

PR #30154 完整报告

sgl-project/sglang

[fix] Reconcile the legacy-getter ratchet baseline after racing merges

合并时间: 2026-07-05 22:06

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30154>

执行摘要

- 一句话: 修复 `get_global_server_args` 棘轮基线竞态合并
- 推荐动作: 可快速合并以解除 CI 阻塞, 无需深入审查。

功能与动机

配置解析系列 PR (#30137) 将 `get_global_server_args` 的精确基线定在 278, 而 #28787 (ROCm RMSNorm 批次不变性修复) 在 `layernorm.py` 中新增了两个调用点 (`rl_on_policy_target` 读取)。这两个 PR 各自在其合并基上通过测试, 但合并后的实际调用点数为 280, 导致 `test_legacy_global_ratchet` 在 main 分支及所有 PR 的合并提交上失败。本 PR 旨在修复这一 CI 故障。

实现拆解

1. 文件 `test/registered/unit/test_legacy_global_ratchet.py` 中, 将 `_RATCHETS` 列表中 `get_global_server_args` 的基线值从 278 更新为 280。
2. 该变更仅涉及一行数值调整, 无其他逻辑修改。

关键文件:

- `test/registered/unit/test_legacy_global_ratchet.py` (模块 棘轮测试; 类别 test; 类型 test-coverage) : 更新 `get_global_server_args` 调用点基线值从 278 到 280, 以反映合并后实际计数。

关键符号: 未识别

关键源码片段

`test/registered/unit/test_legacy_global_ratchet.py`

更新 `get_global_server_args` 调用点基线值从 278 到 280, 以反映合并后实际计数。

```
# 文件 : test/registered/unit/test_legacy_global_ratchet.py
```

```
"""Ratchet guard: legacy global-accessor call-sites may only decrease.
```

```
The process-wide ``ServerArgs`` is owned by the runtime context; the legacy  
``get_global_server_args`` / ``set_global_server_args_for_*`` names survive as  
thin shims for the existing call-sites. New code should use the
```

```

``sglang.srt.runtime_context`` accessors (``get_server_args()`` /
``get_context().set_server_args()``), so the shim call-site counts below must
never grow. When your change removes call-sites, lower the matching baseline
to the new count.
"""

```

```

from sglang.test.ci.ci_register import register_cpu_ci

```

```

register_cpu_ci(est_time=5, suite="base-a-test-cpu")

```

```

import re
import unittest
from pathlib import Path

```

```

import sglang.srt
from sglang.test.test_utils import CustomTestCase

```

```

_SRT_ROOT = Path(next(iter(sglang.srt.__path__)))

```

```

# Baselines counted over python/sglang/srt/**/*.*.py, including each function's
# own def line. Ratchet: decrease-only.

```

```

_RATCHETS = [
    # 更新后的基线值: 从 278 增加为 280, 因为 #30137 和 #28787 合并后新增了两个调用点
    ("get_global_server_args", r"\bget_global_server_args\s*\(", 280),
    (
        "set_global_server_args_for_",
        r"\bset_global_server_args_for_(?:scheduler|tokenizer)\s*\(",
        5,
    ),
]

```

```

class TestLegacyGlobalRatchet(CustomTestCase):
    def test_legacy_accessor_call_sites_match_the_baselines(self):
        # Exact pin, failing in BOTH directions: a grown count means new code
        # bypassed the runtime_context accessors; a shrunk count means a
        # removal forgot to lower the baseline, which would let later changes
        # silently re-add call-sites up to the stale ceiling.
        sources = [
            path.read_text(encoding="utf-8", errors="replace")
            for path in sorted(_SRT_ROOT.rglob("*.py"))
        ]
        for name, pattern, baseline in _RATCHETS:
            count = sum(len(re.findall(pattern, source)) for source in sources)
            if count > baseline:
                self.fail(
                    f"{name} call-sites grew: {count} > baseline {baseline}. "
                    "New code must use the sglang.srt.runtime_context accessors "
                    "(get_server_args() / get_context().set_server_args())."
                )

```

```
)
if count < baseline:
    self.fail(
        f"{name} call-sites shrank: {count} < baseline {baseline}. "
        "Lower the baseline in this file to lock in the progress."
    )
```

评论区精华

无实质性技术讨论。mmangkad 直接批准并合并以解除 CI 阻塞。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅调整测试基线数值，无生产逻辑变更。若未来有其他 PR 减少调用点但未更新基线，此调整可能使棘轮效果延迟，但测试本身仍会在调用点减少时报错。
- 影响：影响范围仅限于修复 CI 中 legacy global accessor 棘轮测试的失败，无用户或系统功能影响。
- 风险标记：低风险

关联脉络

- PR #30137 [refactor] Config resolution pipeline: full-stack review (10-PR series, review only): 该 PR 固定了 get_global_server_args 基线为 278，是本次变更的直接原因之一。
- PR #28787 [AMD] Fix RMSNorm batch-invariance on ROCm under deterministic inference: 该 PR 在 layernorm.py 中新增了两个 get_global_server_args 调用点，导致基线增加 2。