PR #36459 完整报告

sgl-project/sglang

[NPU] Fix evalscope accuracy parsing and add glm5_1 aime26 request timeout

合并时间：2026-08-31 17:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36459

执行摘要

- 一句话：修复 NPU 精度测试解析与超时，并调整显存参数
- 推荐动作：值得快速浏览，不建议精读。对 NPU 测试维护者有参考价值：一是 evalscope 多版本输出格式的兼容处理方式（正则顺序 + finditer + 百分比归一化）；二是 kimi_k2_6 显存参数调参时把 "KV 池上限 vs prefill OOM" 的权衡写进 commit 的做法，是测试配置变更的良好实践。核心逻辑直白，无架构性内容。

功能与动机

PR body 的出发点是一句话："Fix some test cases bug"。具体包括：evalscope 1.11+ 把 accuracy 输出为百分比（如 66.67%），"which was not matched by the existing regex patterns and was not normalized against the baseline"；glm5_1 多节点 aime26 长请求会被路由默认超时切断，需要 "prevent the multi-node aime26 accuracy test from timing out"；kimi_k2_6 64k 性能测试在 0.662 下 KV 池不足（"max allowed length 61050 < 64000"），而 0.7 "causes prefill OOM in the MLP down_proj quantized matmul"，0.68 是平衡点（KV 池约 77000 tokens）。

实现拆解

本 PR 全部为测试文件变更，按以下步骤拆解：

1. 修复 evalscope 精度解析（python/sglang/test/ascend/e2e/test_npu_accuracy_utils.py 的 run_evalscope）：新增针对 evalscope 1.11+ 表格格式的正则 `Accuracy\s*[↑↓]?\s*|\s*[^|]*|\s*\d+\s*|\s*([\d.]+)%?\s*|` 并置于 pattern 列表首位；从 re.findall 切换到 re.finditer，用 group(0) 判断是否含 %、用 group(1) 取值并除以 100.0 归一化到 0-1 区间。这样新格式命中后不会回落到旧 pattern，且归一化后能与 accuracy 基线（如 0.695、0.953）同量纲比较。
2. 清理废弃方法（同一文件）：在 qwen3_next_80b 测试切换到 run_accuracy() 后，run_accuracy_multiple 已无调用方，评审建议删除，实际删除该方法及 n_runs = 3 类属性，净删 67 行，避免死代码残留。
3. glm5_1 aime26 请求超时修复（test/registered/npu/accuracy/glm5_1/test_npu_glm5_1_w4a8_1p1d_32p_in64k_out1k_50ms_aime26.py）：router_args 增加 --request-timeout-secs 7200。aime26 生成配置 max_tokens=65536，在 32 卡 1p1d 分离拓扑下长请求可能超过路由默认超时，显式放大超时窗口避免evalscope请求被提前终止。

4. kimi_k2_6 显存比例调参（test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_16p_in64k_out1k_100ms.py）：--mem-fraction-static从0.662调整为0.68。commit 信息记录了完整调参依据：0.662 下 KV 池最大允许长度 61050 < 64000 输入；0.7 时 MLP down_proj 量化矩阵乘 prefill OOM；0.68 约留 77000 token 的 KV 池并兼顾动态激活空间。
5. qwen3_next_80b 指标口径调整（test/registered/npu/accuracy/qwen3_next_80b_a3b_instruct/test_npu_qwen3_next_80b_w8a8_2p_in6k_out1k5_bs16_aime25.py）：test_aime25 从 run_accuracy_multiple(n_runs=3) 改为 run_accuracy()。后者内部已有 best_metrics 逻辑（多轮取最大值并在达到阈值时提前退出），语义更贴近 " 通过性测试 "，且避免 3 次全量运行带来的时间开销。
6. 范围收敛：提交历史显示 PR 一度包含 nightly-test-npu.yml 的 job 依赖调整与两个 DeepSeek-V4-Flash 性能测试禁用，最终通过 "Restore nightly-test-npu.yml to match community main" 还原，合并内容只保留上述 4 个测试文件。

关键文件：

- python/sglang/test/ascend/e2e/test_npu_accuracy_utils.py（模块 精度测试工具；类别 test；类型 bugfix；符号 run_evalscope, run_accuracy_multiple, TestNpuAccuracyTestCaseBase）：NPU accuracy 测试的公共工具，修复 evalscope 1.11+ 百分比格式解析并删除废弃方法，影响所有下游精度测试用例
- test/registered/npu/accuracy/glm5_1/test_npu_glm5_1_w4a8_1p1d_32p_in64k_out1k_50ms_aime26.py（模块 精度测试；类别 test；类型 configuration；符号 GLM_5_1_PD_SEP_MODEL_CONFIG, TestNPUGLM5_1_W4A8_PD_SEP_AIME2026）：多节点 aime26 精度测试的路由超时修复，避免 evalscope 长请求被提前切断
- test/registered/npu/accuracy/qwen3_next_80b_a3b_instruct/test_npu_qwen3_next_80b_w8a8_2p_in6k_out1k5_bs16_aime25.py（模块 精度测试；类别 test；类型 test-coverage；符号 TestQwen3Next80BA3B_aime25）：精度判定从 3 次平均改为 run_accuracy 的多次取最大值，并触发废弃方法清理
- test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_16p_in64k_out1k_100ms.py（模块 性能测试；类别 test；类型 configuration；符号 OTHER_ARGS）：64k 长输入性能测试的显存比例调优，解决 KV 池不足与 prefill OOM 的两难

关键符号：run_evalscope, run_accuracy, run_accuracy_multiple, test_aime25, test_npu_glm5_1_w4a8_pd_sep_aime2026

关键源码片段

python/sglang/test/ascend/e2e/test_npu_accuracy_utils.py

NPU accuracy 测试的公共工具，修复 evalscope 1.11+ 百分比格式解析并删除废弃方法，影响所有下游精度测试用例

这是 python/sglang/test/ascend/e2e/test_npu_accuracy_utils.py 中 run_evalscope 的 accuracy 解析逻辑（修复后版本），要点是同时兼容 evalscope 1.10 与 1.11+ 的输出格式：

```
# evalscope 1.11+ 的表格里 Accuracy 一行带上升 / 下降箭头（↑ ↓），
```

```
# 数值以百分比呈现（如 66.67%），旧正则无法匹配，导致 baseline 对比失效。
accuracy_patterns = [
    # 优先适配 evalscope 1.11+ 表格式输出，容忍尾部 %
    r"Accuracy\s*[↑ ↓]?\s*|\s*[^|]*|\s*\d+\s*|\s*([\d.]+)%?\s*|",
    # evalscope 1.10 及更早版本的表格格式
    r"mean_acc\s*.*?|\s*\d+\s*|\s*([\d.]+)\s*|",
    # 通用表格列捕获
    r"|\s+([\d.]+)\s+|\s+\S+\s+|\s*$",
    # 键值对输出格式
    r"accuracy\s*[:=]?\s*([\d.]+)",
    r"Accuracy\s*[:=]?\s*([\d.]+)",
    r"score\s*[:=]?\s*([\d.]+)",
]

for pattern in accuracy_patterns:
    # 从 findall 改为 finditer：既保留完整匹配（group(0) 用于判断是否
    # 出现 %），又能通过 group(1) 取出数值本身
    matches = list(re.finditer(pattern, full_output))
    if matches:
        last = matches[-1]
        final_accuracy = float(last.group(1))
        # evalscope 1.11+ 把 accuracy 输出为百分比，归一化到 0-1 后才能
        # 与基线（如 0.953、0.695）进行同量纲比较
        if "%" in last.group(0):
            final_accuracy /= 100.0
        metrics["accuracy"] = final_accuracy
        logger.info(f"The Final Accuracy from output: {final_accuracy}")
        break
```

评论区精华

review 中仅有一条实质评论：

- cherryblo在 test_npu_qwen3_next_80b_w8a8_2p_in6k_out1k5_bs16_aime25.py 的 line 115（run_accuracy_multiple(n_runs=3) 改为 run_accuracy() 之后）回复 "delete method"，指出该方法已无调用方应删除。作者采纳建议，在 test_npu_accuracy_utils.py 中删除整个 run_accuracy_multiple 方法及 n_runs 类属性（净删 67 行）。这是一个典型的 " 改动引发的死代码清理 " 反馈，避免了旧方法残留。
- 此外，提交历史中的 [NPU] Chain PR test jobs and disable two DeepSeek-V4-Flash perf tests、[NPU] Fix nightly-test-npu.yml job deps 与 [NPU] Restore nightly-test-npu.yml to match community main 三条提交说明评审过程中要求收紧 PR 范围、撤回过 CI 工作流改动，最终仅保留纯测试文件。
- Issue 评论中多次出现 /tag-and-rerun-ci 与 /rerun-failed-ci，反映 NPU 测试矩阵在 PR 阶段不够稳定，最终由 sglang-npu-bot 批准合并。
- 废弃 run_accuracy_multiple 方法的清理 (design): 作者接受建议，在 test_npu_accuracy_utils.py 中删除整个 run_accuracy_multiple 方法及其 n_runs 类属性，净删 67 行。

- nightly-test-npu.yml 工作流改动被还原 (other): 最终合并内容仅保留 4 个测试文件，与社区 main 对齐。

风险与影响

- 风险：
  - 解析优先级变更：新正则位于 pattern 列表首位，若旧版 evalscope 输出恰好部分匹配新列布局，解析结果可能变化。由于归一化仅在完整匹配含 % 时触发，旧格式（无 %）数值不会被误除，但匹配优先级变化仍需在后续 evalscope 升级时关注。
  - 单点归一化判断：百分比归一化依赖 '%' in last.group(0) 这一单点判断，如果 evalscope 后续输出改为分数或千分位等格式，解析会再次静默失效（仅走日志，不抛错）。
  - 公共方法删除风险：run_accuracy_multiple 从 test_npu_accuracy_utils.py 删除后，若仓库其他分支或外部脚本仍引用，会触发 AttributeError。当前源码摘要确认仓库内已无其他调用方。
  - 指标口径变化：qwen3_next_80b 从 3 次平均改为多次取最大值，最大值口径天然乐观，可能掩盖偶发的模型精度退化；同时不再产出 accuracy_avg 指标，历史均值口径数据无法直接对比。
  - 显存参数经验性：0.68 是针对当前 2 机 32 卡、DeepEP、DP attention 的 NPU 机型调出的经验值，换机型或 NPU 型号后可能重新失衡。
  - CI 稳定性：多次 rerun 说明 NPU 测试环境仍有偶发失败，本次修复能降低解析与超时类失败，但环境层面的波动仍在。
- 影响：
  - 影响范围：仅限 NPU 测试矩阵（accuracy 与 performance 两套 workflow），无任何产品代码或内核改动，对线上用户零影响。
  - 团队收益：修复 evalscope 版本升级导致的解析失效，避免大量 NPU 精度测试误报失败；glm5_1 多节点长请求测试、kimi_k2_6 64k 性能测试的稳定性提升，减少无效 rerun 与人工排查成本。
  - 数据口径影响：qwen3_next_80b aime25 判定口径从 "3 次平均 " 变为 " 多次取最大 "，后续该用例的指标与历史均值数据不可直接对比，需在结果解读时注意。
  - 维护性：删除死代码方法降低了测试基线的维护负担，pr 提交中记录了显存调参依据，为后续同类调参提供了可复用的推理路径。
  - 风险标记：解析正则优先级变更，百分比归一化依赖单点判断，显存参数依赖特定 NPU 机型，指标口径改为最大值口径，删除公共工具方法，CI 稳定性依赖重跑

关联脉络

- PR #37214 test: re-enable DSV4-Flash W8A8 8p nightly perf cases: 本 PR 提交历史上曾包含 "disable two DeepSeek-V4-Flash perf tests" 的改动（后续被还原），与 #37214 重新启用两个 DSV4-Flash W8A8 8p 夜间性能测试形成一启一停的对照，反映 NPU 性能测试矩阵的启停管理是活跃维护点。