

PR #34356 完整报告

sgl-project/sglang

Add bit-exact hicache logprob-consistency test

合并时间: 2026-08-12 01:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34356>

执行摘要

- 一句话: 新增 HiCache 位精确多轮 logprob 一致性测试
- 推荐动作: 值得精读, 重点学习其测试设计: 用多轮分支 harness 让 cache hit 落在多种前缀长度, 再用 write_through、紧 pool、跨 track decode 的组合把隐蔽状态层缺陷放大到 KL=0 可观测。后续改动 HiCache 状态结构或 Mamba track 记账时, 应保留该用例并保持其位精确语义。

功能与动机

PR body 明确指出: HiCache 把 Mamba 状态搬到 host 内存再搬回, 这一往返不允许改变任何 bit, 所以正确门槛是位精确而非容差; `KL_DIV_THRESHOLD = 1e-9` 只是防 stray ulp 的地板, 松阈值会掩盖这类缺陷。当前 main 在该配置下测不到 0, 两个独立状态层缺陷叠在一个数字下: `#34184` 已合入修复前 96%, 剩余部分由 `#29792` 覆盖, 而这个测试的作用就是让第二个缺陷可见。

实现拆解

1. 引入多轮 harness: 从 `sglang.test.kl_multiturn_utils` 导入 `assert_multiturn_decode_cache_hit` 与 `make_mamba_decode_assert`, 从 `kl_test_utils` 导入 `get_input_ids`, 并保持 `test_` 前缀别名避免被 `pytest` 收集。
2. 新增确定性数据生成: `_random_suffixes` 用固定 seed 生成可复现的随机后缀; `KL_HICACHE_TRACK_INTERVAL = 128` 与页大小对齐, 保证 decode 在单轮内跨过 track 边界。
3. 新增测试类 `TestInklingSmallNvfp4HiCacheDeterministic`: `setUpClass` 以 `--enable-hierarchical-cache`、`--hicache-write-policy write_through`、`--hicache-io-backend direct`、`--hicache-mem-layout page_first_direct` 启动专用 server, 并用 `--mamba-track-interval 128`、紧 pool、`--enable-deterministic-inference` 等配置放大缺陷可观测性。
4. 核心用例 `test_multiturn_decode_cache_hit_over_hicache`: 3 组 × 3 分支共享 512 token 前缀后分叉, 跨两轮追加随机后缀, 使 cache hit 落在多种前缀长度; 同时用 `make_mamba_decode_assert` 断言每次 hit 都落在 track 边界上, logprob 对比与结构断言覆盖不同失败模式。

5. CI 配套: register_cuda_ci 的 est_time 从 1200 提升到 1700, 因为多启动一个 server ; 无 schema、配置或部署改动。

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 _random_suffixes, TestInklingSmallNvfp4HiCacheDeterministic, setUpClass, tearDownClass) : 唯一改动文件: 新增 TestInklingSmallNvfp4HiCacheDeterministic 多轮多分支 HiCache 位精确测试, 并导入 kl_multiturn_utils harness; 该用例把 #29792 覆盖的第二个状态层缺陷暴露出来。

关键符号: _random_suffixes, TestInklingSmallNvfp4HiCacheDeterministic.setUpClass, TestInklingSmallNvfp4HiCacheDeterministic.tearDownClass, TestInklingSmallNvfp4HiCacheDeterministic.test_multiturn_decode_cache_hit_over_hicache

评论区精华

PR 本身没有 review 评论, 核心讨论集中在 PR body 和 Issue 重跑记录。作者强调 $KL_DIV_THRESHOLD = 1e-9$ 是防单个 ULP 的地板而非容差, 并明确该用例依赖 #29792, 合入前会保持红色; 评论区两次 /rerun-test 第一次失败、第二次通过, 最终转绿。

- CI 首次重跑失败与依赖转绿 (other): 第二次重跑通过后合入; 首次失败与目标缺陷或 CI 环境有关, 未能单独定位。

风险与影响

- 风险: 纯测试变更, 对运行时无回归风险, 但有三点值得注意: 一是位精确断言 (KL 阈值 $1e-9$) 对无关 ULP 抖动敏感, 任何浮点环境差异都可能让 extra-b 变红; 二是 est_time 从 1200 提升到 1700, 4-gpu-b200 上的 CI 时长与资源占用增加; 三是首次 rerun 失败、第二次通过, 说明用例对服务器启动或环境状态存在一定敏感性, 后续可能 flaky。另外该用例依赖 #29792, 合入顺序错乱时会持续红色。
- 影响: 影响范围集中在 CI 测试矩阵 extra-b 阶段和 HiCache/Mamba 状态一致性防线: 为 #34184、#29792 修复的两个缺陷建立回归保护, 并对后续 HiCache 状态迁移、Mamba track 记账改动施加位精确约束。用户的推理路径、API 行为、部署配置均不受影响。
- 风险标记: 依赖 #29792 合入, 位精确断言对 ULP 敏感, CI 时长增加 500 秒, 首次重跑失败后通过

关联脉络

- PR #29792 [HiCache] Fix Mamba track-boundary bookkeeping under overlap scheduling: 本 PR 明确依赖它: 该修复合入后, 新增用例的测量值才从 $9.53e-05$ 降到 0.0, 测试的作用是让这个剩余缺陷可见。
- PR #34341 [npu] [bugfix] Fix HiCache MHA backup for NPU: 同为 HiCache 状态备份 / 恢复路径的 bugfix (NPU MHA backup), 与本 PR 守护的 host 往返位精确问题同属 HiCache 回归防线。

- PR #34405 Fix flaky decode cache-hit check in Inkling test: 同文件后续修复 decode cache-hit 检查的 flaky, 说明该测试体系仍在持续打磨, 验证了本 PR 引入的多轮 cache-hit 断言与 CI 稳定性之间的张力。