PR #25831 完整报告

sgl-project/sglang

[Test] Stage-a sanity kits; consolidate core/ + models_e2e/ tests

合并时间：2026-05-20 16:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25831

执行摘要

本次 PR 对 sglang 测试基础设施进行了系统性重构：引入三个可复用检测套件（API 契约、解码正确性、调度压力）替换旧 server_sanity_kit，新增阶段 A 门禁测试，并将模型 e2e 测试归并至 models_e2e/ 目录。重构后测试覆盖更清晰、CI 更稳定，同时减少约 640 行冗余代码。

功能与动机

PR 旨在解决以下痛点：test_srt_backend.py 作为前端 DSL 冒烟测试已经过时且难以维护；dsv4 e2e 测试中存在大量内联重复代码；test/registered/core/ 目录混杂了框架级和模型级测试。通过本次重构，实现测试的模块化、可复用和目录结构统一。

实现拆解

1. 检测套件提取：从旧 ServerSanityMixin 中按关注点拆分出三个独立 mixin — BasicAPIContractMixin（协议层）、BasicDecodeCorrectnessMixin（输出质量）、BasicSchedulerStressMixin（压力 /streaming），每个包含 3-6 条探测。
2. 门禁测试：test_basic_sanity.py 合并三个 mixin 并添加 test_accuracy_floor（hellaswag≥0.60）作为私有准确率门控。
3. dsv4 精简：将各模型类中的内联 _gsm8k_check 统一替换为 GSM8KMixin，并仅保留 BasicDecodeCorrectnessMixin（协议和压力已由门禁覆盖）。
4. core/ 目录清理：删除被门禁替代的文件；裁剪 test_srt_endpoint.py 和 test_srt_engine.py 中冗余用例；共享 Engine 实例减少启动时间。
5. 文件搬迁：将误分类测试移至 attention/、radix_cache/、models/ 等对应子目录，所有模型 e2e 测试统一至 models_e2e/。

test/registered/core/test_basic_sanity.py — 阶段 A 门禁测试的完整实现：

```
"""Basic sanity: small-but-broad server smoke that downstream stages
depend on. Three sanity kits, one shared server, covering protocol
contract, decode correctness, and scheduler stress paths."""

import unittest

from sglang.srt.utils import kill_process_tree
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.kits.basic_api_contract_kit import BasicAPIContractMixin
from sglang.test.kits.basic_decode_correctness_kit import BasicDecodeCorrectnessMixin
```

```
from sglang.test.kits.basic_scheduler_stress_kit import BasicSchedulerStressMixin
from sglang.test.test_utils import (
    DEFAULT_MODEL_NAME_FOR_TEST,
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

register_cuda_ci(est_time=120, stage="base-a", runner_config="1-gpu-small")
register_amd_ci(est_time=120, suite="stage-a-test-1-gpu-small-amd")


class TestBasicSanity(
    BasicAPIContractMixin,
    BasicDecodeCorrectnessMixin,
    BasicSchedulerStressMixin,
    CustomTestCase,
):
    served_model_name = DEFAULT_MODEL_NAME_FOR_TEST

    @classmethod
    def setUpClass(cls):
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            DEFAULT_MODEL_NAME_FOR_TEST,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--cuda-graph-max-bs",
                "4",
                "--mem-fraction-static",
                "0.7",
                "--enable-metrics",
            ],
        )

    @classmethod
    def tearDownClass(cls):
        kill_process_tree(cls.process.pid)

    def test_accuracy_floor(self):
        import sglang as sgl
        from sglang.test.test_programs import test_hellaswag_select

        sgl.set_default_backend(sgl.RuntimeEndpoint(self.base_url))
        try:
            accuracy, _ = test_hellaswag_select()
        finally:
```

```
        sgl.set_default_backend(None)
    self.assertGreater(
        accuracy,
        0.60,
        f"hellaswag accuracy floor breached: {accuracy:.3f}",
    )


if __name__ == "__main__":
    unittest.main()
```

评论区精华

无审核评论；PR 由作者直接合并。

风险与影响

- 测试覆盖风险：删除 test_srt_backend.py 可能移除一些 DSL 特定用例，但门禁已覆盖核心路径。
- 资源竞争风险：test_hidden_states.py 共享 Engine 并设置低 mem_fraction_static 可能导致显存不足，但已通过 flush_cache() 和单独配置缓解。
- 搬迁遗漏：文件移动后若 CI 配置文件未同步更新可能跳过部分测试，但 PR CI 已通过验证。
- 正面影响：开发者编写新模型 e2e 测试时可直接复用 mixin，减少启动时间约 40-60s。

关联脉络

该 PR 属于测试基础设施清理系列，与之前 #25825（通过构造函数传递 PP start_layer）等核心重构无直接关联，但为后续更大规模的 CI 优化奠定了基础。