PR #26610 完整报告

sgl-project/sglang

test/registered: cleanup pure model e2e tests (moves, splits, dedup, kit)

合并时间：2026-05-29 06:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26610

执行摘要

- 一句话：清理模型 E2E 测试目录结构
- 推荐动作：值得精读，展示了如何系统性地组织大型测试套件，包括目录约定、server 启动拆分和公共 mixin 提取。对维护大规模测试套件的团队有参考价值。

功能与动机

为匹配已建立的约定：(1) 纯模型架构 E2E 测试应位于 models_e2e/ 目录；(2) 每个文件只启动一个server；(3)特性测试应位于其特性目录。通过清理减少冗余、提高可维护性和CI效率。引用 PR body：'Mechanical, behavior-preserving cleanup to match the established conventions'。

实现拆解

按以下步骤：

1. 移动纯模型测试文件从 4-gpu-models/、8-gpu-models/、quant/ 等到 models_e2e/，同时重命名以消除冗余标识（如 test_qwen35_fp4_mtp_v2 -> test_qwen35_fp4_mtp）。
2. 拆分多 server 启动的文件：test_gpt_oss_4gpu.py 拆分为 bf16/mxfp4 两个文件；test_mimo_models.py 拆分为 MiMo-V2 和 V2.5；test_nvidia_nemotron_3_super_bf16.py 拆分为 base 和 mtp；test_deepseek_v32_fp4_mtp_4gpu.py 拆分为 dp 和 tp 变体。
3. 删除重复文件 test_qwen35_models.py，其与 test_qwen35_fp4_mtp_v2.py 功能重复。
4. 提取 UnifiedRadixTreeTestMixin 到 python/sglang/test/kits/unified_radix_cache_kit.py，移除原有 TestCase 子类，让 radix_cache 下的各文件从该 mixin 继承并各自管理 server 启动。
5. 将 test_qwen35_hicache.py 移动到 hicache/ 目录，并更新依赖导入路径。

关键文件：

- test/registered/quant/test_deepseek_v32_fp4_mtp_4gpu.py（模块 FP4-MTP；类别 test；类型 deletion；符号 TestDeepseekV32FP4DPSpecV2, TestDeepseekV32FP4TPSpecV2, setUpClass, tearDownClass）：被删除并拆分为 dp/tp 两个新文件，是拆分动作的核心文件

- test/registered/4-gpu-models/test_qwen35_models.py（模块 Qwen35；类别 test；类型 deletion；符号 TestQwen35FP4MTPV2, setUpClass, tearDownClass, test_gsm8k）：被识别为 test_qwen35_fp4_mtp_v2.py 的重复并删除，体现去重清理
- test/registered/models_e2e/test_deepseek_v3_mtp.py（模块 DeepSeek-MTP；类别 test；类型 test-coverage；符号 TestDeepseekV3MTP, test_z_bs_1_speed）：新增文件，模型架构测试移至 models_e2e 并使用 GSM8KMixin + DefaultServerBase 简化
- python/sglang/test/kits/unified_radix_cache_kit.py（模块 基数缓存；类别 test；类型 rename-or-move；符号 UnifiedRadixTreeTestMixin, _random_suffixes）：提取 UnifiedRadixTreeTestMixin 到公共 kit，被多个 radix_cache 测试文件使用
- test/registered/models_e2e/test_nvidia_nemotron_3_super_bf16_mtp.py（模块 Nemotron；类别 test；类型 rename-or-move；符号 TestNvidiaNemotron3SuperBF16MTP, test_gsm8k）：从 8-gpu-models 移入 models_e2e，移除了旧的非 MTP 类，仅保留 MTP 测试

关键符号：UnifiedRadixTreeTestMixin, TestDeepseekV3MTP, TestDeepseekV32FP4DPSpec, TestDeepseekV32FP4TPSpec, TestNvidiaNemotron3SuperBF16, TestNvidiaNemotron3SuperBF16MTP, _run_gsm8k

关键源码片段

python/sglang/test/kits/unified_radix_cache_kit.py

提取 UnifiedRadixTreeTestMixin 到公共 kit，被多个 radix_cache 测试文件使用

```
# 生成随机后缀 token ID 列表
# n: 样本数，length: 每个后缀长度，seed: 随机种子
def _random_suffixes(n, length, seed):
    rng = random.Random(seed)
    return [[rng.randint(1, 30000) for _ in range(length)] for _ in range(n)]


class UnifiedRadixTreeTestMixin:
    """
    Mixin: 提供 GSM8K、MMLU 和多轮 KL 散度测试的方法
    ，供不同注意力模式（full/mamba/swa）的 TestCase 继承使用。
    """
    # 可被子类覆盖的配置属性
    kl_threshold: float = 0.003
    max_new_tokens: int = 512
    num_groups: int = 3
    branches_per_group: int = 3
    prefix_len: int = 512
    prefill_cache_assert = None
    decode_cache_assert = None
    sampling_temperature: float = 1
    decode_hit_request_batch_size: int | None = None
    decode_hit_inter_batch_delay_s: float = 0

    gsm8k_threshold: float = 0.93
```

```
    mmlu_threshold: float = 0.8
    num_gsm8k_questions: int = 200

    def test_gsm8k(self):
        """少样本 GSM8K 数学推理准确率测试。"""
        from sglang.test.few_shot_gsm8k import run_eval as run_few_shot_gsm8k

        url = urlparse(self.base_url)
        args = SimpleNamespace(
            num_shots=10,
            data_path=None,
            num_questions=self.num_gsm8k_questions,
            max_new_tokens=16000,
            parallel=128,
            host=f"http://{url.hostname}",
            port=int(url.port),
        )
        metrics = run_few_shot_gsm8k(args)
        # ...（后续 MMLU 和多轮测试方法省略）
```

评论区精华

无 review 评论。PR 由作者自行合并，说明变更为机械性且已充分验证。

- 暂无高价值评论线程

风险与影响

- 风险：风险很低。关键点：(1) 拆分后每个测试文件独立启动 server，CI 资源使用略微变化，但 runner_config 保持不变，影响有限；(2) 测试路由依赖于 register_cuda_ci 注解而非目录，因此移动后 CI 执行路径不变；(3) 若有不正确移动导致测试不被发现，则会在 CI 中遗漏。通过本地运行和 CI 验证已降低风险。
- 影响：影响范围：主要影响团队测试维护和 CI 执行。正向影响：减少测试文件耦合，提高定位效率；拆分 server 启动可缩短单个测试文件执行时间；删除重复避免混淆。不影响用户和系统功能。
- 风险标记：测试文件移动，CI 路由不变，拆分 server 启动可能影响资源

关联脉络

- 暂无明显关联 PR