# PR #22147 完整报告

sgl-project/sglang

Add dump_metric to MMMU, lm-eval, and NeMo Skills eval paths

合并时间：2026-04-05 18:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22147

## 执行摘要

该 PR 在 MMMU（lmms-eval）、lm-eval harness 和 NeMo Skills（mmmu-pro）三个评估路径中添加了 dump_metric 调用，是评估统一计划的第二阶段。变更完全向后兼容（dump_metric 静默失败），为后续回归检测基础设施提供数据基础。但 review 指出的标签模式不一致问题未解决，可能影响指标聚合。

## 功能与动机

根据 PR body，这是评估统一计划的第二阶段（第一阶段是 #21667 的 GSM8K 统一）。目标是确保所有评估路径都输出 dump_metric 数据，为未来回归检测基础设施奠定基础。所有通过 run_eval.py 的评估路径已经具备 dump_metric，此 PR 覆盖剩余的三个路径。dump_metric 被设计为静默失败，不会影响现有测试功能。

## 实现拆解

在三个文件中添加了 dump_metric 调用：

| 文件 | 修改位置 | 指标名 | 标签 |
| --- | --- | --- | --- |
| python/sglang/test/kits/mmmu_vlm_kit.py | test_mmmu 和 _run_vlm_mmmu_test | "mmmu_score" | model, eval="mmmu", api="lmms-eval" |
| python/sglang/test/kits/lm_eval_kit.py | test_lm_eval 循环 | "{task_name}_{metric_name}" | model, eval="lm-eval", task |
| python/sglang/test/accuracy_test_runner.py | _run_nemo_skills_eval | "{dataset}_score" | model, eval=dataset, api="nemo-skills" |

关键代码示例（来自 mmmu_vlm_kit.py）：

```
dump_metric(
    "mmmu_score",
```

```
        mmmu_accuracy,
        labels={"model": self.model, "eval": "mmmu", "api": "lmms-eval"},
    )
```

## 评论区精华

review 中只有一个来自 gemini-code-assist[bot] 的评论，指出标签模式不一致：

> "In mmmu_vlm_kit.py and accuracy_test_runner.py, the eval label represents the benchmark/dataset name (e.g., "mmmu" or "mmmu-pro") and the api label represents the framework/runner (e.g., "lmms-eval" or "nemo-skills"). In this file, eval is set to "lm-eval" and the benchmark is stored in a separate task label."

建议统一标签模式，但作者未回复，PR 已合并，表明该问题可能被接受或将在后续处理。

## 风险与影响

风险：

1. 标签模式不一致可能影响后续指标聚合和分析。
2. review 问题未完全解决，可能留下技术债务。

影响：

1. 对系统：为评估系统添加统一指标输出，支持未来回归检测。
2. 对用户：无直接影响，测试行为不变。
3. 对团队：提供数据基础，但需要关注标签一致性以确保数据质量。

## 关联脉络

- 这是 #21667（GSM8K 统一）开始的评估统一计划的第二阶段，两者共同构建评估基础设施。
- 从近期历史 PR 看，仓库频繁进行测试和 CI 相关改进（如 #22140、#22139、#22102），该 PR 延续了这一趋势，专注于评估指标收集的统一化。
- 标签模式不一致问题与近期多个 PR 关注的 "consistency" 标签（如 #22148、#22102、#22137）主题相关，但在此 PR 中未完全解决。