PR #22136 完整报告

sgl-project/sglang

[CI] Lower GSM8K baselines for B200 nightly after eval unification

合并时间：2026-04-23 13:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22136

执行摘要

- 一句话：修复 B200 夜间测试因评估统一导致的 GSM8K 准确性基线问题。
- 推荐动作：该 PR 值得快速浏览，以了解评估统一后的测试适配模式；重点关注 AccuracyTestParams 扩展 api 字段的设计，以及如何通过配置修复因评估方法变更导致的测试失败。

功能与动机

PR #21667 统一了 GSM8K 评估路径，但引入了两个问题：

1) FP8 MOE 后端因旧评估包含少量样本泄露导致分数虚高约 4%，新评估分数降至 0.905-0.925，原基线 0.93 过高； 2) FP4 DeepSeek-R1 因新评估默认使用 Chat API 且缺少样本数限制，分数从 0.975 降至 0.86。需调整测试基线以反映真实准确性，避免 CI 失败。

实现拆解

1. 扩展测试框架支持 API 参数：在 python/sglang/test/accuracy_test_runner.py 中，为 AccuracyTestParams 数据类添加 api 字段（可选 "chat" 或 "completion"），并在 _run_simple_eval 函数中新增 api 参数，将其传递给评估后端，以控制评估使用的 API 模式。
2. 调整 FP8 MOE 后端基线阈值：在 test/registered/backends/test_flashinfer_trtllm_gen_moe_backend.py 中，将 test_gsm8k 方法的断言阈值从 0.93 降至 0.89，匹配新评估方法下的实际分数范围（0.905-0.925）。
3. 修复 FP4 DeepSeek-R1 测试配置：在 test/registered/perf/test_dpsk_v3_fp4_4gpu_perf.py 中，为 AccuracyTestParams 添加 num_examples=200 和 api="completion"，确保评估使用正确的 API 并限制样本数，恢复基线 0.935。
4. 测试配套调整：所有变更均为测试文件修改，无源码主路径改动，旨在确保夜间 CI 测试通过，不涉及功能逻辑变更。

关键文件：

- python/sglang/test/accuracy_test_runner.py（模块 测试框架；类别 test；类型 test-coverage；符号 AccuracyTestParams, _run_simple_eval, run_accuracy_test）：核心测试框架扩展，新增 API 参数支持，影响所有准确性测试的配置传递。

- test/registered/perf/test_dpsk_v3_fp4_4gpu_perf.py（模块 性能测试；类别 test；类型 test-coverage；符号 AccuracyTestParams）：修复 DeepSeek-R1 FP4 测试配置，添加 num_examples 和 api 参数以匹配旧评估行为。
- test/registered/backends/test_flashinfer_trtllm_gen_moe_backend.py（模块 后端测试；类别 test；类型 test-coverage；符号 test_gsm8k）：调整 FP8 MOE 后端 GSM8K 测试的断言阈值，反映评估统一后的真实准确性。

关键符号：AccuracyTestParams, _run_simple_eval, run_accuracy_test, test_gsm8k

关键源码片段

python/sglang/test/accuracy_test_runner.py

核心测试框架扩展，新增 API 参数支持，影响所有准确性测试的配置传递。

```
@dataclass
class AccuracyTestParams:
    """Parameters for accuracy testing."""

    dataset: str # e.g., "mgsm_en", "gsm8k", "mmmu", "gpqa"
    baseline_accuracy: float # Required: minimum accuracy threshold
    num_examples: Optional[int] = None
    num_threads: Optional[int] = None
    max_tokens: Optional[int] = None
    return_latency: bool = False
    thinking_mode: Optional[str] = None # e.g., "deepseek-v3"
    temperature: Optional[float] = None
    top_p: Optional[float] = None
    top_k: Optional[int] = None
    repeat: Optional[int] = None
    api: Optional[str] = None # 新增字段：控制评估 API 模式，可选 "chat" 或 "completion"，默认在 run_eval 中为 "chat"


def _run_simple_eval(
    model: ModelLaunchSettings,
    base_url: str,
    dataset: str,
    num_examples: Optional[int] = None,
    num_threads: Optional[int] = None,
    max_tokens: Optional[int] = None,
    return_latency: bool = False,
    thinking_mode: Optional[str] = None,
    temperature: Optional[float] = None,
    top_p: Optional[float] = None,
    top_k: Optional[int] = None,
    repeat: Optional[int] = None,
    api: Optional[str] = None, # 新增参数：接收 API 模式配置
) -> Tuple[bool, Optional[str], Optional[dict]]:
```

```
    """Run evaluation using simple_eval backend (run_eval.py)."""
    process = None
    try:
        process = popen_launch_server(
            model.model_path,
            base_url,
            other_args=model.extra_args,
            timeout=model.launch_timeout or DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            env=model.env,
        )

        args = SimpleNamespace(
            base_url=base_url,
            model=model.model_path,
            eval_name=dataset,
            num_examples=num_examples,
            num_threads=num_threads or 1024,
        )

        if api is not None: # 将 API 参数传递给评估后端
            args.api = api
        # ... 其他参数处理逻辑
```

评论区精华

review 中主要讨论了测试一致性和潜在遗漏：

- gemini-code-assist[bot] 指出：FlashinferTrtllmGenMoeBackendMXFP8Base（使用 mxfp8 量化）可能同样受评估统一影响，建议将其基线也从 0.93 更新至 0.89 以避免 CI 失败。
- 建议增强一致性：推荐在 DeepSeek-R1 测试中显式设置 num_threads=128 以匹配其他后端测试配置。
- 决策结论：PR 已合并，但 FlashinferTrtllmGenMoeBackendMXFP8Base 的基线未调整，可能存在后续 CI 风险。
  - FP8 MOE 后端基线一致性 (correctness): PR 未调整该基线，可能存在后续测试失败风险。
  - 测试参数一致性建议 (testing): PR 未采纳此建议，但核心问题已通过 num_examples 和 api 修复。

风险与影响

- 风险：
  1. 回归风险低：变更仅影响测试断言和配置，不修改核心业务逻辑。
  2. 兼容性风险：新增 api 字段向后兼容，但依赖方需注意默认值从 "chat" 变为可配置。
  3. 测试覆盖风险：FlashinferTrtllmGenMoeBackendMXFP8Base 基线未更新，可能导致未来 CI 失败。

4. 性能风险无：纯测试调整，不影响运行时性能。

- 影响：
  1. 对用户影响：无直接影响，仅夜间测试通过性改善。
  2. 对系统影响：确保 B200 夜间测试准确反映模型性能，避免误报失败。
  3. 对团队影响：减少 CI 噪音，但需关注未调整基线的潜在测试失败。 - 风险标记：测试基线未完全覆盖，配置依赖变更

关联脉络

- PR #21667 [CI] Lower GSM8K baselines for B200 nightly after eval unification: PR body 中提及该 PR 统一了 GSM8K 评估路径，是本次测试失败的根因，直接关联。