

PR #2834 完整报告

radixark/miles

fix(session): use incremental R3 for all non-retract cases, and fix asserts

合并时间: 2026-09-02 03:44

原文链接: <http://prhub.com.cn/radixark/miles/pull/2834>

执行摘要

- 一句话: 修复 session R3 增量逻辑, 让 abort 模式正常工作并规范 partial_rollout 校验
- 推荐动作: 此 PR 是一个高质量、范围明确的 bug 修复。它解决了一个导致特定有效配置 (session-server + abort) 无法运行的关键问题, 逻辑清晰, 测试覆盖全面。其核心改动 (server.py 中一行, arguments.py 中几行) 虽小, 但影响的是会话管理与路由重放的核心路径。对于参与 session 管理、rollout 或参数系统开发的工程师, 建议精读此 PR, 尤其是 use_addition_r3 的派生逻辑变更和 retract 警告的考量。

功能与动机

根据 PR body, 问题源于 Shi Dong 报告 `partial_rollout` 是真正不兼容的 session-server 模式, 而非之前误判的 `abort`。SessionServer 在 `abort` 模式下错误地返回了完整的 R3 数据, 但参数验证却拒绝了这个在 disaggregated serving 中有效的组合。核心目标是修复这一逻辑错误, 使 `abort` 模式在 session-server 下正常工作, 并正确拒绝不支持的 `partial_rollout`。

实现拆解

1. 核心逻辑调整 (`miles/rollout/session/server.py`): 将 SessionServer.__init__ 中 use_addition_r3 的派生逻辑从 `getattr(...) == "in_place"` 修改为 `args.pause_generation_mode != "retract"`。这意味着只要不是 retract 模式 (包括 abort 和 in_place), 都使用增量 R3。此修改是修复的核心, 确保了 abort 模式行为的正确性。
2. 参数校验逻辑重构 (`miles/utils/arguments.py`):
 - 移除了之前针对 abort 的断言 (`--use-session-server is incompatible with --pause-generation-mode=abort`)。
 - 新增了针对所有 session-server 版本 (v1, v2) 的 `--partial-rollout` 禁用断言, 替代了之前仅在 v2 检查列表中的逻辑。
 - 新增了当同时使用 `--use-session-server`、`--use-rollout-routing-replay` 和 `--pause-generation-mode=retract` 时的警告日志, 明确指出 retract 模式下 R3 payload 可能变大且存在 SGLang 已知问题。
3. 测试配套与覆盖扩展 (多个测试文件):
 - 在 `tests/fast/utils/test_arguments.py` 中: 将 `test_session_server_rejects_abort` 改为参数化的 `test_session_server_accepts_abort` (覆盖 colocated 和非 colocated), 新

增 `test_session_server_rejects_partial_rollout` (覆盖所有 `session-server` 版本) 和 `test_retract_r3_warning`。

- 在 `tests/fast/router/test_sessions.py` 和 `tests/fast/router/test_sessions_v2.py` 中: 将仅针对 `in_place` 的偏移测试改为参数化覆盖 `abort` 和 `in_place` 两种增量模式 (`test_incremental_offsets_across_turns_and_rollback` 和 `test_incremental_offsets_across_turns_and_branches`) , 并为测试夹具补充了必要的 `pause_generation_mode` 参数。
- 在其他测试辅助文件 (`session_pretokenized_test_utils.py`, `test_session_race_conditions.py`, `test_session_anthropic.py`, `session_parity_utils.py`) 中补充了 `pause_generation_mode="retract"` 参数, 确保 `SessionServer` 初始化不会因缺少该属性而失败。

4. 文档更新 (<docs/user-guide/agent-ic-rollout.md>) : 更新了关于 `session-server` 与 `--pause-generation-mode` 兼容性的说明, 准确反映 `abort` 模式现在已被支持, 并对 `retract` 模式的行为和警告进行了描述。

关键文件:

- `miles/rollout/session/server.py` (模块 会话服务; 类别 `source`; 类型 `core-logic`; 符号 `SessionServer.init`) : 核心逻辑修改点, 将 `use_addition_r3` 的派生条件从仅 `in_place` 改为所有非 `retract` 模式, 这是修复 `abort` 模式行为的根本所在。
- `miles/utils/arguments.py` (模块 参数校验; 类别 `source`; 类型 `core-logic`; 符号 `miles_validate_args`) : 参数验证逻辑的核心变更, 移除了错误的 `abort` 校验, 新增了正确的 `partial_rollout` 校验和 `retract` 警告。
- `tests/fast/utills/test_arguments.py` (模块 参数校验; 类别 `test`; 类型 `test-coverage`; 符号 `test_session_server_accepts_abort`, `test_session_server_rejects_partial_rollout`, `test_retract_r3_warning`) : 关键的测试文件, 全面覆盖了本次修复涉及的参数校验逻辑, 包括接受 `abort`、拒绝 `partial_rollout` 和 `retract` 警告。

关键符号: `SessionServer.init`, `miles_validate_args`

关键源码片段

`miles/rollout/session/server.py`

核心逻辑修改点, 将 `use_addition_r3` 的派生条件从仅 `in_place` 改为所有非 `retract` 模式, 这是修复 `abort` 模式行为的根本所在。

```
# miles/rollout/session/server.py
```

```
class SessionServer:
```

```
    """Lightweight FastAPI server that manages sessions and proxies inference
    requests through the inference router (sglang or miles)."""
```

```
    def __init__(self, args, backend_url: str):
```

```
        # ... 其他初始化代码 ...
```

```
        # `retract` may recompute earlier rows and must return full R3; all other
```

```
# pause modes preserve prior rows and can request only the appended R3.
# 修复点：将条件从 `== "in_place"` 改为 `!= "retract"`，使 abort 和 in_place 都使用增量 R3。
self.use_addition_r3 = args.pause_generation_mode != "retract"
setup_session_routes(self.app, self, args, use_addition_r3=self.use_addition_r3)
```

miles/utils/arguments.py

参数验证逻辑的核心变更，移除了错误的 abort 校验，新增了正确的 partial_rollout 校验和 retract 警告。

```
# miles/utils/arguments.py (在 miles_validate_args 函数中)
```

```
# 1. 新增全局断言：session-server 不支持 partial-rollout。
# 之前的检查仅在 v2 版本中，现在对所有版本生效。
assert not (
    args.use_session_server and args.partial_rollout
), "--use-session-server does not support --partial-rollout"

# 2. 移除了之前针对 abort 的断言：
# assert not (
#     args.use_session_server and args.pause_generation_mode == "abort"
# ), "--use-session-server is incompatible with --pause-generation-mode=abort"

# 3. 新增警告：当 session-server + routing-replay + retract 模式同时启用时，提示风险。
if args.use_session_server and args.use_rollout_routing_replay and args.pause_generation_mode == "retract":
    logger.warning(
        "--use-session-server with --use-rollout-routing-replay and "
        "--pause-generation-mode=retract returns full R3 data on every turn; "
        "R3 payloads can become very large. TODO: Retract-mode weight updates R3 "
        "have known issues in SGLang and need to be fixed."
    )
```

评论区精华

1. 核心逻辑审查：claude[bot] 在其评审中验证了从 getattr 到直接属性访问的改动（server.py:46）的安全性，确认 --pause-generation-mode 是一个带有默认值 "retract" 的 argparse 标志，因此直接访问不会引发 AttributeError。这消除了潜在的回归风险。
 2. 决策与结论：主要的讨论点和决策已在 PR 中体现并被采纳：确认 partial_rollout 是与 session-server 不兼容的行为，而非 abort；将增量 R3 逻辑泛化；为 retract 模式添加显式警告。作者 guapisolo 通过后续的提交（2975c57, 420acf3）进一步澄清了警告文本和校验范围。
 3. 已解决 / 未解决的疑虑：所有讨论的疑虑均已通过代码改动和测试覆盖得到解决。retract 模式下 R3 payload 过大和 SGLang 已知问题作为已知限制通过警告日志明确告知用户，这是一个有意识的设计权衡，而非未解决的 bug。
- getattr 到直接属性访问的安全性审查 (correctness): 已验证 --pause-generation-mode 是 argparse 定义的带默认值的标志，属性始终存在，此改动安全。

风险与影响

- 风险:

1. 参数校验逻辑变更风险: miles/utils/arguments.py 中校验逻辑的修改 (移除 abort 断言, 新增全局 partial_rollout 断言) 改变了启动时的验证行为。如果存在依赖旧行为 (例如, 依赖 abort 被拒绝来避免特定配置) 的外部脚本或部署流程, 可能会意外开始运行 abort 模式。
2. 核心路径影响: use_addition_r3 的计算逻辑是 SessionServer 处理 R3 数据的核心开关, 直接影响 setup_session_routes 及其下游的路由重放行为。错误的计算可能导致 R3 数据不一致或 OOM。
3. 测试夹具显式化: 此次修复强制所有创建 SessionServer 的测试夹具必须显式提供 pause_generation_mode 参数。虽然这是积极的清理, 但可能暴露了之前未注意的、依赖默认值的测试用例的潜在问题 (本次已全部修复)。

- 影响:

1. 用户影响: 对于使用 --use-session-server 和 --pause-generation-mode=abort 的用户, 此次修复是关键性的, 它使得在 colocated 和 disaggregated serving 两种拓扑下 abort 模式都能正常工作。对于使用 retract 模式的用户, 新增的日志警告提供了重要的运维信息。
2. 系统影响: 影响 session-server (v1 和 v2) 的 R3 数据组装逻辑。修复确保了在非 retract 模式下, 系统只请求和处理增量 R3, 这对于 disaggregated serving 下的网络和内存效率至关重要。
3. 团队影响: 扩展的测试覆盖 (特别是参数化测试) 提高了 session R3 和参数校验逻辑的健壮性和可维护性, 减少了未来回归的风险。- 风险标记: 核心路径变更, 参数校验行为变更, 已知功能限制 (retract 模式)

关联脉络

- PR #2774 feat(session): configure session server workers explicitly: 同属 session-server 功能演进, 涉及 session-server 启动参数的配置优化 (port/workers 解耦), 与本 PR 共同完善 session-server 的配置与校验体系。
- PR #2760 feat(tito): support Qwen3.8 27B and Flash-Next templates: 同属 session 与 rollout 功能线。TITO 模板变更可能影响会话消息的渲染与序列化, 与 R3 数据处理属于不同的子系统但共享会话上下文。
- PR #2794 dashboard: scroll the token strip instead of paging through it: 虽为 dashboard 变更, 但其涉及 session 产生的 token 数据 (可能与 R3 相关) 的展示优化, 反映了对会话数据流向的持续关注。
- PR #2817 dashboard: give a step with no samples the summary schema: 属于 observability 修复, 确保在会话数据处理异常 (如空 step) 时前端不崩溃, 与本 PR 保证核心会话逻辑正确性在系统稳定性上目标一致。