/flexkv_radix_cache.py` - 在 `cache_finished_req` 方法中将 `from sglang.srt.server_args import get_global_server_args` 和 `get_global_server_args()` 分别改为 `from sglang.srt.runtime_context import get_server_args` 和 `get_server_args()`。此举消除了仅剩的遗留访问器调用，使 `test_legacy_global_ratchet` 的计数恢复至基线 280。
2. 修正测试 Mock 配置，避免进入 FlexKV 分支 - 文件: `test/registered/unit/mem_cache/test_registry.py` - 在 `_make_ctx` 辅助函数中，为 `server_args` Mock 显式设置 `enable_flexkv = False`。这样 `default_radix_cache_factory` 在遇到 MagicMock 的 `flexkv_config_file` 时不会进入 FlexKV 分支，从而防止 `os.environ` 赋值崩溃，使得测试正常走 radix cache 回退路径。

关键文件:

- `python/sglang/srt/mem_cache/storage/flexkv/flexkv_radix_cache.py`（模块 缓存层；类别 source；类型 dependency-wiring）：修复 FlexKV 中遗留的 `get_global_server_args()` 调用，替换为 `runtime_context.get_server_args()`，解决了 `test_legacy_global_ratchet` 回归。

- test/registered/unit/mem_cache/test_registry.py (模块测试; 类别 test; 类型 test-coverage) : 在 _make_ctx 中设置 enable_flexkv=False, 避免 Mock 的 server_args 进入 FlexKV 分支导致崩溃, 修复 test_fallback_to_radix_cache。

关键符号: 未识别

关键源码片段

python/sglang/srt/mem_cache/storage/flexkv/flexkv_radix_cache.py

修复 FlexKV 中遗留的 get_global_server_args() 调用, 替换为 runtime_context.get_server_args(), 解决了 test_legacy_global_ratchet 回归。

```
# flexkv_radix_cache.py 中 cache_finished_req 方法的片段
# 原代码使用已废弃的 get_global_server_args(), 导致 legacy 访问器计数器超标。
# 改为从 runtime_context 获取 server_args, 符合最新代码规范。

def cache_finished_req(self, req: Req, is_insert: bool = True) -> None:
    super().cache_finished_req(req, is_insert=is_insert)
    if not is_insert:
        self._load_markers.pop(req.rid, None)
    return

# 使用 runtime_context 获取 server_args, 而非 legacy 的 get_global_server_args()
from sglang.srt.runtime_context import get_server_args
global_server_args = get_server_args()
topk = global_server_args.speculative_eagle_topk
# ... 后续逻辑不变
```

test/registered/unit/mem_cache/test_registry.py

在 _make_ctx 中设置 enable_flexkv=False, 避免 Mock 的 server_args 进入 FlexKV 分支导致崩溃, 修复 test_fallback_to_radix_cache。

```
# test_registry.py 中 _make_ctx 辅助函数片段
# 新增 enable_flexkv=False 以阻止测试走进 FlexKV 分支 (该分支需要真实配置)
```

```
def _make_ctx(
    *,
    backend=None,
    enable_streaming=False,
    enable_lmcache=False,
    # ... 其他参数
):
    server_args = MagicMock()
    server_args.radix_cache_backend = backend
    server_args.enable_streaming_session = enable_streaming
    server_args.enable_lmcache = enable_lmcache
    server_args.enable_flexkv = False # 新增: 确保不进入 FlexKV 分支
    # ... 返回 TreeCacheBuildContext
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 仅涉及两行源码变更和一行测试配置变更，无模型、内核或正向路径修改。风险极低。但需注意 `get_server_args()` 与 `get_global_server_args()` 的行为一致性，依赖 `runtime_context` 模块的按需初始化；在首次调用前若未正确初始化可能返回 `None`。测试中强制关闭 `enable_flexkv` 可能掩盖 FlexKV 分支的实际测试覆盖，长期需补充相关 mock 测试。
- 影响：
 - 影响范围：仅影响两个 trunk 单元测试（`test_legacy_global_ratchet` 和 `test_fallback_to_radix_cache`）。
 - 用户影响：无。
 - 系统影响：恢复 trunk CI 稳定性，确保开发者可以在 main 分支上运行这些测试。
 - 团队影响：低，合入后可立即解除 CI 阻塞。
 - 风险标记：测试配置 Mock 覆盖不全

关联脉络

- PR #29701 Feat/flexkv main connector: 本 PR 修复了 #29701 引入的两项 trunk 测试回归，是 #29701 的跟进修复。