

PR #34768 完整报告

sgl-project/sglang

[AMD] CI: pin antlr4-python3-runtime back after lmms-eval (unblocks ROCm 7.2 stage-b evals)

合并时间: 2026-08-14 08:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34768>

执行摘要

- 一句话: 修复 AMD CI 中 ANTLR 版本冲突, 解锁 ROCm 7.2 stage-b 评测
- 推荐动作: 值得合入, 并可作为 CI 依赖冲突排查的范例阅读。关注三点: 一是在安装脚本中立即钉回版本的修复时机选择; 二是 `--force-reinstall --no-deps` 的取舍理由; 三是对 `install_with_retry` 内部 `eval` 拆词的防御。对普通读者不需要精读, 对 AMD CI 维护者建议顺手补充安装后 `antlr` 版本校验, 防止回归。

功能与动机

34689 修复了 ROCm 7.2 stage-a 的 `sgl-eval` 安装语法问题后, stage-b 首次运行, 所有调用 `run_eval` 的 shard 都报 `ImportError: Unsupported ANTLR version 4.7.2`。PR body 指出这是 `scripts/ci/amd/amd_ci_install_dependency.sh` 中的依赖顺序冲突: `sgl-eval` 先装并带入 `antlr4-python3-runtime 4.9.3`, `lmms-eval v0.4.1` 随后拉入 `latex2sympy2` 并 `pin 4.7.2`、卸载 `4.9.3`。CUDA 路径在 #34477 引入 `sgl-eval` 时已修过同样的冲突, AMD 因 stage-a 前置故障从未得到等价修复。

实现拆解

1. 定位与复现: 根据 PR body 与 `install` 日志, 确认是依赖安装顺序冲突, 即 `lmms-eval v0.4.1` 的 `latex2sympy2` 将 `antlr4-python3-runtime` 从 `4.9.3` 降级为 `4.7.2`, 运行期 `sgl_eval._vendored.nemo_skills.math_grader` 拒绝 `4.7.2`。
2. 选择修复点: 在 `scripts/ci/amd/amd_ci_install_dependency.sh` 中 `lmms-eval` 安装后立即用 `install_with_retry` 执行 `pip install "antlr4-python3-runtime==4.9.3"` `--force-reinstall --no-deps`, 与 `scripts/ci/cuda/ci_install_dependency.sh` 的既有修复保持一致。
3. 放置与细节: `--force-reinstall --no-deps` 只重装 `antlr` 而不动 `lmms-eval` 其他依赖; `pin` 放在 `SKIP_TT_DEPS` 的 `else` 分支内, 因为跳过测试依赖时 `lmms-eval` 根本不会安装; `spec` 不含空格, 可安全穿过 `install_with_retry` 内部的 `eval` (避免 #34689 修过的拆词陷阱)。

4. 验证：PR Test ROCm 7.2 run 31742250192 中 ImportError 出现 0 次，stage-b-1gpu-small 13/14 通过；剩余 5 个失败各有独立原因（HIP graph-capture、perf 阈值、EmbeddingReqInput 属性缺失等），其中 EmbeddingReqInput 由 #34769 单独修复。

5. 配套：无模型或内核代码改动，无单测；验证依赖 run-ci 标签的 ROCm 7.2 流水线。

关键文件：

- scripts/ci/amd/amd_ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure）：唯一变更文件，在 Imms-eval 安装后立即把 antlr4-python3-runtime 重新钉为 4.9.3，修复 ROCm 7.2 stage-b 评测路径的依赖顺序冲突，并对齐 CUDA 安装脚本的既有方案。

关键符号：未识别

关键源码片段

scripts/ci/amd/amd_ci_install_dependency.sh

唯一变更文件，在 Imms-eval 安装后立即把 antlr4-python3-runtime 重新钉为 4.9.3，修复 ROCm 7.2 stage-b 评测路径的依赖顺序冲突，并对齐 CUDA 安装脚本的既有方案。

```
# Imms-eval v0.4.1 会拉入 latex2sympy2，后者把 antlr4-python3-runtime 钉为 4.7.2，
# 并卸载 sgl-eval 依赖的 4.9.3，导致每个调用 sgl-eval run mmlu 的 shard 报错
# "Unsupported ANTLR version 4.7.2"；CUDA 安装脚本早已用同样方式修复，这里对齐。
install_with_retry docker exec -w /Imms-eval ci_sglang pip install --cache-dir=/sgl-data/pip-cache
-e .
```

```
# 紧跟 Imms-eval 安装，因为它是本脚本最后一个会触碰 antlr 的步骤；
# --force-reinstall --no-deps 只重装 antlr，不动 Imms-eval 的其他依赖。
# 该 pin 放在 SKIP_TT_DEPS 的 else 分支内：跳过测试依赖时 Imms-eval 不会安装，
# 也就不存在被覆盖的问题。spec 不含空格，可安全经过 install_with_retry 的 eval。
install_with_retry docker exec ci_sglang pip install --cache-dir=/sgl-data/pip-cache "antlr4-
python3-runtime==4.9.3" --force-reinstall --no-deps
```

评论区精华

Review 阶段没有文字评论，bingxche 直接 APPROVED。主要讨论在 Issue 评论中由 amd-bot 给出：

Merge verdict: fix 验证有效——antlr4-python3-runtime==4.9.3 的 pin 在每个 ROCm 7.2 job 中生效，Unsupported ANTLR 4.7.2 错误完全消失。没有正在执行的失败由本 PR 引起。

amd-bot 同时提醒：stage-b 仍有 5 个与本次 pin 无关的失败，且因 fast-fail，所有 stage-c 任务被跳过未测试，PR 无法单独让 workflow 变绿。作者在 PR body 中逐一给出了剩余失败的成因与关联修复。

- PR CI 未全绿与 stage-c fast-fail 跳过 (testing)：修复已验证有效，但 PR 无法单独让 workflow 变绿，剩余失败需单独修复或 bypass-fastfail 以获取 stage-c 信号。

风险与影响

- 风险:

1. 依赖覆盖风险: `--force-reinstall --no-deps` 会把 `lmms-eval` 侧 `latex2sympy2` 想要的 4.7.2 强制覆盖为 4.9.3, `lmms-eval` 相关功能存在兼容性风险; `CUDA` 路径已长期接受同一取舍, 属已知且可接受的权衡。
2. 防回归缺口: 脚本没有在安装后校验 `antlr` 版本, 未来 `lmms-eval` 或 `latex2sympy2` 升级可能再次引发同类问题, 需要人工从日志发现。
3. 流水线完整性: 本 PR 合入时 `stage-c` 全部未跑、`stage-b` 还有 5 个失败, 虽然与本次修改无关, 但 CI 未处于全绿状态, 回归信号被部分遮蔽。
4. 影响面限制: 仅修改 AMD 安装脚本, `NPU` 不装 `lmms-eval` 不受影响, `CUDA` 已有同款 `pin`, 无运行时影响。- 影响: 对用户与模型无直接影响 (纯 CI 脚本); 对团队的影响是解锁了 AMD ROCm 7.2 `stage-b` 的 `accuracy` 评测, 使 `sgl-eval run mmlu` 相关测试从大量失败恢复到 13/14 通过, 并让 AMD 与 `CUDA` 两条 CI 路径的依赖处理行为对齐。后续 `stage-b` 剩余问题 (#34769 等) 仍会影响 AMD 夜间与 PR CI 的绿灯状态。- 风险标记: 依赖版本强制覆盖, `stage-c` 未验证, 防回归校验缺失

关联脉络

- PR #34689 [AMD] CI: drop the spaces from `SGL_EVAL_SPEC` (fixes ROCm 7.2 `stage-a sgl-eval install`): 前置修复, 解锁 ROCm 7.2 `stage-a`, 使 `stage-b` 首次运行并暴露本 PR 修复的 `ANTLR` 冲突。
- PR #34769 [VLM] Fix `AttributeError` on multimodal embedding requests: 本 PR CI 结果中 `stage-b` 的 `EmbeddingReqInput AttributeError` 失败由该 Issue 修复, 属于同一批 ROCm 7.2 `stage-b` 首跑暴露的问题。