

PR #2526 完整报告

radixark/miles

fix(swe-agent example): stop logging unmeasured agent metrics as zero

合并时间: 2026-08-14 06:01

原文链接: <http://prhub.com.cn/radixark/miles/pull/2526>

执行摘要

- 一句话: SWE-agent 示例不再把未上报的 agent 指标记为 0
- 推荐动作: 值得快速精读: 核心逻辑只有 7 行改动, 但把“缺失”与“真零”的语义区别处理得很干净, 测试五场景覆盖完整, 适合作为指标上报聚合的最佳实践模板; 如果未来把 SWE-agent 示例的指标逻辑提升为通用 agent metrics 模块, 本 PR 的缺失值处理方式可直接复用。示例使用者无需额外操作, 重新拉取代码即可获得正确指标。

功能与动机

PR body 说明了两类失败模式: 不同 agent harness 上报的键集不同, `_collect_values` 使用 `m.get(key, 0)` 后, 只上报 wall-clock 总量的 agent 在 wandb 面板上出现 `agent/tool_calls_mean = 0.0` 等 12 个假零, 而真实轨迹每条含 23–47 次 `bash_command` 调用, 面板却显示 agent 从未调用工具; 混合批次中单个样本缺键会把上报 47 次的样本均值拉低成 23.5。作者强调 *A zero sitting next to real numbers reads as a measurement*, 假零会伪装成真实测量, 这是必须修复的语义问题。

实现拆解

1. 修复聚合入口 `_collect_values` (`examples/swe-agent-harbor-docker/generate.py`): 将 `[m.get(key, 0) for m in all_metrics]` 替换为带海象表达式的列表推导, 只有当键存在且值不为 `None` 时才收集; agent 真正测得的值 0 仍会被保留并正常上报。
2. 依托调用方已有的空列表跳过逻辑: `_agg_mean` 等聚合函数对空列表不产出统计项, 因此 `aggregate_agent_metrics` 输出中未上报的键会整体消失, wandb 直接不显示对应 series, 行为从“伪造零”变为“诚实的缺失”。
3. 新增回归测试 `tests/fast/examples/swe_agent_harbor_docker/test_agent_metrics.py`: 由于目录名含连字符不是合法 Python 标识符, 测试通过 `importlib.util.spec_from_file_location` 按路径加载 `generate.py`; 5 个用例分别验证全上报无回归、未上报键被省略、混合批次均值只统计上报者、真实 0 保留、无指标样本被忽略。作者验证修复前 2 个用例失败、修复后全部通过。
4. 自审打磨: 作者 self-review 指出新 docstring 过长后, 第二个提交 Shorten the `_collect_values` docstring 完成精简, 随后由 nblintao 批准合并。

关键文件:

- `examples/swe-agent-harbor-docker/generate.py` (模块 示例脚本; 类别 `infra`; 类型 `core-logic`; 符号 `_collect_values`, `aggregate_agent_metrics`, `_agg_mean`) : 核心修复点 : `_collect_values` 从 `m.get(key, 0)` 兜底零改为跳过未上报键, 消除假零与混合批次均值偏差, 真实 0 仍保留。
- `tests/fast/examples/swe_agent_harbor_docker/test_agent_metrics.py` (模块 指标测试; 类别 `test`; 类型 `test-coverage`; 符号 `generate_module`, `sample_with`, `test_fully_instrumented_agent_reports_every_metric`, `test_unreported_keys_are_omitted_rather_than_logged_as_zero`) : 新增 5 个用例固化“缺失 $\neq 0$ ”语义, 覆盖全上报、仅总量、混合批次、真 0、空样本, 是本次修复唯一的回归保障。
- `tests/fast/examples/swe_agent_harbor_docker/__init__.py` (模块 测试包; 类别 `test`; 类型 `test-coverage`) : 让新测试目录成为可导入的测试包, 配合按路径加载生成脚本的测试方式。

关键符号: `_collect_values`, `aggregate_agent_metrics`, `_agg_mean`, `sample_with`, `test_mixed_batch_averages_only_over_agents_that_measured_the_key`

关键源码片段

`examples/swe-agent-harbor-docker/generate.py`

核心修复点: `_collect_values` 从 `m.get(key, 0)` 兜底零改为跳过未上报键, 消除假零与混合批次均值偏差, 真实 0 仍保留。

```
def _collect_values(all_metrics: list[dict], key: str) -> list[float]:
    """Values agents reported for ``key``; agents that omit it are skipped.

    Defaulting to 0 would log a fake measurement and drag the batch mean down.
    Callers skip empty lists, so an unreported key stays out of the log.
    """
    # 海象表达式同时完成取值与非 None 判断:
    # 只有键存在且值不为 None 时收集, agent 真正测到的 0 仍会被保留。
    return [value for metrics in all_metrics if (value := metrics.get(key)) is not None]
```

`tests/fast/examples/swe_agent_harbor_docker/test_agent_metrics.py`

新增 5 个用例固化“缺失 $\neq 0$ ”语义, 覆盖全上报、仅总量、混合批次、真 0、空样本, 是本次修复唯一的回归保障。

```
# FULLY_INSTRUMENTED 上报 13 个键 (turns=30、tool_calls=47...) ;
# TOTALS_ONLY 只上报 3 个 wall-clock 键 (eval_time、agent_run_time、total_time) 。

@pytest.fixture(scope="module")
def generate_module() -> ModuleType:
    # 目录名含连字符, 不是合法 Python 标识符, 因此像其它 example 测试
    # 一样用 importlib 按路径加载 generate.py。
    spec = importlib.util.spec_from_file_location("swe_agent_harbor_docker_generate",
        GENERATE_SCRIPT)
    module = importlib.util.module_from_spec(spec)
```

```

spec.loader.exec_module(module)
return module

def sample_with(agent_metrics: dict) -> SimpleNamespace:
    # 构造带 metadata.agent_metrics 的最小样本，供聚合函数直接消费。
    return SimpleNamespace(metadata={"agent_metrics": agent_metrics})

def test_unreported_keys_are_omitted_rather_than_logged_as_zero(generate_module:
ModuleType) -> None:
    # 0 与“未测量”不可混为一谈：未上报的键不应出现在日志里，
    # 否则会伪装成一次真实的零测量。
    metrics = generate_module.aggregate_agent_metrics([sample_with(TOTALS_ONLY)])

    for key in UNREPORTED_BY_TOTALS_ONLY:
        assert key not in metrics, f"{key} was logged despite never being measured"

    # 真正上报的键仍然正常输出。
    assert metrics["agent/eval_time_mean"] == pytest.approx(42.8)
    assert metrics["agent/agent_run_time_mean"] == pytest.approx(486.8)
    assert metrics["agent/total_time_mean"] == pytest.approx(559.3)

def test_mixed_batch_averages_only_over_agents_that_measured_the_key(generate_module:
ModuleType) -> None:
    # 修复前：一个样本缺 tool_calls 会把 47 次调用拉低成 23.5；
    # 修复后：均值只对真正上报该键的 agent 计算。
    metrics = generate_module.aggregate_agent_metrics([sample_with(FULLY_INSTRUMENTED),
sample_with(TOTALS_ONLY)])

    assert metrics["agent/tool_calls_mean"] == 47
    assert metrics["agent/turns_mean"] == 30
    assert metrics["agent/time_per_turn"] == pytest.approx(13.3)
    # 两个 agent 都上报的键，仍然按整个批次求均值。
    assert metrics["agent/total_time_mean"] == pytest.approx((500.0 + 559.3) / 2)

```

评论区精华

唯一的 review 评论是作者 Shi-Dong 的自评，针对 `_collect_values` 的新 docstring：

Shorten the explanation here. This is too long. Make it clear and concise.。作者随即在第二个提交中压缩注释，只保留“跳过未上报键、避免伪造测量”的核心说明。整个过程没有其它争议或未解决问题，nblintao 直接给出 APPROVED，说明修复逻辑本身得到认可，唯一打磨点在可读性。

- 缩短 `_collect_values` 的 docstring (style): 第二个提交 Shorten the `_collect_values` docstring 完成精简，保留核心语义说明。

风险与影响

- 风险：行为变更：aggregate_agent_metrics 输出中未上报键会消失，任何按固定键读取结果的消费方会触发 KeyError；当前 _agg_mean 等调用方已具备空列表跳过逻辑，未发现受影响的下游，但后续新增消费方需记住该语义。测试加载方式存在脆弱性：
test_agent_metrics.py 通过 spec_from_file_location 按路径加载 generate.py，若脚本将来引入依赖同包相对导入的模块，此加载方式会失效，目前验证可正常加载。真实 0 语义依赖 is not None 判断，约定“None = 未测量”自洽；若未来 agent 用 NaN 表达缺失则不会被跳过，需要同步扩展。整体风险低，且影响范围仅限示例脚本。
- 影响：直接受影响的是 examples/swe-agent-harbor-docker 的 wandb 指标输出：只报总时长的 agent 不再出现 12 个假零指标；完全插桩的 agent 全部 17 个指标前后完全一致；混合批次中 tool_calls_mean 从错误的 23.5 修正为 47.0。对使用该示例做 agentic RL / SWE-bench 类实验的团队，面板可读性和可信度提升。改动不触及 miles 核心训练、rollout、dashboard 代码，核心用户无感知；示例维护者获得 5 个可持续回归的测试用例。
- 风险标记：指标键缺失属行为变更，下游消费方需容忍键消失，测试依赖按路径加载脚本

关联脉络

- PR #2280 [example] GLM-5.2 744B-A40B LoRA agentic launcher (TB2 on Daytona): 同属 examples/swe-agent-harbor-docker 示例家族，本 PR 修复的是该目录共享的 agent 指标聚合逻辑。
- PR #2369 fix(rollout): normalize rewards per rollout: 同为统计聚合语义修正：修复混合统计偏差，说明仓库近期在系统化清理聚合缺陷，本 PR 是示例层面的同类修复。