PR #2791 完整报告

radixark/miles

test(tito): use FP8 Qwen3.6 checkpoint

合并时间：2026-09-01 04:13

原文链接：http://prhub.com.cn/radixark/miles/pull/2791

执行摘要

- 一句话：Qwen3.6 会话测试改用 FP8 检查点，验证目标不变
- 推荐动作：该 PR 不值得精读，可快速合并。值得注意的设计原则是“验证目标优先于 artifact 选择”——在 CI 面上用更轻量的检查点保持等价覆盖，而非削弱验证范围。若想理解 TITO Qwen3.6 支持的全貌，建议阅读上游 PR #2759（引入 tito_tokenizer、qwen3.6_fixed.jinja 模板与 test_qwen36.py）。

功能与动机

PR body 明确指出：当前 Qwen3.6 session verifier 加载的是全精度 35B 检查点，而这条 CI 表面验证的是 TITO 聊天序列化与会话复用（原文 "validates TITO chat serialization and session reuse"）。改用匹配的 FP8 检查点可以在保持验证目标的同时避免拉取全精度模型 artifact，降低 CI 的存储、下载与加载成本。Review Focus 要求确认 FP8 检查点仍保留 TITO 所依赖的 tokenizer 与 chat-template 行为。

实现拆解

1. 变更入口与唯一 diff：修改 tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py 中 CONFIG 定义的 model_name 字段，从 Qwen/Qwen3.6-35B-A3B 改为 Qwen/Qwen3.6-35B-A3B-FP8，全 PR 仅此一行改动（+1/-1）。
2. 行为保持：reasoning_parser="qwen3"、tool_call_parser="qwen3_coder"、tito_model="qwen36"、tp_size=2、enable_spec=True、cycles=2、tool_call_failure_mode="append_tool" 全部原样保留；register_cuda_ci 与 register_rocm_ci 两个硬件通道注册，以及 rollout/tito_session_mismatch_rate/v1|v2/assistant_text 两个 metric 门禁均未改动，确保检查点替换不削弱 v1/v2 会话与工具调用覆盖。
3. 测试与验证配套：CI 机器人确认 tests.ci.file_run 将该文件解析到 CUDA 套件 stage-c-4-gpu-h200（runner 标签 h200,4gpu，超时 1800 秒）；实际 GPU 运行在 18m15s 内通过，两个 session-mismatch 门禁均未触发。提交者在本 PR 内用 uvx pre-commit run --files ... 验证了 lint。
4. 其他配套：无源码、配置、文档或部署改动；ROCm 通道仍注册在 nightly 套件 nightly-stage-c-4-gpu-mi350，预计同样受益于更小的 artifact。

关键文件：

- tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py（模块 会话验证；类别 test；类型 test-coverage）：唯一变更文件：把 Qwen3.6 TITO 会话验证器使用的检查点从全精度 35B 换成官方 FP8 版本，同时保持 CUDA/ROCm 通道、四卡拓扑、投机解码、解析器与两份 session-mismatch metric 门禁不变。PR body 的 Review Focus 与 claude[bot] 审查均集中于此文件。

关键符号：test_qwen36, run_both_versions

关键源码片段

tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py

唯一变更文件：把 Qwen3.6 TITO 会话验证器使用的检查点从全精度 35B 换成官方 FP8 版本，同时保持 CUDA/ROCm 通道、四卡拓扑、投机解码、解析器与两份 session-mismatch metric 门禁不变。PR body 的 Review Focus 与 claude[bot] 审查均集中于此文件。

```
# Qwen3.6 TITO 会话验证用例：验证多轮 chat 序列化与会话复用。
# 本文件将检查点从全精度 35B 换成官方 FP8 版本，以减小 CI 加载大权重 artifact 的开销，
# 同时保留既有验证面：CUDA/ROCm 通道、4 卡拓扑、投机解码与解析器均不变。
from tests.ci.ci_register import register_cuda_ci, register_rocm_ci
from tests.ci.metric_history import register_ci_gate
from tests.e2e.sglang.test_session_server_multi_role._common import ModelConfig, run_both_
versions

# CUDA 侧跑在 stage-c-4-gpu-h200，ROCm 侧走 nightly mi350 通道
register_cuda_ci(est_time=800, suite="stage-c-4-gpu-h200", labels=["sglang"])
register_rocm_ci(est_time=500, suite="nightly-stage-c-4-gpu-mi350", labels=["sglang"])
# 两份 metric gate 分别守住 v1/v2 会话不匹配率，检查点替换不得削弱该覆盖
register_ci_gate(metric_key="rollout/tito_session_mismatch_rate/v1/assistant_text")
register_ci_gate(metric_key="rollout/tito_session_mismatch_rate/v2/assistant_text")


CONFIG = ModelConfig(
    # FP8 与全精度共享同一 tokenizer 与 chat template，因此 TITO 序列化行为保持一致；
    # 这是本次变更唯一修改的字段
    model_name="Qwen/Qwen3.6-35B-A3B-FP8",
    reasoning_parser="qwen3",
    tool_call_parser="qwen3_coder",
    tito_model="qwen36",
    tp_size=2,
    enable_spec=True, # 投机解码开关保持不变
    cycles=2, # 多轮会话轮数保持不变
    tool_call_failure_mode="append_tool", # 工具调用失败时继续追加工具消息
)


def test_qwen36():
    # 同一 CONFIG 同时驱动 v1/v2 两个 session 协议版本的验证
    run_both_versions(CONFIG)
```

```
if __name__ == "__main__":
    test_qwen36()
```

评论区精华

整个 review 流程非常轻量：claude[bot] 先按仓库配置提示手动审查模式（@claude review / @claude review always），随后 Shi-Dong 评论 @claude review always 触发 bot 复查；claude[bot] 复查后返回 "Code review found no issues / No high-confidence issues detected in this change"，Shi-Dong 直接 APPROVED（"LGTM!"）。PR body 的 Review Focus 提出唯一需要确认的点——FP8 检查点必须保留 TITO 所需的 tokenizer 与 chat-template 行为，且不得削弱 v1/v2 会话与工具调用覆盖——该疑虑由实际 CI 运行和保持不变的 metric 门禁消除，全程无技术争议。

- FP8 检查点是否保持 TITO 验证行为 (question): FP8 与全精度共享同一 tokenizer/chat template，TITO 序列化行为不变；实际 GPU 运行（stage-c-4-gpu-h200，18m15s）通过，两个 session-mismatch metric 门禁未触发，疑虑消除。

风险与影响

- 风险：
  1. 外部 artifact 依赖：Qwen/Qwen3.6-35B-A3B-FP8 是 HuggingFace 上的外部检查点，若其命名、标签或可用性变化，CI 会在模型加载阶段失败，这是相对全精度版本新增的供应链依赖风险。
  2. 测试有效性边界：FP8 推理与全精度存在数值差异，理论上可能改变模型输出；但该测试面验证的是 TITO 序列化与会话复用，这些行为由 tokenizer 与 chat template 决定而非权重精度，且两个 metric 门禁与实测均通过，风险很低。
  3. 影响范围：仅测试文件改动，不触及 miles/backends、miles/rollout 等生产代码，无业务回归面，也无性能与安全影响。 - 影响：对用户无任何线上功能影响。对 CI 系统而言，该 e2e 用例不再拉取全精度 35B 权重，存储、下载与加载开销下降，ROCm nightly 通道同样受益；对团队而言，验证覆盖保持完整（v1/v2 会话与工具调用），并为后续在 e2e 验证面上使用更轻量 artifact 提供了先例。整体影响程度低，局限于一条测试配置。 - 风险标记：外部检查点依赖，测试有效性依赖 FP8 行为

关联脉络

- PR #2759 feat(tito): support Qwen3.5 and Qwen3.6 templates: 本 PR 修改的 test_qwen36.py 正是 #2759 引入的；#2759 落地了 tito_tokenizer、qwen3.6_fixed.jinja 模板与 TITO 契约，本 PR 是对该验证面的轻量优化。
- PR #2711 test(e2e): verify multi-turn Anthropic sessions: 同属 session server 多角色 e2e 验证基础设施，均依赖 sglang 契约门禁；两者共同构成 session/TITO 契约的 CI 防线。