PR #22282 完整报告

sgl-project/sglang

[tiny] migrate /get_server_info; print accept length in accuracy tests

合并时间：2026-04-08 04:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22282

执行摘要

此 PR 将最后两个使用已废弃 /get_server_info 端点的调用迁移至 /server_info，提升 API 一致性；同时修改精度测试工具包，在所有测试混合类中打印推测解码的接受长度指标，便于性能监控。变更影响有限，主要涉及测试文件，风险较低。

功能与动机

根据 PR 描述，动机是迁移废弃的 /get_server_info 端点，以保持 API 一致性。具体来说：

- 迁移 test/registered/spec/test_ngram_speculative_decoding.py 和 python/sglang/test/kits/eval_accuracy_kit.py 中的端点调用。
- 在精度测试中打印 avg_spec_accept_length，用于监控推测解码性能，对于非推测解码服务器则静默跳过，避免干扰测试。

实现拆解

主要改动集中在两个文件：

1. python/sglang/test/kits/eval_accuracy_kit.py：
   - 重构 _check_accept_length 函数：python def _check_accept_length(test_case, base_url, threshold=None): """Print accept length; optionally assert it exceeds threshold.""" try: server_info = requests.get(base_url + "/server_info").json() val = server_info["internal_states"][0]["avg_spec_accept_length"] except (KeyError, IndexError, requests.RequestException): return print(f"avg_spec_accept_length={val:.4f}") if threshold is not None: test_case.assertGreater(val, threshold)
   - 简化各测试混合类（GSM8K、MMLU、HumanEval、MGSM-EN）的调用，移除条件判断，统一调用 _check_accept_length。
2. test/registered/spec/test_ngram_speculative_decoding.py：
   - 将端点从 /get_server_info 改为 /server_info。

评论区精华

无 review 评论，讨论内容为空。

风险与影响

- 风险：端点迁移可能引入兼容性问题，但已添加异常处理（KeyError、IndexError、requests.RequestException），降低失败风险；测试逻辑变更需确保 CI 通过，PR 已标记 run-ci 标签。
- 影响：对用户无直接影响；系统层面提升 API 一致性；团队层面便于开发者查看推测解码性能指标，辅助调试。

关联脉络

- 与 PR #22199（"[Spec][Ngram] Add output-as-corpus accept length benchmark for external SAM"）相关，同属推测解码领域，且修改了相同文件 test/registered/spec/test_ngram_speculative_decoding.py，显示团队在持续优化推测解码的测试和监控能力。