

SOLAR LLM + FAIL-CLOSED WORKFLOW

AI가 실패해도 업무 사고 없이 끝나는가

문서 구조화, 사람 승인, checkpoint 복구, 재시도와 중복 방지를 하나의 실행 가능한 데모와 검증 evidence로 묶었습니다.

432

고유 시험

24/24

최종 재실행

20,736

matrix testcase

0

중복 효과

Windows · Linux / Python 3.11 · 3.12 · 3.13

합성 데이터 · 로컬 mock tool · 서버 측 API key · 공개용 evidence 요약

WHY RELIABILITY, NOT ONLY MODEL QUALITY

좋은 답을 만드는 것과 안전하게 일을 끝내는 것은 다릅니다

Agent가 도구를 쓰기 시작하면 정확도 외에 승인 경계, 중단 복구, 재시도, 중복 효과, 실행 증거가 제품 품질이 됩니다.

01

환각과 누락

모델 출력이 그럴듯해도 필수값이 없거나 허용되지 않은 필드가 섞일 수 있습니다.

02

승인 없는 실행

설명 생성과 실제 변경 권한을 붙이면 모델의 오류가 곧 업무 사고가 됩니다.

03

응답 유실

저장은 끝났지만 응답만 사라지면 단순 재시도가 같은 기록을 두 번 만들 수 있습니다.

04

중단과 재개

프로세스가 바뀐 뒤 기억만 믿고 재개하면 승인 대상과 실행 인자가 달라질 수 있습니다.

PROJECT QUESTION

“모델이나 시스템이 실패해도 승인되지 않은 작업과 중복 작업을 0으로 유지할 수 있는가?”

SEPARATE GENERATION FROM AUTHORITY

LLM은 읽고, 통제층이 실행합니다

확률적인 생성 결과와 실제 변경 권한을 분리한 fail-closed 구조입니다.

01

Solar LLM

문서를 strict JSON 후보로 구조화

02

Schema + Policy

필수값·허용 필드·업무 범위를 검사

03

Human Approval

case·tool·arguments digest를 승인과 결속

04

Checkpoint Resume

새 runtime이 SQLite 상태에서 재개

05

Idempotent Tool

동일 요청이면 persisted result를 재사용

핵심 원칙 · 승인 증거가 없거나 schema가 맞지 않거나 상태가 불명확하면 실행하지 않습니다.

90-SECOND INTERACTIVE SCENARIO

응답이 사라져도 기록은 한 번

실제 화면에서는 문서 분석, 승인, fault injection, replay를 한 흐름에서 확인합니다.

Agent Reliability Live Demo

샘플 / LIVE 표시

01 합성 지출 문서

공급자: 한빛 데이터 서비스
청구번호: INV-2026-0718-04
금액: 1,280,000
요청자: 김도윤
사용목적: 성능 검증

선택: 저장 후 응답 유실

1. 문서 분석 시작

02 실행 결과

문서 분석

정책 검사

사람 승인

안전 기록

사람 승인이 필요합니다

승인 전 외부 기록 0회

2. 승인하고 checkpoint에서 재개

1
실제 기록

0
중복

1+
결과 재사용

시연 결론

commit 후 응답이 유실되고 같은 요청이 다시 와도 persisted result를 재사용해 실제 효과는 한 번만 남습니다.

공개판은 Cloudflare 서버의 Solar Pro 3를 사용하며 API 키와 입력 원문을 브라우저에 저장하지 않습니다.

04 / 08

FROM PROTOTYPE TO CONTROL PLANE

완성한 것은 ‘챗봇’이 아니라 실행과 복구의 제어층입니다

서로 독립적으로 시험할 수 있는 여섯 개 책임으로 나뉘었습니다.

STATE MACHINE

명시적 상태 전이

불법 전이와 terminal 이후 write를
차단

CHECKPOINT

지속 가능한 재개

프로세스가 바뀌어도 SQLite
state에서 재개

APPROVAL

승인 증거 결속

case·tool·arguments·TTL이 맞을
때만 실행

RETRY

분류된 실패와 budget

일시 오류만 제한 횟수 안에서 재시도

IDEMPOTENCY

중복 효과 방지

같은 key의 committed result를
조회·재사용

EVALUATOR

독립 검증

agent 주장 대신 persisted state와
journal을 판정

표현 계층 추가

Solar strict-output adapter + Cloudflare Pages Functions·D1 공개판 + 반응형 한국어 UI + 로컬 자동 실행기 + demo test

A REAL FAILURE FOUND DURING VALIDATION

Windows PID 재사용이 cleanup 증거를 속일 수 있었습니다

오래 도는 matrix에서 드러난 문제를 identity 결속으로 수정하고 회귀 시험에 고정했습니다.

BEFORE

종료한 child의 PID만 기억

운영체제가 같은 PID를 다른 프로세스에 재사용하면, 새 프로세스를 예전 child로 오인하거나 잘못 정리할 위험이 있었습니다.



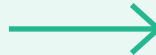
PID 8124

다른 process

AFTER

PID + creation time 결속

프로세스 ID와 생성 시각이 모두 일치할 때만 같은 실행 주체로 인정했습니다. cleanup과 회수 대상의 identity가 고정됩니다.



8124 + 10:42:18.431

identity match

WHY IT MATTERS

Agent 평가에서는 testcase PASS만큼 '실패 후 무엇이 남았는가'가 중요합니다. 이 버그는 orphan·zombie 0이라는 주장 자체가 정확한 identity 검사에 의존한다는 점을 보여줬습니다.

EVIDENCE, NOT CLAIMS

고정된 소스와 후보를 독립 검증으로 다시 확인했습니다

시험 수가 아니라 소스 결속, 실패 주입, cleanup, 재현성까지 판정했습니다.

432/432

manifest와 discovery 일치

고유 시험

14/14

기준·음성대조·독립검증

Authority

10/10

Windows/Linux + 3 Python

Platform

24/24

최종 후보 고정 재실행

Exact-final

20,736

 matrix testcase executions · 432×24 행

failure · error · missing · duplicate · unexpected · orphan · zombie = 각 0

무결성

99 source files · source tree SHA-256 고정 · candidate/result/exact-final digest 결속 · rollback 420 → reapply 432 PASS

WHY THIS FITS AGENT GYM

모델 · 환경 · 평가 · 운영 안전을 하나의 검증 루프로 다뤘습니다

결과물은 PDF 한 장의 주장보다 재현 가능한 환경과 실패 evidence를 우선합니다.

ENVIRONMENT

상태 · 도구 · 승인 경계를 명시한 synthetic environment

EVALUATION

독립 evaluator와 executable ablation, source-bound receipt

ROBUSTNESS

timeout · stall · signal · 응답 유실 · process hazard 주입

REPRODUCIBILITY

Windows/Linux, Python 3.11-3.13 cold/warm matrix

HONEST BOUNDARIES

- 합성 문서와 로컬 mock business tool을 사용합니다.
- idempotency 계약을 검증하며 distributed exactly-once는 주장하지 않습니다.
- 432개 core 검증과 10개 demo · 자동 실행기 시험의 범위를 구분합니다.
- 라이브 Solar E2E는 키 1회 자동 실행 후 별도 PASS receipt로 판정합니다.
- 기술 검증은 운영환경의 완전한 안전성이나 채용 결과를 보장하지 않습니다.

공개 웹주소 접속 → 키 입력 없이 가상 문서 분석 · 승인 · 오류복구 · 중복방지 체형

로컬 실행본도 함께 제공하며, 두 경로 모두 실제 Solar 호출 증거와 비밀정보 비저장을 검증