

PR #31837 完整报告

sgl-project/sglang

Fix MiniMax-M3 crash on ROCm by making its override fields resolvable

合并时间: 2026-07-21 07:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31837>

执行摘要

- 一句话: 修复 MiniMax-M3 在 ROCm 上因字段不可解析导致的启动崩溃
- 推荐动作: 推荐快速合并。这是一个针对特定硬件 (ROCm) 和模型 (MiniMax-M3) 的启动崩溃修复, 变更量小、风险低、已验证。开发者在为模型添加覆盖声明时, 应注意检查所声明的字段是否已标记 `resolvable=True`。

功能与动机

在 ROCm (MI355X) 上启动 MiniMax-M3 时, 如果未传递 `--enable-aiter-allreduce-fusion`, 启动器会在引擎启动前崩溃, 报错 `'disable_custom_all_reduce not model-oververidable'`。该问题源于 `_minimax_m3_overrides` 声明的字段未被标记为 `resolvable=True`, 导致 `validate_declarations` 拒绝。此外, 即使传递该标志规避默认路径崩溃, 标准 EP 路径仍会因 `enable_aiter_allreduce_fusion` 声明而崩溃。详见 PR body。

实现拆解

1. 修改 `server_args.py` 中 `disable_custom_all_reduce` 字段定义: 将简单的 `bool` 描述改为 `Arg(help="...", resolvable=True)` 包装, 使该字段可在模型覆盖中合法声明。
2. 修改 `server_args.py` 中 `enable_aiter_allreduce_fusion` 字段定义: 添加 `Arg(help="...", resolvable=True)` 包装, 同样使其可被模型覆盖声明。
3. 更新测试白名单: 在 `test/registered/unit/test_model_overrides.py` 的 `test_server_args_whitelist_is_exactly_the_migrated_fields` 测试中, 向白名单集合添加 `"disable_custom_all_reduce"` 和 `"enable_aiter_allreduce_fusion"`, 确保覆盖字段与白名单一致。
4. 验证: 在 MI355X 上以 `tp8 mxfp8` 配置启动, 无需 `--enable-aiter-allreduce-fusion` 即可正常启动并提供服务。

关键文件:

- `python/sglang/srt/server_args.py` (模块 启动参数; 类别 `source`; 类型 `core-logic`; 符号 `disable_custom_all_reduce`, `enable_aiter_allreduce_fusion`): 核心修复文件: 为 `disable_custom_all_reduce` 和 `enable_aiter_allreduce_fusion` 两个字段的 `Arg` 添加 `resolvable=True`, 使其可在模型覆盖中合法声明。
- `test/registered/unit/test_model_overrides.py` (模块 覆盖注册; 类别 `test`; 类型 `test-coverage`; 符号 `test_server_args_whitelist_is_exactly_the_migrated_fields`): 测试

配套：将新增的可解析字段加入白名单集合，确保与 server_args.py 变更一致，避免 CI 失败。

关键符号：未识别

关键源码片段

python/sglang/srt/server_args.py

核心修复文件：为 disable_custom_all_reduce 和 enable_aiter_allreduce_fusion 两个字段的 Arg 添加 resolvable=True，使其可在模型覆盖中合法声明。

```
# 修改前 (base)
disable_custom_all_reduce: A[
    bool,
    "Disable the custom all-reduce kernel and fall back to NCCL.",
] = False

# 修改后 (head)：使用 Arg 包装并设置 resolvable=True
# 这样模型覆盖（如 _minimax_m3_overrides）可以合法声明此字段
disable_custom_all_reduce: A[
    bool,
    Arg(
        help="Disable the custom all-reduce kernel and fall back to NCCL.",
        resolvable=True, # 允许模型覆盖声明此字段
    ),
] = False

# 类似地，enable_aiter_allreduce_fusion 也添加了 resolvable=True
enable_aiter_allreduce_fusion: A[
    bool,
    Arg(help="Enable Aiter AllReduce Fusion.", resolvable=True),
] = False
```

test/registered/unit/test_model_overrides.py

测试配套：将新增的可解析字段加入白名单集合，确保与 server_args.py 变更一致，避免 CI 失败。

```
# 在 frozenset 白名单末尾添加两个新字段
frozenset(
    {
        ...
        "flashinfer_allreduce_fusion_backend",
        "fp8_gemm_runner_backend",
        "disable_custom_all_reduce", # 新增
        "enable_aiter_allreduce_fusion", # 新增
    }
),
```

评论区精华

该 PR 审核人为 ch-wan，审核状态为 APPROVED，无评论和 Review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险低。仅修改了两个字段的元数据（添加 `resolvable=True`）和测试白名单，未改变默认值或逻辑行为。潜在风险：若其他模型覆盖也依赖这两个字段但尚未添加 `resolvable`，理论上均会触发类似崩溃，但本次修复已覆盖最常见场景。测试白名单同步更新，确保 CI 检测通过。
- 影响：影响范围小。仅影响 ROCm 平台 MiniMax-M3 模型的启动流程。修复后，用户无需传递 `--enable-aiter-allreduce-fusion` 即可正常启动，包括标准 EP 路径。对其他模型或平台无影响。
- 风险标记：暂无

关联脉络

- PR #31819 docs(cookbook): revert MiniMax-M3 to dev image (model not yet in a release): 同样涉及 MiniMax-M3 模型，前者处理文档回退，本 PR 修复启动崩溃，可视为同一模型的支持流程的后续修复。