

PR #30631 完整报告

sgl-project/sglang

[fix] Repoint the prefetch-dispatch test at the loader's current config binding

合并时间: 2026-07-09 17:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30631>

执行摘要

- 一句话: 修复 prefetch 测试因导入变化导致的回归
- 推荐动作: 值得合并。这是一个简单但必要的回归修复, 确保测试基础设施正确同步, 避免 CI 噪音。

功能与动机

PR #30493 合并后, `sglang.srt.model_loader.loader` 模块不再导出 `get_global_server_args`, 而是通过 `runtime_context.get_server_args` 获取。这导致 #30146 中新增的 `TestPrefetchDispatch` 测试用例在 main 分支上失败, 因为测试代码中的 `patch` 目标 `get_global_server_args` 已不存在。

实现拆解

一个单行变更: 在 `test_prefetch_checkpoints.py` 的 `_patch_dispatch` 方法中, 将 `patch` 目标从 `"sglang.srt.model_loader.loader.get_global_server_args"` 改为 `"sglang.srt.model_loader.loader.get_server_args"`, 以匹配当前模块级别的绑定。由于 `dispatch` 代码现在读取的是 `loader.get_server_args`, 拦截仍然有效。

关键文件:

- `test/registered/unit/model_loader/test_prefetch_checkpoints.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 将 `_patch_dispatch` 中的 `patch` 目标从 `get_global_server_args` 更新为 `get_server_args`, 以匹配 `loader` 模块当前导出的符号。

关键符号: 未识别

关键源码片段

`test/registered/unit/model_loader/test_prefetch_checkpoints.py`

将 `_patch_dispatch` 中的 `patch` 目标从 `get_global_server_args` 更新为 `get_server_args`, 以匹配 `loader` 模块当前导出的符号。

```
# test/registered/unit/model_loader/test_prefetch_checkpoints.py
# ... (class TestPrefetchDispatch)
```

```
def _patch_dispatch(self, prefetch, disable_mmap=False):
    return (
```

```

patch.object(
    DefaultModelLoader,
    "_prepare_weights",
    return_value=("/dummy", ["f.safetensors"], True),
),
# 修复: 将 patch 目标从 "get_global_server_args" (已移除)
# 更新为 "get_server_args", 匹配 loader 模块当前导出的符号
# 这样拦截器仍能正确注入 server args
patch(
    "sglang.srt.model_loader.loader.get_server_args",
    return_value=self._server_args(prefetch, disable_mmap),
),
patch(
    "sglang.srt.model_loader.loader."
    "buffered_multi_thread_safetensors_weights_iterator",
    return_value=iter([]),
),
patch(
    "sglang.srt.model_loader.loader.safetensors_weights_iterator",
    return_value=iter([]),
),
patch("sglang.srt.model_loader.loader.logger.warning"),
)

```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更仅为一字符串替换, 且保证了 patch 目标与实际导入一致。CI 确认 13/13 测试通过 (修复前为 7/13)。
- 影响: 直接影响: 修复 main 分支上 6 个失败的单元测试。间接影响: 确保后续对 prefetch dispatch 逻辑的测试覆盖有效, 防止类似回归。
- 风险标记: 测试基础设施回归

关联脉络

- PR #30146 Disable multi-threaded load by default when prefetch is on: 引入 TestPrefetchDispatch 测试类, 其中使用了旧的 patch 目标。
- PR #30493 (Unknown, inferred from context) Flipped import to runtime_context.get_server_args: 该 PR 改变了配置获取方式, 移除了 get_global_server_args 导出, 导致测试失败。