MoonAgentCheck 项目申报书

基本信息

- **项目名称：** MoonAgentCheck：Agent 工具行为契约与离线回归测试
- **参赛者：** 韦昌斌（GitHub：2515050242）
- **联系方式：** 在报名表填写，公开副本不列出
- **GitHub 仓库链接：** 2515050242/moonagentcheck
- **项目方向：** 章程 4.2「季度优秀社区项目评选方向」，维护已有社区项目
- **是否为移植项目：** 否；原创实现，采用 Apache-2.0 许可证
- **版本状态：** Mooncakes 0.3.0，与仓库版本对齐（2026-10-21）

项目简介

我沿用九月通过初审的 MoonAgentCheck 继续开发。工具把接入方记录的工具行为转换为事件，用 call_id 关联单次调用与结果，用 operation_id 识别同一业务操作，检查结果缺失、失败后继续写入、重试超限和重复副作用，返回规则码与事件位置。核心用 MoonBit 实现，离线运行，不连接模型或真实工具。

十月重点是让 Responses 混合输出和未知结果得到可靠处理，并打通 Python 示例到 MoonBit 的实际调用链。消息与推理项按明确规则跳过，未支持类型保留输入索引并阻止通过。operation_id 与执行结果 outcome 由应用遥测提供，不假定为 API 原生字段。

核心功能范围

- **行为评估：** 提供事件模型、重试与工具白名单策略、调用/结果/写入关联检查，以及 AGC001–AGC018 稳定规则码和 JSON 报告。
- **Responses 适配与 JSON 入口：** 支持工具调用及结果映射、message / reasoning 混合序列；仓库 main 接收 JSON 字符串，实际调用 MoonBit 适配器和评估器。
- **三类离线场景：** 数据导出正常配对时通过；仓库助手重复写入产生 AGC004；工单结果未知产生 AGC018 并令报告 passed: false。
- **真实接入回归：** 5 个 Python fixture 通过子进程调用 MoonBit CLI，断言其报告；Python 只做输入映射与断言，不独立重写评估规则。
- **工程交付：** 以 MoonBit 0.10.16+b1a0e846e 验证，包含 96 项 MoonBit 测试、可运行演示、三项 CI job 和发布包的 2 项独立消费测试。测试数量以对应提交的 CI 为准。

移植或参考说明

本项目为原创实现，非现有项目移植。Responses 适配依据公开字段结构进行映射，不包含上游框架源码。来源与适配边界见 Responses 说明，生态职责对照见 生态定位。当前场景均为合成离线 fixture；不宣称真实 Agent 集成、权限沙箱或性能优势。

开发与验证说明

项目目标、功能边界与验收要求由参赛者提出；AI 辅助实现、测试整理和文档校对。最终申报由参赛者确认并负责，验证结论以可复现命令和公开 CI 为准。验收指南提供运行步骤、预期结果与十月新增提交索引；PDF 副本与本页内容一致。