PR #1921 完整报告

THUDM/slime

Add example for streaming output

合并时间：2026-05-19 10:13

原文链接：http://prhub.com.cn/THUDM/slime/pull/1921

执行摘要

- 一句话：新增流式生成示例，支持 abort 时保留部分状态
- 推荐动作：建议阅读 slime/rollout/sglang_streaming_rollout.py 中基于 base snapshot + chunk delta 的增量状态累加设计，该模式在需要中断恢复的场景下值得推广。

功能与动机

每块 SSE 数据直接写入 sample，中断时已积累部分状态，不依赖 abort 请求返回的收集文本。

实现拆解

1. 创建 slime/rollout/sglang_streaming_rollout.py，定义异步函数 generate_streaming：复用 GenerateState 和 _prepare_prompt_ids，构建 stream=True 的 HTTP 请求，通过 aiter_lines 循环解析 SSE 行，增量累积 call_tokens、call_log_probs、text，最后合并到 sample。
2. 创建 tests/test_qwen3_4B_streaming_partial_rollout.py 集成测试：设置 --custom-generate-function-path 指向新函数，配置 --over-sampling-batch-size > --rollout-batch-size 并启用 --partial-rollout，确保每个 rollout 步骤触发 abort 以检验流式恢复。
3. 在 CI 配置（pr-test.yml 和 pr-test.yml.j2）的测试矩阵中新增该测试条目，分配 8 GPU 自动执行。

关键文件：

- slime/rollout/sglang_streaming_rollout.py（模块 流式生成；类别 source；类型 core-logic；符号 generate_streaming）：核心流式生成实现，定义 generate_streaming 异步函数，替代非流式 generate。
- tests/test_qwen3_4B_streaming_partial_rollout.py（模块 流式测试；类别 test；类型 test-coverage；符号 prepare, execute）：CI 集成测试，验证 streaming + partial rollout 组合在 abort 场景下的正确性和奖励信号。
- .github/workflows/pr-test.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CI 矩阵添加新测试项，保障新功能在 CI 中自动运行。
- .github/workflows/pr-test.yml.j2（模块 CI 配置；类别 infra；类型 infrastructure）：CI 模板增加新测试项入口，保持与生成配置同步。

关键符号：generate_streaming, prepare, execute

关键源码片段

tests/test_qwen3_4B_streaming_partial_rollout.py

CI 集成测试，验证 streaming + partial rollout 组合在 abort 场景下的正确性和奖励信号。

```
"""CI smoke test for the streaming sglang rollout path.

Wires slime.rollout.sglang_streaming_rollout.generate_streaming as the
per-sample generate function, with --over-sampling-batch-size >
--rollout-batch-size and --partial-rollout enabled so the rollout
loop must abort in-flight requests every step — exercising the streaming
abort path (partial state should already be on the sample when the SSE is
cut, then the partial groups get recycled into the data buffer).

Uses Qwen3-4B so responses on dapo-math are long enough to actually
trigger mid-stream aborts.
"""

import os

import slime.utils.external_utils.command_utils as U

TIGHT_HOST_MEMORY = U.get_bool_env_var("SLIME_TEST_TIGHT_HOST_MEMORY", "1")

MODEL_NAME = "Qwen3-4B"
MODEL_TYPE = "qwen3-4B"
NUM_GPUS = 8


def prepare():
    # 准备模型权重、数据集和 checkpoint 转换
    U.exec_command("mkdir -p /root/models /root/datasets")
    U.exec_command(f"hf download Qwen/{MODEL_NAME} --local-dir /root/models/{MODEL_
    NAME}")
    U.hf_download_dataset("zhuzilin/dapo-math-17k")
    U.convert_checkpoint(model_name=MODEL_NAME, megatron_model_type=MODEL_TYPE,
    num_gpus_per_node=NUM_GPUS)


def execute():
    ckpt_args = f"--hf-checkpoint /root/models/{MODEL_NAME}/ " f"--ref-load /root/{MODEL_
    NAME}_torch_dist "

    rollout_args = (
        # 使用流式生成函数作为每个样本的 generate 函数
        "--custom-generate-function-path slime.rollout.sglang_streaming_rollout.generate_
        streaming "
```

```
    "--prompt-data /root/datasets/dapo-math-17k/dapo-math-17k.jsonl "
    "--input-key prompt "
    "--label-key label "
    "--apply-chat-template "
    "--rollout-shuffle "
    "--rm-type deepscaler "
    "--num-rollout 2 "
    "--rollout-batch-size 4 "
    # over-sampling 2x 确保半数 in-flight 组必须 abort，partial-rollout 回收它们
    "--over-sampling-batch-size 8 "
    "--partial-rollout "
    "--mask-offpolicy-in-partial-rollout "
    "--n-samples-per-prompt 4 "
    "--rollout-max-response-len 4096 "
    "--rollout-temperature 0.8 "
    "--global-batch-size 16 "
    "--balance-data "
)

perf_args = (
    "--tensor-model-parallel-size 2 "
    "--sequence-parallel "
    "--pipeline-model-parallel-size 1 "
    "--context-parallel-size 2 "
    "--recompute-granularity full "
    "--recompute-method uniform "
    "--recompute-num-layers 1 "
    "--use-dynamic-batch-size "
    f"--max-tokens-per-gpu {2048 if TIGHT_HOST_MEMORY else 8192} "
)

# ... 其他参数（优化器、sglang、CI 等）拼接后执行
# 完整配置见测试源文件
```

评论区精华

本案无审阅讨论，由 author zhuzilin 独立提交并合并到主分支。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 新函数依赖 sglang SSE 流的累积输出格式，若未来 sglang 切换到增量模式 (--incremental-streaming-output) 或修改 JSON 结构，该函数需适配。
  2. 测试使用 Qwen3-4B 和 8 GPU，资源密集，CI 稳定性依赖环境和模型权重下载。
  3. 该函数目前不是默认 generate 路径，仅通过 --custom-generate-function-path 显式启用，风险隔离。 - 影响：对现有用户无直接影响，需手动配置才启用。系统新增可选流式

生成路径，无额外资源占用。为团队提供了 abort 场景下更可靠的样本状态管理示例，未来可能成为部分 rollout 的推荐配置。 - 风险标记：新模块缺少单元测试，依赖 sglangSSE 累积语义

关联脉络

- PR #1920 Move fully_async example to main codebase: 同为 rollout 路径新增示例架构，提升 abort/partial 场景的灵活性。