Humanity's Last Exam (Diamond) Review Report

Author: Benchmarks' Last Exam Lab 2026-10-30

HLE-Diamond draws 1,000 expert-level questions from Humanity's Last Exam and its held-out pool. Half test specialist knowledge across academic fields, and half test the reasoning needed to reach an answer.

Abstract

We reviewed 100 HLE-Diamond questions: **87 Accepted**, **8 Doubtful**, and **5 Refused**.

The review examines whether prompts, reference answers, and grading agree. Confirmed findings comprise two task-content defects and three incorrect judgments on recorded normal responses.

Outcome	Questions	Meaning
Accepted	**87**	No material defect or concrete unresolved concern was found in this review.
Doubtful	**8**	Specific evidence raises a concern, but the correct interpretation or scoring impact remains unresolved.
Refused	**5**	At least one defect affecting the question's core evaluation was confirmed.
Total	**100**	50 reasoning questions and 50 knowledge questions.

Scope and version

We sampled 50 knowledge questions and 50 reasoning questions from the 1,000-question dataset.

- Dataset: cais/hle-diamond, pinned revision 04eeb7e
- Official evaluation code: centerforaisafety/hle, pinned revision 22ed307
- Review date: November 5, 2026.

Distribution of findings

Issue type	Refused	Doubtful	Total
C1 Incorrect answers or scoring logic	3	2	5
C2 Scoring requirements not stated in the task	2	6	8
C3 Missed critical errors	0	0	0
C4 Scoring bypasses or answer leakage	0	0	0
C5 Other issues	0	0	0
Total	**5**	**8**	**13**

These counts describe the 100-question sample. They should not be read directly as the defect rate of the full HLE-Diamond dataset.

Representative cases

Reflection and rotation

C1 Incorrect answers or scoring logic

Question: 673668e658bad7ba89d4ad54

The question asks for a sequence of array transformations and the missing symbols. It explicitly defines antidiagonal_transpose as reflection across the anti-diagonal, but the reference code implements a 180-degree rotation.

Recomputing all six choices under the stated rule produced no complete match. Even with the symbol-replacement order most favorable to the reference, answer E disagreed with 206 visible cells. Changing only the reflection to the reference code's rotation made every target cell and hidden symbol match.

The sign of a constant

C1 Incorrect answers or scoring logic

Question: 6701d869aee3881b852d40a0

The question asks for a simplified neurobiological model containing variables and a constant ρ, without fixing its definition. The reference uses u(v+ρ); the model defines its constant with the opposite sign and correctly derives u(v−ρ).

We checked the full derivation and confirmed equivalence. The judge rejected the response solely because the signs differed, ignoring the definition supplied in the answer. This recorded normal response therefore lost credit for correct work.

Equivalent loop expressions

C1 Incorrect answers or scoring logic

Question: 671360871b12bd5183fec91a

The question asks for a program-expression graph of a variable inside a loop. The model supplies expression A and its shared node I; substituting I into A gives exactly the reference answer. It also labels a correct loop-exit expression separately.

The judge acknowledges that A matches the reference, yet rejects the whole response because the extra expression contains eval and pass. The prompt does not prohibit that correct explanation. The recorded response completes the core task but is still rejected.

Evidence 1

- Tate et al.'s original paper

Unit and position vectors

C2 Scoring requirements not stated in the task

Question: 673641e0ae322e42828041f2

The question asks for the area, volume, and a limit of a region defined by a trajectory equation. It explicitly defines u as a unit vector. The reference instead uses a position vector, changing the radial term from z′(r) to r·z′(r).

The extra radius changes the region. With v_0=1, n=2, r=3/2, and slope c=−2/3, a reference trajectory leaves a residual of −1/3 in the stated equation. Even if the wording is a typo, obtaining the reference requires changing a core definition. This is a static equation-check finding.

The end of a fugue exposition

C2 Scoring requirements not stated in the task

Question: 6793ce7158d4e9f912bfd94a

The question asks where the exposition ends in Bach's C-minor fugue BWV 847. Open Music Theory places it in bars 1–8, with the following episode beginning in bar 9. The model explains the three subject entries and answers bar 8.

The reference uses the start of bar 9 as the boundary, and the judge rejects the different number. Bar 9 is a defensible boundary description, but the prompt does not require that convention. It cannot justify rejecting the textbook-supported answer of bar 8.

Evidence 1

- Open Music Theory: BWV 847 analysis

Task-by-task findings

Question links point to the pinned data file; locate records by their full IDs below.

Refused: 5 questions

6701d869aee3881b852d40a0

C1 Incorrect answers or scoring logic

The response correctly redefines a free constant, but the judge rejects the equivalent formula because its sign differs from the reference.

671360871b12bd5183fec91a

C1 Incorrect answers or scoring logic

The correct loop expression is supplied. A separately named, correct exit expression nevertheless causes the whole response to be rejected.

673641e0ae322e42828041f2

C2 Scoring requirements not stated in the task

The prompt explicitly specifies a unit vector, while the reference uses a position vector. The extra radius factor changes the core equation and region.

673668e658bad7ba89d4ad54

C1 Incorrect answers or scoring logic

The reference implementation replaces anti-diagonal reflection with a 180-degree rotation. Offline recomputation confirms that the fixed answer fails the stated transformation.

6793ce7158d4e9f912bfd94a

C2 Scoring requirements not stated in the task

A textbook supports the exposition spanning bars 1–8. A normal response using that convention is rejected against a fixed answer of 9.

Doubtful: 8 questions

66ee0b4d95cd6c1ce2259e50

C1 Incorrect answers or scoring logic

The reference's reason for excluding herring is in tension with hearing research. Available frequency, sound-level, and propagation evidence does not establish that another option is correct.

6719e946a81614142be02bec

C2 Scoring requirements not stated in the task

Dose addition is a supported default risk model. The prompt remains ambiguous between prediction under that default and identifying an actual interaction from the supplied data.

6722d9de6bc44598e1fd09d6

C2 Scoring requirements not stated in the task

A technology name does not match the rules examined and may change its level and score. A version-specific alias or recoverable typo has not been ruled out.

6723ac21af0a31f10468be05

C2 Scoring requirements not stated in the task

The reference bit count is defensible under full IEEE rules. The prompt's simplified formula and the zero-exponent exception leave an ambiguity affecting the exact count.

6725c1e0eac3991069717fda

C2 Scoring requirements not stated in the task

The reference uses exponentially distributed waiting times; a fixed-delay model gives a different result. Both have domain support, and a conventional compartmental-model reading remains plausible.

6726e3bf4aeb5af8ca49e254

C1 Incorrect answers or scoring logic

The question asks for a lower bound, but the reference uses upper-bound notation. The recorded correct lower-bound response was accepted; a scoring error caused by the notation has not been confirmed.

672e09b50a85795d0ed2d36e

C2 Scoring requirements not stated in the task

The reference bound holds when rest energy is included. Another common energy convention and the full wording of the options leave a concern, but do not establish that another option must receive credit.

67c1fca02f0a2a1da4ab944c

C2 Scoring requirements not stated in the task

Research supports an alternative mechanism, which the model also selected. The evidence does not yet establish that it explains all experimental conditions, so its rejection is not a confirmed false negative.

Effects on evaluation and follow-up

These issues can deny credit to valid answers or require a solver to depart from explicit instructions to match the reference. We will update the findings when verifiable author clarifications, counterevidence, or revisions become available.

Contact Us

benchmarks-last-exam.com Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.