Humanity's Last Exam:
What Hard Questions Teach Us About AI

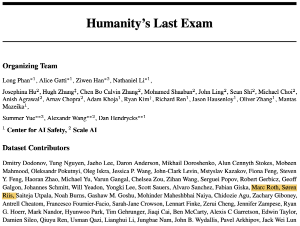
Søren Riis
Queen Mary University of London

November 2025

Humanity's Last Exam in a nutshell

- **Scope:** $\approx$ 2 500 expert-vetted, closed-ended questionsuacross 100 + subjects (41 % mathematics, 10 % computer science / AI, 9–11 % physics or biomedicine).
- **Purpose:** Built by **CAIS + Scale AI** to counter benchmark saturation (e.g. MMLU 90 % accuracy).
- **Evaluation:** Zero-shot, temperature 0; exact-match scoring; tracks **calibration error** (confidence vs correctness).
- **Composition:** Public set + private hold-out; 14 % items are **multimodal** (with figures or diagrams).
- **Human baseline:** 90 % accuracy $\rightarrow$ current LLMs (25 %) still far behind.

HLE at a glance — visuals



Humanity's Last Exam

Organizing Team

Long Phan[*1], Alice Gatti[*1], Ziwen Han[*2], Nathaniel Li[*1],

Josephina Hu[2], Hugh Zhang[‡], Chen Bo Calvin Zhang[2], Mohamed Shaaban[2], John Ling[2], Sean Shi[2], Michael Choi[2], Anish Agrawal[2], Arnav Chopra[2], Adam Khoja[1], Ryan Kim[†], Richard Ren[1], Jason Hausenloy[1], Oliver Zhang[1], Mantas Mazeika[1],

Summer Yue[**2], Alexandr Wang[**2], Dan Hendrycks[**1]

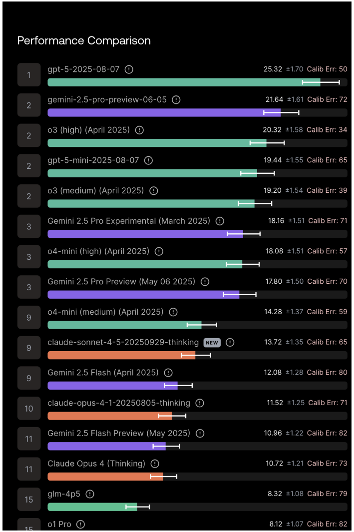
[1] **Center for AI Safety,** [2] **Scale AI**

Dataset Contributors

Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun

Paper cover

arXiv:2501.14249



Official leaderboard

Scale SEAL board

Multimodal items

14% include figures/diagrams

The Humanity's Last Exam Competition

How the challenge worked

- Each contributor submitted one or more original questions with **unique, verifiable answers**.
- The questions were first tested by three frontier AI models under strict zero-shot conditions.
- If **all three models failed**, the problem advanced to a second tier where three even stronger models attempted it.
- A question qualified only if **all six models** failed to solve it correctly.
- Authors then wrote detailed, human-readable **solutions and explanations**, which were reviewed by human referees who graded each question on a 1–5 scale for clarity, originality, and difficulty.
- Questions receiving marks of **4 or 5** progressed to the final prize round.

Outcome

- Around 20 of my questions reached the top group.
- My colleague **Marc Roth** and I won several smaller prizes, and **each received a $5,000 top prize** for one of our own submissions.
- My winning problem was notable for combining **multi-modal visual reasoning**, **mathematical logic** across domains, and a requirement that the AI **write a programme** to obtain the correct answer.

State of the Art: How Good Are the Best AI Models?

Rapid progress at the frontier

- Top language models such as **GPT-5**, **Gemini 2.5**, and **Claude 4.5** have reached **Gold Medal level** in international mathematics competitions such as the IMO and AIME.
- These systems can now solve complex olympiad-style problems, translate natural-language reasoning into formal mathematics, and write high-quality proofs when guided step by step.

Remaining challenges

- Even the best models still **struggle with research-level mathematics**: open-ended conjectures, abstract definitions, or proofs requiring deep insight.
- They often fail when reasoning spans multiple domains or when the problem involves diagrams, geometry, or code generation.

Why my submitted problems cannot be shared

- The full text and solutions of my HLE problems remain **confidential**, as publishing them would risk contamination of future evaluations.

Example 1 – Easy for top LLMs: the Fibonacci pitfall

Python code

```
def testFunction(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    return testFunction(n - 1) + testFunction(n - 2)
```

Task: Compute `testFunction(100)`.

Observation:

- Naive recursion has exponential running time – it would take more than 10,000,000 years, even on a fast computer.
- Models such as Gemini, Claude, and ChatGPT Pro immediately recognise The Fibonacci pattern and create programs that solve problems in less than a millisecond.

Example 2 – Easy for top LLMs: Distinct cubes cover

Question:

*What is the smallest integer N such that every integer n >= N can be written as a sum of **distinct cube numbers**?*

Reasoning:

- Models recognise that this relates to an integer sequence describing numbers *not* representable as sums of distinct cubes.
- They consult (or recall) the **OEIS** – the Online Encyclopedia of Integer Sequences.
- Sequence **A001476** lists the numbers that cannot be expressed as a sum of distinct positive cubes.
- There are exactly 2,788 such numbers; the largest is 12,758.

Answer: N = 12,759.

Example 3 – Borderline-difficulty Chess copycat problem

Problem statement

Starting from the standard initial chess position:
Black copies White's moves exactly, using the same piece type and the mirrored square across the vertical axis.
*Find the **shortest possible game** in which White delivers checkmate under this copy rule.*

Observation

- The puzzle requires reasoning about board symmetry and forced responses. The Scholar's mate does not work!
- **Gemini 2.5** can now solve this puzzle reliably and returns the minimal mate sequence.
- **ChatGPT Pro (GPT-5)** and **Claude 4.5 Sonnet** still fail or produce inconsistent move sequences.
- Demonstrates how small geometric constraints expose weaknesses in spatial and strategic reasoning.

Condorcet Domains and NIM – Reasoning at the Edge (1)

Condorcet Domains (with Bei Zhou)

- *A Heuristic Search Algorithm for Discovering Large Condorcet Domains* (arXiv: 2303.06524); record-breaking domains for $n = 10$ and $n = 11$.
- Published version in *4OR* (2025).
- Explores coherent, peak–pit, and bipartite families with improved lower bounds.
- Using an **AI-inspired search algorithm** running on a high-performance super-cluster, we recently **solved a 30-year-old open problem** in this area, establishing new lower bounds for large Condorcet domains.

Large combinatorial search problems as testbeds for AI-driven discovery.

Condorcet Domains, NIM, and Term Coding – Reasoning at the Edge (2)

NIM and Impartial Games

- *Impartial Games: A Challenge for Reinforcement Learning* (Riis and Zhou, arXiv: 2205.12787).
- *Mastering NIM and Impartial Games with Weak Neural Networks* (Riis, arXiv: 2411.06403).
- Shows that AlphaZero-style reinforcement learning **struggles** on impartial games such as NIM.
- Within **Computational Complexity Theory**, both positive and negative results matter: for some classes of neural networks it can be shown that NIM and, more generally, impartial games **cannot be learnt** using these architectures.

Term Coding (test ground for AI reasoning)

- *Term Coding for Extremal Combinatorics* (Riis, arXiv: 2504.16265): encodes extremal problems as finite systems of term equations (with optional non-equality constraints).
- Unifies perspectives from extremal combinatorics, network/index coding, and finite model theory, yielding crisp optimisation and search tasks – an **excellent test ground for AI reasoning**.
- Includes a **striking theoretical result**: certain decision problems **cannot be solved by any algorithm** (including AI systems), yet a minimal weakening makes them **solvable in polynomial time**.

Connects combinatorial structure, algorithmic search, and reasoning limits.

Take-away and Q&A

Main lessons

- **Humanity's Last Exam (HLE)** exposes a large reasoning gap: top AI models reach about 25% accuracy, while **human domain experts** reach around 90%.
- Frontier models excel at recognising patterns and recalling known facts, but still fail on deep, multi-step reasoning and on tasks requiring logical self-consistency or spatial understanding.
- Benchmarks such as HLE, together with theoretical frameworks from our work on **Condorcet Domains**, **NIM**, **Impartial Games**, and **Term Coding**, provide concrete testbeds for analysing the boundaries of algorithmic reasoning.

Thank you!

Søren Riis – s.riis@qmul.ac.uk