

PR #2873 完整报告

radixark/miles

fix: configure Qwen3.8 Anthropic session verification

合并时间: 2026-09-04 02:16

原文链接: <http://prhub.com.cn/radixark/miles/pull/2873>

执行摘要

- 一句话: 修复 Qwen3.8 Anthropic 会话验证测试配置, 解除 CI 阻塞。
- 推荐动作: 这是一个小而精确的测试配置修复, 旨在解除 CI 阻塞。尽管代码改动极少 (仅增加 4 行), 但它解决了一个明确的阻塞性问题, 并且讨论中清晰地阐明了每个配置项的作用和必要性。对于关注测试基础设施、会话验证或 Qwen3.8 模型集成的工程师, 值得快速浏览以了解配置细节。对于其他开发者, 则无需精读。

功能与动机

驱动本次修复的直接原因是 CI 流水线失败。根据 PR body, `stage-c-2-gpu-h200 (0) / run` 在运行 `test_qwen38small.py` 时, 因触发了 `ValueError: Anthropic per-model verification requires an intermediate-system expectation` 而失败。这表明测试的 `run_both_versions` 函数默认启用了 Anthropic 验证路径, 但对应的 `ModelConfig` 没有提供此路径所必需的中间系统期望配置。修复此问题是为了让 Qwen3.8 模型的多轮会话验证在 Anthropic 端点上能正常运行, 从而修复 CI。

实现拆解

1. 变更入口: `tests/e2e/sglang/test_session_server_multi_role/test_qwen38small.py` 文件中的 `CONFIG` 字典。这是配置端到端测试运行参数的地方。
2. 核心配置改动: 在 `CONFIG` 中新增了两个参数:
 - `anthropic_intermediate_system_expectation="forbidden"`: 明确声明该模型的固定模板不支持在对话中插入中间系统消息, 满足验证器启动的契约要求。
 - `anthropic_assistant_text_threshold=1.0`: 将 Anthropic 端点的助手文本软阈值从默认的 0.2 调整为 1.0。这是因为 Anthropic 端点在工具调用场景下会修改助手消息的序列化格式 (例如 `

与的区别), 这是一种可预期的、端点特定的格式漂移。设置为 1.0 意味着允许 100% 的助手文本差异通过软门控, 而不会影响其他所有硬不匹配门控 (如 `special_token_count`, `non_assistant_text` 等))。`

1. 验证: 提交信息和评论中描述了修复后的验证结果: 测试在 Anthropic 通道上成功运行, 32 个样本全部通过验证, 所有硬不匹配指标为零, 助手文本比例符合预期。这证明了配置的正确性。

关键文件:

- tests/e2e/sglang/test_session_server_multi_role/test_qwen38small.py（模块 会话验证；类别 test；类型 test-coverage）：本次变更的唯一文件。它包含了触发 Anthropic 验证路径的测试配置，修复的核心就在此文件的 CONFIG 字典中。

关键符号：未识别

关键源码片段

tests/e2e/sglang/test_session_server_multi_role/test_qwen38small.py

本次变更的唯一文件。它包含了触发 Anthropic 验证路径的测试配置，修复的核心就在此文件的 CONFIG 字典中。

变更前 CONFIG（核心部分）

```
CONFIG = ModelConfig(  
    model_name="Qwen/Qwen3.8-27B-FP8",  
    reasoning_parser="qwen3",  
    tool_call_parser="qwen3_coder",  
    tito_model="qwen38small",  
    num_gpus=2,  
    kv_cache_dtype="fp8_e4m3",  
    mamba_full_memory_ratio=4.59,  
    cycles=2,  
    tool_call_failure_mode="append_tool",  
)
```

变更后 CONFIG（核心部分）

```
CONFIG = ModelConfig(  
    model_name="Qwen/Qwen3.8-27B-FP8",  
    reasoning_parser="qwen3",  
    tool_call_parser="qwen3_coder",  
    tito_model="qwen38small",  
    num_gpus=2,  
    kv_cache_dtype="fp8_e4m3",  
    mamba_full_memory_ratio=4.59,  
    cycles=2,  
    tool_call_failure_mode="append_tool",  
    # Anthropic tool-call conversion changes raw assistant serialization;  
    # keep this endpoint-only formatting mismatch soft while hard gates stay at 0.  
    anthropic_assistant_text_threshold=1.0,  
    anthropic_intermediate_system_expectation="forbidden",  
)
```

解释：

- # 1. `anthropic\_intermediate\_system\_expectation="forbidden"`:
 - # - 告知验证器，该模型的固定模板不允许在对话轮次间插入系统消息。
 - # - 这是运行 Anthropic 验证路径所必需的启动条件，缺失会导致 ValueError。
- # 2. `anthropic\_assistant\_text\_threshold=1.0`:
 - # - 针对 Anthropic 端点单独设置助手文本的软匹配阈值为 100%。
 - # - 因为 Anthropic 在工具调用时会改变助手消息的原始序列化格式（如注释所述）。

- 这是一个端点特定的配置，不影响全局的硬不匹配门控。

评论区精华

核心讨论集中在配置项是否充分和必要。作者 [guapisolo](#) 在合并后 ([评论](#)) 提供了更详细的 CI 重跑证据，关键结论是：

“On a rerun with only `anthropic_intermediate_system_expectation="forbidden"` applied... The run failed at the default 0.2 soft threshold, while this PR's endpoint-only `anthropic_assistant_text_threshold=1.0` covers the observed normalization without weakening any hard gate.” 这明确证实了仅设置 `forbidden` 不足以通过测试，必须同时将 `anthropic_assistant_text_threshold` 设为 1.0 才能覆盖特定的序列化格式差异。此讨论澄清了为什么需要两个配置项，并强调了调整是“端点特定”的，没有削弱全局的硬不匹配检测。

- 配置的充分性与必要性 (correctness): 通过一次仅应用单个参数的 CI 重跑实验（失败）和应用两个参数的重跑实验（成功）得出结论：两个参数缺一不可。必须同时设置 `forbidden` 以满足启动契约，并将软阈值设为 1.0 以覆盖特定的序列化格式差异。

风险与影响

- 风险：
 1. 测试有效性风险：将 `anthropic_assistant_text_threshold` 设为 1.0 意味着完全容忍 Anthropic 端点的助手文本格式差异。虽然这是基于已知差异做出的有意设计，但需确保此软门控不会意外掩盖新引入的、真正的助手文本解析错误。目前看来风险较低，因为验证通过了所有硬不匹配门控。
 2. 变更范围风险：所有改动严格限制在单个端到端测试文件的配置字典内，不涉及任何生产代码（如会话服务器、验证器核心逻辑、解析器）。因此，对生产系统或用户会话无直接影响，回归风险极小。
- 影响：
 1. 对 CI/CD 的影响：直接修复了 `test_qwen38small.py` 测试在 Anthropic 通道上的失败，恢复了该测试在 CI 流水线中的正常运行，避免了对 `run-ci-sglang` 流程的阻塞。
 2. 对测试覆盖的影响：确保了 Qwen3.8 模型在 Anthropic 多轮会话场景下的验证覆盖得以恢复，避免了该重要模型的一个验证路径因配置问题而被静默跳过或失败。
 3. 对团队的影响：为后续配置类似模型（具有固定模板且不支持中间系统消息）的 Anthropic 会话验证提供了一个清晰、已验证的配置范例。团队其他成员在添加新模型测试时可参考此模式。 - 风险标记：测试软阈值覆盖潜在漂移

关联脉络

- PR #2711 `test(e2e): verify multi-turn Anthropic sessions`: 引入了 Anthropic 多轮会话的验证框架和测试模式，是本次修复所使用的 `run_both_versions` 函数和 Anthropic 验证逻辑的直接来源。

- PR #2760 feat(tito): support Qwen3.8 27B and Flash-Next templates: 为 Qwen3.8 系列添加了模型支持和固定模板, 本次修复的 test_qwen38small.py 测试正是针对该模型新增的验证用例。