

PR #2106 完整报告

THUDM/slime

feat(examples/strands_sglang): update to strands-sglang 0.4.2

合并时间: 2026-06-30 12:13

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2106>

执行摘要

- 一句话: 更新 strands_sglang 示例至 0.4.2 API
- 推荐动作: 建议使用 strands-sglang 集成或参考该示例的用户仔细阅读代码变更, 特别是 model.rollout 的用法和内联解释器的实现。该 PR 展示了减少外部依赖、内联关键组件的设计思路。

功能与动机

strands-sglang 0.4.2 移除了 model.token_manager, 改用 model.rollout 跟踪器。此外, 为使示例更轻量自主, 需要移除 camel-ai 依赖并内联子进程解释器。

实现拆解

实现步骤:

1. 创建子进程解释器— 新增 examples/strands_sglang/subprocess_interpreter.py, 从 camel-ai 项目移植并内联依赖, 移除对 camel-ai 的导入。
2. 修改生成入口— 更新 generate_with_strands.py, 将导入从 camel.interpreters 改为本地 SubprocessInterpreter; 移除 MAX_TOOL_CALLS 参数 (ToolLimiter 不再支持); 将 model.token_manager 替换为 model.rollout, 使用 initial_prompt_length、token_ids、loss_mask、logprobs。
3. 添加包标识— 新建 __init__.py 使目录成为包。
4. 删除依赖文件— 移除 requirements.txt, 因为不再需要 camel-ai。
5. 更新文档— 修改 README.md, 更新安装指令为 pip install strands-sglang==0.4.2, 并调整解释器相关说明。

关键文件:

- examples/strands_sglang/subprocess_interpreter.py (模块 子进程解释器; 类别 source; 类型 dependency-wiring; 符号 InterpreterError, BaseInterpreter, run, supported_code_types): 新增文件, 从 camel-ai 移植子进程解释器并内联依赖, 使示例不再依赖 camel-ai 包。
- examples/strands_sglang/generate_with_strands.py (模块 生成入口; 类别 source; 类型 dependency-wiring; 符号 generate, execute_python_code): 主入口文件, 修改导入和 token 轨迹提取方式, 适配 0.4.2 API。

- examples/strands_sglang/__init__.py (模块 包标识; 类别 source; 类型 core-logic) : 空文件, 使目录成为包, 允许内部导入。
- examples/strands_sglang/README.md (模块 文档; 类别 docs; 类型 documentation) : 更新文档, 反映 API 变化和依赖安装方式。
- examples/strands_sglang/requirements.txt (模块 依赖文件; 类别 docs; 类型 deletion) : 删除不再需要的依赖声明。

关键符号: generate, execute_python_code, SubprocessInterpreter.init, SubprocessInterpreter.run

关键源码片段

examples/strands_sglang/generate_with_strands.py

主入口文件, 修改导入和 token 轨迹提取方式, 适配 0.4.2 API。

```
# 从本地模块导入子进程解释器, 不再依赖 camel-ai
from .subprocess_interpreter import SubprocessInterpreter

# ... 省略其他导入 ...

async def generate(args, sample: Sample, sampling_params) -> Sample:
    """Generate with TITO: tokens captured during generation, no retokenization."""
    assert not args.partial_rollout, "Partial rollout not supported."

    state = GenerateState(args)
    model = SGLangModel(
        tokenizer=state.tokenizer,
        client=get_client_from_slime_args(args, timeout=300.0),
        tool_parser=HermesToolParser(), # tool parsing for wrapped JSON tool calls
        sampling_params=sampling_params,
    )

    tool_limiter = ToolLimiter(max_tool_iters=MAX_TOOL_ITERS)
    agent = Agent(
        model=model,
        tools=[execute_python_code],
        hooks=[tool_limiter],
        callback_handler=None,
        system_prompt=SYSTEM_PROMPT,
    )

    prompt = sample.prompt if isinstance(sample.prompt, str) else sample.prompt[0]["content"]

    try:
        await agent.invoke_async(prompt)
        sample.status = Sample.Status.COMPLETED
    except Exception as e:
        sample.status = Sample.Status.TRUNCATED
```

```
logger.warning(f"TRUNCATED: {type(e).__name__}: {e}")

# 使用新的 rollout API 替代已移除的 token_manager
rollout = model.rollout
prompt_len = rollout.initial_prompt_length # system + user are the first segment
sample.tokens = rollout.token_ids
sample.loss_mask = rollout.loss_mask[prompt_len:]
sample.rollout_log_probs = rollout.logprobs[prompt_len:]
sample.response_length = len(sample.tokens) - prompt_len
sample.response = model.tokenizer.decode(sample.tokens[prompt_len:], skip_special_tokens=False)
sample.tool_iters = tool_limiter.tool_iter_count
sample.tool_calls = tool_limiter.tool_call_count
```

评论区精华

该 PR 没有实质性讨论，reviewer [yitianlian](#) 直接批准（LGTM）。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。改动局限在示例目录，不影响核心库。主要风险包括：
 - 子进程解释器无沙箱隔离（已在文档中明确警示）。
 - 依赖锁定于 strands-sglang==0.4.2，未来版本可能需要再次更新。
 - 如果用户仍使用旧版本 strands-sglang 或依赖 token_manager 则会出错。
 - 影响：仅影响 examples/strands_sglang 示例目录。用户需要更新 strands-sglang 至 0.4.2，并注意子进程解释器的安全说明。不影响其他模块。
 - 风险标记：示例代码无安全沙箱，依赖版本锁定于 0.4.2，旧 API 兼容性缺失

关联脉络

- 暂无明显关联 PR