

PR #35928 完整报告

sgl-project/sglang

Expose the declared sglang env vars of a scheduler in its internal state

合并时间: 2026-08-24 20:20

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35928>

执行摘要

- 一句话: 调度器内部状态支持可选导出声明环境变量
- 推荐动作: 值得精读, 尤其关注安全边界设计: 默认关闭的开关 + 声明式 secret 标记 + 白名单导出, 可作为对外暴露内部信息的参考模式。

功能与动机

提交信息明确说明: Add `SGLANG_EXPOSE_OWN_ENV_VARS` (default off) and, when it is on, attach the environment `Envs` declares as `env_vars` inside the dict returned by `Scheduler.get_internal_state()`, which `/server_info` surfaces per DP scheduler。即让运维人员能够通过现有内部状态接口查看调度器实际声明的环境变量, 同时默认关闭以保护敏感信息。

实现拆解

1. 在 `python/sglang/srt/envron.py` 中为 `EnvField` 增加 `secret` 标记 (默认 `False`) , 并扩展 `_DeprecatedEnvFallback` 透传该参数; 同时新增模块级函数 `exportable_env_vars()` 与 `_exportable_value()`, 前者遍历 `Envs` 类中所有 `EnvField` 实例, 过滤 `secret` 字段并仅导出当前 `os.environ` 中存在的变量, 后者对非 UTF-8 值用 `base64` 包装 (前缀 `base64:`) 。
2. 在 `Envs` 类中新增开关 `SGLANG_EXPOSE_OWN_ENV_VARS = EnvBool(False)`, 并将 `EXA_API_KEY` 标记为 `secret=True`, 作为敏感字段声明示例。
3. 在 `python/sglang/srt/managers/scheduler.py` 的 `get_internal_state()` 中, 当开关开启时调用 `exportable_env_vars()` 并写入 `ret["env_vars"]`。
4. 新增测试文件 `test/registered/unit/managers/test_scheduler_internal_state_env_vars.py`, 覆盖开关默认关闭、关闭时不输出键、开启时输出已声明变量、不输出未声明或 `secret` 变量、非 UTF-8 值编码等场景。CI 方面, 通过 Codex 手动触发测试运行验证该测试在最终 rebase 头通过。

关键文件:

- `python/sglang/srt/envron.py` (模块 环境配置; 类别 `source`; 类型 `core-logic`; 符号 `EnvField.init`, `exportable_env_vars`, `_exportable_value`) : 核心实现: 引入 `secret` 标记、新增导出函数与非 UTF-8 编码处理, 定义开关变量。
- `python/sglang/srt/managers/scheduler.py` (模块 调度器; 类别 `source`; 类型 `dependency-wiring`; 符号 `Scheduler.get_internal_state`) : 将导出功能接入

get_internal_state(), 是内部状态 API 的扩展点。

- test/registered/unit/managers/test_scheduler_internal_state_env_vars.py (模块 单元测试; 类别 test; 类型 test-coverage; 符号 TestSchedulerInternalStateEnvVars, _get_internal_state, test_the_gate_is_declared_off, test_env_vars_absent_when_disabled): 新增测试覆盖开关行为、secret 过滤与非 UTF-8 编码, 验证功能正确性。

关键符号: exportable_env_vars, _exportable_value, Scheduler.get_internal_state

关键源码片段

python/sglang/srt/envron.py

核心实现: 引入 secret 标记、新增导出函数与非 UTF-8 编码处理, 定义开关变量。

```
# python/sglang/srt/envron.py
_NON_UTF8_PREFIX = "base64:"

class EnvField:
    _allow_set_name = True

    def __init__(self, default: Any, secret: bool = False):
        self.default = default
        self._set_to_none = False
        # secret 标记: 声明后该字段不会被 exportable_env_vars() 导出
        self.secret = secret

    # 其余方法 (parse/get 等) 此处省略

def exportable_env_vars() -> dict[str, str]:
    """收集 Envs 中声明且当前已设置、非 secret 的环境变量, 按名称排序输出。"""
    return {
        field.name: _exportable_value(os.environ[field.name])
        for field in sorted(
            (value for value in vars(Envs).values() if isinstance(value, EnvField)),
            key=lambda field: field.name,
        )
        if not field.secret and field.name in os.environ
    }

def _exportable_value(value: str) -> str:
    """环境变量值含非法 UTF-8 字节时, 用 base64 包装, 避免 json 序列化失败。"""
    try:
        value.encode()
    except UnicodeEncodeError:
        return _NON_UTF8_PREFIX + base64.b64encode(
            value.encode(errors="surrogateescape")
        ).decode()
    return value
```

python/sclang/srt/managers/scheduler.py

将导出功能接入 `get_internal_state()`，是内部状态 API 的扩展点。

```
# python/sclang/srt/managers/scheduler.py
def get_internal_state(self, recv_req: GetInternalStateReq):
    # ... 前面已填充 ret 字典的各类内部指标 ...

    # 可观测性：默认关闭，开启后附加声明过的环境变量
    if envs.SGLANG_EXPOSE_OWN_ENV_VARS.get():
        ret["env_vars"] = exportable_env_vars()

    # 这些字段无法 msgpack 序列化（config 对象与 signal handler），丢弃
    ret.pop("model_config", None)
    ret.pop("custom_sigquit_handler", None)

    return GetInternalStateReqOutput(internal_state=msgspec_to_builtins(ret))
```

评论区精华

本 PR 无人工 review 评论。Issue 评论区仅有 Codex 自动代理针对 CI 失败说明与手动触发测试验证的留言，未涉及设计决策讨论。

- 暂无高价值评论线程

风险与影响

- 风险：1) 敏感字段过滤依赖 `secret` 标记，若未来声明新变量时漏标 `secret=True`，开启开关后会泄露；当前仅 `EXA_API_KEY` 被标记，需要审计 `Envs` 中所有可能含凭据的字段。2) 开关默认关闭降低了默认风险，但用户显式开启后需自行评估暴露范围。3) `_exportable_value` 对非 UTF-8 值用 `surrogateescape` 编码回 `environ` 原始字节，可能暴露二进制数据，但 `base64` 前缀使输出可辨识。4) 每次调用 `get_internal_state()` 会遍历 `Envs` 类所有字段，数量级约数百，性能影响可忽略。
- 影响：对用户：默认无感知；开启开关后 `/server_info` 会多返回 `env_vars` 字段。对开发者：新增 `EnvField.secret` 声明约定，声明敏感环境变量时必须设置 `secret=True`。对系统：仅影响 `Scheduler.get_internal_state()` 的返回内容，不影响其他路径。对团队：增强了多 DP 环境差异的可观测性，便于运维排查配置不一致问题。
- 风险标记：默认关闭开关，敏感字段依赖手动标记，内部状态 API 扩展

关联脉络

- PR #35929 Report the whole server's world size in the scheduler's internal state: 同样扩展 `Scheduler.get_internal_state()` 内部状态字段，且同样涉及 `server_args/` 调度器内部状态的可观测性增强，与本 PR 属于同一演进线。