

PR #35143 完整报告

sgl-project/sglang

Add bit-exact class for MTP

合并时间: 2026-08-17 19:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35143>

执行摘要

- 一句话: 为 MTP 投机解码路径新增 bit-exact 测试类并关闭 sheared bias
- 推荐动作: 值得精读。类 docstring 本身就是一份简短设计文档, 展示了 "用 bit-exact 测试作为正确性 oracle 暴露底层 kernel 数值不一致" 的方法论: 测试不是只求通过, 而是把根因修复的验收条件定义为 "移除 override 后测试仍通过"。建议关注两点: 一是 SGLANG_OPT_USE_INKLING_SHEARED_BIAS=0 的撤销时机 (issue 34899); 二是后续是否将 MTP 场景扩展到 multiturn 与更多状态保存路径。

功能与动机

MTP 目前没有 bit-exact 覆盖: PR body 指出 "MTP has no bit-exact coverage today. #29792 was a wrong-slot pick in the non-speculative decode track save; the speculative verify writes its checkpoint from its own path, and nothing exercises that path, so a matching defect there reaches production unguarded." 即投机 verify 路径的 checkpoint 保存是独立缺陷面, 普通 decode 路径的测试无法保护它。关联 issue 34899 的 Roadmap 也明确把 MTP 列为待覆盖项, 并确立了 "Any nonzero KL is a KV/state reuse bug" 的验收契约。

实现拆解

1. 测试类入口: 在 test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hybrid_bitexact.py 文件末尾追加 TestUnifiedHybridMTPBitExact(CustomTestCase), 复用同文件已有的 KL_DIV_THRESHOLD = 1e-9、MAX_NEW_TOKENS = 1024、_base_args() 和 kill_process_tree 等基建。
2. 服务器启动配置: setUpClass 以 --speculative-algorithm EAGLE 加 --enable-multi-layer-eagle 启动服务器; --speculative-num-steps 2 精确匹配 checkpoint 携带的两个 MTP 头 (第三个 step 无权重, loader 拒绝启动而非运行错误校准的 draft); --speculative-eagle-topk 1 与 --speculative-num-draft-tokens 3 固定 draft 行为; --chunked-prefill-size 16384 控制 prefill 分块。
3. 确定性环境: 通过进程级环境变量注入 SGLANG_ENABLE_UNIFIED_RADIX_TREE=1 开启统一 radix 树, SGLANG_OPT_USE_INKLING_SHEARED_BIAS=0 关闭 sheared relative-bias 路径。类 docstring 明确该 override 是 workaround 而非偏好, 根因 (shear 依赖 max_seqlen_q 导致 verify/prefill 落在不同 tile) 跟踪在 issue 34899。

4. 校验用例：_run 统一封装 32 样本、1024 new tokens 的断言入口；
test_logprobs_match 调用 assert_logprobs_match 验证 prefill 与 decode 的 logprob 位级一致，test_decode_cache_hit 调用 assert_decode_cache_hit 验证 decode 缓存命中路径一致。
5. CI 配套：register_cuda_ci 的 est_time 从 790 上调至 1150，反映新增 server 启动与两个用例的开销，仍注册在 base-b 阶段、1-gpu-large runner。

关键文件：

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hybrid_bitexact.py (模块 位精确测试；类别 test；类型 test-coverage；符号 TestUnifiedHybridMTPBitExact, setUpClass, tearDownClass, _run)：本 PR 唯一变更文件，新增 TestUnifiedHybridMTPBitExact 测试类，以 EAGLE 投机解码模式启动 Inkling 服务器，为 MTP 的 checkpoint 保存路径建立 bit-exact 护栏；文档化 sheared bias 数值发散的量化结果与 workaround 理由。

关键符号：TestUnifiedHybridMTPBitExact.setUpClass,
TestUnifiedHybridMTPBitExact.tearDownClass, TestUnifiedHybridMTPBitExact._run,
TestUnifiedHybridMTPBitExact.test_logprobs_match,
TestUnifiedHybridMTPBitExact.test_decode_cache_hit

关键源码片段

[test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hybrid_bitexact.py](#)

本 PR 唯一变更文件，新增 TestUnifiedHybridMTPBitExact 测试类，以 EAGLE 投机解码模式启动 Inkling 服务器，为 MTP 的 checkpoint 保存路径建立 bit-exact 护栏；文档化 sheared bias 数值发散的量化结果与 workaround 理由。

```
# 关键：这是 MTP 投机解码路径的 bit-exact 测试类。  
# 背景：MTP 的 checkpoint 由投机 verify 路径写入，此前无任何测试覆盖该路径，  
# #29792 这类 wrong-slot bug 因而可能流入生产。本类把 prefill 与 decode 的  
# logprob 一致性当作 oracle：任何非 0 KL 都视为状态复用缺陷。
```

```
class TestUnifiedHybridMTPBitExact(CustomTestCase):  
    # 在 setUpClass 中启动一个专用服务器，配置是 load-bearing 的：  
    # --speculative-num-steps 2 匹配该 checkpoint 的两个 MTP 头，  
    # 第三个 step 没有权重，loader 会拒绝启动而非运行错误校准的 draft。  
    @classmethod  
    def setUpClass(cls):  
        cls.model = _MODEL_PATH  
        cls.base_url = DEFAULT_URL_FOR_TEST  
  
        other_args = _base_args() + [  
            '--speculative-algorithm', 'EAGLE',  
            '--enable-multi-layer-eagle',  
            '--speculative-num-steps', '2',  
            '--speculative-eagle-topk', '1',  
            '--speculative-num-draft-tokens', '3',
```

```

        '--chunked-prefill-size', '16384',
    ]
    if _MODEL_REVISION:
        other_args += ['--revision', _MODEL_REVISION]

    cls.process = popen_launch_server(
        cls.model,
        cls.base_url,
        timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
        other_args=other_args,
        env={
            **os.environ,
            'SGLANG_ENABLE_UNIFIED_RADIX_TREE': '1',
            # 关闭 sheared relative-bias 路径：该路径按 max_seqlen_q 做 shear,
            # verify pass (多个 query) 与 prefill pass 的同一 (q, k) 对会落在
            # 不同 tile 上。实测 12/16 tokens 发散 (最大 ldlogprob 8.1e-03) ,
            # 关闭后 0/16、精确 0.0。这是 workaround 而非偏好,
            # 根因修复跟踪在 issue 34899, 达标后应移除本行。
            'SGLANG_OPT_USE_INKLING_SHEARED_BIAS': '0',
        },
    )

    @classmethod
    def tearDownClass(cls):
        if getattr(cls, 'process', None) is not None:
            kill_process_tree(cls.process.pid)

    # 两个用例共用 32 样本 / 1024 new tokens 的校验入口,
    # 仅 helper 不同：一个验证 logprob 位级一致，一个验证 decode 缓存命中路径。
    def _run(self, helper):
        helper(
            self.base_url,
            {self.model: {'kl_div': KL_DIV_THRESHOLD}},
            self.model,
            max_samples=32,
            max_new_tokens=MAX_NEW_TOKENS,
            trust_remote_code=True,
        )

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

    def test_decode_cache_hit(self):
        self._run(assert_decode_cache_hit)

```

评论区精华

本 PR 没有实质 review 线程（仅有 1 条 [/rerun-test](#) 命令与对应 CI 结果），核心设计讨论由作者在 PR body 与测试 docstring 中完成：

- sheared bias 的数值发散：作者指出 "The sheared relative-bias path shears on max_seqlen_q, so a verify pass (a few queries) lands the same absolute (q, k) pair on a different tile than a prefill pass does, and one output element of the attention differs", 并给出实测对比：默认开启时 12/16 tokens 发散、最大 ldlogprob 8.1e-03；关闭后 0/16、精确 0.0。
- 配置约束的取舍：--speculative-num-steps 2 是 load-bearing 配置，与 checkpoint 的 MTP 头数绑定；服务配方中的 --speculative-use-rejection-sampling 被有意省略，因为与 --enable-deterministic-inference 互斥。
 - sheared relative-bias 导致 verify/prefill 数值发散 (correctness): 以 SGLANG_OPT_USE_INKLING_SHEARED_BIAS=0 作为 workaround 达成 bit-exact；根因修复跟踪在 issue 34899，修复后应移除 override，并由本测试类证明 query-count 不变性。
 - --speculative-num-steps 必须匹配 MTP 头数 (question): 固定 --speculative-num-steps 2，与 checkpoint 权重结构绑定，属于 load-bearing 配置。
 - rejection sampling 与确定性推理互斥 (question): 省略该参数以保持确定性推理，保证 bit-exact 测试前提成立。

风险与影响

- 风险：
 1. workaround 依赖：SGLANG_OPT_USE_INKLING_SHEARED_BIAS=0 是绕过而非修复。一旦 issue 34899 的 shear 修复落地，必须同步移除该环境变量，否则测试将不再覆盖真实默认路径，产生虚假安全感。
 2. 覆盖边界有限：只覆盖 single-turn 的 logprob 一致与 decode cache hit；MTP 场景的 multiturn、分支切换（同文件其他类覆盖的是非 MTP 场景）以及 DSpark/DFlash 路径仍无 bit-exact 测试。
 3. CI 时长增加：est_time 从 790 增至 1150（约 45%），base-b 的 1-gpu-large 流水线被显著拉长。
 4. 外部模型依赖：测试强依赖 thinkingmachines/Inkling 的 test revision 与两个 MTP 头的权重结构，模型 revision 变化时 --speculative-num-steps 2 的合法性需要重新确认。
 - 影响：对用户无运行时影响（纯测试变更）。对 CI 系统，base-b 阶段新增约 6 分钟的服务器生命周期开销，est_time 调整为 1150。对团队而言，本 PR 堵住了 MTP 投机验证路径 checkpoint 保存缺陷流入生产的窗口（#29792 同类问题），并把 sheared bias 的 query 数依赖发散固化为可回归验证的显式约束；同时为 issue 34899 roadmap 中的 DSpark/DFlash 等后续覆盖项提供了可复用的测试模板。
 - 风险标记：sheared bias 依赖环境变量绕过，测试覆盖路径有限（无 multiturn），CI 时长增加约 45%，强依赖模型 revision 与 MTP 头配置

关联脉络

- PR #35042 Fix sconv track refresh on graph capture: 本 PR 的前置依赖：修复 CUDA graph 捕获下 sconv track 索引刷新崩溃，否则 MTP 与 CUDA graph 同时开启时 server 无法启动。

- PR #29792 Wrong-slot pick in non-speculative decode track save: 本 PR 防护的缺陷原型: 非投机 decode 路径的 checkpoint 保存存在 wrong-slot 选择, 投机 verify 路径存在同类风险的未被测试覆盖。
- PR #34607 Per-commit bit-exact guards: 同一 bit-exact KL 一致性方法论系列, issue 34899 中列为支持性工作, 本 PR 兑现了 roadmap 中的 MTP 覆盖项。
- PR #34168 Real-weight consistency coverage: 同一 bit-exact 覆盖系列的另一环节, 与真实权重下的 KL 一致性测试方法论一脉相承。