

PR #2242 完整报告

THUDM/slime

fix: honor every eval.defaults key and restore per-dataset stop / min_new_tokens

合并时间: 2026-08-12 14:07

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2242>

执行摘要

- 一句话: 修复 eval.defaults 丢键与 per-dataset stop 失效
- 推荐动作: 值得精读。这是一个高信号的配置解析 bugfix: slime/utils/eval_config.py 中反射校验 + setdefault 兜底的做法, 以及 slime/rollout/sglang_rollout.py 中采样参数的三级回退, 都是可复用的防御式配置设计。建议重点阅读 build_eval_dataset_configs 的完整实现、DATASET_RUNTIME_SPECS 的新增条目, 以及 tests/test_eval_config.py 对契约的表述; 合并后顺手核对仓库内现有 eval.defaults 配置是否存在拼写错误键。该 PR 也是很好的教学案例: 说明 spec 表模式在字段漏配时如何静默失败, 以及如何用反射消除硬编码清单。

功能与动机

PR body 描述了两个相互独立但同源的配置解析洞: 其一, eval.defaults 仅通过 `_pick_from_mapping(defaults, spec['default_keys'])` 读取 spec 表内 14 个字段, 其余字段被静默丢弃——'rm_type and repetition_penalty come out None — the eval runs with no reward model, silently'; 其二, EvalDatasetConfig 接受 stop / stop_token_ids / min_new_tokens, 但采样参数硬编码训练期 args.rollout_stop, 'An eval config that sets stop: ['</answer>'] for one dataset silently evaluates with the training stop strings instead — changing eval scores with no warning', 且 --eval-min-new-tokens 在仓库中无读取者。README 文档承诺 eval.defaults 是每个数据集共享的推理参数、可被数据集内覆盖, 两个 bug 都违背了这一契约。

实现拆解

1. 扩展 spec 表: 在 slime/utils/eval_config.py 的 DATASET_RUNTIME_SPECS 中新增 stop、stop_token_ids、min_new_tokens 三个条目, 分别映射命令行 fallback 为 rollout_stop、rollout_stop_token_ids、eval_min_new_tokens, 使这三个字段与 temperature 等兄弟字段走同一套 dataset > defaults > args 三级解析。
2. 未知键校验: build_eval_dataset_configs 开头用 dataclasses.fields(EvalDatasetConfig) 反射出全部数据类字段, 与 spec 表的 default_keys 取并集作为合法键集合, defaults 中出现集合外键立即抛 ValueError 并列出合法键, 使 defaults 与 dataset entry 的严格度对齐。反射而非硬编码清单, 避免以后新增字段时再次漏配。

3. 非 spec 字段兜底: spec pass 之后, 对 defaults 中不属于 spec 的键 (如 `rm_type`、`repetition_penalty`、`eval_task_timeout`、`app_service` 等) 逐项执行 `cfg_dict.setdefault(key, value)`, dataset entry 显式值优先、defaults 补齐其余, 消除静默丢弃。
4. 采样参数接线: `slime/rollout/sglang_rollout.py` 的 `eval_rollout_single_dataset` 中, `stop / stop_token_ids` 从硬编码 args 改为 `dataset_cfg` 优先、args 兜底; 新增 `min_new_tokens` 解析 (先 `dataset_cfg.min_new_tokens`, 再 `args.eval_min_new_tokens`, 均无则不写入), 复活 `--eval-min-new-tokens`。
5. 测试与 CI 配套: 新增 `tests/test_eval_config.py` 五个用例, 覆盖非 spec 默认值透传到每个数据集、dataset 覆盖优先级、三个恢复字段的 dataset > defaults > args 解析、未知键拒绝、spec 字段 args 兜底; 并通过 `.github/workflows/pr-test.yml` 及 j2 模板注册进 `cpu-unittest` job (`num_gpus` 为 0)。

关键文件:

- `slime/utils/eval_config.py` (模块 配置解析; 类别 source; 类型 core-logic; 符号 `build_eval_dataset_configs`, `DATASET_RUNTIME_SPECS`): 核心修复所在: `DATASET_RUNTIME_SPECS` 扩展三个字段, `build_eval_dataset_configs` 新增未知键校验与非 spec 字段 `setdefault` 兜底, 统一 eval 配置解析语义。
- `slime/rollout/sglang_rollout.py` (模块 评测采样; 类别 source; 类型 core-logic; 符号 `eval_rollout_single_dataset`): 采样参数接线修复: `stop / stop_token_ids` 从硬编码 args 改为 `dataset_cfg` 优先, 并补上 `min_new_tokens` 解析, 使 per-dataset 配置真正生效。
- `tests/test_eval_config.py` (模块 单元测试; 类别 test; 类型 test-coverage; 符号 `_args`, `test_non_spec_defaults_reach_every_dataset`, `test_dataset_entry_overrides_default`, `test_stop_fields_resolve_dataset_then_default_then_args`): 新增 5 个 CPU 单元测试, 直接钉住本次修复的配置契约 (非 spec 默认值透传、dataset 覆盖优先级、stop 字段三级解析、未知键拒绝、args 兜底), 其中 4 个在 main 上会失败。
- `.github/workflows/pr-test.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure): 将新增的 `test_eval_config.py` 注册进 `cpu-unittest` job (`num_gpus` 为 0), 保证该测试在 CI 中运行。
- `.github/workflows/pr-test.yml.j2` (模块 CI 配置; 类别 infra; 类型 infrastructure): CI 工作流的 Jinja 模板同步注册 `test_eval_config.py`, 保证模板生成后的 workflow 与手写版本一致。

关键符号: `build_eval_dataset_configs`, `eval_rollout_single_dataset`

关键源码片段

`slime/utils/eval_config.py`

核心修复所在: `DATASET_RUNTIME_SPECS` 扩展三个字段, `build_eval_dataset_configs` 新增未知键校验与非 spec 字段 `setdefault` 兜底, 统一 eval 配置解析语义。

```
# slime/utils/eval_config.py (head 版本关键片段)
```

```
# DATASET_RUNTIME_SPECS 中新增的三个字段:
```

与 temperature / top_p 等字段走同一套 dataset > defaults > args 解析管线。

```
'stop': {
    'dataset_keys': ('stop',),
    'default_keys': ('stop',),
    # 命令行 fallback: 训练侧的停止符参数
    'arg_attrs': ('rollout_stop',),
},
'stop_token_ids': {
    'dataset_keys': ('stop_token_ids',),
    'default_keys': ('stop_token_ids',),
    'arg_attrs': ('rollout_stop_token_ids',),
},
'min_new_tokens': {
    'dataset_keys': ('min_new_tokens',),
    'default_keys': ('min_new_tokens',),
    # 注意: 这里走 eval 专用参数, 此前它在仓库里没有任何读取者
    'arg_attrs': ('eval_min_new_tokens',),
},
```

```
def build_eval_dataset_configs(
    args: Any,
    raw_config: Iterable[dict[str, Any]],
    defaults: dict[str, Any],
) -> list[EvalDatasetConfig]:
    defaults = defaults or {}
    # 合并运行时参数与采样参数两张 spec 表, 统一解析入口
    combined_specs = {**DATASET_RUNTIME_SPECS, **DATASET_SAMPLE_SPECS}
```

与 dataset entry 的从严校验对齐: defaults 里的未知键不再静默忽略。

合法键 = 数据类所有字段 U spec 表中声明的 default_keys。

用反射而非硬编码清单, 避免以后新增字段时再次漏配。

```
valid_default_keys = {f.name for f in fields(EvalDatasetConfig)} | {
    key for spec in combined_specs.values() for key in spec['default_keys']}
}
```

unknown_keys = set(defaults) - valid_default_keys

if unknown_keys:

```
    raise ValueError(
        f'Unknown key(s) in eval.defaults: {sorted(unknown_keys)}. '
        f'Valid keys: {sorted(valid_default_keys)}.'
    )
```

datasets: list[EvalDatasetConfig] = []

for cfg in raw_config:

cfg_dict = dict(cfg or {})

先走 spec 表: dataset > defaults > args 三级优先级

_apply_dataset_field_overrides(args, cfg_dict, defaults, combined_specs)

再兜底非 spec 字段 (rm_type / repetition_penalty / app_service ...):

setdefault 保证 dataset entry 显式给出的值不被 defaults 覆盖,

```

# 未给出的键则由 defaults 补齐，与文档契约一致。
for key, value in defaults.items():
    if key not in combined_specs:
        cfg_dict.setdefault(key, value)
dataset = EvalDatasetConfig(**cfg_dict)
datasets.append(dataset)
return datasets

```

slime/rollout/sclang_rollout.py

采样参数接线修复：stop / stop_token_ids 从硬编码 args 改为 dataset_cfg 优先，并补上 min_new_tokens 解析，使 per-dataset 配置真正生效。

slime/rollout/sclang_rollout.py (head 版本关键片段)
eval_rollout_single_dataset 内部：构建 eval 采样的基础参数。

```

base_sampling_params = dict(
    temperature=dataset_cfg.temperature,
    top_p=dataset_cfg.top_p,
    top_k=dataset_cfg.top_k,
    max_new_tokens=dataset_cfg.max_response_len,
    # 修复前这里直接写死 args.rollout_stop, per-dataset 的 stop 配置是死配置；
    # 现在 dataset 配置优先、训练侧参数兜底。
    stop=dataset_cfg.stop if dataset_cfg.stop is not None else args.rollout_stop,
    stop_token_ids=(
        dataset_cfg.stop_token_ids
        if dataset_cfg.stop_token_ids is not None
        else args.rollout_stop_token_ids
    ),
    skip_special_tokens=(
        dataset_cfg.skip_special_tokens
        if dataset_cfg.skip_special_tokens is not None
        else args.rollout_skip_special_tokens
    ),
    no_stop_trim=dataset_cfg.no_stop_trim if dataset_cfg.no_stop_trim is not None else True,
    spaces_between_special_tokens=False,
)
if dataset_cfg.repetition_penalty is not None:
    base_sampling_params['repetition_penalty'] = dataset_cfg.repetition_penalty

# min_new_tokens 此前完全没有接入：--eval-min-new-tokens 在仓库里没有读取者。
# 现在 dataset 配置优先，其次命令行参数；两者都未设置时不写入采样参数。
min_new_tokens = dataset_cfg.min_new_tokens
if min_new_tokens is None:
    min_new_tokens = getattr(args, 'eval_min_new_tokens', None)
if min_new_tokens is not None:
    base_sampling_params['min_new_tokens'] = min_new_tokens

```

评论区精华

本 PR 没有任何 review 评论或审核线程，唯一相关交互是合并前 zhuzilin 让 @copilot 解决 main 合入的冲突（第二次 commit 为 merge main）。因此没有公开的取舍交锋可供提炼；设计权衡主要来自 PR body 的自述：spec 表迁移是 #1005 回归的根源，而本 PR 选择用反射校验 + setdefault 兜底，而不是把全部字段都搬进 spec 表，保留了既有解析结构的稳定性。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 行为收紧：eval.defaults 中未知键从静默忽略改为抛 ValueError，存量配置若有拼写错误或自定义扩展键，升级后启动即失败；报错信息会列出合法键便于修复，但仍是 breaking change，需要通知使用方排查。
2. 采样行为变化：per-dataset stop / stop_token_ids / min_new_tokens 重新生效，此前实际使用训练 stop 字符串的 eval 会改变生成结果与分数，部分用户可能观察到分数波动（这是修复意图，但需核对配置）。
3. setdefault 边界：若 dataset entry 显式写 null，该键已存在，defaults 不会兜底，与 spec 字段 _first_not_missing 把 None 视为缺失的语义不完全一致，属于未覆盖的边界情况。
4. 影响面：涉及 eval 配置解析与 slang eval 采样路径，覆盖所有使用 eval.datasets / eval.defaults / --eval-min-new-tokens 的评测场景；无性能与安全影响。- 影响：影响范围集中在 eval 配置层与 slang eval rollout：所有依赖 eval.defaults 共享参数、per-dataset stop / min_new_tokens、以及 --eval-min-new-tokens 参数的评测任务都会受到影响，原本静默失效的配置开始按文档契约生效。对团队而言，CI cpu-unitest 新增测试文件，后续改动 eval 配置解析有回归保护；对用户而言，配置错误从无声失败变为显式报错，可诊断性显著提升。该改动不涉及训练路径、模型权重、API 兼容性，对 rollout 的额外开销只有一个 setdefault 循环，可忽略。- 风险标记：defaults 未知键升级为报错，eval 采样行为回归修复，setdefault 对显式 null 不兜底，--eval-min-new-tokens 重新生效

关联脉络

- PR #1005 Fix evaluation parameter parsing: PR body 明确指出该回归源于 #1005 重构：spec 表迁移时删除了 stop / stop_token_ids / min_new_tokens 的解析行，本 PR 是这条配置解析重构线的收尾。
- PR #2261 fix(rollout): restore partial continuation token budget: 与 #2261 同时修改 slime/rollout/slang_rollout.py，同属 eval/rollout 采样参数正确性修复线，合并阅读可相互印证。