

PR #29686 完整报告

sgl-project/sglang

Update GLM tests to 5.2 and delete redundant tests

合并时间: 2026-06-30 14:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29686>

执行摘要

- 一句话: 清理 DeepSeek V3.2 测试并升级 GLM 测试至 5.2
- 推荐动作: 本 PR 属于测试清理维护, 技术含量较低, 但体现了团队及时清理过时测试资源的良好实践。若关注 GLM 或 DeepSeek 测试架构, 可了解测试注册模式 (register_cuda_ci)。

功能与动机

根据 PR 描述, 删除已请求移除的 CUDA DeepSeek V3.2 夜间测试和 e2e 测试注册, 删除 GB300 GLM5 FP8 夜间测试及其过时套件配置, 并将剩余 GLM 测试更新至 GLM-5.2 checkpoint 以确保测试覆盖最新模型版本。

实现拆解

1. 删除 DeepSeek V3.2 相关测试文件: test/registered/8-gpu-models/test_deepseek_v32.py (200 行)、test/registered/models_e2e/test_deepseek_v32_fp4_mtp_tp.py (73 行)、test/registered/models_e2e/test_deepseek_v32_fp4_mtp_dp.py (68 行), 以及两个 dsa 变体测试文件 test_dsa_dsv32_dp_mtp.py 和 test_dsa_dsv32_tp_mtp.py。
2. 删除 GB300 GLM5 FP8 测试文件 test/registered/gb300/test_glm5_fp8.py (72 行), 该测试被标记为 "not needed"。
3. 将 GLM5 FP8 测试文件 test_glm_51_fp8.py 重命名为 test_glm52_fp8.py, 更新类名从 TestGlm51Fp8 到 TestGlm52Fp8, 方法名从 test_glm51_fp8 到 test_glm52_fp8, 模型路径从 zai-org/GLM-5.1-FP8 改为 zai-org/GLM-5.2-FP8。
4. 将 GLM5 NVFP4 测试文件 test_glm5_nvfp4.py 重命名为 test_glm52_nvfp4.py, 类似更新模型路径和类 / 方法名。
5. 将其他 GLM5 相关测试 (HiSparse、DP MTP、TP MTP、PCG FP4) 也相应重命名为 GLM52 版本, 更新符号和引用。
6. 调整 HiCache 内存分片参数 (mem_fraction_static) 以适应新模型。

关键文件:

- test/registered/8-gpu-models/test_deepseek_v32.py (模块 V32 测试; 类别 test; 类型 deletion; 符号 TestDeepseekV32, test_deepseek_v32_all_variants, test_deepseek_v32_dsa_backends, test_deepseek_v32_b200): 删除的最大文件 (200 行), 包含 DeepSeek V3.2 的主要性能和精度测试, 是本次变更的核心删除目标。

- test/registered/8-gpu-models/test_glm52_fp8.py (模块 GLM 测试; 类别 test; 类型 rename-or-move; 符号 TestGlm51Fp8, TestGlm52Fp8, test_glm51_fp8, test_glm52_fp8) : 从 GLM-5.1 升级到 GLM-5.2 的关键测试文件, 展示了模型路径、类名、方法名和配置的更新。
- test/registered/models_e2e/test_deepseek_v32_fp4_mtp_tp.py (模块 FP4 测试; 类别 test; 类型 deletion; 符号 TestDeepseekV32FP4TPSpec, test_z_bs_1_speed) : 删除的 DeepSeek V3.2 FP4 MTP TP 测试, 包含 GSM8K 精度和速度测试, 是重要清理目标。
- test/registered/gb300/test_glm52_nvfp4.py (模块 NVFP4 测试; 类别 test; 类型 rename-or-move; 符号 TestGlm5Nvfp4, TestGlm52Nvfp4, test_glm5_nvfp4, test_glm52_nvfp4) : 从 GLM-5 NVFP4 升级到 GLM-5.2 NVFP4 的测试文件, 更新了模型路径和类 / 方法名。
- test/registered/gb300/test_glm5_fp8.py (模块 GLM5 测试; 类别 test; 类型 deletion; 符号 TestGlm5Fp8, test_glm5_fp8) : 删除的 GB300 GLM5 FP8 测试, 标记为 "not needed", 清理过时硬件测试。

关键符号: TestGlm52Fp8.test_glm52_fp8, TestGlm52Nvfp4.test_glm52_nvfp4, TestGLM52HiSparse, TestPCGlm52Fp4, TestDeepseekV32

关键源码片段

test/registered/8-gpu-models/test_glm52_fp8.py

从 GLM-5.1 升级到 GLM-5.2 的关键测试文件, 展示了模型路径、类名、方法名和配置的更新。

```
# 该文件从 test_glm_51_fp8.py 重命名并更新为 GLM-5.2
import unittest

from sglang.test.accuracy_test_runner import AccuracyTestParams
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.performance_test_runner import PerformanceTestParams
from sglang.test.run_combined_tests import run_combined_tests
from sglang.test.test_utils import ModelLaunchSettings

# 在 nightly-8-gpu-common suite 中运行
register_cuda_ci(est_time=1800, suite="nightly-8-gpu-common", nightly=True)

# 更新模型路径为 GLM-5.2-FP8
GLM_52_FP8_MODEL_PATH = "zai-org/GLM-5.2-FP8"

COMMON_ARGS = [
    "--trust-remote-code",
    "--reasoning-parser=glm45",
    "--tool-call-parser=glm47",
    "--mem-fraction-static=0.85",
    "--enable-metrics",
]

MTP_ARGS = [
```

```

"--speculative-algorithm=EAGLE",
"--speculative-num-steps=3",
"--speculative-eagle-topk=1",
"--speculative-num-draft-tokens=4",
]

# 类名和方法名从 GLM-5.1 更新为 GLM-5.2
class TestGlm52Fp8(unittest.TestCase):
    """GLM-5.2 FP8 on H200/B200 (8-GPU, tp=8)."""

    def test_glm52_fp8(self):
        dp_args = ["--dp=8", "--enable-dp-attention"]

        variants = [
            ModelLaunchSettings(
                GLM_52_FP8_MODEL_PATH,
                tp_size=8,
                extra_args=COMMON_ARGS,
                variant="TP8",
            ),
            ModelLaunchSettings(
                GLM_52_FP8_MODEL_PATH,
                tp_size=8,
                extra_args=COMMON_ARGS + dp_args,
                variant="TP8+DP8",
            ),
            ModelLaunchSettings(
                GLM_52_FP8_MODEL_PATH,
                tp_size=8,
                extra_args=COMMON_ARGS + dp_args + MTP_ARGS,
                variant="TP8+DP8+MTP",
            ),
        ]

        run_combined_tests(
            models=variants,
            test_name="GLM-5.2-FP8", # 测试名称同步更新
            accuracy_params=AccuracyTestParams(dataset="gsm8k", baseline_accuracy=0.92),
            performance_params=PerformanceTestParams(
                profile_dir="performance_profiles_glm_52_fp8", # profile 目录更新
            ),
        )

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

```

评论区精华

PR 没有实质性技术讨论，仅包含作者触发的 rerun-test 命令和 CI 结果。作者自行合并。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于测试覆盖的减少：删除了 DeepSeek V3.2 的准确性和性能测试（包括 GSM8K 基准测试和速度测试），可能遗漏回归。但推测这些测试已被其他测试替代或模型已不再维护。GLM 测试从 5.1 升级到 5.2，若新 checkpoint 有行为差异可能导致测试失败，但从 CI 结果看已通过验证。
- 影响：对用户无直接影响；对团队：减少了测试维护量，统一了 GLM 测试版本，降低了 CI 耗时。删除的 DeepSeek V3.2 测试如果仍有价值，需要后续补充。
- 风险标记：测试覆盖减少，模型版本升级

关联脉络

- 暂无明显关联 PR