PR #34168 完整报告

sgl-project/sglang

Add deterministic logprob-consistency test for inkling-small nvfp4

合并时间：2026-08-09 22:22

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34168

执行摘要

- 一句话：新增 Inkling-Small 确定性 logprob 一致性测试
- 推荐动作：值得精读。该 PR 展示了如何利用确定性推理将 KL 测试从统计容差变为精确断言（阈值 1e-9），并覆盖普通、prefill cache-hit、decode cache-hit 三种关键路径，设计思路对状态复用类 bug 的防护很有借鉴意义。关注点：_run 封装与三个 helper 的阈值传递方式、cache-hit 变体对滑动窗口交接的覆盖、以及 GSM8K 阈值放宽与 CI 耗时的权衡。

功能与动机

PR body 指出既有 KL 覆盖只跑缩小版 checkpoint 且 tp=1，'never reaches the all-reduce and never scores the real FP4 weights'，真正值得抓的失败（radix cache 状态复用、conv/mamba checkpoint 复位）会被宽松阈值淹没在浮点噪声中。随着 #34159 合入，路径上每个 kernel 都是 batch-invariant，prefill 与 decode 逐 bit 一致，因此可以做出精确断言（三个 helper 实测均为 0），阈值只需作为 stray-ulp 地板。

实现拆解

1. 引入 kl_test_utils 三个 helper 并别名化：从 sglang.test.kl_test_utils 导入 test_input_output_logprobs_match_helper、..._prefill_cache_hit_helper、...decode_cache_hit_helper，使用 as 别名改为 assert_logprobs_match*，避免被 pytest 收集为独立 test_ 用例。
2. 调整原 accuracy 测试：GSM8K_THRESHOLD 从 0.85 降至 0.80（按 ~4.5 sigma 的采样噪声设置），max_new_tokens 从 16000 降至 512，以压缩整个文件的 CI 耗时；register_cuda_ci 的 est_time 从 600 提升到 1200，覆盖新增确定性测试的开销。
3. 新增确定性测试类 TestInklingSmallNvfp4Deterministic：独立启动一个 server，复用同一组生产启动参数（tp=4、modelopt_fp4、fa4、flashinfer_trtllm 等），额外追加 --enable-deterministic-inference 和 --disable-prefill-cuda-graph，以确保数值路径完全确定。
4. 三个测试方法分别调用普通、prefill cache-hit、decode cache-hit 三种 helper，统一使用 _run 封装，传入 KL_DIV_THRESHOLD=1e-9 与 KL_MAX_NEW_TOKENS=1024，其中 KL_MAX_NEW_TOKENS 特意超过 512 token 滑动窗口，让 decode 承担窗口交接以暴露状态恢复问题。

5. 配套说明：本次纯测试变更，无生产代码改动，CI 运行在 4-gpu-b200 runner 上，属于 extra-b 阶段。

关键文件：

- test/registered/models_e2e/test_inkling_small_nvfp4.py（模块 模型测试；类别 test；类型 test-coverage；符号 TestInklingSmallNvfp4Deterministic, setUpClass, tearDownClass, _run）：唯一变更文件，新增 TestInklingSmallNvfp4Deterministic 类，复用 kl_test_utils 三个 helper 在确定性推理下做精确 logprob 一致性断言，并收紧 GSM8K 阈值与 token 数。

关键符号：test_input_output_logprobs_match, test_input_output_logprobs_match_prefill_cache_hit, test_input_output_logprobs_match_decode_cache_hit, _run, setUpClass, tearDownClass

关键源码片段

test/registered/models_e2e/test_inkling_small_nvfp4.py

唯一变更文件，新增 TestInklingSmallNvfp4Deterministic 类，复用 kl_test_utils 三个 helper 在确定性推理下做精确 logprob 一致性断言，并收紧 GSM8K 阈值与 token 数。

```
# 三个 helper 都来自 kl_test_utils，导入时用别名，避免 pytest 将其收集为独立测试用例。
from sglang.test.kl_test_utils import (
    test_input_output_logprobs_match_helper as assert_logprobs_match,
    test_input_output_logprobs_match_prefill_cache_hit_helper as assert_logprobs_match_prefill_cache_hit,
    test_input_output_logprobs_match_decode_cache_hit_helper as assert_logprobs_match_decode_cache_hit,
)

# 实测该配置下三个指标都恰好为 0（逐 bit 一致），因此阈值只是防 stray-ulp 的地板；
# 真正的状态复用 bug 产生的分歧会高出数个数量级。
KL_DIV_THRESHOLD = 1e-9

# 生成长度超过 512 token 的滑动窗口，让 decode 阶段完整承担窗口从 prompt token
# 到生成 token 的交接，从而暴露 conv/mamba checkpoint 或 radix cache 前缀恢复问题。
KL_MAX_NEW_TOKENS = 1024


class TestInklingSmallNvfp4Deterministic(CustomTestCase):
    """确定性推理下 prefill 与 decode 必须逐 bit 一致，否则即为状态复用 bug。

    单独启动 server：accuracy 用例必须保持生产数值路径，不能共享该配置。
    """

    @classmethod
    def setUpClass(cls):
        cls.model = _MODEL_PATH
        cls.base_url = DEFAULT_URL_FOR_TEST
```

```
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--tp", "4",
                "--quantization", "modelopt_fp4",
                "--attention-backend", "fa4",
                "--enable-deterministic-inference",
                "--disable-prefill-cuda-graph", # 确定性推理要求关闭 prefill CUDA graph
            ],
            env={**os.environ, "SGLANG_ENABLE_UNIFIED_RADIX_TREE": "1"},
        )

    def _run(self, helper):
        # 三个 helper 均接受 (base_url, model_thresholds, model, ...)
        helper(
            self.base_url,
            {self.model: {"kl_div": KL_DIV_THRESHOLD}},
            self.model,
            max_samples=32,
            max_new_tokens=KL_MAX_NEW_TOKENS,
            trust_remote_code=True,
        )

    def test_input_output_logprobs_match(self):
        self._run(assert_logprobs_match)

    def test_input_output_logprobs_match_prefill_cache_hit(self):
        # prefill cache-hit：前缀复用 radix cache 后返回，验证状态恢复一致性
        self._run(assert_logprobs_match_prefill_cache_hit)

    def test_input_output_logprobs_match_decode_cache_hit(self):
        # decode cache-hit：窗口滑过 512 token 后再次命中缓存，验证 mamba/conv 状态交接
        self._run(assert_logprobs_match_decode_cache_hit)
```

评论区精华

该 PR 没有 review 评论，评论区（Issue comments）集中在 CI 重跑：作者四次发起 /rerun-test，前三次在 4-gpu-b200 上失败，第四次通过并出现 🤔。失败原因未在评论中说明，但最终被判定为可接受（作者以 bypass-fastfail 标签处理）。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. CI 时长与资源：est_time 从 600 提到 1200 秒，且新增一个独立 server 在 4 卡 B200 上运行，显著增加 CI 队列负担。

2. GSM8K 阈值放宽：从 0.85 降到 0.80，可能降低对真实数值回退的敏感度；作者用 4.5 sigma 论证，但实际模型精度为 0.90，仍有约 10% 的余量，若精度缓慢退化可能漏报。
3. 确定性依赖：精确断言依赖 #34159 使所有 kernel batch-invariant。若未来某个 kernel（如 flashinfer fp4 gemm）失去该性质，测试会直接失败，这是预期行为，但可能造成误报。
4. 阈值 1e-9 的 flaky 风险：实测为 0，但若确定性推理存在偶发 ulp 差异（如不同驱动版本），可能产生偶发失败。
5. max_new_tokens 降至 512：对长答案 GSM8K 问题的覆盖可能不足，但该用例本身是 few-shot completion，512 token 通常足够。 - 影响：对用户无直接影响；对 CI 系统增加了一个 4 卡 B200 的长时间运行测试（约 20 分钟量级）；对团队而言，该测试为 Inkling-Small-NVFP4 提供了精确的状态复用回归防护，能捕获 radix cache 前缀恢复、conv/mamba checkpoint 复位等仅靠精度阈值难以发现的 bug，并可作为后续其他模型（如 Kimi-K3）确定性测试的模板。 - 风险标记：CI 时长翻倍，GSM8K 阈值放宽，依赖确定性推理保证

关联脉络

- PR #32402 Switch inkling per-commit test to nvfp4: 该 PR 创建了本文件（Inkling-Small-NVFP4 真实 checkpoint 的 CI 精度测试），本次在此基础上扩展确定性一致性覆盖，属于同一功能线的演进。
- PR #34159 Fix deterministic inference all-reduce for tp>1: PR body 明确依赖 #34159 使路径上 kernel batch-invariant，prefill 与 decode 逐 bit 一致，这是本测试精确断言的前提。