PR #25348 完整报告

sgl-project/sglang

[UnifiedTree]: Add nightly hicache ci for dsa model

合并时间：2026-05-15 16:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25348

PR 分析报告：[UnifiedTree]: Add nightly hicache ci for dsa model

执行摘要

本次 PR 为 GLM-5.1-FP8 模型在 UnifiedRadixTree + HiCache L3 文件后端场景下新增了一个夜间 CI 测试用例，通过 GSM8K 双通道验证缓存正确性，填补了该模型在 HiCache 回归测试中的空白。

功能与动机

HiCache 是 SGLang 的分层缓存 offload 方案，此前已在 DeepSeek V4 等模型上测试，但 GLM-5 作为新支持模型，其 L3 文件后端场景缺乏自动化回归覆盖。本 PR 通过添加一个复用 GSM8K 混合测试方法的测试类，确保 GLM-5 在 HiCache 下的输出精度在两轮推理间保持稳定（精度差 ≤ 2%）。

实现拆解

1. 新增 GSM8KTwoPassMixin 类(同一文件)：封装了两轮 GSM8K 评测并比较精度差的通用逻辑，便于其他模型复用。它通过 _run_gsm8k 调用 few-shot 评测脚本，通过 _flush_cache 清理服务器端缓存，test_gsm8k_two_passes 是测试入口。
2. 新增 TestGLM5HiCacheL3GSM8K 测试类(同一文件)：继承以上 Mixin 和 CustomTestCase，在 setUpClass 中以 8 卡 TP 启动 GLM-5.1-FP8，配置 --hicache-io-backend file 等 L3 参数，并创建临时缓存目录。tearDownClass 负责清理进程和目录。
3. 注册 Nightly CI：通过 register_cuda_ci(suite="nightly-8-gpu-h200", nightly=True) 将测试纳入 8×H200 夜间套件。
4. 小范围配置调整：给已有的 Mamba 测试增加了 --max-mamba-cache-size 500 参数，以保持配置一致性。

请参见 key_files[0].annotated_snippet_markdown 中的完整代码片段。核心逻辑在 GSM8KTwoPassMixin.test_gsm8k_two_passes 中：两轮推理之间调用 _flush_cache，确保缓存不影响第二次准确率，并比较两次结果的差异。

评论区精华

本 PR 没有实质性的 review 讨论。作者通过 /rerun-test 触发了两次 CI 执行，均在 8×H200 上通过，并贴出了测试结果截图（准确率约 0.93），验证了测试有效性。

风险与影响

- 风险：测试依赖 8 卡 GPU 集群且需较长时间（约 15 分钟），若 CI 资源紧张可能超时；临时目录清理逻辑在异常退出时可能泄露。
- 影响：只有测试管线受影响，用户无感知。提高了 GLM-5 + HiCache L3 场景的回归保障，为后续模型扩展提供示例。

关联脉络

本 PR 是 HiCache 功能在更多模型上落地的持续工作。此前 PR #24691 在 UnifiedRadixTree 中加入了 HiCache 支持，PR #25369 更新了 DeepSeek V4 的部署文档。本 PR 将 HiCache 回归覆盖扩展到 GLM-5，表明团队正在有意识地扩展 HiCache 的兼容模型列表。