

PR #24689 完整报告

sgl-project/sglang

[NPU] Add GitHub test summary and deduplicate test code. Part 2

合并时间: 2026-06-08 23:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/24689>

执行摘要

- 一句话: NPU 测试代码去重与重构, 使用 Mixin 消除重复 GSM8K 测试逻辑
- 推荐动作: 值得精读, 展示了如何通过 Mixin 模式重构大型测试套件中的重复代码, 是测试基础设施优化的良好实践。需特别关注 Review 中尚未完全解决的属性名不匹配问题, 以及权重路径是否使用缓存。

功能与动机

来自 PR Body: 'There is a lot of duplicated code because, instead of using a GSM8KAscendMixin common class, the GSM8K model tests were added by copy-pasting.' 目标是消除重复代码, 统一测试配置方式, 并在 CI 中输出测试摘要。

实现拆解

1. 创建 Mixin 基类: 在 `python/sglang/test/ascend/gsm8k_ascend_mixin.py` 中定义 `GSM8KAscendMixin`, 封装 GSM8K 评估的公共逻辑 (启动服务、执行评估、断言)。在 `python/sglang/test/ascend/test_mmlu.py` 中定义 `TestMMLU` 基类, 封装 MMLU 评估逻辑。基类通过读取子类的类属性 (`model`、`other_args`、`env`、`accuracy` 等) 自动完成服务启动和测试执行。
2. 重构各测试类: 修改 6 个 DeepEP 和注意力并行测试文件, 使其继承 `GSM8KAscendMixin` 和 `TestMMLU` (部分类多重继承)。将原来的 `setUpClass/tearDownClass` 方法替换为声明式类属性, 移除了大量重复的 `popen_launch_server` 调用和 `kill_process_tree` 逻辑。测试方法 (如 `test_gsm8k`、`test_mmlu`) 由基类统一实现。
3. 添加 GitHub Summary 输出: 在基类和部分测试中集成 GitHub Actions Step Summary, 将准确率结果以 JSON 格式输出, 方便 CI 流水线解析。
4. 更新权重路径: 在 `test_ascend_utils.py` 中添加新的模型权重常量 `REDHATAI_QWEN2_5_0_5B_INSTRUCT_QUANTIZED_W8A8_WEIGHTS_PATH`, 用于量化测试。
5. 修复配置问题: 根据 Review 反馈, 修正了 `other_args` 误定义为元组的问题、属性名不匹配问题 (`accuracy_mmlu` → `accuracy_mmlu_threshold`), 以及权重路径缺少 `MODEL_WEIGHTS_DIR` 前缀的问题。

6. 更新 CODEOWNERS: 修改 [.github/CODEOWNERS](https://github.com/Ascend/CodeOwners), 将 Ascend 测试的维护者指向新团队。

关键文件:

- test/registered/ascend/llm_models/test_npu_qwen3_30b_attn_cp.py (模块 注意力并行; 类别 test; 类型 test-coverage; 符号 TestQwen330BAttnCP, setUpClass, tearDownClass, test_gsm8k_accuracy): 核心测试文件, 重构后使用 GSM8KAscendMixin, 减少约 70 行重复代码, 展示了从 setUpClass 到声明式配置的转变。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_d eepep_auto_qwen3_480b.py (模块 专家并行; 类别 test; 类型 test-coverage; 符号 TestDeepEpQwen, setUpClass, tearDownClass, test_mmlu): 典型的 DeepEP 自动模式测试文件, 重构后同时使用 GSM8KAscendMixin 和 TestMMLU, 显著减少重复代码。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_d eepep_low_latency_qwen3_480b.py (模块 专家并行; 类别 test; 类型 test-coverage; 符号 TestDeepEpQwen, setUpClass, tearDownClass, test_mmlu): DeepEP 低延迟模式测试, 重构后同样减少约 100 行重复代码。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_d eepep_low_latency_qwen3_next.py (模块 专家并行; 类别 test; 类型 test-coverage; 符号 TestQwen3Next, setUpClass, tearDownClass, test_mmlu): Qwen3-Next 低延迟模式测试, 重构后减少约 100 行重复代码。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_d eepep_auto_qwen3_next.py (模块 专家并行; 类别 test; 类型 test-coverage; 符号 TestQwen3Next, setUpClass, tearDownClass, test_mmlu): Qwen3-Next 自动模式测试, 重构后减少约 96 行重复代码。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_d eepep_low_latency_deepseek_v3_2_w8a8.py (模块 专家并行; 类别 test; 类型 test-coverage; 符号 TestDeepEpDeepseekV32, setUpClass, tearDownClass, test_mmlu): DeepSeek V3.2 低延迟模式测试, 重构后减少约 87 行重复代码。
- test/registered/ascend/basic_function/quant/test_npu_w4a4_quantization.py (模块 量化测试; 类别 test; 类型 test-coverage; 符号 TestAscendW4A4, setUpClass, tearDownClass, test_gsm8k): 量化测试文件, 同样引入 GSM8KAscendMixin 以消除重复的 setUpClass 和 test_gsm8k。
- test/registered/ascend/basic_function/quant/test_npu_w8a8_quantization.py (模块 量化测试; 类别 test; 类型 test-coverage; 符号 TestAscendW8A8CompressedTensors, setUpClass, tearDownClass, test_gsm8k): W8A8 量化测试, 同样采用 Mixin 去掉重复样板代码。
- python/sglang/test/ascend/test_ascend_utils.py (模块 测试工具; 类别 test; 类型 test-coverage): 工具文件, 新增模型权重路径常量, 并修复了路径前缀问题。
- python/sglang/test/ascend/test_mmlu.py (模块 测试基类; 类别 test; 类型 test-coverage): MMLU 测试基类, 引入 TestMMLU 供其他测试继承, 并在此 PR 中调整了 Summary 输出格式。

- python/sglang/test/ascend/gsm8k_ascend_mixin.py (模块 测试基类; 类别 test; 类型 test-coverage) : 核心 Mixin 文件, 定义了 GSM8KAscendMixin, 封装了服务启动、评估执行和 Summary 输出逻辑。
- .github/CODEOWNERS (模块 仓库配置; 类别 infra; 类型 infrastructure) : 更新 Ascend 测试目录的负责人, 将维护者指向新团队。

关键符号: GSM8KAscendMixin, TestMMLU, setUpClass, test_gsm8k, test_mmlu

关键源码片段

test/registered/ascend/llm_models/test_npu_qwen3_30b_attn_cp.py

核心测试文件, 重构后使用 GSM8KAscendMixin, 减少约 70 行重复代码, 展示了从 setUpClass 到声明式配置的转变。

```
# test/registered/ascend/llm_models/test_npu_qwen3_30b_attn_cp.py
# 重构后: 使用 Mixin + 类属性, 无需重复编写服务启动和测试方法
import os
import unittest
from sglang.test.ascend.gsm8k_ascend_mixin import GSM8KAscendMixin # 引入共享 Mixin
from sglang.test.ascend.test_ascend_utils import QWEN3_30B_A3B_WEIGHTS_PATH
from sglang.test.ci.ci_register import register_npu_ci
from sglang.test.test_utils import CustomTestCase
```

```
register_npu_ci(est_time=500, suite="nightly-4-npu-a3", nightly=True)
```

```
class TestQwen330BAttnCP(GSM8KAscendMixin, CustomTestCase):
    """GSM8K accuracy test for Qwen3-30B-A3B mixed deployment on 4 NPUs."""
    model = QWEN3_30B_A3B_WEIGHTS_PATH # 类属性: 模型路径
    other_args = [ # 类属性: 额外启动参数
        "--trust-remote-code",
        "--mem-fraction-static", "0.7",
        "--max-running-requests", "32",
        "--attention-backend", "ascend",
        "--tp-size", "4",
        "--moe-dp-size", "2",
        "--attn-cp-size", "2",
        "--cuda-graph-max-bs", "32",
        "--enable-prefill-context-parallel",
    ]
    env = {**os.environ, "ASCEND_USE_FIA": "1"} # 类属性: 环境变量
    accuracy = 0.92 # GSM8K 最低准确率
    gsm8k_parallel = 32
    num_questions = 100
    gsm8k_num_shots = 5

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

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepest_auto_qwen3_480b.py

典型的 DeepEP 自动模式测试文件，重构后同时使用 GSM8KAscendMixin 和 TestMMLU，显著减少重复代码。

```
# test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepest_
auto_qwen3_480b.py
# 重构后：多重继承 Mixin，同时获得 GSM8K 和 MMLU 测试能力
import os
import unittest
from sglang.test.ascend.gsm8k_ascend_mixin import GSM8KAscendMixin
from sglang.test.ascend.test_ascend_utils import QWEN3_CODER_480B_A35B_INSTRUCT_
W8A8_QUAROT_WEIGHTS_PATH
from sglang.test.ascend.test_mmlu import TestMMLU # 引入 MMLU 测试基类
from sglang.test.ci.ci_register import register_npu_ci
from sglang.test.test_utils import CustomTestCase

register_npu_ci(est_time=200, suite="nightly-16-npu-a3", nightly=True)

class TestDeepEpQwen(GSM8KAscendMixin, TestMMLU, CustomTestCase):
    """Test with DeepEP auto mode"""
    model = QWEN3_CODER_480B_A35B_INSTRUCT_W8A8_QUAROT_WEIGHTS_PATH
    other_args = [
        "--trust-remote-code", "--nnodes", "1", "--node-rank", "0",
        "--attention-backend", "ascend", "--device", "npu",
        "--quantization", "modelslim", "--max-running-requests", 96,
        "--context-length", 8192, "--dtype", "bfloat16",
        "--chunked-prefill-size", 28672, "--max-prefill-tokens", 458880,
        "--disable-radix-cache", "--moe-a2a-backend", "deepest",
        "--deepest-mode", "auto", "--tp-size", 16, "--dp-size", 4,
        "--enable-dp-attention", "--enable-dp-lm-head",
        "--mem-fraction-static", 0.7, "--cuda-graph-bs", 16, 20, 24,
    ]
    env = {
        "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
        "SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT": "600",
        "HCCL_BUFFSIZE": "2100", "HCCL_OP_EXPANSION_MODE": "AIV",
        "TRANSFORMERS_VERBOSITY": "error", **os.environ,
    }
    mmlu_num_examples = 8
    accuracy_mmlu_threshold = 0.61 # MMLU 准确性阈值
    accuracy = 0.91 # GSM8K 准确性阈值
    num_questions = 200
    gsm8k_num_shots = 8

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

评论区精华

Review 主要指出三个问题：

- 属性名不匹配：test_mmlu.py 中 getattr(self, 'accuracy_mmlu') 与子类定义的 accuracy_mmlu_threshold 不一致，会导致 Summary 显示 'N/A'。
- 权重路径缺少前缀：新增的 REDHATAI_QWEN2_5_0_5B_INSTRUCT_QUANTIZED_W8A8_WEIGHTS_PATH 未使用 os.path.join(MODEL_WEIGHTS_DIR, ...)，可能导致文件未找到。作者回应：“Since the test did not specify an absolute path, I copied it here the same way.” 后改为使用缓存。
- other_args 类型错误：多个文件将 other_args 定义为 ([...]) 即元组包裹列表，应改为纯列表。

最终 reviewer ping1jing2 批准了 PR，但部分评论指出问题可能未完全修复（如 accuracy_mmlu 仍残留），需注意。

- 属性名不匹配导致 Summary 显示 N/A (correctness): 作者未明确回复，但后续提交可能已修复。最终需验证。
- 新权重路径缺少 MODEL_WEIGHTS_DIR 前缀 (correctness): 作者在后续提交中修复，使用了缓存路径。
- other_args 定义为元组而非列表 (correctness): 作者应已修复（相关文件最终版本显示为列表格式）。

风险与影响

- 风险：
 1. 基类兼容性风险：GSM8KAscendMixin 和 TestMMLU 的多重继承 MRO 可能导致预期外的行为，例如若两个基类都定义了 setUpClass，子类调用 super() 时顺序可能非预期。但当前实现中 Mixin 只提供类属性，不重写 setUpClass，风险较低。
 2. 配置遗漏风险：声明式类属性可能遗漏某些测试类特定的配置（如超时时间、环境变量），例如 test_npu_deepest_low_latency_deepseek_v3_2_w8a8.py 中增加了 timeout_for_server_launch = 6000，但其他测试未显式设置，可能使用基类默认值，需确认基类是否妥善处理。
 3. 属性名不一致：Review 指出的 accuracy_mmlu 与 accuracy_mmlu_threshold 不匹配问题，在最终代码中可能仍存在，若未修复会导致 Summary 中 MMLU 阈值显示异常，但不影响测试执行。
 4. 回归风险：修改涉及 12 个文件，删除约 600 行代码，若基类实现有缺陷可能影响所有 NPU 测试，需依赖 CI 验证。- 影响：范围：仅影响 NPU 测试套件（Ascend 平台），不涉及任何生产代码。程度：显著降低测试代码重复度，提升可维护性。新增的 GitHub Summary 输出便于 CI 结果查看。未来新增类似测试只需继承 Mixin 并配置少量属性。对用户和推理服务无影响。
- 风险标记：属性名不一致残留，多重继承 MRO 可能非预期，基类默认配置可能不适合所有测试

关联脉络

- PR #24689 [NPU] Add GitHub test summary and deduplicate test code. Part 2: 自身, 作为上下文参考。