PR #25737 完整报告

sgl-project/sglang

Reduce excessively long logs caused by transformer version updates.

合并时间：2026-05-20 11:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25737

执行摘要

- 一句话：限制 NPU 测试中 transformers 的日志级别
- 推荐动作：该 PR 属于常规维护性质，值得快速合入以改善测试体验。如果有更多 NPU 测试文件出现类似日志问题，建议统一纳入。

功能与动机

当 pipeline 执行测试用例时，打印了过多的日志，影响了执行时间和失败用例的定位。根本原因是 transformers 版本更新导致调用相关方法时打印兼容性警告。PR body 原文："When the pipeline executes test cases, too many logs are printed, affecting the execution time and locating the failure cause of failed test cases."

实现拆解

实现方式非常简单，在所有 5 个 NPU 测试文件中添加 TRANSFORMERS_VERBOSITY 环境变量设置，值为 "error"，使得只有 error 级别的 transformers 日志才会输出。

1. 在测试类级别设置环境变量：在 test_npu_qwq_32b_w8a8.py 和 test_npu_hierarchical_cache_mla.py 中，通过 os.environ["TRANSFORMERS_VERBOSITY"] = "error" 在类定义或测试方法内设置环境变量。
2. 在启动服务器的 env 参数中设置：在 test_npu_memory_consumption.py 中，通过 popen_launch_server 的 env 字典参数传递环境变量，仅对子进程生效。
3. 在 extra_envs 字典中添加：在 test_npu_deepep.py 和 test_npu_deepep_auto_deepseek_v3_2_w8a8.py 中，将 TRANSFORMERS_VERBOSITY 加入 extra_envs 字典后通过 os.environ.update() 统一设置。

关键文件：

- test/registered/ascend/llm_models/test_npu_qwq_32b_w8a8.py（模块 NPU 测试；类别 test；类型 test-coverage）：主要修改：在测试类中添加 os.environ["TRANSFORMERS_VERBOSITY"] = "error" 以抑制 transformers 兼容性警告日志。
- test/registered/ascend/basic_function/HiCache/test_npu_hierarchical_cache_mla.py（模块 NPU 测试；类别 test；类型 test-coverage）：在测试方法 test_no_chunked_prefill_without_radix_cache 中设置环境变量，抑制日志。

- test/registered/ascend/test_npu_memory_consumption.py（模块 NPU 测试；类别 test；类型 test-coverage）：通过 popen_launch_server 的 env 参数设置环境变量，确保仅子进程生效。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep.py（模块 NPU 测试；类别 test；类型 test-coverage）：在 extra_envs 字典中添加环境变量，通过 os.environ.update() 统一设置。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_deepseek_v3_2_w8a8.py（模块 NPU 测试；类别 test；类型 test-coverage）：在 extra_envs 字典中添加环境变量，与上一个文件类似。

关键符号：未识别

关键源码片段

test/registered/ascend/llm_models/test_npu_qwq_32b_w8a8.py

主要修改：在测试类中添加 os.environ["TRANSFORMERS_VERBOSITY"] = "error" 以抑制 transformers 兼容性警告日志。

```
import os
import unittest

from sglang.test.ascend.gsm8k_ascend_mixin import GSM8KAscendMixin
from sglang.test.ascend.test_ascend_utils import QWQ_32B_W8A8_WEIGHTS_PATH
from sglang.test.ci.ci_register import register_npu_ci
from sglang.test.test_utils import CustomTestCase

register_npu_ci(est_time=400, suite="nightly-2-npu-a3", nightly=True)

class TestQWQ32BW8A8(GSM8KAscendMixin, CustomTestCase):
    # ... 其他代码 ...
    os.environ["TRANSFORMERS_VERBOSITY"] = "error" # 仅输出 error 级别日志

if __name__ == "__main__":
    unittest.main()
```

test/registered/ascend/basic_function/HiCache/test_npu_hierarchical_cache_mla.py

在测试方法 test_no_chunked_prefill_without_radix_cache 中设置环境变量，抑制日志。

```
import os
import unittest
# ... 其他导入 ...

class TestNpuHierarchicalCacheMla(CustomTestCase):
    def test_no_chunked_prefill_without_radix_cache(self):
        # ... 设置 common_args ...
        os.environ["TRANSFORMERS_VERBOSITY"] = "error" # 抑制 transformers 警告
        for common_arg in common_args:
```

```
        # ... 执行测试 ...
```

test/registered/ascend/test_npu_memory_consumption.py

通过 popen_launch_server 的 env 参数设置环境变量，确保仅子进程生效。

```
# 在 test_memory_consumption 方法中
process = popen_launch_server(
    model,
    base_url,
    timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    other_args=[
        # ... 参数 ...
    ],
    env={
        "TRANSFORMERS_VERBOSITY": "error", # 仅对子进程生效
    },
)
```

评论区精华

无 review 评论，PR 获得两个 approve（adarshxs 和 sglang-npu-bot），没有发现争议或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。该变更仅在测试环境中设置环境变量，不影响生产代码。如果 future transformers 版本改变了 verbosity 的 key 名称，此设置将失效但不会导致错误。另需注意 os.environ 的设置是进程全局的，可能会影响同一进程内其他模块的 transformers 日志行为（但在测试场景中无影响）。
- 影响：影响范围限定在 NPU 测试流水线中的 5 个测试文件，不涉及任何产品代码或用户可见行为。变更后这些测试的运行日志量将显著减少，有助于更快定位真正的失败原因，且不会改变测试的语义或结果。
- 风险标记：暂无

关联脉络

- PR #25483 [codex] Update Wan2.2 ModelOpt CI checkpoints: 该 PR 也涉及 transformers 模型配置更新，但非直接关联。