

PR #33277 完整报告

sgl-project/sglang

[CI] Fix stale CPU test fixtures

合并时间: 2026-08-02 22:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33277>

执行摘要

- 一句话: 修复两个过期 CPU 测试夹具, 字段对齐新配置
- 推荐动作: 不值得精读, 但值得作为「跨 PR fixture 漂移」案例被记录: 它和 #33254 一起展示了当多个功能 PR 并行演进配置结构时, 手写测试桩需要及时同步。若后续有梳理测试基建的计划, 可考虑把这类 SimpleNamespace 桩替换为真实配置工厂方法, 从根上避免同类 CI 失败。

功能与动机

PR body 明确指出这两个 fixture 已过期: 一是 reasoning 配置 fixture 在 #33025 与 #30177 交叉合入后过期, 二是 #25545 引入了 draft attention backend 字段但测试桩未同步。过期字段会导致 base-a-test-cpu 套件中的单测因缺少属性或字段名不匹配而失败, 属于典型的跨 PR 演进带来的测试漂移。

实现拆解

1. 补全 `test_model_overrides.py` 的 `_view` 测试桩: 在 `test_page_constraint_passes_at_callable_level` 内的 `_view` 辅助函数默认字典中新增 `speculative_draft_attention_backend=None`。该字段由 #25545 (Gemma 4 MTP draft attention backend 支持) 引入到 `ResolvedView`, 缺少它时 `_mla_backend_page_constraints` 等 page 约束校验在读取 `ResolvedView` 属性时会抛 `AttributeError`。
2. 更新 `test_batch_result_processor_hidden_states.py` 的 `model_config` 字段名: 将 `SimpleNamespace(think_end_id=None)` 改为 `SimpleNamespace(think_end_ids=None)`。这是 reasoning 配置在 #33025 (Kimi K3 reasoning parser) 与 #30177 (`return_hidden_states="last"`) 交叉演进后的字段名变更, `SchedulerBatchResultProcessor` 构造时按新字段名读取, 旧名称已失效。
3. 验证与收尾: 作者先在本机运行两个测试文件确认通过, 再发起 `/rerun-test` 让 CI 在 `ubuntu-latest` 上分别执行 `test_batch_result_processor_hidden_states.py` 与 `test_model_overrides.py`, 两个测试均通过; 随后 `/tag-and-rerun-ci` 触发完整 CI 并由维护者 BBuf 批准合入。

配套改动仅限测试文件本身, 无生产代码、配置或部署变更。

关键文件:

- test/registered/unit/test_model_overrides.py (模块 模型覆盖; 类别 test; 类型 test-coverage; 符号 _view, test_page_constraint_passes_at_callable_level) : 在 _view 测试桩默认参数中补上 #25545 引入的 speculative_draft_attention_backend 字段, 修复 test_page_constraint_passes_at_callable_level 因 ResolvedView 缺属性而失败的问题。
- test/registered/unit/managers/test_batch_result_processor_hidden_states.py (模块 结果处理器; 类别 test; 类型 test-coverage; 符号 _make_processor) : 将 model_config 的 think_end_id 改为 think_end_ids, 适配 #33025 与 #30177 交叉后 reasoning 配置字段名的演进, 修复 _make_processor 构造失败。

关键符号: _view, _make_processor, test_page_constraint_passes_at_callable_level

关键源码片段

test/registered/unit/test_model_overrides.py

在 _view 测试桩默认参数中补上 #25545 引入的 speculative_draft_attention_backend 字段, 修复 test_page_constraint_passes_at_callable_level 因 ResolvedView 缺属性而失败的问题。

```
def _view(**kw):
    # 手写测试桩: 用 SimpleNamespace 模拟 ResolvedView 的字段
    defaults = dict(
        attention_backend=None,
        decode_attention_backend=None,
        prefill_attention_backend=None,
        # #25545 引入 draft attention backend 后, ResolvedView 新增该字段;
        # 若不补上, page 约束校验在读取属性时会抛 AttributeError
        speculative_draft_attention_backend=None,
        page_size=1,
    )
    defaults.update(kw)
    return ResolvedView(SimpleNamespace(**defaults))
```

test/registered/unit/managers/test_batch_result_processor_hidden_states.py

将 model_config 的 think_end_id 改为 think_end_ids, 适配 #33025 与 #30177 交叉后 reasoning 配置字段名的演进, 修复 _make_processor 构造失败。

```
def _make_processor(server_mode: str = "full") -> SchedulerBatchResultProcessor:
    metrics_reporter = Mock()
    metrics_reporter.num_generated_tokens = 0
    metrics_reporter.forward_ct_decode = 0
    return SchedulerBatchResultProcessor(
        is_generation=True,
        disaggregation_mode=None,
        enable_overlap=False,
        enable_overlap_mlx=False,
        server_args=SimpleNamespace(
```

```
enable_metrics=False,
enable_hispase=False,
enable_return_hidden_states=True,
return_hidden_states_mode=server_mode,
),
# #33025 与 #30177 交叉后, reasoning 配置字段已从 think_end_id
# 演进为复数 think_end_ids, 旧字段名会导致构造失败
model_config=SimpleNamespace(think_end_ids=None),
token_to_kv_pool_allocator=Mock(),
tree_cache=None,
hispase_coordinator=None,
req_to_token_pool=None,
decode_offload_manager=None,
metrics_collector=None,
metrics_reporter=metrics_reporter,
draft_worker=None,
model_worker=Mock(),
logprob_result_processor=None,
output_streamer=Mock(),
abort_request=lambda *args, **kwargs: None,
)
```

评论区精华

本 PR 没有实质性的技术争论。Review 层面 BBuf 直接 APPROVED 且未留评论；Issue 评论中主要是自动化流程：作者发起 `/rerun-test` 指定两个单测文件，GitHub Actions bot 返回两个测试均通过的运行结果；另有 `gemini-code-assist[bot]` 的停用提示，说明该 bot 已停止代码审查活动，不影响本 PR。整体属于低分歧的纯测试修复。

- CI 失败定位与 `/rerun-test` 验证 (test): 两个受影响测试文件在本地与 CI 均通过，无代码分歧，可直接合入。
- Review 批准 (other): 无实质技术讨论，作为低风险测试修复直接合入。

风险与影响

- 风险：风险极低：改动仅涉及两个测试文件内的测试桩字段，无任何生产代码路径变更。唯一值得注意的隐患是这些 fixture 仍然依赖手写的 SimpleNamespace 模拟配置对象，字段列表与真实配置演进完全靠人工同步——本次与 #33254 都是在 #25545 等 PR 之后出现同类漂移，说明这种手写桩在字段频繁演进时容易再次过期。若后续配置对象能改为从真实 `server_args` 派生，可从根本上消除这类 CI 噪声。
- 影响：影响范围限定在 CI 测试领域：修复 `base-a-test-cpu` 套件中两个单测的失败，恢复 CPU 测试流水线的稳定性，减少维护者处理 CI 噪声的时间。对生产用户、推理服务运行无任何影响；对团队而言，这类小修是多个功能 PR 并行合入后必要的配套维护，且通过 `/rerun-test` 精准验证，展示了高效的验证流程。
- 风险标记：纯测试文件变更，手写测试桩易随配置演进再次漂移

关联脉络

- PR #25545 [Spec] Add trtllm_mha support for Gemma 4 MTP draft attention backend: 本 PR 补上的 speculative_draft_attention_backend 字段正是 #25545 引入的，属于同一配置结构的测试桩同步。
- PR #30177 [Feature] Support return_hidden_states="last": 与 #33025 交叉导致 reasoning 配置 fixture 过期，本 PR 将 think_end_id 更新为 think_end_ids。
- PR #33025 [Kimi K3] Add reasoning, tool-call, and OpenAI serving support: Kimi K3 reasoning parser 相关配置字段演进，是 hidden states 测试桩更新 think_end_ids 的直接原因之一。
- PR #33254 [CI] Add speculative_draft_attention_backend to the page-constraint test view: 同因 #25545 导致的同类测试桩过期修复，且同样涉及 test_model_overrides.py，体现跨 PR fixture 漂移的重复模式。