PR #32402 完整报告

sgl-project/sglang

Switch inkling per-commit test to nvfp4

合并时间：2026-08-09 16:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32402

执行摘要

- 一句话：新增 Inkling-Small-NVFP4 真实 checkpoint 的 CI 精度测试
- 推荐动作：值得精读。该 PR 展示了如何为真实量化模型设计 CI 精度门控：使用真实 checkpoint 而非缩小版、基于抽样噪声设置合理阈值、选择与部署一致的硬件和后端参数。对于维护量化模型测试或 CI 门控的工程师有参考价值，但需注意其外部依赖和资源成本。

功能与动机

原 test_inkling.py 启动一个缩小的 checkpoint，只能保证代码路径可运行，无法对答案质量把关；而真实的 NVFP4 checkpoint 才能捕获权重加载或 FP4 内核回归，这类回归会让服务器保持健康但输出错误。PR 标题明确为“Switch inkling per-commit test to nvfp4”，目标是建立针对真实模型和 FP4 数值链路的 per-commit 回归保护。

实现拆解

1. 新增测试文件 test/registered/models_e2e/test_inkling_small_nvfp4.py，通过 register_cuda_ci(est_time=600, stage="extra-b", runner_config="4-gpu-b200") 注册到 CI 的 extra-b 阶段，使用 4 卡 B200 运行。
2. 实现 TestInklingSmallNvfp4 测试类：setUpClass 使用 popen_launch_server 启动真实模型 thinkingmachines/Inkling-Small-NVFP4，配置 --quantization modelopt_fp4、--attention-backend fa4、--fp4-gemm-backend flashinfer_trtllm、--moe-runner-backend flashinfer_trtllm_routed 等关键参数，并设置 Mamba 缓存与显存策略；tearDownClass 通过 kill_process_tree 清理进程。
3. 核心测试 test_gsm8k 调用 sglang.test.few_shot_gsm8k.run_eval 执行 10-shot、200 题、并行度 128 的 GSM8K 评估，断言准确率不低于 GSM8K_THRESHOLD（0.85）。
4. 提交历史中有两个临时 debug 提交（打印 w13 load shapes 和 fp4 strategy），最终未保留；并合并 origin/main 解决了 fused_moe_triton/layer.py 的冲突。
5. CI 验证：原 test_inkling.py 在 1-gpu-h100 上多次失败，新测试在 4-gpu-b200 上通过，确认了切换的合理性。

关键文件：

- test/registered/models_e2e/test_inkling_small_nvfp4.py（模块 回归测试；类别 test；类型 test-coverage；符号 TestInklingSmallNvfp4, setUpClass, tearDownClass, test_gsm8k）：本 PR 的唯一变更文件，新增了基于真实 NVFP4 checkpoint 的

end-to-end 精度测试，是 FP4 数值回归保护的核心载体。

关键符号：test_gsm8k, setUpClass, tearDownClass

关键源码片段

test/registered/models_e2e/test_inkling_small_nvfp4.py

本 PR 的唯一变更文件，新增了基于真实 NVFP4 checkpoint 的 end-to-end 精度测试，是 FP4 数值回归保护的核心载体。

```
# test/registered/models_e2e/test_inkling_small_nvfp4.py
"""Per-commit accuracy + logprob-consistency test for Inkling-Small-NVFP4.

``test_inkling.py`` boots a shrunken checkpoint, so it can only guard that the
code paths run -- an undertrained model has no answer quality to gate on. This
one serves the real NVFP4 checkpoint, which is what catches a weight-load or
FP4-kernel regression that keeps the server healthy while the outputs go wrong.
"""

import os
import unittest
from types import SimpleNamespace
from urllib.parse import urlparse

from sglang.srt.utils import kill_process_tree
from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.test_utils import (
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

# 注册到 CI 测试，预计耗时 600 秒，放入 extra-b 阶段，运行在 4 卡 B200 上
register_cuda_ci(est_time=600, stage="extra-b", runner_config="4-gpu-b200")

# 默认加载真实 FP4 量化 checkpoint，允许用环境变量覆盖
_MODEL_PATH = os.environ.get(
    "INKLING_SMALL_TEST_MODEL_PATH", "thinkingmachines/Inkling-Small-NVFP4"
)

# 实测 10-shot GSM8K 准确率为 0.900，阈值 0.85 约在 2.5 倍抽样标准差以下
# 这样既能捕获真实精度崩塌，又不会被 200 题抽样波动误触发
GSM8K_THRESHOLD = 0.85


class TestInklingSmallNvfp4(CustomTestCase):
    @classmethod
    def setUpClass(cls):
```

```
    cls.model = _MODEL_PATH
    cls.base_url = DEFAULT_URL_FOR_TEST
    # 启动真实服务器的关键参数：modelopt_fp4 量化、FA4 注意力、
    # TRTLLM FP4 GEMM/MoE 后端，确保走的是生产推理路径
    cls.process = popen_launch_server(
        cls.model,
        cls.base_url,
        timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
        other_args=[
            "--tp", "4",
            "--trust-remote-code",
            "--quantization", "modelopt_fp4",
            "--attention-backend", "fa4",
            "--page-size", "128",
            "--fp4-gemm-backend", "flashinfer_trtllm",
            "--moe-runner-backend", "flashinfer_trtllm_routed",
            "--mamba-radix-cache-strategy", "extra_buffer",
            "--swa-full-tokens-ratio", "0.1",
            "--mamba-full-memory-ratio", "0.1",
            "--mem-fraction-static", "0.85",
        ],
        env={**os.environ, "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1"},
    )

@classmethod
def tearDownClass(cls):
    # 确保测试结束时清理服务器进程，避免资源泄漏
    if getattr(cls, "process", None) is not None:
        kill_process_tree(cls.process.pid)

def test_gsm8k(self):
    """Answer quality on the real checkpoint: guards the modelopt_fp4 weight
    load and the FP4 GEMM/MoE kernels against changes that keep the server
    healthy but corrupt the numerics."""
    from sglang.test.few_shot_gsm8k import run_eval as run_few_shot_gsm8k

    url = urlparse(self.base_url)
    metrics = run_few_shot_gsm8k(
        SimpleNamespace(
            num_shots=10,
            data_path=None,
            num_questions=200,
            max_new_tokens=16000,
            parallel=128,
            host=f"http://{url.hostname}",
            port=int(url.port),
        )
    )
    print(f"[{self.__class__.__name__}] gsm8k: {metrics['accuracy']:.3f}")
```

```
        self.assertGreaterEqual(metrics["accuracy"], GSM8K_THRESHOLD)


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

评论区精华

该 PR 没有 review 评论。评论区主要记录了 CI 重跑过程：作者三次重跑 test/registered/models/test_inkling.py 在 1-gpu-h100 上均失败，随后运行新增的 test_inkling_small_nvfp4.py 在 4-gpu-b200 上通过。这反映出原测试在新环境或真实 checkpoint 下不稳定，新测试更贴合实际部署形态。

- 原 Inkling 测试在 H100 上反复失败 (testing): 新测试采用真实 NVFP4 checkpoint 和 B200 4 卡环境，验证通过，替代原缩水模型测试。

风险与影响

- 风险：
  1. 测试依赖外部 HuggingFace checkpoint（thinkingmachines/Inkling-Small-NVFP4），若网络或仓库不可达会导致 CI 失败。
  2. 新增 CI 阶段占用 4 卡 B200 约 10 分钟（est_time 600 秒），增加 GPU 资源消耗和排队时间。
  3. GSM8K_THRESHOLD 0.85 基于当前模型实测 0.900，若模型版本更新导致分数自然下降可能误报。
  4. 测试仅覆盖 few-shot completion 模式，不经过 thinking 路径，无法捕获推理链相关数值问题。
  5. 服务器启动参数全部硬编码，若默认配置或相关默认值变化，可能导致测试启动失败或不再代表真实路径。 - 影响：用户：无任何运行时行为影响。 系统：CI 流水线新增一个 4 卡 B200 的端到端测试阶段，预计耗时约 10 分钟，增加 GPU 资源占用和排队延迟。 团队：为 Inkling-Small-NVFP4 提供 per-commit 精度回归保护，可在 FP4 权重加载或内核改动引入静默错误时快速告警，降低线上事故风险。
- 风险标记：外部 checkpoint 依赖，新增 CI GPU 资源消耗，阈值可能随模型更新漂移，仅覆盖 completion 不覆盖 thinking 路径

关联脉络

- PR #33913 [inkling] Let Anthropic thinking=disabled map to reasoning effort "none": 同为 Inkling 模型相关的变更，涉及 serving_chat 的 reasoning 映射，与本次测试无代码关联，但同属于 Inkling 模型线。