Humanity's Last Exam (Diamond) Review Report

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

HLE-Diamond 从 Humanity's Last Exam 的题库及保留题中精选了 1,000 道跨学科难题。知识题和推理题各 500 道，分别考查模型对专业知识的掌握和推导答案的能力。

摘要

我们审查了 HLE-Diamond 的 100 道题，其中 **87 道 Accepted**、**8 道 Doubtful**、**5 道 Refused**。

审查关注题面、参考答案和评分是否一致。确认的问题包括两道题目内容缺陷，以及三道正常作答中的实际误判。

定档	数量	含义
Accepted	**87**	本轮未发现实质缺陷或具体未解决疑点。
Doubtful	**8**	有具体证据提示问题，但正确解释或实际评分影响尚待确认。
Refused	**5**	至少一项影响核心评测的缺陷得到确认。
合计	**100**	推理题、知识题各 50 道。

审查范围与版本

本次从 1,000 道题中抽取知识题、推理题各 50 道。

- 数据集：cais/hle-diamond，固定版本 04eeb7e
- 官方评测代码：centerforaisafety/hle，固定版本 22ed307
- 审核日期：2026 年 11 月 5 日。

问题分布

问题类型	Refused	Doubtful	合计
C1 答案或评分逻辑错误	3	2	5
C2 存在题面未说明的评分要求	2	6	8
C3 关键错误漏判	0	0	0

问题类型	Refused	Doubtful	合计
C4 评分可绕过 / 答案泄露	0	0	0
C5 其他问题	0	0	0
合计	**5**	**8**	**13**

这些数量描述本次 100 题样本，不能直接作为整个 HLE-Diamond 的缺陷率。

代表性案例

反对角线镜像与旋转

C1 答案或评分逻辑错误

题目：673668e658bad7ba89d4ad54

题目要求按一串命令变换符号数组，并找出隐藏符号。其中，antidiagonal_transpose 明确定义为沿反对角线镜像，参考代码却实现成了旋转 180°。

按题面规则重算全部六个选项，没有一个完整匹配目标。采用对参考最有利的字符替换顺序，固定答案 E 仍有 206 个可见格不符；只把镜像换成参考代码的旋转，目标格和隐藏符号便全部吻合。

常数的符号约定

C1 答案或评分逻辑错误

题目：6701d869aee3881b852d40a0

题目要求简化神经生物学模型，得到含变量和常数 ρ 的表达式，但没有固定 ρ 的定义。参考答案写成 u(v+ρ)，模型将常数定义为相反数，正确推得 u(v−ρ)。

我们核对了完整推导，两种表达等价。judge 却只因加减号不同判错，忽略了作答中已明确说明的常数定义。这次正常作答因此未获正确性认可。

循环表达式的等价写法

C1 答案或评分逻辑错误

题目：671360871b12bd5183fec91a

题目要求给出循环中变量的程序表达式图。模型分别写出循环表达式 A 和它引用的节点 I；将 I 代入 A，恰好得到标准答案。模型还单独标注了一个正确的退出循环表达式。

judge 承认 A 与标准答案一致，却因为附加表达式含有 eval、pass 操作而整体判错。题面没有禁止这些正确说明；本次正常作答的核心答案已完成，却仍被拒绝。

证据 1

- Tate 等原始论文

单位向量与位置向量

C2 存在题面未说明的评分要求

题目：673641e0ae322e42828041f2

题目要求根据轨迹方程计算区域的面积、体积和极限，并明确规定 u 是单位向量。参考解却将 u 当成位置向量：径向计算由 $z'(r)$ 变成 $r{\cdot}z'(r)$，多出一个随位置变化的半径。

这会改变所求区域。取 $v_0=1$、$n=2$、$r=3/2$、斜率 $c=-2/3$，参考轨迹代入公开方程的残差为 $-1/3$，并不满足题面。即使原文是笔误，也必须修改核心定义才能得到参考解。该结论来自静态方程核对。

赋格呈示部的小节边界

C2 存在题面未说明的评分要求

题目：6793ce7158d4e9f912bfd94a

题目问巴赫 C 小调赋格 BWV 847 的呈示部在哪一小节结束。Open Music Theory 将呈示部划为第 1–8 小节，后续插部从第 9 小节开始。模型按三次主题进入的位置解释后，回答第 8 小节。

参考答案以第 9 小节开头作为结束边界，judge 因数字不同判错。第 9 小节作为边界可以成立，但题面未限定这种编号口径，不能据此拒绝有教材依据的第 8 小节答法。

证据 1

- Open Music Theory：BWV 847 分析

逐题审查结果

题目链接指向固定版本数据文件，可按下表的完整 ID 定位。

Refused：5 道题

6701d869aee3881b852d40a0

C1 答案或评分逻辑错误

真实作答正确重定义常数，judge 却仅因正负号不同判错。

671360871b12bd5183fec91a

C1 答案或评分逻辑错误

真实答案的循环图正确，额外且独立标注的正确退出表达式却导致整体拒绝。

673641e0ae322e42828041f2

C2　存在题面未说明的评分要求

题面单位向量与参考的位置向量相差变量半径，核心方程明确冲突。

673668e658bad7ba89d4ad54

C1　答案或评分逻辑错误

参考实现将反对角线镜像写成 180 度旋转；独立离线复算确认目标数组不匹配。

6793ce7158d4e9f912bfd94a

C2　存在题面未说明的评分要求

原始教材支持第 1–8 小节口径，真实作答按此回答却被固定第 9 小节判错。

Doubtful：8 道题

66ee0b4d95cd6c1ce2259e50

C1　答案或评分逻辑错误

参考解释排除鲱鱼的理由与已查听觉研究存在张力，但现有频率范围、声强与传播证据不足以证明另一个选项正确。

6719e946a81614142be02bec

C2　存在题面未说明的评分要求

剂量加和是有依据的默认风险模型；题面究竟要求采用默认模型预测，还是从数据识别真实交互效应，仍有歧义。

6722d9de6bc44598e1fd09d6

C2　存在题面未说明的评分要求

题面的科技名称与查到的规则名称不对应，可能影响等级和得分；尚未排除版本别名或可恢复笔误。

6723ac21af0a31f10468be05

C2　存在题面未说明的评分要求

按完整 IEEE 规则，参考比特数有合理依据；题面显式简化公式与零指数例外之间仍存在影响精确计数的歧义。

6725c1e0eac3991069717fda

C2　存在题面未说明的评分要求

参考采用指数等待时间，固定等待时间模型则给出不同结果。两者都有领域依据，尚不能排除题目默认采用常见舱室模型的解释。

6726e3bf4aeb5af8ca49e254

C1　答案或评分逻辑错误

问题要求下界，参考却使用上界记号。本次正确的下界回答获得通过，尚未确认该记号问题造成实际误判。

672e09b50a85795d0ed2d36e

C2　存在题面未说明的评分要求

包含静质量能时参考上界成立；另一种常用能量约定与选项完整措辞留下具体歧义，尚不能确认另一选项应当获分。

67c1fca02f0a2a1da4ab944c

C2　存在题面未说明的评分要求

存在有研究依据的另一种机制解释，模型也实际选择了它，但现有证据不足以证明它能解释全部实验条件，因此不能将判错直接算作误杀。

评测影响与后续更新

上述问题会使符合题意的答案失分，或要求解题者违背明确题面才能匹配参考答案。我们将根据可核查的作者澄清、反证和版本修订更新结论。

联系我们

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

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