PR #24581 完整报告

sgl-project/sglang

[CI] pin NeMo-Skills install to known-good SHA in accuracy_test_runner

合并时间：2026-05-07 13:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24581

执行摘要

- 一句话：修复 NeMo-Skills 安装因上游改动的依赖冲突
- 推荐动作：简单但必要的 CI 稳定性修复，值得快速合并。后续可考虑更稳健的依赖管理策略，如使用 lock 文件或定期自动更新固定 commit。

功能与动机

GB300 夜间 CI 中 mmmu-pro 评估分片（Qwen3.5 / Kimi-K2.5）因 NeMo-Skills 安装依赖冲突失败：litellm==1.83.14 depends on httpx==0.28.1 ... and all versions of nemo-run are incompatible。未固定版本导致下次夜间构建立即失效，且失败后 venv 未正确安装包，后续分片报 ModuleNotFoundError: No module named 'nemo_skills'。

实现拆解

修改 python/sglang/test/accuracy_test_runner.py 中 _get_nemo_venv 函数，将 NeMo-Skills 安装源的 URL 从无版本固定的 git+https://github.com/NVIDIA/NeMo-Skills.git 改为固定到已知正常 commit 589294c 的 git+https://github.com/NVIDIA/NeMo-Skills.git@589294c。同时更新了注释，说明固定原因。

关键文件：

- python/sglang/test/accuracy_test_runner.py（模块 测试脚本；类别 test；类型 test-coverage）：核心变更文件，修改 NeMo-Skills 安装的版本固定逻辑。

关键符号：_get_nemo_venv

关键源码片段

python/sglang/test/accuracy_test_runner.py

核心变更文件，修改 NeMo-Skills 安装的版本固定逻辑。

```
# python/sglang/test/accuracy_test_runner.py 第 202-221 行

def _get_nemo_venv() -> Tuple[str, dict]:
    # ... 前面的 venv 创建代码 ...

    # Install nemo_skills.
    # Pinned: NeMo-Skills main after PR #1433 pins litellm==1.83.14 (httpx==0.28.1),
```

```
    # which is unsatisfiable against nemo-run's transitive leptonai dep.
    nemo_skills_ref = "589294c"
    print(f"Installing nemo_skills (pinned to {nemo_skills_ref})...")
    pip_result = subprocess.run(
        [
            "uv",
            "pip",
            "install",
            "--python",
            f"{_nemo_venv_dir}/venv/bin/python",
            f"git+https://github.com/NVIDIA/NeMo-Skills.git@{nemo_skills_ref}",
        ],
        capture_output=True,
        text=True,
        timeout=300,
    )
    if pip_result.returncode != 0:
        raise RuntimeError(f"Failed to install nemo_skills: {pip_result.stderr[-500:]}")

    print("NeMo Skills installed successfully")
    return _get_nemo_venv()
```

评论区精华

无 review 评论，PR 作者自行合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅将依赖从滚动最新版本固定到指定 commit，不影响任何核心逻辑。若上游修复后需更新固定 commit，有良好注释指引。
- 影响：直接修复 GB300 夜间 CI 中 mmmu-pro 评估分片的安装失败问题，确保相关评估（Qwen3.5、Kimi-K2.5）能正常运行。对其他 CI 分片或用户无影响。
- 风险标记：暂无

关联脉络

- PR #1433 [Feature] Support LoRA path renaming and add LoRA serving benchmarks: 上游 NeMo-Skills 的 PR，引入了导致依赖冲突的 litellm 版本固定。