PR #37431 完整报告

sgl-project/sglang

test(npu): add DSV4-Flash / GLM-5.2 / Kimi-K3 gpqa accuracy cases

合并时间：2026-09-01 22:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37431

执行摘要

- 一句话：为 NPU 平台上的 DeepSeek-V4-Flash、GLM-5.2 和 Kimi-K3 添加 GPQA 准确率测试用例。
- 推荐动作：这是一个 重要且值得精读的测试 PR。其核心价值在于为 SGLang NPU 支持的三个关键 MoE 模型建立了标准化的、覆盖多种拓扑的准确性基准。审查者应重点关注：
  1. 各测试文件中启动参数与环境变量的配置是否合理、完备。
  2. GLM-5.2 测试配置的调整（如启用 DP、调整内存比例）是否基于有效的性能调优结果。
  3. CI 工作流中对 Kimi-K3 源码构建的配置是否正确，以及是否对整体 CI 时间有显著影响。
  4. 共享工具常量的重命名是否已完成对所有消费者的更新。

功能与动机

实现拆解

1. 新增 DeepSeek-V4-Flash W8A8 准确性测试：在 test/registered/npu/accuracy/deepseek_v4_flash/ 下新增 test_npu_deepseek_v4_flash_w8a8_8p_gpqa.py。测试类 TestNPUDeepSeekV4FlashW8A88PGPQA 继承自 TestNpuAccuracyTestCaseBase，配置了包括 --speculative-algorithm DSPARK、--dp-size 16、--tp-size 16 等启动参数和多个环境变量（如 SGLANG_DSPARK_FAST_KERNEL），并将目标准确率设为 0.874。该测试被注册到 CI 套件 nightly-acc-16-npu-a3。
2. 新增 Kimi-K3 W4A8 准确性测试：在 test/registered/npu/accuracy/kimi_k3/ 下新增 test_npu_kimi_k3_w4a8_32p_gpqa.py。测试类 TestNPUKimiK3_W4A8_32P_GPQA 继承自 TestNpuAccuracyMultiNodePdMixTestCaseBase，配置了四节点（--nnodes 4）、--tp-size 64、--dp-size 4 的大规模拓扑，并启用了 --speculative-algorithm DSPARK。该测试在 CI 工作流中被添加到多节点混合测试列表，并特别设置了 install_sglang_from_source: true，表明其需要从源码构建运行。
3. 更新 GLM-5.2 W4A8 准确性测试配置：修改现有文件 test/registered/npu/accuracy/glm5_2/test_npu_glm_5_2_w4a8_16p_gpqa.py。主要变更是优化启动配方：启用了数据并行（--dp-size 8， --enable-dp-attention）、调整了内存比例（0.70 -> 0.76）、添加了 --context-length 135000 和 --disable-shared-experts-fusion，并新增了 DeepEP 长序列优化相关的环境变量（DEEPEP_NORMAL_COMBINE_ENABLE_LONG_SEQ 等）。

4. 更新共享测试工具常量：在 python/sglang/test/ascend/e2e/test_npu_performance_utils.py 中：a) 将 DEEPSEEK_V4_FLASH_W8A8_MTP_MODEL_PATH 重命名为 DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH 并更新了路径，指向新的 0731 版本模型；b) 新增 KIMI_K3_W4A8_MODEL_PATH 和 KIMI_K3_DSPARK_MODEL_PATH 常量。
5. 更新现有性能测试引用：同步更新了 test/registered/npu/performance/deepseek_v4_flash/ 下三个性能测试文件，将它们导入的模型路径从旧的 DEEPSEEK_V4_FLASH_W8A8_MTP_MODEL_PATH 统一为新的 DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH。
6. 注册新测试到 CI 工作流：在 .github/workflows/nightly-test-npu.yml 中，为 Kimi-K3 测试添加了条目，指定需要 4 节点 (node_size: 4) 并关联到新的测试脚本。

关键文件：

- test/registered/npu/accuracy/kimi_k3/test_npu_kimi_k3_w4a8_32p_gpqa.py（模块 NPU Kimi-K3 测试；类别 test；类型 test-coverage；符号 TestNPUKimiK3_W4A8_32P_GPQA）：新增的 Kimi-K3 多节点准确性测试，配置复杂，是 PR 的核心交付物之一。
- test/registered/npu/accuracy/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_gpqa.py（模块 NPU DSV4 测试；类别 test；类型 test-coverage；符号 TestNPUDeepSeekV4FlashW8A88PGPQA）：新增的 DeepSeek-V4-Flash 准确性测试，使用 DSPARK 和大 DP 规模，是另一个核心交付物。
- python/sglang/test/ascend/e2e/test_npu_performance_utils.py（模块 测试工具常量；类别 test；类型 test-coverage）：共享测试工具文件，新增了模型路径常量并重命名了 DeepSeek 常量，影响所有引用它的测试。
- .github/workflows/nightly-test-npu.yml（模块 CI 配置；类别 config；类型 infrastructure）：CI 工作流配置，注册新的多节点测试任务，直接影响测试的自动化执行。

关键符号：TestNPUKimiK3_W4A8_32P_GPQA.test_npu_kimi_k3_w4a8_32p_gpqa, TestNPUDeepSeekV4FlashW8A88PGPQA.test_npu_deepseek_v4_flash_w8a8_8p_gpqa, TestNPUGLM_5_2_W4A8_16P_GPQA.test_npu_glm_5_2_w4a8_16p_gpqa

关键源码片段

test/registered/npu/accuracy/kimi_k3/test_npu_kimi_k3_w4a8_32p_gpqa.py

新增的 Kimi-K3 多节点准确性测试，配置复杂，是 PR 的核心交付物之一。

```
# 新增的 Kimi-K3 W4A8 32p 准确性测试类配置
# 该测试在四节点上运行，使用 DSPARK 推测解码算法。
class TestNPUKimiK3_W4A8_32P_GPQA(TestNpuAccuracyMultiNodePdMixTestCaseBase):
    """Test NPU accuracy for Kimi-K3-w4a8 32p four nodes on gpqa_diamond"""
    benchmark_tool = BENCHMARK_TOOL_DEFAULT
    # 关联到包含启动参数的配置字典
    model_config = KIMI_K3_W4A8_32P_MODEL_CONFIG # 包含 model_path, other_args, node_envs
    accuracy = 0.935 # 目标准确率
    datasets = ["gpqa_diamond"]
```

```
    few_shot_num = 0
    eval_batch_size = 32
    generation_config = {
        "max_tokens": 131072,
        "temperature": 1.0,
        "top_p": 0.95,
        # 启用推理模式，可能影响生成内容
        "extra_body": {"reasoning_effort": "max"},
    }
    timeout = 10000 # 超时时间（秒）
    seed = 42

    def test_npu_kimi_k3_w4a8_32p_gpqa(self):
        """Run NPU accuracy test for Kimi-K3-w4a8 32p four nodes on gpqa_diamond"""
        self.run_accuracy() # 调用基类的通用测试运行方法

# 关键的启动参数配置，定义了四节点、大并行度、以及复杂的推测解码环境
KIMI_K3_W4A8_32P_OTHER_ARGS = [
    "--nnodes", 4,
    "--tp-size", 64, # 总张量并行度为 64
    "--enable-dp-attention",
    "--dp-size", 4, # 数据并行度为 4，实现 TP64/DP4 混合拓扑
    "--speculative-algorithm", "DSPARK",
    # ... 其他参数
]
```

test/registered/npu/accuracy/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_gpqa.py

新增的 DeepSeek-V4-Flash 准确性测试，使用 DSPARK 和大 DP 规模，是另一个核心交付物。

```
# DeepSeek-V4-Flash W8A8 8p 准确性测试配置
# 这是一个单节点测试，但启用了大规模数据并行。
DEEPSEEK_V4_FLASH_W8A8_DSPARK_8P_OTHER_ARGS = [
    "--tp-size", 16, # 张量并行度 16
    "--dp-size", 16, # 数据并行度 16，总并行度 256
    "--enable-dp-attention",
    "--speculative-algorithm", "DSPARK",
    "--speculative-draft-model-path", DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH,
    "--speculative-num-draft-tokens", 6,
    "--speculative-dspark-block-size", 5,
    # ... 其他启动参数
]

# 测试类，继承自单节点测试基类
class TestNPUDeepSeekV4FlashW8A88PGPQA(TestNpuAccuracyTestCaseBase):
    """Test NPU accuracy for DeepSeek-V4-Flash W8A8 8p DSPARK GPQA."""
    model = DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH # 使用更新后的模型路径
    # ... 其他配置属性
    def test_npu_deepseek_v4_flash_w8a8_8p_gpqa(self):
```

```
        """Run NPU accuracy test for DeepSeek-V4-Flash W8A8 8p DSPARK GPQA."""
        self.run_accuracy()
```

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

共享测试工具文件，新增了模型路径常量并重命名了 DeepSeek 常量，影响所有引用它的测试。

```
# 测试工具文件中新增和更新的模型路径常量
# 这些常量被多个测试脚本引用，确保版本统一。

# 重命名并更新 DeepSeek-V4-Flash 模型路径，指向 0731 版本
DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH = (
    "/root/.cache/modelscope/hub/models/Eco-Tech/DeepSeek-V4-Flash-0731-w8a8"
)

# 新增 Kimi-K3 主模型和推测解码模型路径
KIMI_K3_W4A8_MODEL_PATH = "/root/.cache/modelscope/hub/models/sgl-npu/Kimi-K3-W4A8"
KIMI_K3_DSPARK_MODEL_PATH = "/root/.cache/modelscope/hub/models/RadixArk/Kimi-K3-
DSpark"
```

评论区精华

本 PR 未提供具体的 review 讨论内容。但从变更本身和提交历史（3 次提交）可以推断出一些隐含的决策和考量：

1. 模型版本统一：第三个提交明确将所有 DeepSeek-V4-Flash 性能测试统一到新的 DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH，表明团队在推进使用最新的 0731 版本模型进行基准测试。
2. 构建策略差异：Kimi-K3 测试被特别配置为 install_sglang_from_source: true，而其他测试使用镜像。这反映了不同模型测试对构建环境有不同要求，可能 Kimi-K3 需要最新的源码特性或定制编译。
3. 配置精细化调整：第二个提交专门调整了 Kimi-K3 的配置参数（如 mem-fraction 0.72, chunked-prefill 8192），并引用了 PR #33465 的启动配方，说明团队在努力使测试配置与已知的最佳实践对齐。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 测试配置正确性风险：新增和修改的测试用例包含大量复杂的环境变量和启动参数（如 HCCL 设置、DeepEP 配置、推测解码参数）。任何配置错误都可能导致测试无法启动、结果不稳定或无法反映真实性能。
  2. 资源占用风险：Kimi-K3 测试需要 4 个节点，GLM-5.2 需要 2 个节点。这些多节点测试会显著增加 CI 的资源消耗和运行时间，可能影响其他测试任务的排期。
  3. 模型路径与版本风险：虽然统一了 DeepSeek 模型路径，但依赖于远端模型仓库（/root/.cache/modelscope/hub/...）中的模型文件。如果模型文件更新、删除或版本不匹配，会导致测试失败。

4. 向后兼容性：此 PR 修改了共享工具文件 test_npu_performance_utils.py 中的常量名称（DEEPSEEK_V4_FLASH_W8A8_MTP_MODEL_PATH -> DEEPSEEK_V4_FLASH_0731_W8A8_MODEL_PATH）。所有引用该常量的其他代码必须同步更新，否则会引发导入错误。本 PR 已在内部完成了更新。 - 影响：用户影响：无直接影响。此变更为内部测试性质。系统影响：增强了 NPU 平台关键模型的端到端准确性验证能力，有助于在发布前捕获模型在 NPU 上的精度回归问题。引入的多个多节点测试会增加 CI 基础设施的负载。团队影响：为 NPU 后端团队提供了更完善的自动化测试套件，便于验证新特性或修复。统一的测试配置和常量管理有助于维护测试代码的一致性。

- 风险标记：测试配置复杂度高，依赖远端模型资源，多节点 CI 资源消耗大

关联脉络

- PR #33465 [NPU] Add Kimi-K3 W4A8 performance case: 提交信息中提到 'align Kimi-K3 case with the PR 33465 launch recipe'，表明本 PR 的 Kimi-K3 测试配置参考了该历史 PR 的启动配方。
- PR #37047 fix(vlm): contain multimodal feature transport failures: 近期合并的 PR 涉及 NPU 和运行时稳定性，本 PR 补充的测试有助于验证在更复杂模型配置下的系统健壮性。