# Beyond Pass Rates

통과율은 결과를 가립니다.
실행 기록은 그 결과에 가려진 뒷면을 추적하게 합니다.

Tracing GEODE's Execution Quality

Harbor · Terminal-Bench 2.1 · native Codex



# 점수 뒤의 실행 경로를 고칩니다

하나의 순위 대신, 수정할 수 있는 두 문제를 추적합니다.

**01 · 정답이 아닌데 왜 judge는 완료를 수락했을까?**

02 · 입력은 적은데 왜 완료시간은 줄지 않았을까?

설계 → 관측 → 진단 → 수정 후보 → 같은 기준으로 재검증

# 종료 규칙부터 고정한 뒤 점수를 읽습니다

Cell = task × repetition × arm. Pair = 같은 task·반복의 GEODE와 Codex.

**445** → **435** → **429**

planned pairs → after exclusions → common-valid

bn-fit-modify · tune-mjcf: 20 excluded cells / 6 native-invalid cells

**Task-owned tests + 정해진 종료 조건 → selected reward**

정상 실패·canonical timeout = 0 / 인프라 무효·미실행 ≠ 0

전체 890-cell primary 불성립. 이후 수치는 공통 유효 셀의 보조 관측입니다.

# Task는 격리했지만 런타임 배치는 달랐습니다

Historical: Harbor 0.22.0 · gpt-5.6-sol / subscription / max · b549f3e

**arm64 host / Docker / amd64 emulation**

**GEODE host loop**

terminal_exec



**Task container**

shell + workspace

**Task container**

Codex CLI + workspace

**Harbor → isolated trial → termination → task-owned verifier**

Task별 image·CPU/RAM·timeout 동결 / arms 순차 / c1 또는 calibrated c2

격리는 파일 상태의 경계입니다. 서로 다른 런타임 배치까지 동등하다는 뜻은 아닙니다.

# 통과 8건의 차이보다, 불일치 98쌍을 읽습니다

공통 유효 429쌍: GEODE 339건 · Codex 331건. 어느 경로를 고칠지 찾는 출발점입니다.

**286 both pass · 53 GEODE only · 45 Codex only · 45 neither**

G-code historical: GEODE 0/5 · Codex 4/5 → 개발 사례로 선택

Full-suite primary 불성립 / 공식 순위·전체 우위의 근거 아님

arm64 호스트에서 측정 · amd64 task container는 Rosetta로 실행

# 정답을 놓친 판정, 근거 전달부터 고쳤습니다

저장·readback만으로는 내용의 정답을 확인할 수 없었습니다.

**INTERNAL JUDGE**

**5/5**

**TASK VERIFIER**

**2/5**

**FALSE ACCEPTANCES**

**3**

**IMAGE EVIDENCE**

7 · 9 · 9 images observed

≤2 images selected → **이미지 포함 호출을 별도로 최대 12개 선택**

**WORKER MODEL**

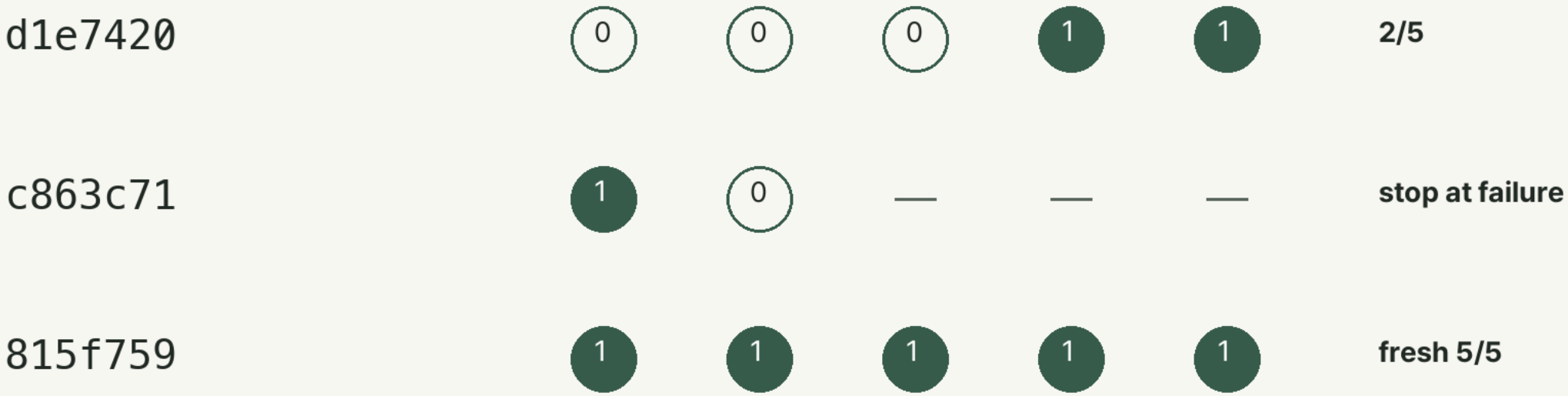
3/3 failed before inference

model/provider mismatch → **model 생략 시 parent model 상속**

후보 주장보다 관측 근거를 먼저 전달했습니다. 여러 수정을 함께 적용한 개발 후보입니다.

# 수정 후보는 새 5회를 모두 통과했습니다

5/5 · invalid 0 · trial 내부 repair 0. 실패한 후보와 중단한 시도도 남겼습니다.

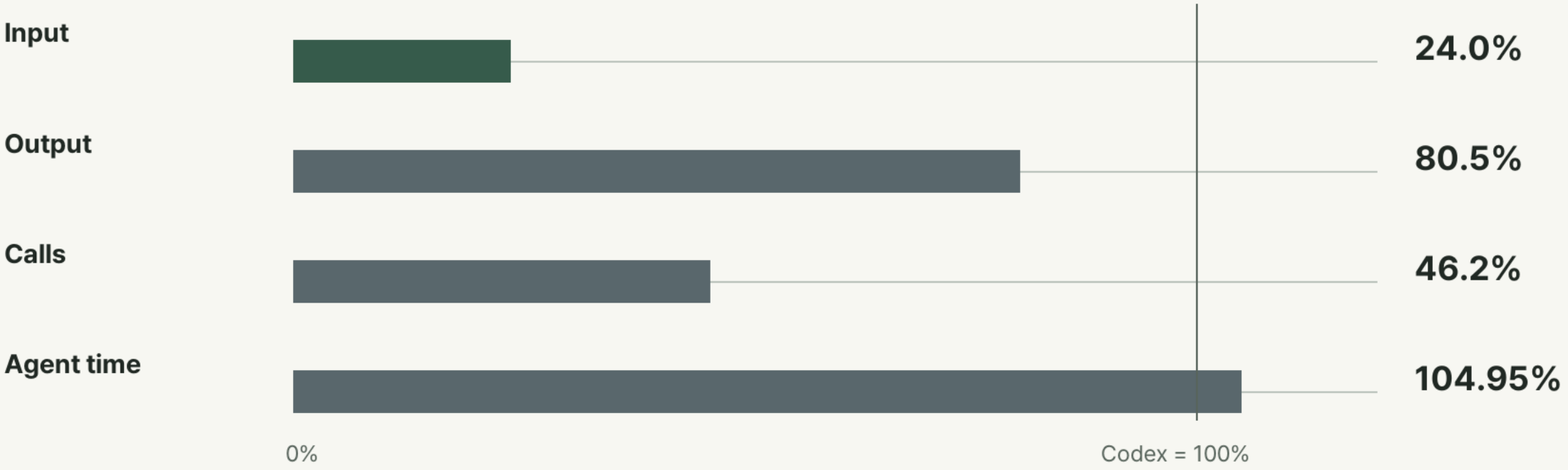



— = unexecuted · r2 SHM custody issue remains open; see NARRATIVE

Sources: G · NARRATIVE.md

# 적은 입력으로 성공한 작업, 줄지 않은 완료시간

공동 성공 286쌍의 reported totals. GEODE / Codex 비율이며, 과제별 비율의 평균이나 청구액이 아닙니다.



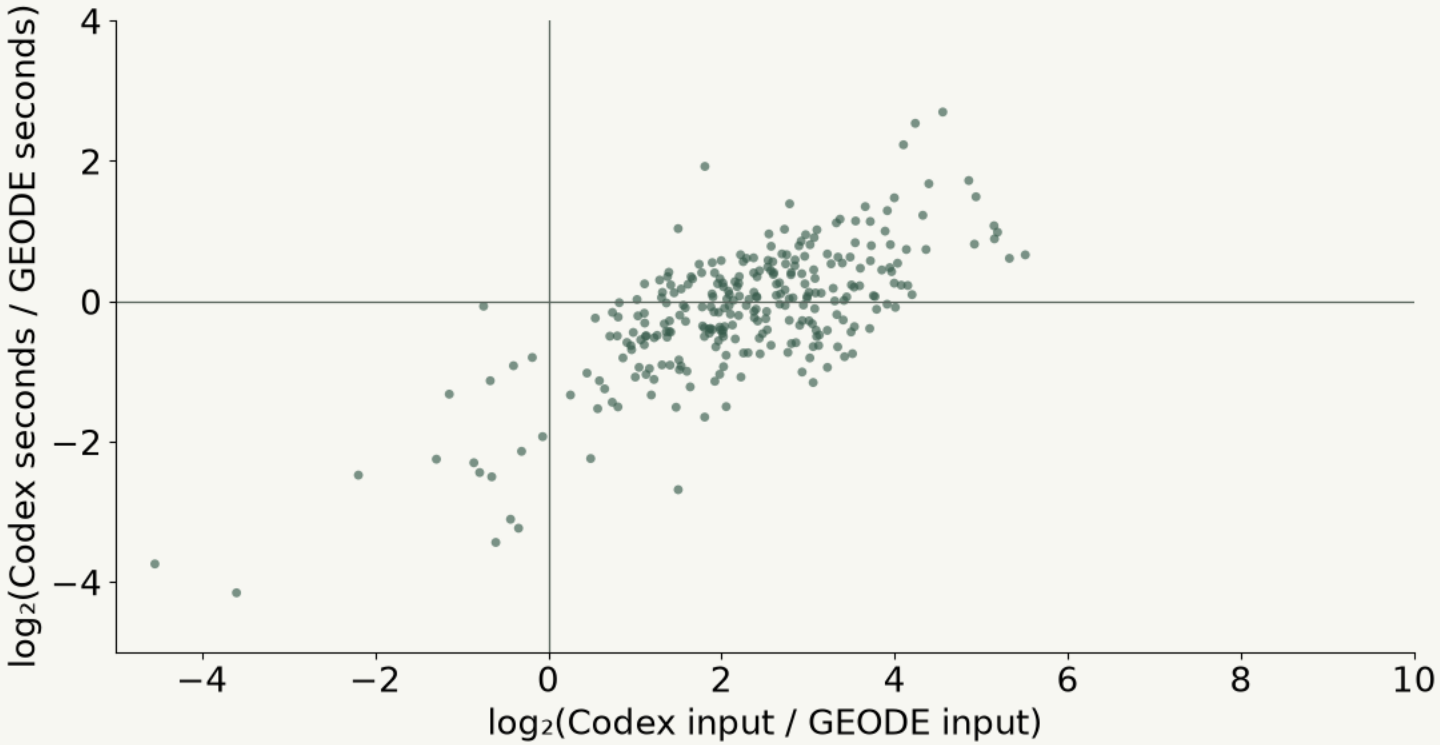

Input 24.03% · Output 80.49% · Calls 46.19% · Agent elapsed 104.95%

arm64 호스트에서 측정 · amd64 task container는 Rosetta로 실행

Sources: T · NARRATIVE.md

# 적은 입력이 빠른 완료로 이어지지 않은 133쌍이 있습니다

점 하나는 공동 성공 task–repetition 쌍입니다. 모든 286점을 표시합니다.




오른쪽: GEODE input이 적음 · 위쪽: GEODE가 빠름

Input이 적고 빠름 136쌍 · 적지만 느림 133쌍

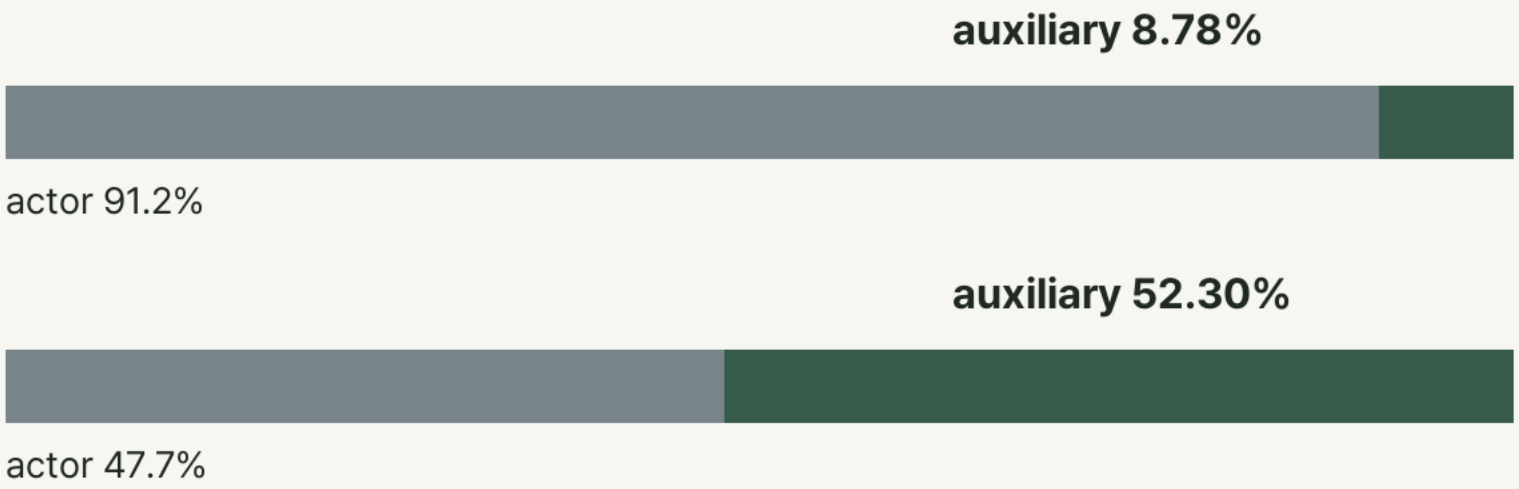
**호출 수 46.19% × 호출당 입력 52.03% ≈ 총 입력 24.03%. 경과시간은 별도 관측입니다.**

arm64 호스트에서 측정 · amd64 task container는 Rosetta로 실행

# 보조 호출까지 관측하니 출력의 절반이 보였습니다

별도 1-task smoke · 7c0e927d · modernize-scientific-stack · full runtime in container

**31 requests, all max: actor 16 / reflection 8 / text 4 / judge 3**

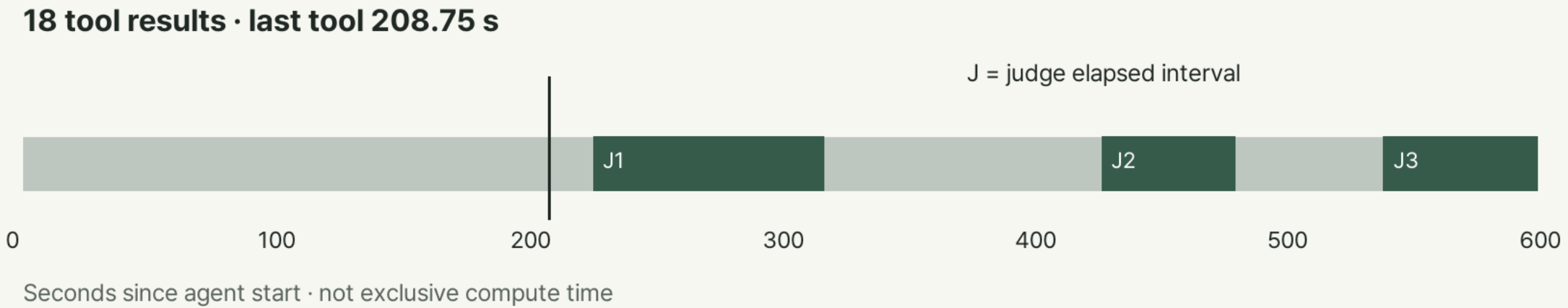



1 CPU / 2 GiB / 10 GiB declared / 600 s / c1 / retries 0

Usage 30/31: 취소된 호출은 null / cache-read 35,712는 관측된 부분합

# 수정 판단 뒤에 다음 행동이 이어지지 않았습니다

마지막 tool 208.75 s → revise 두 번 → 세 번째 judge 취소 → 600 s 제한 소진




## 종료 후 task verifier raw 1 / canonical selected 0 · valid failed 0/1

Judge elapsed 205.35 s / agent phase의 34.18% · CPU 점유율 아님

중간 작업물은 별도 채점하지 않아 false rejection 여부는 알 수 없습니다.

# 같은 기록을 점수와 Replay로 다르게 투영합니다

공개용 정제는 원본을 대체하지 않습니다. 새 실행도 기존 셀을 덮어쓰지 않습니다.

| PRIVATE SOURCE | CANONICAL | PUBLIC DERIVATIVE |
| --- | --- | --- |
| `raw/<arm>/<job>/<trial>/` | `attempts.jsonl / analysis.json` | `trajectory / recording.cast` |
| result / verifier / sessions.db | source SHA + validity + reward | sanitize → scan → readback |





```
Local root: artifacts/eval/runs/<run_id>/

Public: geode-eval-artifacts/terminal-bench/<run_id>/
```

공개 cast는 ATIF-derived replay. Source/export 일치는 전체 물리 호출 포착을 뜻하지 않습니다.

Proposed experiment · not executed

# 다음 비교는 판단이 아닌 완료의 개선을 검증합니다

새 G-code 후보의 5회 통과를 확인했습니다. 다음은 동일 조건에서 완료시간과 비용을 비교할 차례입니다.

**FIX THE PRODUCER**

먼저: native field presence·cache 계약을 고치고 source/export 대조

**FREEZE THE CONTRAST**

이후: baseline 고정 → revise 정책 하나 변경 → fresh matched trials

**MEASURE COMPLETION**

Judge 입력·다음 tool·남은 시간·canonical reward·usage를 함께 비교

Sources: G / S / C / O · NARRATIVE.md

# 수치는 어떤 실행에서 나왔는가

결과마다 population·revision·파일 hash를 연결했습니다.

**H / T**

### Frozen comparison · b549f3e

429 공통 유효 쌍 / usage 분석은 공동 성공 286쌍

**G**

### G-code candidate · 815f759

새 5회 통과는 개발 게이트입니다. Historical 점수를 대체하지 않습니다.

**S / C**

### Full-runtime smoke · 7c0e927d

31 requests와 종료 지연의 근거입니다. 전체 비교와 합치지 않습니다.

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

# Harbor 규격과 채점의 근거를 구분합니다

공식 문서는 구조를 설명합니다. 실제 버전과 값은 실행 산출물에서 확인합니다.

**P1 / P2**

**Terminal-Bench paper (2026) · Harbor**

Task·trial·job의 정의 / 2.1 실행 버전은 동결 manifest로 확인

**P3**

**ATIF · Core Models / Validation**

tool_call_id ↔ source_call_id / 현행 v1.8을 과거 규격으로 대체하지 않음

**M2**

**Anthropic · Demystifying evals (2026)**

Transcript의 주장과 환경 outcome을 구분 / judge와 task verifier 대조

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

# 방법론은 적용한 지점까지 인용합니다

설계 참고, 해석 기준, 향후 연구를 같은 성과처럼 섞지 않았습니다.

**M1**

**Reflexion · Shinn et al. (2023), §3**

피드백을 수정 지시로 전달 / 이번 5/5는 repair 효과의 측정이 아님

**M3**

**AI Agents That Matter · Kapoor et al. (2024)**

정확도·비용·holdout을 함께 검토 / token 비율은 달러 비용이 아님

**M4**

**Darwin Gödel Machine · Zhang et al. (2025)**

후보 계보를 보존·탐색하는 향후 방향 / 자율 탐색 실험은 미실행

Pillow · Matplotlib · FFmpeg / Pretendard · Source Han Serif KR · STIX Two Text

# 이제 원본 비교의 445쌍을 다시 읽습니다

이후 영상은 변경하지 않은 derived Replay입니다. 새 G-code 실행이 아닙니다.

**835 ATIF-derived · 35 receipt-only · 20 excluded cells**

왼쪽 GEODE / 오른쪽 Codex · tool-event 순서 정렬

Raw PTY 아님 · 실제 동시 실행 아님 · UTC 보존 / KST 표시