PR #26811 完整报告

sgl-project/sglang

Add the KV-canary mock-model end-to-end test harness

合并时间：2026-05-31 09:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26811

执行摘要

- 一句话：添加 KV-canary mock 模型端到端测试框架
- 推荐动作：值得精读，尤其是 utils.py 中的 run_mock_model_bench_serving 和 perturb_e2e_base.py 中的 MockModelPerturbE2EBase，它们定义了 KV-canary 测试的标准模式。对于要编写新 KV-canary 测试的开发者是必读材料。

功能与动机

需要一套统一的端到端测试框架来验证 KV-canary 各功能组件（如扰动注入、token 校验、EAGLE 集成等）的正确性。PR body 指出：‘Add the mock-model end-to-end test harness that the KV-canary self-tests build on: a tiny mock model plus the shared perturb-e2e base and utilities, and the generic tp/pp/pd e2e smoke tests.’目的是为后续特性测试提供可复用的基础。

实现拆解

1. 在 python/sglang/test/mock_model/utils.py 中定义了 mock 模型启动参数生成函数 mock_model_server_args、环境变量 mock_model_server_env 以及通用基准测试运行器 run_mock_model_bench_serving，它封装了服务器启动、benchmark 执行和结果验证。
2. 在 python/sglang/test/mock_model/perturb_e2e_base.py 中创建了 MockModelPerturbE2EBase 基类，继承自 CapturedServerE2EBase，为扰动测试提供服务器启动和并行请求发送能力。
3. 在 python/sglang/test/server_fixtures/disaggregation_fixture.py 中扩展了 PDDisaggregationServerBase，新增了 capture_per_side_logs 选项，允许捕获 prefill/decode 端的 stdout/stderr，以便在 PD 场景下检测 canary 违规。
4. 在 test/registered/mock_model/ 下添加了三个冒烟测试文件：test_e2e_tp.py、test_e2e_pp.py、test_e2e_pd.py，分别对 tensor parallel、pipeline parallel 和 PD 分离并行进行基本的 canary 无违规验证。
5. 添加了 python/sglang/test/mock_model/__init__.py 使 mock_model 成为包。

关键文件：

- python/sglang/test/mock_model/utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 MockModelBenchResult, log_text, log_tail, mock_model_server_args）：核心工具库，定义了 mock 模型服务器参数生成、环境配置和通用基准测试运行器，是所有 e2e

测试的基础依赖。

- python/sglang/test/mock_model/perturb_e2e_base.py（模块 扰动测试；类别 test；类型 test-coverage；符号 MockModelPerturbE2EBase, setUpClass, send_parallel_requests）：扰动端到端测试基类，提供服务器启动和并行请求发送的模板，供后续扰动特性测试继承。
- test/registered/mock_model/test_e2e_pd.py（模块 测试用例；类别 test；类型 test-coverage；符号 _send_parallel_requests, _one, _make_input_ids, _MockModelPDBase）：PD 分离并行场景的冒烟测试，验证预填充 / 解码分离时 canary 无违规。
- test/registered/mock_model/test_e2e_pp.py（模块 测试用例；类别 test；类型 test-coverage；符号 TestE2EPipelineParallel, test_pp_no_canary_violation）：Pipeline parallel 冒烟测试，验证 PP 2 下 canary 无违规。
- test/registered/mock_model/test_e2e_tp.py（模块 测试用例；类别 test；类型 test-coverage；符号 TestE2ETensorParallel, test_tp_no_canary_violation）：Tensor parallel 冒烟测试，验证 TP 2 下 canary 无违规。
- python/sglang/test/server_fixtures/disaggregation_fixture.py（模块 测试固件；类别 test；类型 test-coverage）：修改现有 PD 测试基类，新增 capture_per_side_logs 支持，以便在 PD 测试中捕获并检查 canary 违规。
- python/sglang/test/mock_model/__init__.py（模块 包初始化；类别 test；类型 test-coverage）：使 mock_model 成为 Python 包，允许导入。

关键符号：mock_model_server_args, mock_model_server_env, run_mock_model_bench_serving, MockModelPerturbE2EBase.setUpClass, MockModelPerturbE2EBase.send_parallel_requests, _send_parallel_requests, _make_input_ids, TestPdTransferCanaryClean.test_pd_transfer_canary_clean, TestE2ETensorParallel.test_tp_no_canary_violation, TestE2EPipelineParallel.test_pp_no_canary_violation

评论区精华

该 PR 无实质性 review 讨论，只有 gemini-code-assist 的配额警告。PR 由作者自行合并。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是测试框架不够健壮可能造成假阴性或假阳性，尤其是 assert_no_canary_violation 依赖日志扫描，若日志格式变化可能导致误判。此外，修改 disaggregation_fixture.py 增加了 capture_per_side_logs 的类变量，若其他测试未正确重置可能引入副作用。
- 影响：对用户无直接影响，仅影响内部测试。对团队的影响是标准化了 KV-canary 测试的编写方式，后续所有 KV-canary 特性测试都应继承此框架。
- 风险标记：日志依赖断言，测试基础设施变更

关联脉络

- PR #26816 Add the KV-canary perturb framework for fault-injection self-tests: 该 PR 添加的扰动 *e2e* 测试（如 test_self_unit_perturb）依赖于本 PR 提供的 MockModelPerturbE2EBase 和 test *e2e* harness。
- PR #26815 Add *a deterministic token oracle and production write-input assertion*: Token oracle 的 mock 模型测试（如 test_self_unit_oracle）使用本 PR 的 mock_model 基础设施。
- PR #26818 Add token-id verification to the KV-canary: Token ID 验证的 *e2e* 测试（如 test_scatter_req_token_ids）依赖本 PR 的测试框架运行。
- PR #26813 Support EAGLE *speculative decoding in the* KV-canary: EAGLE 集成的 *e2e* 测试（如 test_self_unit_runner_per_forward）构建在本 PR 的测试 harness 之上。