Article

A benchmark of expert-level academic questions to assess AI capabilities

https://doi.org/10.1038/s41586-025-09962-4

Received: 7 May 2025

Accepted: 25 November 2025

Published online: 28 January 2026

Open access

Check for updates

Center for AI Safety*, Scale AI* & HLE Contributors Consortium*

Benchmarks are important tools for tracking the rapid advancements in large language model (LLM) capabilities. However, benchmarks are not keeping pace in difficulty: LLMs now achieve more than 90% accuracy on popular benchmarks such as Measuring Massive Multitask Language Understanding[1], limiting informed measurement of state-of-the-art LLM capabilities. Here, in response, we introduce Humanity's Last Exam (HLE), a multi-modal benchmark at the frontier of human knowledge, designed to be an expert-level closed-ended academic benchmark with broad subject coverage. HLE consists of 2,500 questions across dozens of subjects, including mathematics, humanities and the natural sciences. HLE is developed globally by subject-matter experts and consists of multiple-choice and short-answer questions suitable for automated grading. Each question has a known solution that is unambiguous and easily verifiable but cannot be quickly answered by internet retrieval. State-of-the-art LLMs demonstrate low accuracy and calibration on HLE, highlighting a marked gap between current LLM capabilities and the expert human frontier on closed-ended academic questions. To inform research and policymaking upon a clear understanding of model capabilities, we publicly release HLE at https://lastexam.ai.

The capabilities of large language models (LLMs) have advanced markedly, exceeding human performance across a diverse array of tasks. To systematically measure these capabilities, LLMs are evaluated on benchmarks: collections of questions that assess model performance on tasks such as math, programming or biology. However, state-of-the-art LLMs[2–6] now achieve more than 90% accuracy on popular benchmarks such as Measuring Massive Multitask Language Understanding (MMLU)[1], which were once challenging frontiers for LLMs. The saturation of existing benchmarks, as shown in Fig. 1, limits our ability to precisely measure artificial intelligence (AI) capabilities and calls for more challenging evaluations that can meaningfully assess the rapid improvements in LLM capabilities at the frontiers of human knowledge.

To address this gap, we introduce HLE (originally defined as Humanity's Last Exam, although we will use the term HLE for this paper), a benchmark of 2,500 challenging questions from dozens of subject areas, designed to assess LLM capabilities at an expert level in broad academic subjects. HLE is developed by academics and domain experts, providing a precise measure of capabilities as LLMs continue to improve (see section 'Collection'). HLE is multi-modal, featuring questions that are either text-only or accompanied by an image reference and includes both multiple-choice and exact-match questions for automated answer verification. Questions are original, precise, unambiguous and resistant to simple internet lookup or database retrieval. Among the diversity of questions in the benchmark, HLE emphasizes world-class mathematics problems aimed at testing deep reasoning skills broadly applicable across multiple academic areas.

We use a multi-stage review process to thoroughly ensure question difficulty and quality (see section 'Review'). Before submission, each question is tested against state-of-the-art LLMs to verify its difficulty—questions are rejected if LLMs can answer them correctly. Questions submitted are then processed through a two-stage reviewing process: (1) an initial feedback round with multiple graduate-level reviewers and (2) an approval of organizer and expert reviewer, ensuring quality and adherence to our submission criteria. Following the release, we conducted a public review period, welcoming community feedback to correct any points of concern in the dataset.

Frontier LLMs consistently demonstrate low accuracy across all models, highlighting a marked gap between current capabilities and expert-level academic performance (see section 'Evaluation'). Models also provide incorrect answers with high confidence rather than acknowledging uncertainty on these challenging questions, with most models exhibiting root mean square (RMS) calibration errors above 70%.

As AI systems approach human expert performance in many domains, precise measurement of their capabilities and limitations is essential for informing research, governance and the broader public. High performance on HLE would suggest expert-level capabilities on closed-ended academic questions. To establish a common reference point for assessing these capabilities, we publicly release a large number of 2,500 questions from HLE to enable this precise measurement, while maintaining a private test set to assess potential model overfitting.

*Lists of authors and their affiliations appear at the end of the paper.

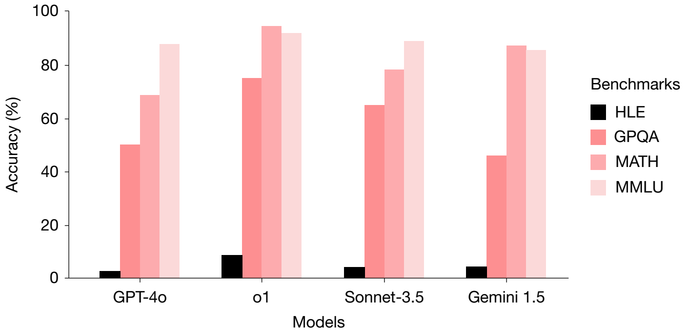


Fig. 1 | Performance of frontier LLMs on popular benchmarks and HLE. Compared with the saturation of other popular capability benchmarks, HLE accuracy remains low across several frontier models, demonstrating its effectiveness for measuring advanced, closed-ended, academic capabilities.

Dataset

Collection

HLE consists of 2,500 challenging questions across over a hundred subjects. A high-level summary is provided in Fig. 2. HLE is a global collaborative effort, with questions from nearly 1,000 subject expert contributors affiliated with more than 500 institutions across 50 countries—comprised mostly of professors, researchers and graduate degree holders. Examples of the diverse and challenging questions submitted to HLE are shown in Fig. 3.

Question style. HLE contains two question formats: exact-match questions (models provide an exact string as output) and multiple-choice questions (the model selects one of five or more answer choices). HLE is a multi-modal benchmark, with around 14% of questions requiring comprehending both text and an image; 24% of questions are multiple-choice, with the remainder being exact match.

Each question submission includes several required components: the question text itself, answer specifications (either an exact-match answer or multiple-choice options with the correct answer marked), detailed rationale explaining the solution, academic subject and name of the contributor and institutional affiliation to maintain accountability and accuracy.

Submission format. To ensure question quality and integrity, we enforce strict submission criteria. Questions should be precise, unambiguous, solvable and non-searchable, ensuring models cannot rely on memorization or simple retrieval methods. All submissions must be original work or non-trivial syntheses of published information, although contributions from unpublished research are acceptable. Questions typically require graduate-level expertise or test knowledge of highly specific topics (for example, precise historical details, trivia and local customs) and have specific, unambiguous answers accepted by domain experts. When LLMs provide correct answers with faulty reasoning, authors are encouraged to modify question parameters, such as the number of answer choices, to discourage false positives. We require clear English with precise technical terminology, supporting LaTeX notation wherever necessary. Answers are kept short and easily verifiable for exact-match questions to support automatic grading. We prohibit open-ended questions, subjective interpretations, and content related to weapons of mass destruction. Finally, every question is accompanied by a detailed solution to verify accuracy. More details about guidelines for contributors can be found in Supplementary Information section 1.

Prize pool. To attract high-quality submissions, we establish a USD$500,000 prize pool, with prizes of USD$5,000 for each of the top 50 questions and USD$500 for each of the next 500 questions, as determined by organizers. This incentive structure, combined with the opportunity for paper co-authorship for anyone with an accepted question in HLE, draws participation from qualified experts, particularly those with advanced degrees or notable technical experience in their fields.

Review

LLM difficulty check. To ensure question difficulty, each question is first validated against several frontier LLMs before submission (Methods). If the LLMs cannot solve the question (or, in the case of multiple choices, if the models on average do worse than random guessing), the question proceeds to the next stage: human expert review. In total, we logged more than 70,000 attempts, resulting in approximately 13,000 questions, which stumped LLMs that were forwarded to expert human review.

Expert review. Our human reviewers possess a graduate degree (for example, master's, PhD and JD) in their fields. Reviewers select submissions in their domain, grading them against standardized rubrics and offering feedback when applicable. There are two rounds of reviews. The first round focuses on iteratively refining submissions, with each question receiving between one and three reviews. The primary goal

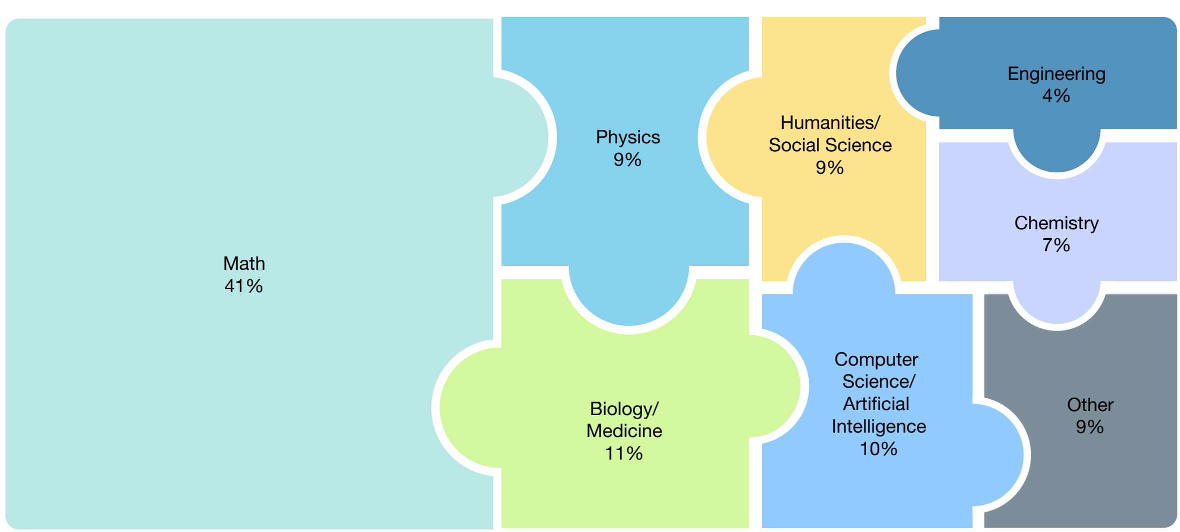


Fig. 2 | Distribution of HLE questions across categories. HLE consists of 2,500 exam questions in over a hundred subjects, grouped into eight high-level categories.

Classics

Question:

Here is a representation of a Roman inscription, orginally found on a tombstone. Provide a translation for the Palmyrene script.
RGYNᵓ BT HRY BR ᶜTᵓ HBL

Hernry T
Merton College, Oxford

Ecology

Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Edward V
Massachusetts Institute of Technology

Mathematics

Question:

The set of natural transformations between two functors $F, G : C \to D$ can be expressed as the end

$$Nat(F, G) \cong \int_A Hom_D(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$CoNat(F, G) \cong \int^A Hom_D(F(A), G(A)).$$

Let:

- $F = B_\bullet(\Sigma_4)_{*/}$ be the under ∞ -category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0 -simplex $*$ of $B_\bullet \Sigma_4$.
- $G = B_\bullet(\Sigma_7)_{*/}$ be the under ∞ -category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0 -simplex$*$ of $B_\bullet \Sigma_7$.

How many natural cotransformations are there between F and G? Here ∞ -categories are modelled as quasicategories, and F and G are functors from the opposite of the simplicial category to the category of sets

Emily S
University of São Paulo

Computer Science

Question:

Let G be a graph. An edge-indicator of G is a function $a : \{0,1\} \to V(G)$ such that $\{a(0), a(1)\} \in E(G)$.

Consider the following Markov Chain $M = M(G)$:
The statespace of M is the set of all edge-indicators of G, and the transitions are defined as follows:

Assume $M_t = a$.
1. pick $b \in \{0,1\}$u.a.r.
2. pick $v \in N(a(1-b))$ u.a.r. (here $N(v)$ denotes the open neighbourhood of v)
3. set $a'(b) = v$ and $a'(1-b) = a(1-b)$
4. Set $M_{t+1} = a'$

We call a class of graphs $\mathcal{G}$ well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

Which of the following graph classes is well-behaved?

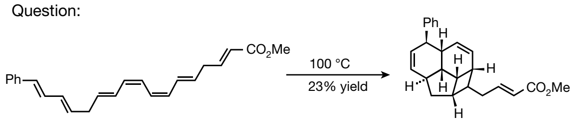
Answer Choices:

A. The class of all non-bipartite regular graphs
B. The class of all connected cubic graphs
C. The class of all connected graphs
D. The class of all connected non-bipartite graphs
E. The class of all connected bipartite graphs.

Marc R
Queen Mary University of London

Chemistry

Question:



endiandric acid B methyl ester

The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endiandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nπ]-con or [nπ]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

Noah B
Stanford University

Linguistics

Question:

I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

מִן־גַּעֲרָתְךָ יְנוּסוּן מִן־קוֹל רַֽעַמְךָ יֵחָפֵזוּן (Psalms 104:7) ?

Lina B
University of Cambridge

Fig. 3 | Example questions from HLE. Samples of the diverse and challenging questions submitted to HLE.

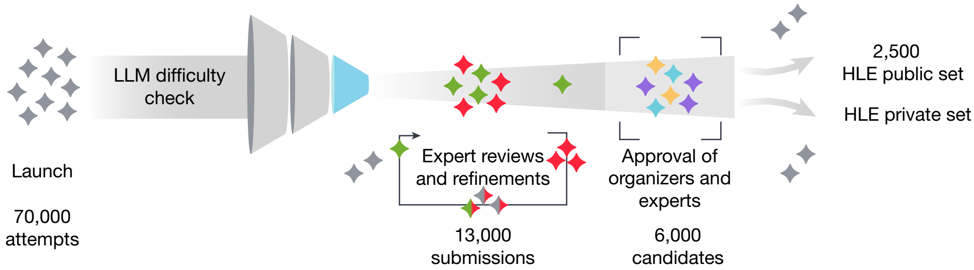


Fig. 4 | HLE dataset creation pipeline. We accept questions that make frontier LLMs fail, then iteratively refine them with the help of expert peer reviewers. Each question is then manually approved by organizers or expert reviewers trained by organizers. A private held-out set is kept apart from the public set to assess model overfitting and gaming on the public benchmark.

is to help the question contributors (who are primarily academics and researchers from a wide range of disciplines) better design questions that are closed-ended, robust and of high quality for AI evaluation. In the second round, good and outstanding questions from the first round are identified and approved by organizers and reviewers to be included in the final HLE dataset. Details, instructions and rubrics for both rounds can be found in Supplementary Information section 2. Figure 4 shows our full process.

Evaluation

We evaluate the performance of state-of-the-art LLMs on HLE and analyse their capabilities across different question types and domains. We describe our evaluation setup (see section 'Setup') and present several quantitative results on metrics that track model performance (see section 'Quantitative results').

Setup

After data collection and review, we evaluated our final HLE dataset on additional frontier multi-modal LLMs. We use a standardized system prompt that structures model responses into explicit reasoning followed by a final answer. As the question–answers are precise and close-ended, we use o3-mini as a judge to verify answer correctness against model predictions while accounting for equivalent formats (for example, decimals compared with fractions or estimations). Evaluation prompts are detailed in the Methods.

Quantitative results

Accuracy. All frontier models achieve low accuracy on HLE (Table 1), highlighting substantial room for improvement in narrowing the gap between current LLMs and expert-level academic capabilities on closed-ended questions. These low scores are partially by design the dataset collection process attempts to filter out questions that existing models can answer correctly. Nevertheless, we notice on evaluation that models exhibit non-zero accuracy. This is due to inherent noise in model inference—models can inconsistently guess the right answer or guess worse than random chance for multiple-choice questions. We notice an elevated accuracy on multiple-choice questions compared with exact-answer questions in Extended Data Table. 3. We choose to leave these questions in the dataset as a natural component instead of strongly adversarially filtering. However, we stress that the true capability floor of frontier models on the dataset will remain an open question, and small inflections close to zero accuracy are not strongly indicative of progress.

Calibration error. Given low performance on HLE, models should be calibrated, recognizing their uncertainty rather than confidently provide incorrect answers. To measure calibration, we prompt models to provide both an answer and their confidence from 0% to 100% (Methods), using the setup from[7]. The implementation of our RMS calibration error is from ref. 8. The stated confidence of a well-calibrated model should match its actual accuracy, for example, achieving 50% accuracy on questions, in which it claims 50% confidence. Table 1 shows poor calibration across all models, reflected in high RMS calibration error scores. Models frequently provide incorrect answers with high confidence on HLE, failing to recognize when questions exceed their capabilities.

Inference time computation. Reasoning models are designed to spend extra compute thinking before answering: they generate intermediate reasoning tokens and then produce the final response, which means substantially more tokens must be decoded at inference time[5,6]. To shed light on this in our evaluation, we analyse the compute-intensive scaling of output tokens (including reasoning tokens) across several state-of-the-art reasoning models in Fig. 5. Through binning output lengths with a $\log_2$ scale, we observe a log-linear scaling of accuracy with more reasoning tokens; however, this trend reverses after 2^{14} tokens, highlighting that a larger reasoning budget is not always optimal. The observation that accuracy benefits diminish beyond a certain threshold suggests that future models should improve not only their raw accuracy on HLE but also their computational efficiency.

Discussion

Limitations

Although present-day LLMs achieve very low accuracy on HLE, recent history shows benchmarks are quickly saturated—with models

Table 1 | Accuracy and RMS calibration error of different models on HLE, demonstrating low accuracy and high calibration error across all models

Model	Accuracy (%) ↑	Calibration error (%) ↓
GPT-4o	2.7 ± 0.6	89
Claude 3.5 Sonnet	4.1 ± 0.8	84
Gemini 1.5 Pro	4.6 ± 0.8	88
o1	8.0 ± 1.1	83
DeepSeek R1[a]	8.5 ± 1.2	73
Post-release models		
Claude 4 Sonnet	7.8 ± 1.1	75
Gemini 2.5 Pro	21.6 ± 1.6	72
GPT-5	25.3 ± 1.7	50

The most updated evaluations are hosted on https://lastexam.ai. Post-release models are released after HLE was open-sourced; we separate them as model builders have access to the HLE dataset. We report a breakdown of the text-only subset and other categories in Extended Data Tables. 1 and 2.
[a]Model is not multi-modal, evaluated on a text-only subset.

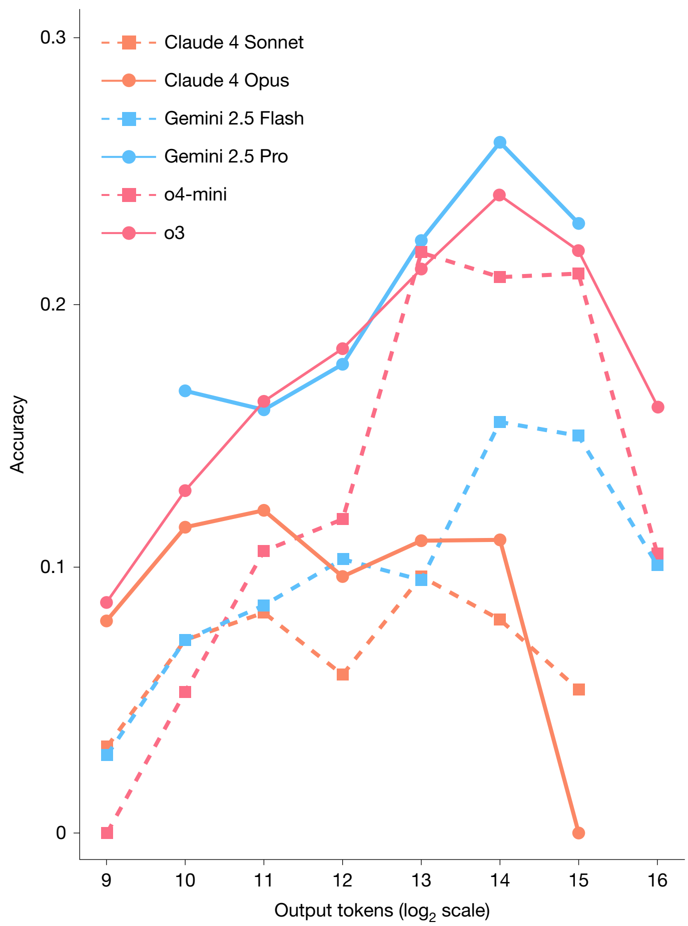


Fig. 5 | Accuracy compared with reasoning token budget. Accuracy binned by the total number of generated output tokens, showing a log-linear increase in accuracy peaking around 2^{14} tokens before reversing.

markedly progressing from near-zero to near-perfect performance in a short timeframe[9,10]. High accuracy on HLE would demonstrate expert-level performance on closed-ended, verifiable questions and cutting-edge scientific knowledge, but it would not alone suggest autonomous research capabilities or artificial general intelligence[11]. HLE tests structured academic problems rather than open-ended research or creative problem-solving abilities, making it a focused measure of technical knowledge and reasoning across a diverse range of subjects, albeit with a stronger representation in math and STEM (science, technology, engineering and mathematics) disciplines, as shown in Fig. 2. By pushing the limits of established closed-ended benchmarks, HLE is intended to hasten the transition towards a new class of benchmarks focused on more dynamic and open-ended AI capabilities.

Impact

By providing a clear measure of AI progress, HLE creates a common reference point for scientists and policymakers to assess AI capabilities. This enables more informed discussions about development trajectories, potential risks and necessary governance measures.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09962-4.

1. Hendrycks, D. et al. Measuring massive multitask language understanding. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=d7KBjmI3GmQ (ICLR, 2021).
2. Gemini Team Google. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at https://arxiv.org/abs/2403.05530 (2024).
3. OpenAI et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2024).
4. *The Claude 3 Model Family: Opus, Sonnet, Haiku* (Anthropic, 2024).
5. *OpenAI o1 System Card* (OpenAI, 2024).
6. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
7. Wei, J. et al. Measuring short-form factuality in large language models. Preprint at https://arxiv.org/abs/2411.04368 (2024).
8. Hendrycks, D. et al. PixMix: Dreamlike pictures comprehensively improve safety measures. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16783–16792 (IEEE/CVF, 2022).
9. Rein, D. et al. GPQA: A graduate-level Google-proof Q&A benchmark. In *Proc. First Conference on Language Modeling (COLM)* https://openreview.net/forum?id=Ti67584b98 (COLM, 2024).
10. Chollet, F., Knoop, M., Kamradt, G. & Landers, B. ARC prize 2024: technical report. Preprint at https://arxiv.org/abs/2412.04604 (2024).
11. Hendrycks, D. et al. A definition of AGI. Preprint at https://arxiv.org/abs/2510.18212 (2025).

Center for AI Safety

Long Phan[1]✉, Alice Gatti[1], Nathaniel Li[1], Adam Khoja[1], Ryan Kim[1], Richard Ren[1], Jason Hausenloy[1], Oliver Zhang[1], Mantas Mazeika[1] & Dan Hendrycks[1]✉

[1]Center for AI Safety, San Francisco CA, USA. ✉e-mail: agibenchmark@safe.ai; dan@safe.ai

Scale AI

Ziwen Han[2], Josephina Hu[2], Hugh Zhang[2], Chen Bo Calvin Zhang[2], Mohamed Shaaban[2], John Ling[2], Sean Shi[2], Michael Choi[2], Anish Agrawal[2], Arnav Chopra[2], Aakaash Nattanmai[2], Gordon McKellips[2], Anish Cheraku[2], Asim Suhail[2], Ethan Luo[2], Marvin Deng[2], Jason Luo[2], Ashley Zhang[2], Kavin Jindel[2], Jay Paek[2], Kasper Halevy[2], Allen Baranov[2], Michael Liu[2], Advaith Avadhanam[2], David Zhang[2], Vincent Cheng[2], Brad Ma[2], Evan Fu[2], Liam Do[2], Joshua Lass[2], Hubert Yang[2], Surya Sunkari[2], Vishruth Bharath[2], Violet Ai[2], James Leung[2], Rishit Agrawal[2], Alan Zhou[2], Kevin Chen[2], Tejas Kalpathi[2], Ziqi Xu[2], Gavin Wang[2], Tyler Xiao[2], Erik Maung[2], Sam Lee[2], Ryan Yang[2], Roy Yue[2], Ben Zhao[2], Julia Yoon[2], Xiangwan Sun[2], Aryan Singh[2], Clark Peng[2], Tyler Osbey[2], Taozhi Wang[2], Daryl Echeazu[2], Timothy Wu[2], Spandan Patel[2], Vidhi Kulkarni[2], Vijaykaarti Sundarapandiyan[2], Andrew Le[2], Zafir Nasim[2], Srikar Yalam[2], Ritesh Kasamsetty[2], Soham Samal[2], David Sun[2], Nihar Shah[2], Abhijeet Saha[2], Alex Zhang[2], Leon Nguyen[2], Laasya Nagumalli[2], Kaixin Wang[2], Aidan Wu[2], Anwith Telluri[2], Summer Yue[2] & Alexandr Wang[2]

[2]Scale AI, San Francisco CA, USA.

HLE Contributors Consortium

Dmitry Dodonov[3], Tung Nguyen[4], Jaeho Lee[5], Daron Anderson[3], Mikhail Doroshenko[3], Alun Cennyth Stokes[3], Mobeen Mahmood[6], Oleksandr Pokutnyi[7,8], Oleg Iskra[9], Jessica P. Wang[10], John-Clark Levin[11], Mstyslav Kazakov[12], Fiona Feng[13], Steven Y. Feng[14], Haoran Zhao[15], Michael Yu[3], Varun Gangal[3], Chelsea Zou[14], Zihan Wang[16], Serguei Popov[17], Robert Gerbicz[18], Geoff Galgon[19], Johannes Schmitt[20], Will Yeadon[21], Yongki Lee[22], Scott Sauers[23], Alvaro Sanchez[3], Fabian Giska[3], Marc Roth[24], Søren Riis[24], Saiteja Utpala[25], Noah Burns[14], Gashaw M. Goshu[3], Mohinder Maheshbhai Naiya[26], Chidozie Agu[27], Zachary Giboney[3], Antrell Cheatom[28], Francesco Fournier-Facio[11], Sarah-Jane Crowson[29], Lennart Finke[20], Zerui Cheng[30], Jennifer Zampese[31], Ryan G. Hoerr[32], Mark Nandor[3], Hyunwoo Park[9], Tim Gehrunger[20], Jiaqi Cai[33], Ben McCarty[34], Alexis C. Garretson[35,36], Edwin Taylor[3], Damien Sileo[37], Qiuyu Ren[38], Usman Qazi[39,40], Lianghui Li[41], Jungbae Nam[42], John B. Wydallis[3], Pavel Arkhipov[43], Jack Wei Lun Shi[44], Aras Bacho[45], Chris G. Willcocks[21], Hangrui Cao[9], Sumeet Motwani[46], Emily de Oliveira Santos[47], Johannes Veith[48,49], Edward Vendrow[33], Doru Cojoc[50], Kengo Zenitani[3], Joshua Robinson[51], Longke Tang[30], Yuqi Li[52], Joshua Vendrow[33], Natanael Wildner Fraga[3], Vladyslav Kuchkin[53],

Andrey Pupasov Maksimov[54], Pierre Marion[41], Denis Efremov[55], Jayson Lynch[33], Kaiqu Liang[30], Aleksandar Mikov[41], Andrew Gritsevskiy[56], Julien Guillod[57,58], Gözdenur Demir[3], Dakotah Martinez[3], Ben Pageler[3], Kevin Zhou[38], Saeed Soori[59], Ori Press[60], Henry Tang[46], Paolo Rissone[61], Sean R. Green[3], Lina Brüssel[11], Moon Twayana[62], Aymeric Dieuleveut[63], Joseph Marvin Imperial[64,65], Ameya Prabhu[60], Jinzhou Yang[66], Nick Crispino[67], Arun Rao[68], Dimitri Zvonkine[69,70], Gabriel Loiseau[37], Mikhail Kalinin[71], Marco Lukas[72], Ciprian Manolescu[14], Nate Stambaugh[73], Subrata Mishra[74], Tad Hogg[75], Carlo Bosio[38], Brian P. Coppola[76], Julian Salazar[77], Jaehyeok Jin[50], Rafael Sayous[69], Stefan Ivanov[11], Philippe Schwaller[41], Shaipranesh Senthilkumar[41], Andres M. Bran[41], Andres Algaba[78], Kelsey Van den Houte[78,79], Lynn Van Der Sypt[78,79], Brecht Verbeken[78], David Noever[80], Alexei Kopylov[3], Benjamin Myklebust[3], Bikun Li[81], Lisa Schut[46], Evgenii Zheltonozhskii[82], Qiaochu Yuan[3], Derek Lim[33], Richard Stanley[33,83], Tong Yang[9], John Maar[84], Julian Wykowski[11], Mart Oller[11], Anmol Sahu[3], Cesare Giulio Ardito[85], Yuzheng Hu[86], Ariel Ghislain Kemogne Kamdoum[87], Alvin Jin[33], Tobias Garcia Vilchis[88], Yuexuan Zu[33], Martin Lackner[89], James Koppel[3], Gongbo Sun[90], Daniil S. Antonenko[91], Steffi Chern[9], Bingchen Zhao[92], Pierrot Arsene[93], Joseph M. Cavanagh[38], Daofeng Li[67], Jiawei Shen[67], Donato Crisostomi[61], Wenjin Zhang[67], Ali Dehghan[3], Sergey Ivanov[3], David Perrella[94], Nurdin Kaparov[95], Allen Zang[81], Ilia Sucholutsky[96], Arina Kharlamova[97], Daniil Orel[97], Vladislav Poritski[3], Shalev Ben-David[98], Zachary Berger[33], Parker Whitfill[33], Michael Foster[3], Daniel Munro[16], Linh Ho[3], Shankar Sivarajan[99], Dan Bar Hava[100], Aleksey Kuchkin[3], David Holmes[101], Alexandra Rodriguez-Romero[3], Frank Sommerhage[102], Anji Zhang[33], Richard Moat[103], Keith Schneider[3], Zakayo Kazibwe[104], Don Clarke[105], Dae Hyun Kim[106], Felipe Meneguitti Dias[47], Sara Fish[107], Veit Elser[108], Tobias Kreiman[38], Victor Efren Guadarrama Vilchis[109], Immo Klose[50], Ujjwala Anantheswaran[110], Adam Zweiger[33], Kaivalya Rawal[46], Jeffery Li[33], Jeremy Nguyen[111], Nicolas Daans[112], Haline Heidinger[113,114], Maksim Radionov[115], Václav Rozhoň[116], Vincent Ginis[78,107], Christian Stump[117], Niv Cohen[96], Rafał Poświata[118], Josef Tkadlec[119], Alan Goldfarb[38], Chenguang Wang[67], Piotr Padlewski[3], Stanislaw Barzowski[3], Kyle Montgomery[67], Ryan Stendall[120], Jamie Tucker-Foltz[107], Jack Stade[121], T. Ryan Rogers[122], Tom Goertzen[123], Declan Grabb[14], Abhishek Shukla[124], Alan Givré[125], John Arnold Ambay[126], Archan Sen[38], Muhammad Fayez Aziz[86], Mark H. Inlow[127], Hao He[128], Ling Zhang[128], Younesse Kaddar[46], Ivar Ängquist[129], Yanxu Chen[130], Harrison K. Wang[107], Kalyan Ramakrishnan[46], Elliott Thornley[46], Antonio Terpin[20], Hailey Schoelkopf[3], Eric Zheng[9], Avishy Carmi[131], Ethan D. L. Brown[132], Kelin Zhu[99], Max Bartolo[133], Richard Wheeler[92], Martin Stehberger[3], Peter Bradshaw[86], JP Heimonen[134], Kaustubh Sridhar[135], Ido Akov[136], Jennifer Sandlin[110], Yury Makarychev[137], Joanna Tam[138], Hieu Hoang[139], David M. Cunningham[3], Vladimir Goryachev[3], Demosthenes Patramanis[46], Michael Krause[140], Andrew Redenti[50], David Aldous[38], Jesyin Lai[141], Shannon Coleman[39], Jiangnan Xu[142], Sangwon Lee[3], Ilias Magoulas[143], Sandy Zhao[3], Ning Tang[38], Michael K. Cohen[38], Orr Paradise[38], Jan Hendrik Kirchner[144], Maksym Ovchynnikov[145], Jason O. Matos[138], Adithya Shenoy[3], Michael Wang[38], Yuzhou Nie[146], Anna Sztyber-Betley[147], Paolo Faraboschi[148], Robin Riblet[93], Jonathan Crozier[149], Shiv Halasyamani[150], Shreyas Verma[3], Prashant Joshi[151], Eli Meril[152], Ziqiao Ma[76], Jérémy Andréoletti[57], Raghav Singhal[97], Jacob Platnick[153], Volodymyr Nevirkovets[154], Luke Basler[155], Alexander Ivanov[117], Seri Khoury[116], Nils Gustafsson[129], Marco Piccardo[156], Hamid Mostaghimi[87], Qijia Chen[107], Virendra Singh[157], Tran Quoc Khánh[158], Paul Rosu[159], Hannah Szlyk[67], Zachary Brown[33], Himanshu Narayan[3], Aline Menezes[3], Jonathan Roberts[11], William Alley[3], Kunyang Sun[38], Arkil Patel[6,160], Max Lamparth[14], Anka Reuel[14], Linwei Xin[81], Hanmeng Xu[91], Jacob Loader[11], Freddie Martin[3], Zixuan Wang[30], Andrea Achilleos[161], Thomas Preu[162], Tomek Korbak[163], Ida Bosio[164], Fereshteh Kazemi[3], Ziye Chen[165], Biró Bálint[3], Eve J. Y. Lo[166], Jiaqi Wang[15], Maria Inês S. Nunes[167], Jeremiah Milbauer[9], M. Saiful Bari[168], Zihao Wang[81], Behzad Ansarinejad[3], Yewen Sun[169], Stephane Durand[170], Hossam Elgnainy[171], Guillaume Douville[3], Daniel Tordera[172], George Balabanian[135], Hew Wolff[3], Lynna Kvistad[173], Hsiaoyun Milliron[174], Ahmad Sakor[72], Murat Eron[3], D. O. Andrew Favre[175], Shailesh Shah[176], Xiaoxiang Zhou[48], Firuz Kamalov[177], Sherwin Abdoli[3], Tim Santens[11], Shaul Barkan[178], Allison Tee[14], Robin Zhang[33], Alessandro Tomasiello[179], G. Bruno De Luca[14], Shi-Zhuo Looi[45], Vinh-Kha Le[38], Noam Kolt[178], Jiayi Pan[38], Emma Rodman[180], Jacob Drori[3], Carl J. Fossum[181], Niklas Muennighoff[14], Milind Jagota[38], Ronak Pradeep[98], Honglu Fan[182], Jonathan Eicher[3], Michael Chen[45], Kushal Thaman[14], William Merrill[96], Moritz Firsching[183], Carter Harris[184], Stefan Ciobâcă[185], Jason Gross[3], Rohan Pandey[3], Ilya Gusev[3], Adam Jones[3], Shashank Agnihotri[186], Pavel Zhelnov[59], Mohammadreza Mofayezi[59], Alexander Piperski[187], David K. Zhang[14], Kostiantyn Dobarskyi[3], Roman Leventov[3], Ignat Soroko[62], Joshua Duersch[188], Vage Taamazyan[189], Andrew Ho[190], Wenjie Ma[38], William Held[14,153], Ruicheng Xian[86], Armel Randy Zebaze[37], Mohanad Mohamed[191], Julian Noah Leser[89], Michelle X. Yuan[3], Laila Yacar[125], Johannes Lengler[20], Katarzyna Olszewska[3], Claudio Di Fratta[192], Edson Oliveira[193], Joseph W. Jackson[194], Andy Zou[9,195], Muthu Chidambaram[159], Timothy Manik[3], Hector Haffenden[3], Dashiell Stander[196], Ali Dasouqi[197], Alexander Shen[198], Bita Golshani[3], David Stap[130], Egor Kretov[199], Mikalai Uzhou[200], Alina Borisovna Zhidkovskaya[201], Nick Winter[3], Miguel Orbegozo Rodriguez[20], Robert Lauff[84], Dustin Wehr[3], Colin Tang[9], Zaki Hossain[11], Shaun Phillips[3], Fortuna Samuele[202], Fredrik Ekström[3], Angela Hammon[3], Oam Patel[107], Faraz Farhidi[203], George Medley[3], Forough Mohammadzadeh[3], Madellene Peñaflor[204], Haile Kassahun[6], Alena Friedrich[205], Rayner Hernandez Perez[81], Daniel Pyda[206], Taom Sakal[146], Omkar Dhamane[207], Ali Khajegili Mirabadi[39], Eric Hallman[3], Kenchi Okutsu[208], Mike Battaglia[3], Mohammad Maghsoudimehrabani[209], Alon Amit[210], Dave Hulbert[3], Roberto Pereira[211], Simon Weber[20], Handoko[3], Anton Peristyy[3], Stephen Malina[212], Mustafa Mehkary[59,213], Rami Aly[11], Frank Reidegeld[3], Anna-Katharina Dick[60], Cary Friday[214], Mukhwinder Singh[215], Hassan Shapourian[216], Wanyoung Kim[3], Mariana Costa[3], Hubeyb Gurdogan[68], Harsh Kumar[217], Chiara Ceconello[3], Chao Zhuang[3], Haon Park[218,219], Micah Carroll[38], Andrew R. Tawfeek[15], Stefan Steinerberger[15], Daattavya Aggarwal[11], Michael Kirchhof[60], Linjie Dai[33], Evan Kim[33], Johan Ferret[77], Jainam Shah[3], Yuzhou Wang[153], Minghao Yan[90], Krzysztof Burdzy[15], Lixin Zhang[3], Antonio Franca[11], Diana T. Pham[220], Kang Yong Loh[14], Joshua Robinson[221], Abram Jackson[3], Paolo Giordano[222], Philipp Petersen[222], Adrian Cosma[223], Jesus Colino[3], Colin White[224], Jacob Votava[30], Vladimir Vinnikov[3], Ethan Delaney[225], Petr Spelda[120], Vit Stritecky[120], Syed M. Shahid[226], Jean-Christophe Mourrat[70,227], Lavr Vetoshkin[228], Koen Sponselee[229], Renas Bacho[230], Zheng-Xin Yong[5], Florencia de la Rosa[231], Nathan Cho[14], Xiuyu Li[38], Guillaume Malod[58,232], Orion Weller[197], Guglielmo Albani[233], Leon Lang[130], Julien Laurendeau[41], Dmitry Kazakov[107], Fatimah Adesanya[3], Julien Portier[11], Lawrence Hollom[11], Victor Souza[11], Yuchen Anna Zhou[234], Julien Degorre[3], Yiğit Yalın[235], Gbenga Daniel Obikoya[3], Rai Michael Pokorny[236], Filippo Bigi[41], M. C. Boscá[237], Oleg Shumar[3], Kaniuar Bacho[92], Gabriel Recchia[238], Mara Popescu[239], Nikita Shulga[240], Ngefor Mildred Tanwie[241], Thomas C. H. Lux[3], Ben Rank[3], Colin Ni[68], Matthew Brooks[3], Alesia Yakimchyk[242], Huanxu Quinn Liu[243], Stefano Cavalleri[3], Olle Häggström[244], Emil Verkama[129], Joshua Newbould[21], Hans Gundlach[33], Leonor Brito-Santana[245], Brian Amaro[14], Vivek Vajipey[14], Rynaa Grover[153], Ting Wang[67], Yosi Kratish[154], Wen-Ding Li[108], Sivakanth Gopi[25], Andrea Caciolai[61], Christian Schroeder de Witt[46], Pablo Hernández-Cámara[172], Emanuele Rodolà[61], Jules Robins[3], Dominic Williamson[123], Brad Raynor[3], Hao Qi[165], Ben Segev[50], Jingxuan Fan[107], Sarah Martinson[107], Erik Y. Wang[107], Kaylie Hausknecht[107], Michael P. Brenner[107], Mao Mao[165], Christoph Demian[48], Peyman Kassani[246], Xinyu Zhang[165], David Avagian[186], Eshawn Jessica Scipio[247], Alon Ragoler[248], Justin Tan[11], Blake Sims[3], Rebeka Plecnik[3], Aaron Kirtland[5], Omer Faruk Bodur[3], D. P. Shinde[3], Yan Carlos Leyva Labrador[249], Zahra Adoul[250], Mohamed Zekry[251], Ali Karakoc[252], Tania C. B. Santos[3], Samir Shamseldeen[253], Loukmane Karim[213], Anna Liakhovitskaia[254], Nate Resman[255], Nicholas Farina[3], Juan Carlos Gonzalez[256], Gabe Maayan[165], Earth Anderson[257], Rodrigo De Oliveira Pena[258], Elizabeth Kelley[255], Hodjat Mariji[3], Rasoul Pouriamanesh[3], Wentao Wu[39], Ross Finocchio[3], Ismail Alarab[259], Joshua Cole[260], Danyelle Ferreira[3], Bryan Johnson[261], Mohammad Safdari[262], Liangti Dai[46], Siriphan Arthornthurasuk[3], Isaac C. McAlister[3], Alejandro José Moyano[263], Alexey Pronin[264], Jing Fan[239], Angel Ramirez-Trinidad[3], Yana Malysheva[67], Daphiny Pottmaier[265], Omid Taheri[266], Stanley Stepanic[267], Samuel Perry[3], Luke Askew[268], Raúl Adrián Huerta Rodrguez[3], Ali M. R. Minissi[269], Ricardo Lorena[270], Krishnamurthy Iyer[23], Arshad Anil Fasiludeen[11], Ronald Clark[46], Josh Ducey[271], Matheus Piza[272], Maja Somrak[3], Eric Vergo[3], Juehang Qin[273], Benjámin Borbás[274], Eric Chu[77], Jack Lindsey[144], Antoine Jallon[3], I. M. J. McInnis[3], Evan Chen[33], Avi Semler[46], Luk Gloor[3], Tej Shah[275], Marc Carauleanu[276], Pascal Lauer[128,277], Tran Duc Huy[278], Hossein Shahrtash[279], Emilien Duc[20], Lukas Lewark[20], Assaf Brown[178], Samuel Albanie[3], Brian Weber[280], Warren S. Vaz[3], Pierre Clavier[281], Yiyang Fan[3], Gabriel Poesia Reis e Silva[14], Long Tony Lian[38], Marcus Abramovitch[3], Xi Jiang[81], Sandra Mendoza[282,283], Murat Islam[284], Juan Gonzalez[3], Vasilios Mavroudis[285], Justin Xu[46], Pawan Kumar[286], Laxman Prasad Goswami[124], Daniel Bugas[3], Nasser Heydari[3], Ferenc Jeanplong[3], Thorben Jansen[287], Antonella Pinto[3], Archimedes Apronti[288], Abdallah Galal[289], Ng Ze-An[290], Ankit Singh[291], Tong Jiang[107], Joan of Arc Xavier[3], Kanu Priya Agarwal[3], Mohammed Berkani[292], Gang Zhang[3], Zhehang Du[135], Benedito Alves de Oliveira Junior[47], Dmitry Malishev[3], Nicolas Remy[293], Taylor D. Hartman[294], Tim Tarver[295], Stephen Mensah[3], Gautier Abou Loume[241], Wiktor Morak[3], Farzad Habibi[296], Sarah Hoback[107], Will Cai[38], Javier Gimenez[3], Roselynn Grace Montecillo[297], Jakub Łucki[20], Russell Campbell[298], Asankhaya Sharma[299], Khalida Meer[3], Shreen Gul[300], Daniel Espinosa Gonzalez[146], Xavier Alapont[3], Alex Hoover[81], Gunjan Chhablani[153], Freddie Vargus[301], Arunim Agarwal[1], Yibo Jiang[81], Deepakkumar Patil[302], David Outevsky[3], Kevin Joseph Scaria[110], Rajat Maheshwari[303], Abdelkader Dendane[3], Priti Shukla[3], Ashley Cartwright[304], Sergei Bogdanov[281], Niels Mündler[20], Sören Möller[305], Luca Arnaboldi[41], Kunvar Thaman[306], Muhammad Rehan Siddiqi[307], Prajvi Saxena[308], Himanshu Gupta[110], Tony Fruhauff[3], Glen Sherman[3], Mátyás Vincze[309,310], Siranut Usawasutsakorn[311], Dylan Ler[3], Anil Radhakrishnan[149], Innocent Enyekwe[3], Sk Md Salauddin[312], Jiang Muzhen[3], Aleksandr Maksapetyan[3], Vivien Rossbach[3], Chris Harjadi[14], Mohsen Bahaloohoreh[3], Claire Sparrow[81], Jasdeep Sidhu[3], Sam Ali[51], Song Bian[90], John Lai[3], Eric Singer[313], Justine Leon Uro[3], Greg Bateman[3], Mohamed Sayed[3], Ahmed Menshawy[314], Darling Duclosel[315], Dario Bezzi[316], Yashaswini Jain[317], Ashley Aaron[3], Murat Tiryakioglu[3], Sheeshram Siddh[3], Keith Krenek[3], Imad Ali Shah[225], Jun Jin[3], Scott Creighton[3], Denis Peskoff[30], Zienab EL-Wasif[269], Ragavendran P V[3], Michael Richmond[3], Joseph McGowan[59], Tejal Patwardhan[236], Hao-Yu Sun[318], Ting Sun[86], Nikola Zubić[162], Samuele Sala[319], Stephen Ebert[68], Jean Kaddour[161], Manuel Schottdorf[320], Dianzhuo Wang[107], Gerol Petruzella[321], Alex Meiburg[98,322], Tilen Medved[323], Ali ElSheikh[154], S. Ashwin Hebbar[30], Lorenzo Vaquero[309], Xianjun Yang[146], Jason Poulos[324], Vilém Zouhar[20], Sergey Bogdanik[3], Mingfang Zhang[325], Jorge Sanz-Ros[14], David Anugraha[59], Yinwei Dai[30], Anh N. Nhu[99], Xue Wang[197], Ali Anil Demircali[326], Zhibai Jia[108], Yuyin Zhou[327], Juncheng Wu[327], Mike He[30], Nitin Chandok[3], Aarush Sinha[328], Gaoxiang Luo[23], Long Le[51], Mickaël Noyé[329], Michał Perełkiewicz[118], Ioannis Pantidis[330], Tianbo Qi[331], Soham Sachin Purohit[76], Letitia Parcalabescu[332], Thai-Hoa Nguyen[333], Genta Indra Winata[3], Edoardo M. Ponti[92], Hanchen Li[81], Kaustubh Dhole[143], Jongee Park[334], Dario Abbondanza[335], Yuanli Wang[165], Anupam Nayak[9], Diogo M. Caetano[270], Antonio A. W. L. Wong[39], Maria del Rio-Chanona[161,336], Dániel Kondor[336], Pieter Francois[46,285], Ed Chalstrey[161], Jakob Zsambok[336], Dan Hoyer[336], Jenny Reddish[336], Jakob Hauser[336], Francisco-Javier Rodrigo-Ginés[337], Suchandra Datta[3], Maxwell Shepherd[197], Thom Kamphuis[338], Qizheng Zhang[14], Hyunjun Kim[339], Ruiji Sun[38], Jianzhu Yao[30], Franck Dernoncourt[340], Satyapriya Krishna[107], Sina Rismanchian[296], Bonan Pu[3], Francesco Pinto[81], Yingheng Wang[108], Kumar Shridhar[20], Kalon J. Overholt[33], Glib Briia[341], Hieu Nguyen[342], David Quod Soler Bartomeu[343], Tony CY Pang[123,344], Adam Wecker[3], Yifan Xiong[25], Fanfei Li[266], Lukas S. Huber[60,345], Joshua Jaeger[345], Romano De Maddalena[346], Xing Han Lù[6], Yuhui Zhang[14], Claas Beger[108], Patrick Tser Jern Kon[76], Sean Li[94], Vivek Sanker[14], Ming Yin[30], Yihao Liang[30], Xinlu Zhang[146], Ankit Agrawal[347], Li S. Yifei[135], Zechen Zhang[107], Mu Cai[90], Yasin Sonmez[38], Costin Cozianu[25], Changhao Li[33], Alex Slen[135], Shoubin Yu[348], Hyun Kyu Park[349], Gabriele Sarti[350], Marcin Briański[351], Alessandro Stolfo[20], Truong An Nguyen[352], Mike Zhang[353], Yotam Perlitz[354], Jose Hernandez-Orallo[355], Runjia Li[46], Amin Shabani[356], Felix Juefei-Xu[3], Shikhar Dhingra[357], Orr Zohar[14], My Chiffon Nguyen[3], Alexander Pondaven[46], Abdurrahim Yilmaz[326], Xuandong Zhao[38], Chuanyang Jin[197], Muyan Jiang[38], Stefan Todoran[15], Xinyao Han[33], Jules Kreuer[60], Brian Rabern[92], Anna Plassart[103], Martino Maggetti[358], Luther Yap[30], Robert Geirhos[60], Jonathon Kean[359], Dingsu Wang[3], Sina Mollaei[14], Chenkai Sun[86], Yifan Yin[197], Shiqi Wang[331], Rui Li[14], Yaowen Chang[86], Anjiang Wei[14], Alice Bizeul[20], Xiaohan Wang[14], Alexandre Oliveira Arrais[3], Kushin Mukherjee[14], Jorge Chamorro-Padial[360], Jiachen Liu[76], Xingyu Qu[97], Junyi Guan[97], Adam Bouyamourn[38], Shuyu Wu[76], Martyna Plomecka[162], Junda Chen[16], Mengze Tang[90], Jiaqi Deng[153], Shreyas Subramanian[361], Haocheng Xi[38], Haoxuan Chen[14], Weizhi Zhang[28], Yinuo Ren[14],

Haoqin Tu[327], Sejong Kim[339], Yushun Chen[362], Sara Vera Marjanović[121], Junwoo Ha[363], Grzegorz Luczyna[3], Jeff J. Ma[76], Zewen Shen[59], Dawn Song[38], Cedegao E. Zhang[33], Zhun Wang[38], Gaël Gendron[364], Yunze Xiao[9], Leo Smucker[59], Erica Weng[9], Kwok Hao Lee[44], Zhe Ye[38], Stefano Ermon[14], Ignacio D. Lopez-Miguel[89], Theo Knights[81], Anthony Gitter[90,365], Namkyu Park[366], Boyi Wei[30], Hongzheng Chen[108], Kunal Pai[367], Ahmed Elkhanany[368], Han Lin[348], Philipp D. Siedler[332], Jichao Fang[294], Ritwik Mishra[369], Károly Zsolnai-Fehér[370], Xilin Jiang[50], Shadab Khan[371], Jun Yuan[372], Rishab Kumar Jain[107], Xi Lin[76], Mike Peterson[3], Zhe Wang[373], Aditya Malusare[374], Maosen Tang[108], Isha Gupta[143], Ivan Fosin[3], Timothy Kang[3], Barbara Dworakowska[326], Kazuki Matsumoto[375], Guangyao Zheng[197], Gerben Sewuster[376], Jorge Pretel Villanueva[377], Ivan Rannev[378], Igor Chernyavsky[85], Jiale Chen[101], Deepayan Banik[59], Ben Racz[9], Wenchao Dong[379], Jianxin Wang[196], Laila Bashmal[3], Duarte V. Gonçalves[17], Wei Hu[86], Kaushik Bar[380], Ondrej Bohdal[92], Atharv Singh Patlan[30], Shehzaad Dhuliawala[20], Caroline Geirhos[381], Julien Wist[382], Yuval Kansal[30], Bingsen Chen[96], Kutay Tire[383], Atak Talay Yücel[383], Brandon Christof[13], Veerupaksh Singla[374], Zijian Song[367], Sanxing Chen[159], Jiaxin Ge[38], Kaustubh Ponkshe[97], Isaac Park[96], Tianneng Shi[38], Martin Q. Ma[9], Joshua Mak[384], Sherwin Lai[14], Antoine Moulin[385], Zhuo Cheng[9], Zhanda Zhu[59], Ziyi Zhang[81], Vaidehi Patil[348], Ketan Jha[386], Qiutong Men[96], Jiaxuan Wu[90], Tianchi Zhang[81], Bruno Hebling Vieira[162], Alham Fikri Aji[97], Jae-Won Chung[76], Mohammed Mahfoud[160], Ha Thi Hoang[3], Marc Sperzel[3], Wei Hao[50], Kristof Meding[60], Sihan Xu[76], Vassilis Kostakos[387], Davide Manini[82], Yueying Liu[86], Christopher Toukmaji[296], Eunmi Yu[388], Arif Engin Demircali[389], Zhiyi Sun[76], Ivan Dewerpe[342], Hongsen Qin[45], Roman Pflugfelder[390,391], James Bailey[392], Johnathan Morris[9], Ville Heilala[393], Sybille Rosset[394], Zishun Yu[28], Peter E. Chen[6], Woongyeong Yeo[339], Eeshaan Jain[41], Sreekar Chigurupati[395], Julia Chernyavsky[3], Sai Prajwal Reddy[395], Subhashini Venugopalan[342], Hunar Batra[46], Core Francisco Park[107], Hieu Tran[99], Guilherme Maximiano[3], Genghan Zhang[14], Yizhuo Liang[51], Hu Shiyu[396], Rongwu Xu[15], Rui Pan[30], Siddharth Suresh[90], Ziqi Liu[90], Samaksh Gulati[362], Songyang Zhang[159], Peter Turchin[336], Christopher W. Bartlett[169], Christopher R. Scotese[154], Phuong M. Cao[86], Ben Wu[397], Jacek Karwowski[46] & Davide Scaramuzza[162]

[3]Independent Researcher, https://lastexam.ai. [4]Texas A&M University, College Station, TX, USA. [5]Brown University, Providence, RI, USA. [6]McGill University, Montreal Quebec, Canada. [7]Institute of Mathematics of NAS of Ukraine, Kyiv, Ukraine. [8]Kiev School of Economics, Kyiv, Ukraine. [9]Carnegie Mellon University, Pittsburgh, PA, USA. [10]RWTH Aachen University, Aachen, Germany. [11]University of Cambridge, Cambridge, UK. [12]Kyiv Polytechnic Institute, Kyiv, Ukraine. [13]Queen's University, Kingston Ontario, Canada. [14]Stanford University, Stanford, CA, USA. [15]University of Washington, Seattle, WA, USA. [16]University of California San Diego, San Diego, CA, USA. [17]University of Porto, Porto, Portugal. [18]ELTE, Budapest, Hungary. [19]Nimbus AI, Honolulu, HI, USA. [20]ETH Zürich, Zurich, Switzerland. [21]Durham University, Durham, UK. [22]Georgia Southern University, Statesboro, GA, USA. [23]University of Minnesota, Minneapolis, MN, USA. [24]Queen Mary University of London, London, UK. [25]Microsoft, Redmond, WA, USA. [26]Auckland University of Technology, Auckland, New Zealand. [27]Alberta Health Services, Edmonton Alberta, Canada. [28]University of Illinois Chicago, Chicago, IL, USA. [29]Hereford College of Arts, Hereford, UK. [30]Princeton University, Princeton, NJ, USA. [31]University of Canterbury, Christchurch, New Zealand. [32]Metropolitan State University of Denver, Denver, CO, USA. [33]Massachusetts Institute of Technology, Cambridge, MA, USA. [34]Accenture Labs, Washington, DC, USA. [35]Tufts University, Medford, MA, USA. [36]The Jackson Laboratory, Bar Harbor, ME, USA. [37]INRIA, Paris, France. [38]University of California, Berkeley, Berkeley, CA, USA. [39]University of British Columbia, Vancouver British Columbia, Canada. [40]Ross University School of Medicine, Bridgetown, Barbados. [41]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [42]Concordia University, Montreal Quebec, Canada. [43]Institute of Science and Technology Austria, Klosterneuburg, Austria. [44]National University of Singapore, Singapore, Singapore. [45]California Institute of Technology, Pasadena, CA, USA. [46]University of Oxford, Oxford, UK. [47]University of São Paulo, São Paulo, Brazil. [48]Humboldt-Universität zu Berlin, Berlin, Germany. [49]Charité – Universitätsmedizin, Berlin, Germany. [50]Columbia University, New York, NY, USA. [51]University of Southern California, Los Angeles, CA, USA. [52]C. N. Yang Institute for Theoretical Physics, Stony Brook, NY, USA. [53]University of Luxembourg, Luxembourg City, Luxembourg. [54]Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil. [55]Rockwell Automation, Milwaukee, WI, USA. [56]Contramont Research, San Francisco, CA, USA. [57]École Normale Supérieure, Paris, France. [58]Sorbonne Université, Paris, France. [59]University of Toronto, Toronto Ontario, Canada. [60]University of Tübingen, Tübingen, Germany. [61]Sapienza University of Rome, Rome, Italy. [62]University of North Texas, Denton, TX, USA. [63]Institut Polytechnique de Paris, Palaiseau, France. [64]National University Philippines, Manila, The Philippines. [65]University of Bath, Bath, UK. [66]Maastricht University, Maastricht, The Netherlands. [67]Washington University, St Louis, MO, USA. [68]University of California, Los Angeles, Los Angeles, CA, USA. [69]Université Paris-Saclay, Gif-sur-Yvette, France. [70]CNRS, Paris, France. [71]Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany. [72]Leibniz University Hannover, Hannover, Germany. [73]Diverging Mathematics, Boston, MA, USA. [74]Indian Institute of Technology Bombay, Mumbai, India. [75]Institute for Molecular Manufacturing, Palo Alto, CA, USA. [76]University of Michigan, Ann Arbor, MI, USA. [77]Google DeepMind, London, UK. [78]Vrije Universiteit Brussel, Brussels, Belgium. [79]UZ Brussel, Brussels, Belgium. [80]PeopleTec, Huntsville, AL, USA. [81]University of Chicago, Chicago, IL, USA. [82]Technion – Israel Institute of Technology, Haifa, Israel. [83]University of Miami, Coral Gables, FL, USA. [84]Technische Universität Berlin, Berlin, Germany. [85]University of Manchester, Manchester, UK. [86]University of Illinois Urbana-Champaign, Urbana-Champaign, IL, USA. [87]University of Calgary, Calgary Alberta, Canada. [88]Universidad Iberoamericana, Mexico City, Mexico. [89]TU Wien, Vienna, Austria. [90]University of Wisconsin-Madison, Madison, WI, USA. [91]Yale University, New Haven, CT, USA. [92]University of Edinburgh, Edinburgh, UK. [93]École Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France. [94]University of Western Australia, Perth Western Australia, Australia. [95]Snorkel AI, Redwood City, CA, USA. [96]New York University, New York, NY, USA. [97]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. [98]University of Waterloo, Waterloo Ontario, Canada. [99]University of Maryland, College Park, MD, USA. [100]Manhattan School of Music, New York, NY, USA. [101]Universiteit Leiden, Leiden, The Netherlands. [102]Synbionix, Casselberry, FL, USA. [103]The Open University, Milton Keynes, UK. [104]Corteva Agriscience, Indianapolis, IN, USA. [105]Sanford Burnham Prebys, La Jolla, CA, USA. [106]Yonsei University, Seoul, South Korea. [107]Harvard University, Cambridge, MA, USA. [108]Cornell University, Ithaca, NY, USA. [109]University of Leeds, Leeds, UK. [110]Arizona State University, Tempe, AZ, USA. [111]Swinburne University of Technology, Melbourne Victoria, Australia. [112]KU Leuven, Leuven, Belgium. [113]St Petersburg College, St Petersburg, FL, USA. [114]La Molina National Agrarian University, Lima, Peru. [115]Brandenburg University of Technology, Cottbus, Germany. [116]INSAIT, Sofia, Bulgaria. [117]Ruhr University Bochum, Bochum, Germany. [118]National Information Processing Institute, Warsaw, Poland. [119]Charles University, Prague, Czech Republic. [120]Cranfield University, Cranfield, UK. [121]University of Copenhagen, Copenhagen, Denmark. [122]TRR Designs, Fayetteville, AR, USA. [123]The University of Sydney, Sydney New South Wales, Australia. [124]Indian Institute of Technology Delhi, New Delhi, India. [125]Universidad de Buenos Aires, Buenos Aires, Argentina. [126]University of Technology Sydney, Sydney New South Wales, Australia. [127]Indiana State University, Terre Haute, IN, USA. [128]Australian National University, Canberra Australian Capital Territory, Australia. [129]KTH Royal Institute of Technology, Stockholm, Sweden. [130]University of Amsterdam, Amsterdam, The Netherlands. [131]Ben-Gurion University, Beersheba, Israel. [132]Donald and Barbara Zucker School of Medicine, Hempstead, NY, USA. [133]Cohere, Toronto Ontario, Canada. [134]Siili Solutions, Helsinki, Finland. [135]University of Pennsylvania, Philadelphia, PA, USA. [136]Aalto University, Espoo, Finland. [137]Toyota Technological Institute at Chicago, Chicago, IL, USA. [138]Northeastern University, Boston, MA, USA. [139]Case Western Reserve University, Cleveland, OH, USA. [140]University of Windsor, Windsor, Ontario, Canada. [141]St. Jude Children's Research Hospital, Memphis, TN, USA. [142]Rochester Institute of Technology, Rochester, NY, USA. [143]Emory University, Atlanta, GA, USA. [144]Anthropic, San Francisco, CA, USA. [145]CERN, Geneva, Switzerland. [146]University of California Santa Barbara, Santa Barbara, CA, USA. [147]Warsaw University of Technology, Warsaw, Poland. [148]Hewlett Packard Enterprise, San Francisco, CA, USA. [149]North Carolina State University, Raleigh, NC, USA. [150]University of Houston, Houston, TX, USA. [151]All India Institute of Medical Sciences, New Delhi, India. [152]Tel Aviv University, Tel Aviv, Israel. [153]Georgia Institute of Technology, Atlanta, GA, USA. [154]Northwestern University, Evanston, IL, USA. [155]University of Arizona, Tucson, AZ, USA. [156]Universidade de Lisboa, Lisbon, Portugal. [157]Indian Institute of Technology Kharagpur, Kharagpur, India. [158]Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. [159]Duke University, Durham, NC, USA. [160]Mila - Québec AI Institute, Montreal Quebec, Canada. [161]University College London, London, UK. [162]University of Zurich, Zurich, Switzerland. [163]UK AI Safety Institute, London, UK. [164]University of Padua, Padua, Italy. [165]Boston University, Boston, MA, USA. [166]Royal Veterinary College, London, UK. [167]Instituto Superior Técnico, Lisbon, Portugal. [168]SDAIA, Riyadh, Saudi Arabia. [169]The Ohio State University, Columbus, OH, USA. [170]University of Montreal, Montreal Quebec, Canada. [171]Cairo University Specialized Pediatric Hospital, Cairo, Egypt. [172]Universidad de Valencia, Valencia, Spain. [173]Monash University, Melbourne Victoria, Australia. [174]Van Andel Institute, Grand Rapids, MI, USA. [175]Larkin Community Hospital, South Miami, FL, USA. [176]The University of Texas at Dallas, Richardson, TX, USA. [177]Canadian University Dubai, Dubai, UAE. [178]The Hebrew University of Jerusalem, Jerusalem, Israel. [179]Università di Milano-Bicocca, Milan, Italy. [180]University of Massachusetts Lowell, Lowell, MA, USA. [181]Virginia Tech, Blacksburg, VA, USA. [182]University of Geneva, Geneva, Switzerland. [183]Google Research, Mountain View, CA, USA. [184]Cal Poly San Luis Obispo, San Luis Obispo, CA, USA. [185]Alexandru Ioan Cuza University, Iasi, Romania. [186]University of Mannheim, Mannheim, Germany. [187]Stockholm University, Stockholm, Sweden. [188]College of Eastern Idaho, Idaho Falls, ID, USA. [189]Intrinsic Innovation, Mountain View, CA, USA. [190]Ivy Natal, San Francisco, CA, USA. [191]King Saud University, Riyadh, Saudi Arabia. [192]SAMPE Switzerland, Zurich, Switzerland. [193]CERo Therapeutics Holdings, South San Francisco, CA, USA. [194]University of Tennessee, Knoxville, TN, USA. [195]Gray Swan AI, Pittsburgh, PA, USA. [196]EleutherAI, Washington, DC, USA. [197]Johns Hopkins University, Baltimore, MD, USA. [198]University of Montpellier, Montpellier, France. [199]Fraunhofer IMTE, Lübeck, Germany. [200]HomeEquity Bank, Toronto Ontario, Canada. [201]Materials Platform for Data Science, Tallinn, Estonia. [202]University of Pisa, Pisa, Italy. [203]Georgia State University, Atlanta, GA, USA. [204]Polytechnic University of the Philippines, Manila, The Philippines. [205]University of Oregon, Eugene, OR, USA. [206]Drexel University, Philadelphia, PA, USA. [207]University of Mumbai, Mumbai, India. [208]Gakushuin University, Tokyo, Japan. [209]University of Guelph, Guelph Ontario, Canada. [210]Intuit, Mountain View, CA, USA. [211]CTTC / CERCA, Castelldefels, Spain. [212]Dyno Therapeutics, Watertown, MA, USA. [213]The Hospital for Sick Children, Toronto Ontario, Canada. [214]Temple University, Philadelphia, PA, USA. [215]Saint Mary's University, Halifax Nova Scotia, Canada. [216]Cisco, San Jose, CA, USA. [217]Indian Institute of Technology (BHU), Varanasi, India. [218]AIM Intelligence, Seoul, South Korea. [219]Seoul National University, Seoul, South Korea. [220]The University of Texas at Arlington, Arlington, TX, USA. [221]The Hartree Centre, Daresbury, UK. [222]University of Vienna, Vienna, Austria. [223]POLITEHNICA Bucharest National University of Science and Technology, Bucharest, Romania. [224]Abacus.AI, San Francisco, CA, USA. [225]University of Galway, Galway, Ireland. [226]Eastern Institute of Technology (EIT), Napier, New Zealand. [227]ENS Lyon, Lyon, France. [228]Czech Technical University in Prague, Prague, Czech Republic. [229]University of Hamburg, Hamburg, Germany. [230]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany. [231]Universidad de Morón, Morón, Argentina. [232]Université Paris Cité, Paris, France. [233]Politecnico di Milano, Milan, Italy. [234]The New School, New York, NY, USA. [235]Max Planck Institute for Software Systems, Saarbrücken, Germany. [236]OpenAI, San Francisco, CA, USA. [237]Universidad de Granada, Granada, Spain. [238]Modulo Research, Cambridge, UK. [239]Heidelberg University, Heidelberg, Germany. [240]La Trobe University, Melbourne Victoria, Australia. [241]University of Yaoundé I, Yaoundé, Cameroon. [242]University of Innsbruck, Innsbruck, Austria. [243]Nabu Technologies, San Francisco, CA, USA. [244]Chalmers University of Technology, Gothenburg, Sweden. [245]Unidade Local de Saúde de Lisboa Ocidental, Lisbon, Portugal. [246]Children's Hospital of Orange County, Orange, CA, USA. [247]The Future Paralegals of America, New York, NY, USA. [248]Eastlake High School, Sammamish, WA, USA. [249]Center for Scientific Research and Higher Education at Ensenada (CICESE), Ensenada, Mexico. [250]University of Bradford, Bradford, UK. [251]Beni Suef University, Beni Suef, Egypt. [252]Bogazici University, Istanbul, Turkey. [253]Mansoura University, Mansoura, Egypt. [254]University of Bristol, Bristol, UK. [255]University of Oklahoma, Norman, OK, USA. [256]Jala University, Honolulu, HI, USA. [257]University of Arkansas, Fayetteville, AR, USA. [258]Florida Atlantic University, Boca Raton, FL, USA. [259]Bournemouth University, Bournemouth, UK. [260]University of Warwick, Coventry, UK. [261]University of Alabama Huntsville, Huntsville, AL, USA. [262]University of Hertfordshire, Hatfield, UK. [263]OncoPrecision, New York, NY, USA. [264]Central College, Pella, IA, USA. [265]Nottingham Trent University, Nottingham, UK. [266]Max Planck Institute for Intelligent Systems, Stuttgart, Germany. [267]University of Virginia, Charlottesville, VA, USA. [268]Dartmouth College, Hanover, NH, USA. [269]Cairo University, Giza, Egypt. [270]INESC Microsistemas e Nanotecnologias, Lisbon, Portugal. [271]James Madison University, Harrisonburg, VA, USA. [272]Instituto Gonçalo Moniz, Salvador, Brazil. [273]Rice University,

Houston, TX, USA. [274]HUN-REN, Budapest, Hungary. [275]Rutgers University, New Brunswick, NJ, USA. [276]AE Studio, Marina Del Rey, CA, USA. [277]Saarland University, Saarbrücken, Germany. [278]HUTECH, Ho Chi Minh City, Vietnam. [279]Pennsylvania College of Technology, Williamsport, PA, USA. [280]Intelligent Geometries, Front Royal, VA, USA. [281]École Polytechnique, Palaiseau, France. [282]CONICET, Buenos Aires, Argentina. [283]Universidad Tecnológica Nacional, Buenos Aires, Argentina. [284]John Crane UK, Slough, UK. [285]Alan Turing Institute, London, UK. [286]Pondicherry Engineering College, Puducherry, India. [287]Leibniz Institute for Science and Mathematics Education, Kiel, Germany. [288]Royal Holloway, University of London, Egham, UK. [289]Tanta University, Tanta, Egypt. [290]University of Malaya, Kuala Lumpur, Malaysia. [291]Hemwati Nandan Bahuguna Garhwal University, Srinagar, India. [292]University Mohammed I, Oujda, Morocco. [293]LGM, Paris, France. [294]Northern Illinois University, DeKalb, IL, USA. [295]Bethune-Cookman University, Daytona Beach, FL, USA. [296]University of California, Irvine, Irvine, CA, USA. [297]Central Mindanao University, Maramag, The Philippines. [298]University of the Fraser Valley, Abbotsford British Columbia, Canada. [299]Patched Codes, San Francisco, CA, USA. [300]Missouri University of Science and Technology, Rolla, MO, USA. [301]Quotient AI, Boston, MA, USA. [302]CSMSS Chh. Shahu College of Engineering, Aurangabad, India. [303]Genomia Diagnostics Research, New Delhi, India. [304]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [305]Forschungszentrum Jülich, Jülich, Germany. [306]Standard Intelligence, San Francisco, CA, USA. [307]RMIT University, Melbourne Victoria, Australia. [308]German Research Center for Artificial Intelligence, Kaiserslautern, Germany. [309]Fondazione Bruno Kessler, Trento, Italy. [310]University of Trento, Trento, Italy. [311]Chulalongkorn University, Bangkok, Thailand. [312]Aligarh Muslim University, Aligarh, India. [313]Happy Technologies LLC, Arlington, VA, USA. [314]Menoufia University, Shebin El Kom, Egypt. [315]Instituto Politécnico Nacional, Mexico City, Mexico. [316]University of Bologna, Bologna, Italy. [317]Manipal University Jaipur, Jaipur, India. [318]The University of Texas at Austin, Austin, TX, USA. [319]Murdoch University, Perth Western Australia, Australia. [320]University of Delaware, Newark, DE, USA. [321]Williams College, Williamstown, MA, USA. [322]Perimeter Institute for Theoretical Physics, Waterloo Ontario, Canada. [323]University of Maribor, Maribor, Slovenia. [324]Brigham and Women's Hospital, Boston, MA, USA. [325]The University of Tokyo, Tokyo, Japan. [326]Imperial College London, London, UK. [327]University of California Santa Cruz, Santa Cruz, CA, USA. [328]Vellore Institute of Technology, Vellore, India. [329]CHRU de Nancy, Nancy, France. [330]Delft University of Technology, Delft, The Netherlands. [331]Scripps Research, La Jolla, CA, USA. [332]Aleph Alpha, Heidelberg, Germany. [333]George Mason University, Fairfax, VA, USA. [334]Atilim University, Ankara, Turkey. [335]Leonardo Labs, Rome, Italy. [336]Complexity Science Hub, Vienna, Austria. [337]Universidad Nacional de Educación a Distancia, Madrid, Spain. [338]Saxion University, Enschede, The Netherlands. [339]Korea Advanced Institute of Science and Technology, Daejeon, South Korea. [340]Adobe Research, San Jose, CA, USA. [341]National Aerospace University 'Kharkiv Aviation Institute', Kharkiv, Ukraine. [342]Google, Mountain View, CA, USA. [343]Hexworks, Barcelona, Spain. [344]Westmead Hospital, Sydney New South Wales, Australia. [345]University of Bern, Bern, Switzerland. [346]Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Kaiserslautern, Germany. [347]SUMM AI, Munich, Germany. [348]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [349]Konkuk University, Seoul, South Korea. [350]University of Groningen, Groningen, The Netherlands. [351]Jagiellonian University, Kraków, Poland. [352]Minerva University, San Francisco, CA, USA. [353]Aalborg University, Aalborg, Denmark. [354]IBM Research, Givatayim, Israel. [355]Universitat Politecnica de Valencia, Valencia, Spain. [356]RBC Borealis, Toronto Ontario, Canada. [357]Mayo Clinic, Rochester, MN, USA. [358]University of Lausanne, Lausanne, Switzerland. [359]Dalhousie University, Halifax Nova Scotia, Canada. [360]Universitat de Lleida, Lleida, Spain. [361]Amazon, Seattle, WA, USA. [362]Dell Technologies, Round Rock, TX, USA. [363]University of Seoul, Seoul, South Korea. [364]University of Auckland, Auckland, New Zealand. [365]Morgridge Institute for Research, Madison, WI, USA. [366]Korea University of Technology and Education, Cheonan, South Korea. [367]University of California Davis, Davis, CA, USA. [368]Baylor College of Medicine, Houston, TX, USA. [369]Indraprastha Institute of Information Technology Delhi, New Delhi, India. [370]Two Minute Papers, Pécs, Hungary. [371]ADIA Lab, Abu Dhabi, UAE. [372]New Jersey Institute of Technology, Newark, NJ, USA. [373]Novo Nordisk, Bagsværd, Denmark. [374]Purdue University, West Lafayette, IN, USA. [375]Gakugei Shuppan-sha, Kyoto, Japan. [376]Universiteit Utrecht, Utrecht, The Netherlands. [377]T-Systems Iberia, Madrid, Spain. [378]University of Klagenfurt, Klagenfurt, Austria. [379]Max Planck Institute for Security and Privacy, Bochum, Germany. [380]InxiteOut, Bangalore, India. [381]Goethe Universität Frankfurt, Frankfurt am Main, Germany. [382]Universidad del Valle, Cali, Colombia. [383]Bilkent University, Ankara, Turkey. [384]Trinity School, New York, NY, USA. [385]Universitat Pompeu Fabra, Barcelona, Spain. [386]Brighton Law School, Brighton, UK. [387]University of Melbourne, Melbourne, Australia. [388]Ankara University, Ankara, Turkey. [389]Dr. Siyami Ersek Thoracic, Cardiovascular and Vascular Surgery Training and Research Hospital, Istanbul, Turkey. [390]AIT Austrian Institute of Technology, Vienna, Austria. [391]Technical University of Munich, Munich, Germany. [392]Providence College, Providence, RI, USA. [393]University of Jyväskylä, Jyväskylä, Finland. [394]Weizmann Institute of Science, Rehovot, Israel. [395]Indiana University, Bloomington, IN, USA. [396]Nanyang Technological University, Singapore, Singapore. [397]University of Sheffield, Sheffield, UK.

Methods

Related works

LLM benchmarks. Benchmarks are important tools for tracking the rapid advancement of LLM capabilities, including general and scientific knowledge[1,10,12–15] and mathematical reasoning[16–21], code generation[22–28] and general-purpose human assistance[7,29–35]. Owing to their objectivity and ease of automated scoring at scale, evaluations commonly include multiple-choice and short-answer questions[31,36–39], with benchmarks such as MMLU[1] also spanning a broad range of academic disciplines and levels of complexity.

Saturation and frontier benchmark design. However, state-of-the-art models now achieve nearly perfect scores on many existing evaluations, obscuring the full extent of current and future frontier AI capabilities[40–43]. This has motivated the development of more challenging benchmarks that test for multi-modal capabilities[17,22,24,44–50], strengthen existing benchmarks[32,44,45,51,52], filter questions over multiple stages of review[9,12,19,42,53,54] and use experts to write tests for advanced academic knowledge[9,12,19,54–56]. HLE combines these approaches: the questions are developed by subject-matter experts and undergo multiple rounds of review, while preserving the broad subject-matter coverage of MMLU. As a result, HLE provides a clear measurement of the gap between current AI capabilities and human expertise on closed-ended academic tasks, complementing other assessments of advanced capabilities in open-ended domains[57,58].

Dataset

Submission process. To ensure question difficulty, we automatically check the accuracy of frontier LLMs on each question before submission. Our testing process uses multi-modal LLMs for text-and-image questions (GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet and o1) and adds two non-multi-modal models (o1-mini and o1-preview) for text-only questions. We use different submission criteria by question type: exact-match questions must stump all models, whereas multiple-choice questions must stump all but one model to account for potential lucky guesses. Users are instructed to submit only questions that meet these criteria. We note that due to non-determinism in models and a non-zero floor in multiple-choice questions, further evaluation on the dataset exhibits some low but non-zero accuracy.

Post-release. Late contributions. In response to research community interest, we opened the platform for late contributors after the initial release, resulting in thousands of submissions. Each submission was manually reviewed by organizers. The new questions are of similar difficulty and quality to our initial dataset, resulting in a second held-out private set, which will be used in future evaluations.

Refinement. Community feedback: owing to the advanced, specialized nature of many submissions, reviewers were not expected to verify the full accuracy of each provided solution rationale, instead focusing on whether the question aligns with guidelines. Given this limitation in the review process, we launched a community feedback bug bounty program following the initial release of the dataset to identify and eliminate the main errors in the dataset, namely, label errors and other errors in the statement of the question. Each error report was manually verified by the organizers with feedback from the original author of the question when appropriate.

Searchable questions: a question is potentially searchable if a model with search tools answered correctly, but answered incorrectly without search. Each of these potentially searchable questions was then manually audited, removing any that were easily found using web search. We used GPT-4o mini/GPT-4o search and Perplexity Sonar models in this procedure. We observe that current frontier model performance on HLE after applying this procedure is similar to the performance on HLE before applying this procedure.

Expert disagreement rate. Before release, we conducted two main rounds of auditing, each on a sample of 200 questions. We recruited students from top universities in the United States to fully solve a sample of questions from HLE. Errors flagged were routed between organizers, original question authors and auditors until consensus was reached. We used data from these audits to further refine our dataset. The first round aimed to identify common categories of imprecise questions, such as open-ended formats, reliance on rounded numerical values or submissions from authors with low acceptance rates. Based on these signals, we manually removed or revised potential questions with similar issues before conducting a second audit on a new sample of 200 questions. This iterative process yielded a final estimated expert disagreement rate of 15.4% for the public set. This level of expert disagreement is in line with what is observed in other well-known machine learning benchmarks[59–62].

Disagreement rates are often higher in domains such as health and medicine. A targeted peer review on a biology, chemistry and health subset, proposed in ref. 63, found an expert disagreement rate of approximately 18%. This is also observed in other similarly expert-grade work; for example[64], notes that disagreement among expert physicians is frequent on complex health topics. To aid future community efforts in identifying other potential dataset errors, we outline several key factors that contribute to the complexity of these audits below:

- The need for multiple experts: our multi-reviewer process highlighted the complexity of these questions. In several cases, a reviewer identified an important piece of information, such as a decades-old paper or a foundational concept not immediately apparent to others, that was essential to confirming the validity of an answer. To illustrate, if we were to adopt a single-reviewer methodology in which a question is flagged based on just one dissenting expert, the disagreement rate on the aforementioned health-focused subset jumps from 18% to 25%, which is close to the approximate numbers and method from ref. 63. This discrepancy highlights the importance of a standard peer-review process, complete with multiple reviewers and author rebuttal, for HLE questions.
- Questions from research experience: HLE is intentionally designed to include questions based on insights from the direct, hands-on experiments of its contributors. This design captures knowledge gained from direct research experiences, which is often difficult to verify through standard literature searches or by external reviewers. This was done to test model knowledge beyond what is readily indexed on the internet.
- Understanding question design: designing challenging closed-ended research questions is difficult. Consequently, the objective for some HLE multiple-choice questions is to identify the most plausible answer among the provided options. Some external reviewers, unfamiliar with these design principles, sought to find external sources to support an open-ended answer rather than evaluating the best choice among the given options.

HLE-Rolling. Inspired by these valuable community discussions and researcher interest across disciplines in contributing to the dataset, and as part of our commitment to continual improvement, we will introduce a dynamic fork of the dataset post-release: HLE-Rolling. This version will be regularly updated to address community feedback and integrate new questions. Information about the updates will be made publicly available at https://lastexam.ai. Our goal is to provide a seamless migration path for researchers once frontier models begin to hit the noise ceiling performance on the original HLE dataset.

Prompts. We use the following system prompt for evaluating LLMs on HLE questions. For models that do not support a system prompt, we add it as a separate user prompt.

Your response should be in the following format:
Explanation: {your explanation for your answer choice}
Answer: {your chosen answer}
Confidence: {your confidence score between 00% and 100% for your answer}

We use the following system prompt to judge the model answers against the correct answers for our evaluations in Table 1. We used o3-mini-2025-01-31 with structured decoding enabled to get an extracted_final_answer, reasoning, correct, confidence extraction for each output. An example of a structured response using an LLM judge is shown in Extended Data Fig. 1.

Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below.
[question]: {question}
[response]: {response}
Your judgement must be in the format and criteria specified below:
extracted_final_answer: The final exact answer extracted from the [response]. Put the extracted answer as 'None' if there is no exact, final answer to extract from the response.
[correct_answer]: {correct_answer}
reasoning: Explain why the extracted_final_answer is correct or incorrect based on [correct_answer], focusing only on if there are meaningful differences between [correct_answer] and the extracted_final_answer. Do not comment on any background to the problem, do not attempt to solve the problem, do not argue for any answer different than [correct_answer], focus only on whether the answers match.
correct: Answer 'yes' if extracted_final_answer matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer 'no' otherwise, i.e. if there if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect.
confidence: The extracted confidence score between 0|%| and 100|%| from [response]. Put 100 if there is no confidence score available.

Data availability

The HLE dataset is open-source and available at https://huggingface.co/datasets/cais/hle. Important updates to the project and dataset will be announced at https://lastexam.ai.

Code availability

The inference script for benchmarking AI systems on HLE is available at GitHub (https://github.com/centerforaisafety/hle).

12. Li, N. et al. The WMDP benchmark: measuring and reducing malicious use with unlearning. In *Proc. 41st International Conference on Machine Learning (ICML)*, 28713–28738 (PMLR, 2024).
13. Laurent, J. M. et al. LAB-bench: measuring capabilities of language models for biology research. Preprint at https://arxiv.org/abs/2407.10362 (2024).
14. Srivastava, A. et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=uyTL5Bvosj (2023).
15. Zhong, W. et al. Agieval: a human-centric benchmark for evaluating foundation models. Preprint at https://arxiv.org/abs/2304.06364 (2023).
16. Hendrycks, D. et al. Measuring mathematical problem solving with the MATH dataset. In Proc. *35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track* https://openreview.net/forum?id=7Bywt2mQsCe (NeurIPS, 2021).
17. Lu, P. et al. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=KUNzEQMWU7 (ICLR, 2024).
18. Cobbe, K. et al. Training verifiers to solve math word problems. Preprint at https://arxiv.org/abs/2110.14168 (2021).
19. Glazer, E. et al. FrontierMath: a benchmark for evaluating advanced mathematical reasoning in AI. Preprint at https://arxiv.org/abs/2411.04872 (2024).
20. He, C. et al. OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 3828–3850 (ACL, 2024).
21. Gao, B. et al. Omni-MATH: A universal Olympiad level mathematic benchmark for large language models. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=yaqPf0KAlN (ICLR, 2025).
22. Chan, J. S. et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=6s5uXNWGIh (ICLR, 2025).
23. Zhang, A. K. et al. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=tc90LV0yRL (ICLR, 2025).
24. Jimenez, C. E. et al. SWE-bench: Can language models resolve real-world Github issues? In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=VTF8yNQM66 (ICLR, 2024).
25. Chen, M. et al. Evaluating large language models trained on code. Preprint at https://arxiv.org/abs/2107.03374 (2021).
26. Hendrycks, D. et al. Measuring coding challenge competence with APPS. In *Proc. 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track* https://openreview.net/forum?id=sD93GOzH3i5 (NeurIPS, 2021).
27. Bhatt, M. et al. Purple Llama CyberSecEval: a secure coding benchmark for language models. Preprint at https://arxiv.org/abs/2312.04724 (2023).
28. Austin, J. et al. Program synthesis with large language models. Preprint at https://arxiv.org/abs/2108.07732 (2021).
29. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint at https://arxiv.org/abs/2204.05862 (2022).
30. Perez, E. et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434 (ACL, 2023).
31. Rajpurkar, P. et al. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2383–2392 (EMNLP, 2016).
32. Rajpurkar, P. et al. Know what you don't know: Unanswerable questions for SQuAD. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 784–789 (ACL, 2018).
33. Bajaj, P. et al. MS MACRO: a human generated machine reading comprehension dataset. Preprint at https://arxiv.org/abs/1611.09268 (2018).
34. Hendrycks, D. et al. What would Jiminy Cricket do? Towards agents that behave morally. In *Proc. 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track* https://openreview.net/forum?id=G1muTb5zuO7 (NeurIPS, 2021).
35. Phan, L., Mazeika, M., Zou, A. & Hendrycks, D. Textquests: how good are LLMs at text-based video games? Preprint at https://arxiv.org/abs/2507.23701 (2025).
36. Wang, A. et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=rJ4km2R5t7 (ICLR, 2019).
37. Wang, A. et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 32, 3261–3275 (NeurIPS, 2019).
38. Yang, Z. et al. HotpotQA: A dataset for diverse, explainable multihop question answering. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2369–2380 (EMNLP, 2018).
39. Dua, D. et al. DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs. Preprint at https://arxiv.org/abs/1903.00161 (2019).
40. Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J. & Samwald, M. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nat. Commun.* **13**, 6793 (2022).
41. Owen, D. How predictable is language model benchmark performance? Preprint at https://arxiv.org/abs/2401.04757 (2024).
42. Kiela, D. et al. Dynabench: Rethinking benchmarking in NLP. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4110–4124 (NAACL, 2021).
43. McIntosh, T. R. et al. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Trans. Artif. Intell.* https://doi.org/10.1109/TAI.2025.3569516 (2025).
44. Wang, Y. et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, article no. 3018 (NeurIPS, 2024).
45. Taghanaki, S. A., Khani, A. & Khasahmadi, A. MMLU-Pro+: evaluating higher-order reasoning and shortcut learning in LLMS. Preprint at https://arxiv.org/abs/2409.02257 (2024).
46. Yao, S. et al. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=roNSXZpUDN (ICLR, 2025).
47. Andriushchenko, M. et al. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *Proc. International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=AC5n7xHuR1 (ICLR, 2025).
48. Kumar, P. et al. Refusal-trained LLMS are easily jailbroken as browser agents. Preprint at https://arxiv.org/abs/2410.13886 (2024).
49. Yan, F. et al. *Berkeley Function Calling Leaderboard* https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html (2024).
50. Srinivasan, V. K. et al. NexusRaven: a commercially-permissive language model for function calling. In *NeurIPS 2023 Foundation Models for Decision Making Workshop* (NeurIPS, 2023).
51. Hosseini, A., Sordoni, A., Toyama, D., Courville, A. & Agarwal, R. Not all LLM reasoners are created equal. Preprint at https://arxiv.org/abs/2410.01748 (2024).
52. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

53. Nie, Y. et al. Adversarial NLI: A new benchmark for natural language understanding. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4885–4901 (ACL, 2020).
54. Götting, J. et al. Virology capabilities test (VCT): a multimodal virology Q&A benchmark. Preprint at https://arxiv.org/abs/2504.16137 (2025).
55. Phuong, M. et al. Evaluating frontier models for dangerous capabilities. Preprint at https://arxiv.org/abs/2403.13793 (2024).
56. *Anthropic's Responsible Scaling Policy Updates* https://www.anthropic.com/rsp-updates (Anthropic, 2024).
57. Mazeika, M. et al. Remote labor index: measuring AI automation of remote work. Preprint at https://arxiv.org/abs/2510.26787 (2025).
58. Patwardhan, T. et al. GDPval: evaluating ai model performance on real-world economically valuable tasks. Preprint at https://arxiv.org/abs/2510.04374 (2025).
59. Kwiatkowski, T. et al. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **7**, 452–466 (2019).
60. Antol, S. et al. VQA: Visual question answering. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2425–2433 (IEEE, 2015).
61. Reddy, S., Chen, D. & Manning, C. D. CoQA: a conversational question answering challenge. *Trans. Assoc. Comput. Linguist.* **7**, 249–266 (2019).
62. Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. A large annotated corpus for learning natural language inference. In *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* (eds Màrquez, L. et al.), 632–642 (ACL, 2015).
63. Skarlinski, M., Laurent, J., Bou, A. & White, A. About 30% of Humanity's Last Exam chemistry/biology answers are likely wrong. *FutureHouse* https://www.futurehouse.org/research-announcements/hle-exam (2025).
64. Arora, R. K. et al. HealthBench: evaluating large language models towards improved human health. Preprint at https://arxiv.org/abs/2505.08775 (2025).

Acknowledgements The research is supported by Center for AI Safety and Scale AI.

Author contributions All authors have contributed to the dataset creation process. The Center for AI Safety and Scale AI consortia jointly designed the dataset premise and pipeline; operated the data collection platform (https://lastexam.ai); and provided funding, inference infrastructure for LLMs and review/auditing resources. The authors in the HLE Contributors Consortium contributed to the dataset in various ways, including submitting at least one accepted question to one of the dataset versions, contributing to dataset refinement or assisting with evaluations. In the Center for AI Safety and Scale AI, Long Phan, Alice Gatti, Ziwen Han and Nathaniel Li led the project, and Summer Yue, Alexandr Wang and Dan Hendrycks provided senior supervision.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09962-4.

Correspondence and requests for materials should be addressed to Long Phan or Dan Hendrycks.

Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Structured Judge Response Example

[correct_answer]: $\frac{\cos\left(\frac{\pi}{n}\right)}{2\left(1+\cos\left(\frac{\pi}{n}\right)\right)}$

[extracted_model_answer]: $\frac{\cot\left(\frac{\pi}{n}\right)}{2\cot\left(\frac{\pi}{2n}\right)}$

[reasoning]: Using the half-angle identity $\tan\left(\frac{\pi}{2n}\right) = \frac{\sin\left(\frac{\pi}{n}\right)}{1+\cos\left(\frac{\pi}{n}\right)}$ and recalling that $\cot x = \frac{\cos x}{\sin x}$, we express $\cot\left(\frac{\pi}{n}\right) = \frac{\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}$ and $\cot\left(\frac{\pi}{2n}\right) = \frac{1+\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}$. Forming the quotient $\frac{\cot\left(\frac{\pi}{n}\right)}{2\cot\left(\frac{\pi}{2n}\right)} = \frac{\frac{\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}}{2\cdot\frac{1+\cos\left(\frac{\pi}{n}\right)}{\sin\left(\frac{\pi}{n}\right)}} = \frac{\cos\left(\frac{\pi}{n}\right)}{2(1+\cos\left(\frac{\pi}{n}\right))}$, where the common $\sin\left(\frac{\pi}{n}\right)$ cancels out. This detailed inline simplification shows that the model answer is equivalent to the correct answer.

[correct]: yes

Extended Data Fig. 1 | Example of a structured response using an LLM judge. Exact-match answers in HLE sometimes require several reasoning steps to compare the AI's final answer with the correct answer; therefore, a capable LLM judge with reasoning capabilities is necessary.

Extended Data Table 1 | Accuracy and RMS Calibration error of frontier LLMs on the text-only questions of HLE

Model	Accuracy (%) ↑	Calibration Error (%) ↓
GPT-4o	2.3 ± 0.6	88
Claude 3.5 Sonnet	4.3 ± 0.9	83
Gemini 1.5 Pro	4.6 ± 0.9	87
Gemini 2.0 Flash Thinking	6.6 ± 1.0	82
o1	7.8 ± 1.1	84
DeepSeek-R1	8.5 ± 1.2	73
o3-mini	13.4 ± 1.4	80
Post-Release Models		
Llama 4 Maverick	5.3 ± 1.0	84
Claude 4 Sonnet	7.6 ± 1.1	76
Gemini 2.5 Flash	12.6 ± 1.4	81
Claude 4 Opus	10.8 ± 1.3	73
o4-mini	18.9 ± 1.7	58
o3	20.6 ± 1.7	36
Gemini 2.5 Pro	22.1 ± 1.8	72
GPT-5	26.3 ± 1.9	50

Extended Data Table 2 | Category-wise breakdown of frontier LLMs performance on HLE

Model	Math	Bio/Med	Physics	CS/AI	Humanities	Chemistry	Engineering	Other
	Text-Only							
GPT-4o	2.3	5.0	1.5	0.9	2.6	2.0	1.6	2.3
Grok 2	3.2	5.4	4.5	3.6	1.0	1.0	4.8	1.1
Claude 3.5 Sonnet	3.8	5.9	4.5	2.2	6.7	5.0	9.7	2.9
Gemini 1.5 Pro	5.3	5.4	2.0	4.0	3.6	6.0	3.2	3.4
Gemini 2.0 Flash Thinking	8.1	7.7	4.5	4.9	6.2	5.0	4.8	2.9
o1	7.4	8.1	6.9	8.4	8.8	10.0	4.8	8.0
DeepSeek-R1	9.1	9.0	5.4	7.5	10.4	5.0	14.5	7.4
o3-mini	18.6	10.0	15.3	8.4	5.2	9.0	6.5	6.9
Post-Release Models								
Llama 4 Maverick	5.1	5.9	5.9	4.9	6.2	8.9	6.2	2.3
Claude 4 Sonnet	8.1	7.2	5.9	7.1	10.9	5.0	14.1	3.4
Claude 4 Opus	10.8	15.3	8.9	11.2	12.4	5.9	10.9	8.0
Gemini 2.5 Flash	14.5	13.1	13.9	8.9	11.4	3.0	10.9	9.1
o4-mini	24.5	14.0	19.8	19.6	9.8	6.9	10.9	10.2
o3	23.6	17.1	19.3	19.6	17.6	16.8	17.2	16.5
Gemini 2.5 Pro	26.2	17.1	20.8	17.0	25.4	22.8	15.6	11.4
	Full Dataset							
GPT-4o	2.3	6.4	1.7	0.8	3.2	3.6	1.8	2.6
Grok 2	3.0	4.6	3.9	3.3	1.4	2.4	3.6	1.7
Claude 3.5 Sonnet	4.0	4.6	3.9	2.5	5.9	4.2	7.2	2.2
Gemini 1.5 Pro	5.2	5.4	3.0	3.7	4.1	6.1	3.6	3.4
Gemini 2.0 Flash Thinking	8.0	8.2	4.8	4.5	6.4	5.5	6.3	3.0
o1	7.4	10.4	7.0	8.2	8.7	9.7	6.3	7.3
Post-Release Models								
Llama 4 Maverick	5.1	6.1	5.7	5.0	7.3	10.9	6.3	3.0
Claude 4 Sonnet	8.3	8.2	6.1	6.6	11.0	6.7	10.8	3.9
Claude 4 Opus	10.5	15.4	10.0	10.4	12.8	7.3	9.0	8.6
Gemini 2.5 Flash	14.3	12.1	13.0	9.1	10.5	6.7	11.7	8.2
o4-mini	24.1	15.4	18.7	19.5	9.1	8.5	11.7	9.9
o3	23.4	18.9	18.7	20.7	17.8	16.4	17.1	15.9
Gemini 2.5 Pro	25.8	18.6	20.4	17.0	23.7	23.6	18.0	11.6

Extended Data Table 3 | Accuracy across multi-modal only, exact answer, and multiple-choice splits of HLE

Model	Multi-Modal Only	Exact Match Only	Multiple-Choice Only
GPT-4o	5.3	1.8	5.6
Grok 2	2.3	2.2	5.8
Claude 3.5 Sonnet	2.6	3.1	6.9
Gemini 1.5 Pro	5.0	3.8	7.1
Gemini 2.0 Flash Thinking	6.7	5.2	10.8
o1	9.4	6.7	12.0
DeepSeek-R1*	-	6.9	13.8
o3-mini*	-	12.9	14.6
Post-Release Models			
Llama 4 Maverick	7.9	4.2	10.5
Claude 4 Sonnet	8.8	6.1	13.0
Gemini 2.5 Flash	9.1	10.1	17.8
Claude 4 Opus	10.2	8.4	18.1
o4-mini	12.9	17.5	19.5
o3	19.0	18.9	24.7
Gemini 2.5 Pro	19.0	19.6	28.1

*Text-only models.