

도달 판단 오류의 진단과 수정

로봇 벽 오르기 · 합성 사례 Full 샘플

실제 관측 · 도달일 통과 수

33 / 68 → 68 / 68

같은 도달 가능 사례 68개 · 변경 전 → 변경 후

01	02	03	04	05
의도 정의	시험 설계	실패 진단	한정된 수정	재검증
무엇을 확인할까	기준을 사례로	어떤 조건일까	어디를 바꿀까	같은 시험으로

결과 요약

낮에 도달한 뒤에도 밤의 하강을 실행하는 순서를 수정했습니다. 고정된 도달 가능 68개에서 실패 35개가 통과로 전환됐고 기존 통과 33개가 유지됐습니다. 별도 가드 4개의 원시 동작에는 변화가 없었습니다.

읽는 순서	확인할 내용
02 의도·판단 기준	어떤 동작을 정답으로 합의했는가
03 시험 설계	왜 이 사례들을 골랐는가
04-06 실패 구조	어떤 조건을 우선 조사했는가
07 코드·패치	어디를 바꾸고 무엇을 예상했는가
08 재검증	개선·회귀·가드를 어떻게 확인했는가
09-10 사례별 근거	72개 전체의 전후 관측과 종료 상태

합성 사례 · 실제 고객 프로젝트가 아닙니다.

대상은 결정적 로컬 Python 프로그램입니다. 이 결과는 LLM API 품질, 다른 제품·환경 또는 사업 성과를 입증하지 않습니다.

01 / DEFINITION

도달 기준과 확인 범위

낮 상승 직후 목표에 도달하면 즉시 종료합니다.

대표 문제 · Q011

로봇이 32m 벽을 낮에 7m 올라가고, 밤마다 2m 미끄러집니다.
처음 꼭대기에 도달하는 날은 언제인가요?

합의된 의미: 낮 상승 직후 32m 이상이면 즉시 종료합니다. 이후 밤 하강은 실행하지 않습니다.

범위와 입출력

항목	기준
시작	벽 높이·낮 상승·밤 하강을 포함한 한국어 문장 입력
종료	처음 도달한 날과 일치하는 결과 메시지 반환
유효 입력	높이·상승은 양의 정수, 하강은 0 이상의 정수
출력	answer: 도달일 정수 / unit: day / message: 결과 설명
변경 경계	도달 판정과 실행 순서. 정답표·시험 Oracle은 변경하지 않음

다섯 판단 기준

ID	확인할 동작
AC-01	answer가 첫 도달일과 일치한다.
AC-02	상승 직후, 하강 전에 도달 여부를 확인한다.
AC-03	도달한 날에는 하강·다음 날 전이를 실행하지 않는다.
AC-04	다른 숫자에도 동일한 규칙을 적용한다.
AC-05	결과 메시지가 answer와 모순되지 않는다.

대표 입력 32·7·2에 답 6을 하드코딩하거나, 문제별 정답만 교체하는 수정을 금지했습니다.

근거 S1 · Definition Contract E1-v1.2. 기준의 존재가 기준별 독립 채점 100%를 의미하지 않습니다.

02 / TEST DESIGN

72개 사례로 구성된 고정 시험

노출·대조 쌍과 경계·별도 회귀 사례로 시험을 구성했습니다.

분석용 48 결함 노출·은폐 24쌍	경계 확인 12 도달 가능 8 별도 가드 4	별도 회귀 12 학습에 쓰지 않은 별도 숫자군
--	---	--

다섯 변형 축

축	포함한 조건
도달 여유	목표선에 정확히 도달 / 목표선을 초과
하강 후 위치	목표선 아래로 내려감 / 목표선 이상을 유지
도달 날짜	첫날 / 여러 낮·밤 뒤
하강 규칙	하강 없음 / 양의 하강 / 도달 불가
숫자 규모	작은 정수 / 대표 규모 / 더 큰 정수

독립 정답과 전후 비교 기준

Agent와 독립적인 두 정답 계산 방식이 72개 모두에서 일치하는지 확인했습니다. 시험과 정답을 고정한 뒤 변경 전·후 버전에 같은 사례 ID를 사용했습니다.

분모	의미
68개	도달 가능한 사례의 도달일 정답 판정
4개	도달 불가 가드. 고객용 출력 계약 미정으로 정답률 분모에서 제외
144회	변경 전 72회 + 변경 후 72회. 개발 중 추가 실행은 포함하지 않음

72개는 이 사례의 시험 설계입니다. 고객마다 필요한 조건·사례·반복 실행·예산을 별도로 합의하며, 일반적인 통계적 충분성을 뜻하지 않습니다.

근거 S2 · Frozen Test Pack / coverage_report. 지원 문장과 동결 입력 범위 안의 시험입니다.

03 / FAILURE STRUCTURE

실패가 모이는 조건

변경 전 결과와 STRIDE 기여 분석

변경 전 · 관측

33 통과

35 실패

변경 전 도달 가능 68개 중 33개 통과, 35개 실패. 분석용 Core 48개의 통과·실패를 함께 사용했습니다.

도달 경계·하강 조건의 기여가 가장 큼니다



순위 안정성

6 / 6

Core fold 재적합 모두에서
도달 조건 그룹이 1위

직접 관측한 대조

밤 뒤 목표선 아래: 35/35 실패

목표선 이상 유지: 0/33 실패

고정 시험 분포 안의 관계

기여값은 이 표현·시험 분포의 함수적 성분입니다. SHAP 값·확률·인과 원인·패치 효과가 아닙니다. 이 단계에서는 “도달 시점의 실행 순서를 우선 조사한다”까지 해석합니다.

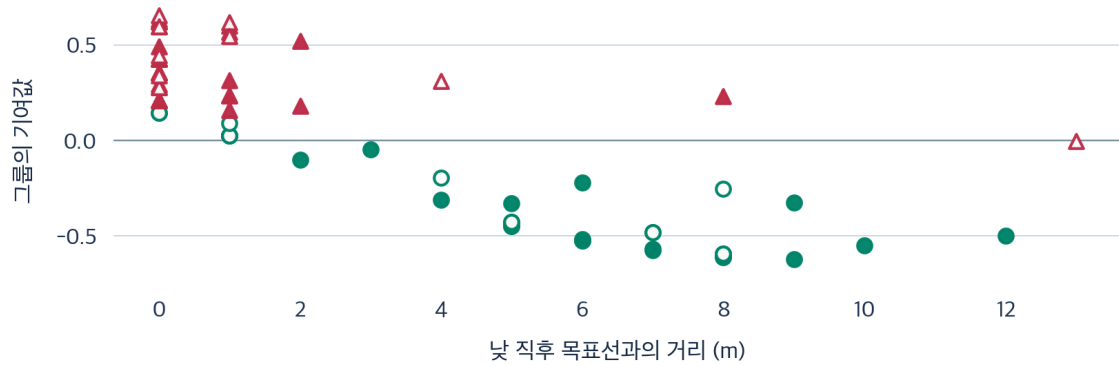
근거 S3-S4 · Baseline / component_summary / 6-fold refit. 분석용 Core 48개로 학습·설정 선택, Holdout 12개와 도달 가능 Boundary 8개는 별도 확인.

03 / FIVE CONDITION VIEWS · 1/2

조건별 기여도 분포

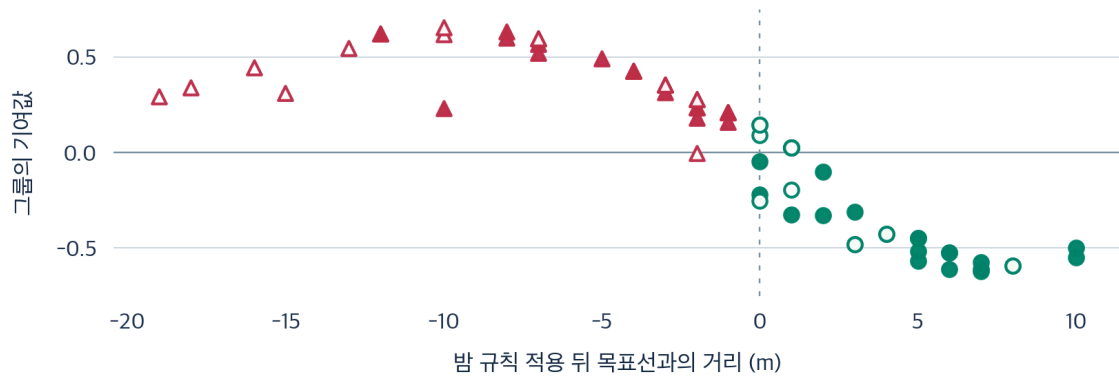
● 통과 · ▲ 실패 / 채운 점: Core · 빈 점: Holdout·Boundary

01 · 낮 직후 목표선과의 거리 (m)



낮에 막 도달한 조건부터 살펴봅니다.

02 · 밤 규칙 적용 뒤 목표선과의 거리 (m)



밤 뒤 목표선 아래가 되는 조건이 실패와 연결됩니다.

Y축은 해당 의미 그룹의 signed contribution입니다. 양수는 실패 surrogate를 높이고 음수는 낮춥니다. 도달 여유·하강 후 위치·하강량은 같은 도달 조건 그룹의 기여를 공유하므로, 각각 독립적인 중요도로 더하지 않습니다. 근거 S4.

03 / FIVE CONDITION VIEWS · 2/2

조건별 기여도 분포

● 통과 · ▲ 실패 / 채운 점: Core · 빈 점: Holdout · Boundary

03 · 도착 전에 완료된 밤의 수 (회)



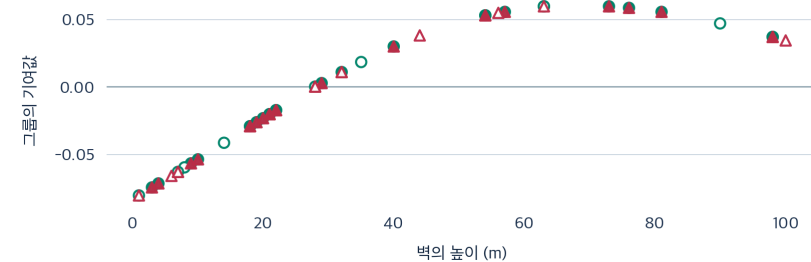
진행 길이의 기여는 상대적으로 작습니다.

04 · 밤마다 미끄러지는 거리 (m)



하강량은 도달 전후 위치와 함께 해석합니다.

05 · 벽의 높이 (m)



04 / CHANGE SURFACE

도달 직후 종료하도록 수정

조건식 신호 → 실행 기록 → 한정된 코드 변경

Q011 · 실패 노출

6일째 시작 25m → 상승 후 32m

도달했지만 하강 → 30m

다음 반복까지 이어져 7일 반환

Q006 · 결함 은폐

6일째 시작 30m → 상승 후 37m

하강 후에도 36m ≥ 32m

같은 순서 결함에도 6일 반환

변경 전과 선택한 Patch B

BEFORE

```
while position < height:
    day += 1
    position += climb
    position -= slip
```

PATCH B

```
while position < height:
    day += 1
    position += climb
    if position >= height:
        break
    position -= slip
```

설명용 코드 발췌 · 원본 days_to_reach 25-29행 부근. 원본 보존 후 별도 수정 버전을 생성했습니다.

후보별 실행 전 예상

후보	변경 내용	실행 전 예상
A · 특정 답 교체	32·7·2만 6일로 하드코딩	34/68 통과 · 50.0%
B · 실행 순서 수정	상승 직후 도달 확인, 미도달일 때만 하강	68/68 통과 · 100.0%

후보 점수는 실행 전 예상입니다. Patch B만 적용·재검증했으며 A는 실행하지 않았습니다. 정답 계산 기준과 정답표는 유지했습니다.

근거 S5 · 코드·Trace / candidate_interventions / patch_b_selection. 결과 확인은 다음 쪽 S6.

변경 후 결과와 회귀 확인

같은 72개를 수정 버전에 실행해 사례별로 대조했습니다.

변경 전 · 관측

33 통과	35 실패
-------	-------

Patch B 적용 후 · 관측

68 통과 · 0 실패

구분	변경 전 관측	B 실행 전 예상	변경 후 관측
분석용	24/48	48/48	48/48
도달 가능 경계	3/8	8/8	8/8
별도 회귀	6/12	12/12	12/12
도달일 통과 합계	33/68	68/68	68/68

개선·유지·가드 결과

점검	관측 결과	해석
기존 실패	35개 → 통과	동일 사례에서 개선
기존 통과	33개 유지 / 새 실패 0	해당 기존 통과 집합에서 회귀 미관측
도달 불가 가드	4개 원시 동작 변화 0	올바른 고객용 오류 응답을 입증한 것은 아님
실행 무효	0개	환경 오류를 실패 점수로 섞지 않음

적용 범위와 미확인 사항

이번 코드 버전과 동결 시험 묶음에서는 개선을 확인했습니다. 새 입력·다른 문장·다른 제품 기능·실제 고객 환경은 별도 검증이 필요합니다. 도달 불가의 고객용 출력 계약도 별도로 정해야 합니다.

근거 S6 · E5-PATCH-B-PAIRED-ATTEMPT-001.

사례별 전후 관측 01-36

입력은 높이 / 낮 상승 / 밤 하강 (m), 답은 도달일입니다.

ID	입력 H/C/S	구분	전 답	후 답	전환
Q001	100/16/11	경계	20	18	개선
Q002	73/13/0	분석	6	6	유지
Q003	63/13/3	회귀	7	6	개선
Q004	97/3/12	경계	—	—	가드
Q005	98/19/9	분석	10	9	개선
Q006	32/7/1	분석	6	6	유지
Q007	73/13/1	분석	7	6	개선
Q008	76/16/4	분석	7	6	개선
Q009	39/4/5	경계	—	—	가드
Q010	14/3/0	회귀	5	5	유지
Q011	32/7/2	분석	7	6	개선
Q012	35/9/3	회귀	6	6	유지
Q013	28/13/10	회귀	10	6	개선
Q014	20/14/8	분석	4	2	개선
Q015	29/18/0	분석	2	2	유지
Q016	22/10/1	분석	3	3	유지
Q017	6/6/7	경계	—	1	개선
Q018	40/15/2	분석	4	3	개선
Q019	54/16/3	분석	5	4	개선
Q020	81/15/4	분석	8	7	개선
Q021	54/20/3	분석	4	3	개선
Q022	22/8/1	분석	4	3	개선
Q023	63/17/8	회귀	7	7	유지
Q024	3/2/0	분석	2	2	유지
Q025	54/20/0	분석	3	3	유지
Q026	54/16/0	분석	4	4	유지
Q027	76/16/3	분석	6	6	유지
Q028	98/15/2	분석	8	8	유지
Q029	1/2/0	경계	1	1	유지
Q030	81/15/2	분석	7	7	유지
Q031	19/15/4	분석	2	2	유지
Q032	28/19/14	회귀	6	3	개선
Q033	18/15/6	분석	2	2	유지
Q034	8/16/0	경계	1	1	유지
Q035	21/11/0	분석	2	2	유지
Q036	4/4/12	분석	—	1	개선

—는 도달일 반환값 없음. 가드 Q004·Q009·Q043·Q050은 전후 AGENT_EXCEPTION, Q017·Q036·Q037·Q054·Q063은 변경 전 AGENT_NO_RETURN에서 변경 후 RETURNED로 전환됐습니다.
그 외 도달 가능 사례는 전후 RETURNED입니다.
전체 원시 종료 상태·판정은 동봉한 72개 전후 비교 CSV에 포함했습니다. 근거 S6.

사례별 전후 관측 37-72

입력은 높이 / 낮 상승 / 밤 하강 (m), 답은 도달일입니다.

ID	입력 H/C/S	구분	전 답	후 답	전환
Q037	9/17/18	분석	—	1	개선
Q038	4/4/0	분석	1	1	유지
Q039	20/14/0	분석	2	2	유지
Q040	28/3/1	회귀	14	14	유지
Q041	21/11/1	분석	3	2	개선
Q042	19/15/9	분석	4	2	개선
Q043	11/1/2	경계	—	—	가드
Q044	90/1/0	경계	90	90	유지
Q045	44/20/19	경계	44	25	개선
Q046	22/8/0	분석	3	3	유지
Q047	22/10/4	분석	4	3	개선
Q048	57/18/5	분석	5	4	개선
Q049	7/20/15	회귀	2	1	개선
Q050	100/1/19	경계	—	—	가드
Q051	21/9/1	분석	3	3	유지
Q052	29/18/7	분석	3	2	개선
Q053	21/15/2	분석	2	2	유지
Q054	1/1/2	경계	—	1	개선
Q055	57/18/2	분석	4	4	유지
Q056	3/2/1	분석	3	2	개선
Q057	10/8/4	분석	3	2	개선
Q058	10/8/0	분석	2	2	유지
Q059	21/9/3	분석	4	3	개선
Q060	32/20/18	경계	16	7	개선
Q061	18/15/12	분석	6	2	개선
Q062	21/15/8	분석	3	2	개선
Q063	7/11/19	회귀	—	1	개선
Q064	18/9/3	분석	3	3	유지
Q065	18/9/4	분석	4	3	개선
Q066	7/12/1	회귀	1	1	유지
Q067	56/17/16	회귀	56	40	개선
Q068	98/15/3	분석	9	8	개선
Q069	40/15/0	분석	3	3	유지
Q070	63/15/4	회귀	6	6	유지
Q071	98/19/8	분석	9	9	유지
Q072	9/17/1	분석	1	1	유지

—는 도달일 반환값 없음. 가드 Q004·Q009·Q043·Q050은 전후 AGENT_EXCEPTION, Q017·Q036·Q037·Q054·Q063은 변경 전 AGENT_NO_RETURN에서 변경 후 RETURNED로 전환됐습니다.

그 외 도달 가능 사례는 전후 RETURNED입니다.

전체 원시 종료 상태·판정은 동봉한 72개 전후 비교 CSV에 포함했습니다. 근거 S6.

APPENDIX / EVIDENCE INDEX

근거 목록과 해석 범위

공개 합성 사례의 저장 실행 자료와 해석 범위

근거	파일·기록	이 보고서에서의 역할
S1	definition_contract.json E1-DEFINITION-002 / E1-v1.2	의도·판단 기준·입출력
S2	coverage_report.json E2 first-reach-test-pack	72개 구성·변형 축·독립 정답 일치
S3	aggregate_report.json E3-1-BASELINE-ATTEMPT-001	변경 전 관측
S4	component_summary.json row_decomposition.jsonl E3-2-FAILURE-STRUCTURE-ATTEMPT-001	기여 순위·5개 조건 그래프
S5	candidate_interventions.json E4-CHANGE-SURFACE-ATTEMPT-001	수정 후보와 실행 전 예상
S6	paired_case_results.jsonl E5-PATCH-B-PAIRED-ATTEMPT-001	72개 전후 관측·회귀·가드

기여 분석의 오차

실패 0/1을 근사하는 연속 surrogate와 관측값 사이의 오차가 남습니다. Core 교차검증 MAE는 0.1975, 별도 Holdout은 0.2648, 도달 가능 Boundary는 0.2642입니다. 학습 적합도와 별도 확인 성능을 구분하며, 설명되지 않은 잔차를 특정 원인에 억지로 배분하지 않습니다.

도달일 정답 판정과 실행 순서 확인은 서로 다른 증거입니다. 68/68이라는 점수만으로 모든 판단 기준·모든 입력·모든 기능이 검증됐다고 해석하지 않습니다.

함께 제공하는 자료

72개 전후 비교 CSV: 입력·관측 답·판정·종료 상태·전환을 포함합니다.
원본 근거의 파일 경로와 SHA-256은 내부 근거 관리 목록으로 보관합니다.

데모 Full v9 / Hero v4와 같은 저장 실행 자료를 사용했습니다. 실제 고객 납품·성과 증명서가 아닙니다.