PR #26033 完整报告

sgl-project/sglang

Reduce excessively long logs caused by transformer version updates.

合并时间：2026-05-23 17:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26033

执行摘要

- 一句话：抑制 NPU 测试中 transformer 版本兼容性警告日志
- 推荐动作：仅作简单说明即可，无需深入代码逻辑。如果团队在其他测试文件中也遇到类似 transformer 日志泛滥，可参考此做法统一加环境变量。

功能与动机

PR body 指出："When the pipeline executes test cases, too many logs are printed, affecting the execution time and locating the failure cause of failed test cases." 根因是 transformer 版本升级后，相关方法调用会触发兼容性警告。通过在测试环境变量中设置 TRANSFORMERS_VERBOSITY=error，仅保留 error 级别日志，从而减少无用输出。

实现拆解

在 test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_480b.py 的 setUpClass 的 env 字典中，新增一行键值对 "TRANSFORMERS_VERBOSITY": "error"，插入位置在 "HCCL_OP_EXPANSION_MODE": "AIV" 之后、**os.environ 之前。该环境变量为 transformer 库的日志级别控制，设为 error 后会抑制 warning 和 info 级别的日志，仅打印 error。由于 **os.environ 在最后展开，若系统环境已有同名变量，则字典中的值会被 os.environ 覆盖，因此优先级并未特意调高，但通常 CI 环境中不会设置此变量。

关键文件：

- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_480b.py（模块 测试；类别 test；类型 test-coverage）：本 PR 唯一修改的文件，在 setUpClass 的环境变量字典中增加 TRANSFORMERS_VERBOSITY=error，用于抑制 transformer 版本兼容性警告日志，减少 CI 测试输出噪音。

关键符号：未识别

关键源码片段

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_480b.py

本 PR 唯一修改的文件，在 setUpClass 的环境变量字典中增加 TRANSFORMERS_VERBOSITY=error，用于抑制 transformer 版本兼容性警告日志，减少 CI 测试输出噪音。

```
# test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_480b.py
# 在 setUpClass 中启动 server 时传入的环境变量字典
env={
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "600",
    "HCCL_BUFFSIZE": "2100",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    # 新增：将 transformer 日志级别设为 error，过滤因版本兼容性产生的 warning/info 日志
    "TRANSFORMERS_VERBOSITY": "error",
    **os.environ, # 注意：os.environ 展开在字典最后，会覆盖前面同名键
},
```

评论区精华

本 PR 无 review 评论（review_comments_count = 0）。审核由 sglang-npu-bot 自动批准，无人工讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改一个测试文件的环境变量，且是抑制日志输出，不影响任何模型推理逻辑或数值精度。唯一潜在影响是：若测试中真的发生了 error 级别日志，现在会被保留；若原本 warning 中包含有用调试信息，则会被隐藏。但在自动化测试场景中，隐藏 warning 通常可接受。
- 影响：影响范围极小：仅针对 NPU 上 DeepEP 自动模式 Qwen3-480B 的端到端测试用例。对用户无直接影响，仅改善 CI 日志可读性。属于测试配套优化。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR