
Humanity’s Last Exam

Organizing Team

Long Phan^{*1}, Alice Gatti^{*1}, Ziwen Han^{*2}, Nathaniel Li^{*1},

Josephina Hu², Hugh Zhang[†], Sean Shi², Michael Choi², Anish Agrawal², Arnab Chharia², Adam Khoja¹, Ryan Kim[†], Richard Ren¹, Jason Hausenloy¹, Oliver Zhang¹, Mantas Mazeika¹,

Summer Yue^{**2}, Alexandr Wang^{**2}, Dan Hendrycks^{**1}

¹ Center for AI Safety, ² Scale AI

Dataset Contributors

Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney, Gashaw M. Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John Wydallis, Mark Nandor, Ankit Singh, Tim Gehringer, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jungbae Nam, Jennifer Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Lianghai Li, Sumeet Motwani, Christian Schroeder de Witt, Edwin Taylor, Johannes Veith, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav Kuchkin, Ng Ze-An, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy, Dakota Martinez, Ben Pageler, Nick Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Subrata, Ariel Ghislain Kemogne Kamdoun, Tobias Kreiman, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Tim Tarver, Haline Heiding, Rafael Sayous, Stefan Ivanov, Joseph M Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Ali Dehghan, Andres Algaba, Brecht Verbeken, David Noever, Ragavendran P V, Lisa Schut, Iliia Sucholutsky, Evgenii Zheltonozhskii, Derek Lim, Richard Stanley, Shankar Sivarajan, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Jennifer Sandlin, Anmol Sahu, Yuzheng Hu, Sara Fish, Nasser Heydari, Archimedes Apronti, Kaivalya Rawal, Tobias Garcia Vilchis, Yuxuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Alan Goldfarb, Sergey Ivanov, Rafat Poświata, Chenguang Wang, Daofeng Li, Donato Crisostomi, Andrea Achilleos, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Elliott Thornley, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Keith Schneider, Zakayo Kazibwe, Nate Stambaugh, Mukhwinder Singh, Ilias Magoulas, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Veit Elser, Kanu Priya Agarwal, Victor Efrén Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, Václav Rozhoň, Ziqiao Ma, Christian Stump, Mohammed Berkani, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccard, Ferenc Jeanplong, Niv Cohen, Varun Gangal, Josef Tkadlec, Paul Rosu, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stadel, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, John Arnold Ambay, Abhishek Shukla, Yan Carlos Leyva Labrador, Alan Givré, Hew Wolff, Vivien Rossbach, Muhammad Fayeaz Aziz, Younesse Kaddar, Yanxu Chen, Robin Zhang, Jiayi Pan, Antonio Terpin, Niklas Muennighoff, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Adam Jones, Jainam Shah, Ethan D. L. Brown, Keliang Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor Kretov, Kaustubh Sridhar, Zienab EL-Wasif, Anji Zhang, Daniel Pyda, Joanna Tam, David M.

^{*}Co-first Authors. ^{**} Senior Authors. [†] Work conducted while at Center for AI Safety. [‡] Work conducted while at Scale AI. Complete list of author affiliations in Appendix A. Correspondence to agibenchmark@safe.ai.

Cunningham, Demosthenes Patramanis, Michael Krause, Andrew Redenti, Daniel Bugas, David Aldous, Jesyin Lai, Shannon Coleman, Mohsen Bahaloo, Jiangnan Xu, Sangwon Lee, Sandy Zhao, Ning Tang, Michael K. Cohen, Micah Carroll, Orr Paradise, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Benedito Alves de Oliveira Junior, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Szyber-Betley, Priti Shukla, Jonathan Crozier, Antonella Pinto, Shreyas Verma, Prashant Joshi, Zheng-Xin Yong, Allison Tee, Jérémy Andréoletti, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Khánh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Jonathan Roberts, William Alley, Kunyang Sun, Ryan Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Sreenivas Goud Raparathi, Pablo Hernández-Cámara, Freddie Martin, Dmitry Malishev, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ziyi Chen, Biró Bálint, M Saiful Bari, Peyman Kassani, Zihao Wang, Behzad Ansarinejad, Laxman Prasad Goswami, Yewen Sun, Hossam Elgnainy, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro José Moyano, Rajat Maheshwari, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Javier Gimenez, Innocent Enyekwe, Andrew Favre D.O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Khalida Meer, Harrison K Wang, Evan Chen, Alessandro Tomasiello, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mündler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, Tej Shah, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Carter Harris, Jason Gross, Ilya Gusev, Asankhaya Sharma, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayez, Sergei Bogdanov, Alexander Piperski, Marc Carauleanu, David K. Zhang, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Wiktor Morak, Wenjie Ma, William Held, Tran Duc Huy, Ruicheng Xian, Arnel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Emilien Duc, Bitu Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, Mátyás Vincze, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Jiang Muzhen, Fredrik Ekström, Angela Hammon, Oam Patel, Nicolas Remy, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Mike Battaglia, Mohammad Maghsoudimehrabani, Hieu Hoang, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Stephen Mensah, Nathan Andre, Anton Peristyy, Chris Harjadi, Himanshu Gupta, Stephen Malina, Samuel Albanie, Will Cai, Mustafa Mehkary, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Shreen Gul, Gunjan Chhablani, Zhehang Du, Adrian Cosma, Colin White, Robin Riblet, Prajvi Saxena, Jacob Votava, Vladimir Vinnikov, Shiv Halasyamani, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Renas Bacho, Vincent Ginis, Aleksandr Maksapetyan, Florencia de la Rosa, Xiuyu Li, Guillaume Malod, Leon Lang, Julien Laurendeau, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Yiğit Yalın, Gbenga Daniel Obikoya, Luca Arnaboldi, Rai (Michael Pokorny), Filippo Bigi, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C.H. Lux, Ben Rank, Colin Ni, Alesia Yakimchyk, Huanxu (Quinn) Liu, Olle Häggström, Emil Verkama, Himanshu Narayan, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Łucki, Wen-Ding Li, Justin Xu, Kevin Joseph Scaria, Freddie Vargus, Farzad Habibi, Long (Tony) Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Declan Grabb, Ida Bosio, Tony Fruhauff, Ido Akov, Eve J. Y. Lo, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Yibo Jiang, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Muhammad Rehan Siddiqi, Alon Ragoler, Justin Tan, Deepakkumar Patil, Rebeka Plecnik, Aaron Kirtland, Roselynn Grace Montecillo, Stephane Durand, Omer Faruk Bodur, Zahra Adoul, Mohamed Zekry, Guillaume Douville, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Hodjat Marji, Rasoul Pouriamanesh, Wentao Wu, Gözdenur Demir, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Hsiaoyn Milliron, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Abdelkader Dendane, Ricardo Lorena, Krishnamurthy Iyer, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Russell Campbell, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámín Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I.M.J. McInnis, Alex Hoover, Sören Möller, Tejal Patwardhan

Co-author list in progress. HUMANITY'S LAST EXAM is still accepting new questions. New questions can be submitted at lastexam.ai/submit for co-authorship in this section, but are not eligible for the prize pool.

Abstract

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 over 90% accuracy on popular benchmarks like MMLU, limiting informed measurement of state-of-the-art LLM capabilities. In response, we introduce HUMANITY’S LAST EXAM (HLE), a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage. HLE consists of 3,000 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 via internet retrieval. State-of-the-art LLMs demonstrate low accuracy and calibration on HLE, highlighting a significant 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>.

1 Introduction

The capabilities of large language models (LLMs) have progressed dramatically, exceeding human performance across a diverse array of tasks. To systematically measure these capabilities, LLMs are evaluated upon *benchmarks*: collections of questions which assess model performance on tasks such as math, programming, or biology. However, state-of-the-art LLMs [3, 14, 16, 34, 37, 49, 56] now achieve over 90% accuracy on popular benchmarks such as MMLU [21], which were once challenging frontiers for LLMs. The saturation of existing benchmarks, as shown in Figure 1, limits our ability to precisely measure 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 HUMANITY’S LAST EXAM (HLE), a benchmark of 3,000 extremely challenging questions from dozens of subject areas, designed to be the final closed-ended benchmark of broad academic capabilities. HLE is developed by academics and domain experts, providing a precise measure of capabilities as LLMs continue to improve (Section 3.1). 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. Amongst 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 employ a multi-stage review process to thoroughly ensure question difficulty and quality (Section 3.2). 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 then proceed through a two-stage reviewing process: (1) an initial feedback round with multiple graduate-level reviewers and (2) organizer and expert reviewer approval, ensuring quality and adherence to our submission criteria. Following release, we plan to further conduct a public review period, welcoming community feedback to correct any points of concern in the dataset.

Frontier LLMs consistently demonstrate low accuracy (less than 10%) across all models, highlighting a significant gap between current capabilities and expert-level academic performance (Section 4). Models also show provide incorrect answers with high confidence rather than acknowledging uncertainty on these challenging questions, with RMS calibration errors above 80% across all models.

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 3,000 questions from HLE to enable this precise measurement, while maintaining a private test set to assess potential model overfitting.

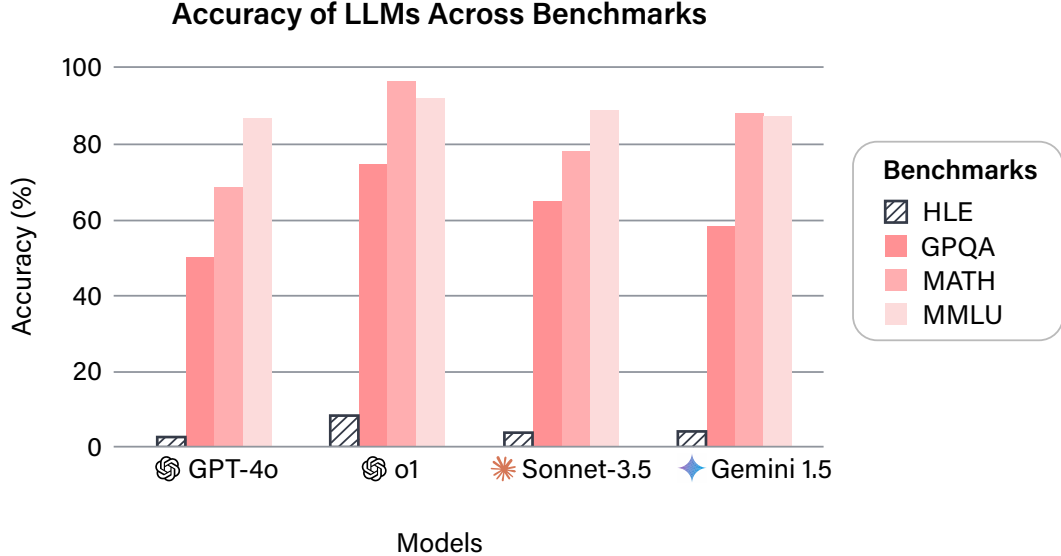


Figure 1: Compared against the saturation of some existing benchmarks, HUMANITY’S LAST EXAM accuracy remains low across several frontier models, demonstrating its effectiveness for measuring advanced, closed-ended, academic capabilities. The sources for our evaluation metrics are detailed in Appendix C.5. We further evaluate more frontier models on HLE in Table 1.

2 Related Work

LLM Benchmarks. Benchmarks are important tools for tracking the rapid advancement of LLM capabilities, including scientific [10, 12, 21, 29, 30, 44, 47, 53, 61] and mathematical reasoning [13, 17–19, 22, 31, 45, 50], code generation [6, 9–11, 20, 26, 60], and general-purpose human assistance [1, 7, 8, 25, 40, 42, 43, 47, 54]. Due to their objectivity and ease of automated scoring at scale, evaluations commonly include multiple-choice and short-answer questions [15, 42, 51, 52, 58], with benchmarks such as MMLU [21] 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 [3, 14, 16, 34, 37, 49, 56], obscuring the full extent of current and future frontier AI capabilities [27, 32, 38, 39]. This has motivated the development of more challenging benchmarks which test for multi-modal capabilities [2, 10, 26, 28, 31, 46, 48, 53, 57, 59], strengthen existing benchmarks [24, 43, 45, 48, 53], filter questions over multiple stages of review [18, 27, 30, 33, 44], and employ experts to write tests for advanced academic knowledge [5, 18, 30, 34, 41, 44]. 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 [10, 35, 36, 55].

3 Dataset

HUMANITY’S LAST EXAM (HLE) consists of 3,000 challenging questions across over a hundred subjects across. A high level summary is provided in Figure 3. We publicly release these questions, while maintaining a private test set of held out questions to assess model overfitting.

3.1 Collection

HLE is a global collaborative effort, with questions from nearly 1000 subject expert contributors affiliated with over 500 institutions across 50 countries – comprised mostly of professors, researchers, and graduate degree holders.

Classics

Question:

Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script.

A transliteration of the text is provided: RGYN^o BT HRY BR ^oT^o HBL

✎ Henry 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 $F, G : C \rightarrow D$ 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:

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

How many natural cotransformations are there between F and G ?

✎ Emily S
 University of São Paulo

Computer Science

Question:

Let G be a graph. An edge-indicator of G is a function $a : 0, 1 \rightarrow V(G)$ 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:

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 [m]π-con or [m]π-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.

וַיִּשְׁלַח יְהוָה מִן־הַשָּׁמַיִם אֵשׁ וַיִּבְרָן יְהוָה מִן־הַשָּׁמַיִם יִסְדּוּן (Psalms 104:7) ?

✎ Lina B
 University of Cambridge

Figure 2: Samples of the diverse and challenging questions submitted to HUMANITY’S LAST EXAM.

5

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 10% of questions requiring comprehending both text and an image reference. 80% of questions are exact-match with the remainder being multiple-choice.

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 contributor name 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, though contributions from unpublished research are acceptable. Questions typically require graduate-level expertise or test knowledge of highly specific topics (e.g., precise historical details, trivia, 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.

Prize Pool. To attract high-quality submissions, we establish a \$500,000 USD prize pool, with prizes of \$5,000 USD for each of the top 50 questions and \$500 USD 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 significant technical experience in their fields.

3.2 Review

LLM Difficulty Check To ensure question difficulty, each question is first validated against several frontier LLMs prior to submission (Appendix B.1). 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 over 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 (eg. Master’s, PhD, JD, etc.) in their fields. Reviewers select submissions in their domain, grading them against standardized rubrics

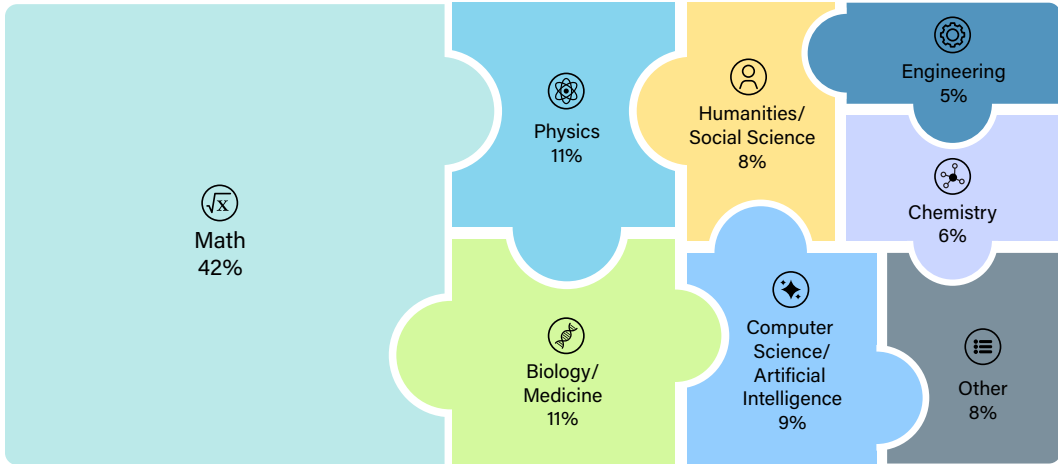


Figure 3: HLE consists of 3,000 exam questions in over a hundred subjects, grouped into high level categories here. We provide a more detailed list of subjects in Appendix B.3.

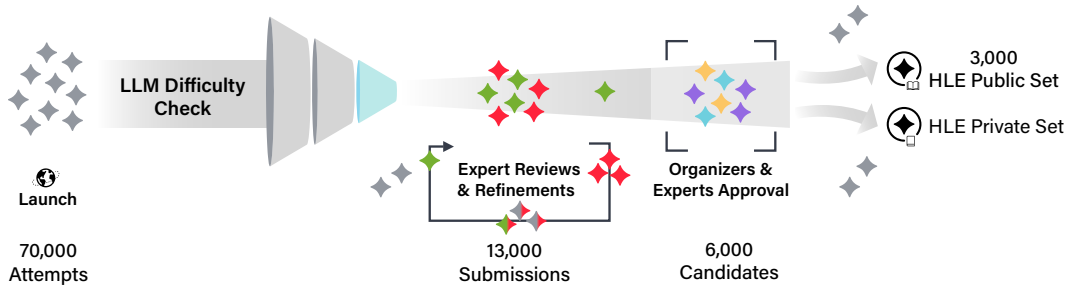


Figure 4: Dataset creation pipeline. We accept questions that make frontier LLMs fail, then iteratively refine them with the help of expert peer reviewers. Each question is then manually approved by organizers or expert reviewers trained by organizers. A private held-out set is kept in addition to the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

and offering feedback when applicable. There are two rounds of reviews. The first round focuses on iteratively refining submissions, with each question receiving between 1-3 reviews. 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 Appendix B.2. Figure 4 details our full process.

Due to the advanced, specialized nature of many submissions, reviewers were not expected to verify the full accuracy of each provided solution rationale if it would take more than five minutes, instead focusing on whether the question aligns with guidelines. Given this limitation in the review process, we welcome community feedback. After initial release, we plan to conduct a public feedback period and periodically update the dataset, assessing any points of concern from the research community.

4 Evaluation

We evaluate the performance of state-of-the-art LLMs on HLE and analyze their capabilities across different question types and domains. We describe our evaluation setup (Section 4.1) and present several quantitative results on metrics that track model performance (Section 4.2).

4.1 Setup

After data collection and review, we evaluated our final HLE dataset on additional frontier multi-modal LLMs. We employ 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 GPT-4o as a judge to verify answer correctness against model predictions while accounting for equivalent formats (e.g., decimals vs. fractions or estimations). Evaluation prompts are detailed in Appendix C.1.1, and exact model versions are provided in Appendix C.4.

4.2 Quantitative Results

Accuracy. All frontier models achieve low accuracy on HLE (Table 1), highlighting significant 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 (Section 3.1) attempts to filter out questions that existing models can answer correctly. Nevertheless, we notice upon evaluation, 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 choose to leave these questions in the dataset as a natural component instead of strongly adversarially filtering. However, we stress 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, indicative of confabulation/hallucination. To measure calibration, we prompt models to provide both an answer and their confidence

Model	Accuracy (%) \uparrow	Calibration Error (%) \downarrow
GPT-4o	3.3	92.5
GROK 2	3.8	93.2
CLAUDE 3.5 SONNET	4.3	88.9
GEMINI 1.5 PRO	5.0	93.1
GEMINI 2.0 FLASH THINKING	6.2	93.9
O1	9.1	93.4
DEEPSEEK-R1*	9.4	81.8

Table 1: Accuracy and RMS calibration error of different models on HLE, demonstrating low accuracy and high calibration error across all models, indicative of hallucination. *Model is not multi-modal, evaluated on text-only subset. We report text-only results on all models in Appendix C.2.

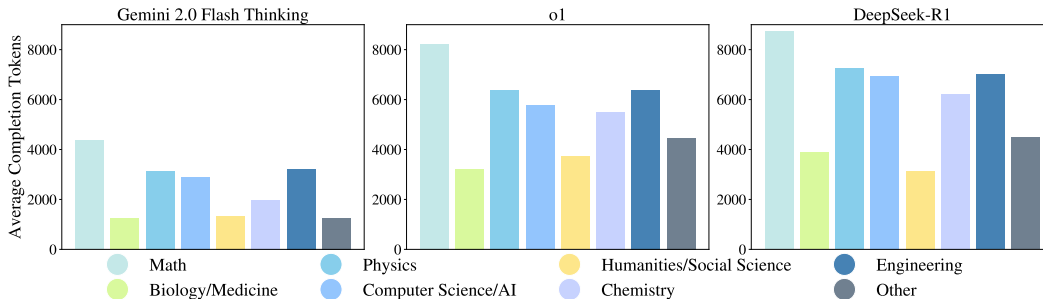


Figure 5: Average completion token counts of reasoning models tested, including both reasoning and output tokens. We also plot average token counts for non-reasoning models in Appendix C.3.

from 0% to 100% (Appendix C.1.1), employing the setup from Wei et al. [54]. The implementation of our RMS calibration error is from Hendrycks et al. [23]. A well-calibrated model’s stated confidence should match its actual accuracy – for example, achieving 50% accuracy on questions where it claims 50% confidence. Table 1 reveals 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.

Token Counts. Models with reasoning require substantially more inference time compute. To shed light on this in our evaluation, we analyze the number of completion tokens used across models. As shown in Figure 5, all reasoning models require generating significantly more tokens compared to non-reasoning models for an improvement in performance (Appendix C.3). We emphasize that future models should not only do better in terms of accuracy, but also strive to be compute-optimal.

5 Discussion

Future Model Performance. While current LLMs achieve very low accuracy on HLE, recent history shows benchmarks are quickly saturated – with models dramatically progressing from near-zero to near-perfect performance in a short timeframe [12, 44]. Given the rapid pace of AI development, it is plausible that models could exceed 50% accuracy on HLE by the end of 2025. 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.” 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. HLE may be the last academic exam we need to give to models, but it is far from the last benchmark for AI.

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.

References

- [1] C. Alberti, K. Lee, and M. Collins. A bert baseline for the natural questions, 2019. URL <https://arxiv.org/abs/1901.08634>.
- [2] M. Andriushchenko, A. Souly, M. Dziemian, D. Duenas, M. Lin, J. Wang, D. Hendrycks, A. Zou, Z. Kolter, M. Fredrikson, E. Winsor, J. Wynne, Y. Gal, and X. Davies. Agentharm: A benchmark for measuring harmfulness of llm agents, 2024. URL <https://arxiv.org/abs/2410.09024>.
- [3] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL <https://api.semanticscholar.org/CorpusID:268232499>.
- [4] Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet, 2024. URL <https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf>.
- [5] Anthropic. Responsible scaling policy updates, 2024. URL <https://www.anthropic.com/rsp-updates>.
- [6] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton. Program synthesis with large language models, 2021. URL <https://arxiv.org/abs/2108.07732>.
- [7] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- [8] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL <https://arxiv.org/abs/1611.09268>.
- [9] M. Bhatt, S. Chennabasappa, C. Nikolaidis, S. Wan, I. Evtimov, D. Gabi, D. Song, F. Ahmad, C. Aschermann, L. Fontana, S. Frolov, R. P. Giri, D. Kapil, Y. Kozyrakis, D. LeBlanc, J. Milazzo, A. Straumann, G. Synnaeve, V. Vontimitta, S. Whitman, and J. Saxe. Purple llama cyberseceval: A secure coding benchmark for language models, 2023. URL <https://arxiv.org/abs/2312.04724>.
- [10] J. S. Chan, N. Chowdhury, O. Jaffe, J. Aung, D. Sherburn, E. Mays, G. Starace, K. Liu, L. Maksin, T. Patwardhan, L. Weng, and A. Madry. Mle-bench: Evaluating machine learning agents on machine learning engineering, 2024. URL <https://arxiv.org/abs/2410.07095>.
- [11] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021. URL <https://arxiv.org/abs/2107.03374>.
- [12] F. Chollet, M. Knoop, G. Kamradt, and B. Landers. Arc prize 2024: Technical report, 2024. URL <https://arxiv.org/abs/2412.04604>.
- [13] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.

- [14] DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://github.com/deepseek-ai/DeepSeek-V3/blob/main/DeepSeek_V3.pdf.
- [15] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 2019. URL <https://arxiv.org/abs/1903.00161>.
- [16] A. Dubey et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- [17] B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu, and B. Chang. Omnimath: A universal olympiad level mathematic benchmark for large language models, 2024. URL <https://arxiv.org/abs/2410.07985>.
- [18] E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. de Oliveira Santos, O. Järvinen, M. Barnett, R. Sandler, J. Sevilla, Q. Ren, E. Pratt, L. Levine, G. Barkley, N. Stewart, B. Grechuk, T. Grechuk, and S. V. Enugandla. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL <https://arxiv.org/abs/2411.04872>.
- [19] C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL <https://arxiv.org/abs/2402.14008>.
- [20] D. Hendrycks, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, and J. Steinhardt. Measuring coding challenge competence with apps, 2021. URL <https://arxiv.org/abs/2105.09938>.
- [21] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- [22] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL <https://arxiv.org/abs/2103.03874>.
- [23] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures, 2022. URL <https://arxiv.org/abs/2112.05135>.
- [24] A. Hosseini, A. Sordoni, D. Toyama, A. Courville, and R. Agarwal. Not all llm reasoners are created equal, 2024. URL <https://arxiv.org/abs/2410.01748>.
- [25] A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, C. Saroufim, C. Fry, D. Marcus, D. Kukliansky, G. S. Tomar, J. Swirhun, J. Xing, L. W. and Madhu Gurumurthy, M. Aaron, M. Ambar, R. Fellingner, R. Wang, R. Sims, Z. Zhang, S. Goldstein, and D. Das. Facts leaderboard. <https://kaggle.com/facts-leaderboard>, 2024. Google DeepMind, Google Research, Google Cloud, Kaggle.
- [26] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL <https://arxiv.org/abs/2310.06770>.
- [27] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking benchmarking in nlp, 2021. URL <https://arxiv.org/abs/2104.14337>.
- [28] P. Kumar, E. Lau, S. Vijayakumar, T. Trinh, S. R. Team, E. Chang, V. Robinson, S. Hendryx, S. Zhou, M. Fredrikson, S. Yue, and Z. Wang. Refusal-trained llms are easily jailbroken as browser agents, 2024. URL <https://arxiv.org/abs/2410.13886>.

- [29] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodrigues. Lab-bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.
- [30] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Khoja, Z. Zhao, A. Herbert-Voss, C. B. Breuer, S. Marks, O. Patel, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Lin, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, R. Kaplan, I. Steneker, D. Campbell, B. Jokubaitis, A. Levinson, J. Wang, W. Qian, K. K. Karmakar, S. Basart, S. Fitz, M. Levine, P. Kumaraguru, U. Tupakula, V. Varadharajan, R. Wang, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, and D. Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL <https://arxiv.org/abs/2403.03218>.
- [31] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL <https://arxiv.org/abs/2310.02255>.
- [32] T. R. McIntosh, T. Susnjak, N. Arachchilage, T. Liu, P. Watters, and M. N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024. URL <https://arxiv.org/abs/2402.09880>.
- [33] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020. URL <https://arxiv.org/abs/1910.14599>.
- [34] OpenAI. Openai o1 system card, 2024. URL <https://cdn.openai.com/o1-system-card-20240917.pdf>.
- [35] OpenAI. Openai and los alamos national laboratory announce bio-science research partnership, 2024. URL <https://openai.com/index/openai-and-los-alamos-national-laboratory-work-together/>.
- [36] OpenAI. Introducing swe-bench verified, 2024. URL <https://openai.com/index/introducing-swe-bench-verified/>.
- [37] OpenAI et al. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- [38] S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, and M. Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1): 6793, 2022.
- [39] D. Owen. How predictable is language model benchmark performance?, 2024. URL <https://arxiv.org/abs/2401.04757>.
- [40] E. Perez, S. Ringer, K. Lukošiušė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. El Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL <https://arxiv.org/abs/2212.09251>.
- [41] M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Deletang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, and T. Shevlane. Evaluating frontier models for dangerous capabilities, 2024. URL <https://arxiv.org/abs/2403.13793>.

- [42] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL <https://arxiv.org/abs/1606.05250>.
- [43] P. Rajpurkar, R. Jia, and P. Liang. Know what you don’t know: Unanswerable questions for squad, 2018. URL <https://arxiv.org/abs/1806.03822>.
- [44] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>.
- [45] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023.
- [46] V. K. Srinivasan, Z. Dong, B. Zhu, B. Yu, H. Mao, D. Mosk-Aoyama, K. Keutzer, J. Jiao, and J. Zhang. Nexusraven: A commercially-permissive language model for function calling. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. URL <https://openreview.net/forum?id=51cPe6DqfI>.
- [47] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shueb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askeel, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. Dai, A. La, A. Lampinen, A. Zou, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL <https://arxiv.org/abs/2206.04615>.
- [48] S. A. Taghanaki, A. Khani, and A. Khasahmadi. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms, 2024. URL <https://arxiv.org/abs/2409.02257>.
- [49] G. Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL <https://arxiv.org/abs/2403.05530>.
- [50] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, and S. Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition, 2024. URL <https://arxiv.org/abs/2407.11214>.
- [51] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL <https://arxiv.org/abs/1804.07461>.
- [52] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGlue: A stickier benchmark for general-purpose language understanding systems, 2020. URL <https://arxiv.org/abs/1905.00537>.
- [53] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark (published at neurips 2024 track datasets and benchmarks), 2024. URL <https://arxiv.org/abs/2406.01574>.
- [54] J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models, 2024. URL <https://arxiv.org/abs/2411.04368>.
- [55] H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, E. Ericeva, K. Garcia, B. Goodrich, N. Jurkovic, M. Kinniment, A. Lajko, S. Nix, L. Sato, W. Saunders, M. Taran, B. West, and E. Barnes. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts, 2024. URL <https://arxiv.org/abs/2411.15114>.
- [56] xAI. Grok-2 beta release, 2024. URL <https://x.ai/blog/grok-2>.

- [57] F. Yan, H. Mao, C. C.-J. Ji, T. Zhang, S. G. Patil, I. Stoica, and J. E. Gonzalez. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html, 2024.
- [58] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL <https://arxiv.org/abs/1809.09600>.
- [59] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. τ -bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL <https://arxiv.org/abs/2406.12045>.
- [60] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, J. W. Lin, E. Jones, C. Menders, G. Hussein, S. Liu, D. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, M. Yang, T. Zhang, R. Alluri, N. Tran, R. Sangpisit, P. Yiorkadjis, K. Osele, G. Raghupathi, D. Boneh, D. E. Ho, and P. Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models, 2024. URL <https://arxiv.org/abs/2408.08926>.
- [61] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. URL <https://arxiv.org/abs/2304.06364>.

A Authors

We offered optional co-authorship to all question submitters with an accepted question in HUMANITY'S LAST EXAM (including both public and private splits). All potential co-authors with an accepted question were contacted directly. Authorship order is ranked based on the number of accepted questions in HUMANITY'S LAST EXAM.

As we give co-authors the time and freedom to choose between opting-in or staying anonymous, we will periodically update this list. We further note that this list only represents a subset of our participating institutions and authors, many chose to remain anonymous.

A.1 Data Contributors & Affiliations

In progress. Sorted in descending order by number of accepted questions.

Authors Daron Anderson³, Tung Nguyen⁴, Mobeen Mahmood⁵, Fiona Feng⁶, Steven Y. Feng⁷, Haoran Zhao⁸, Michael Yu³, Chelsea Zou⁷, Zihan Wang⁹, Jessica P. Wang¹⁰, Pawan Kumar¹¹, Oleksandr Pokutnyi¹², Robert Gerbicz¹³, Serguei Popov¹⁴, John-Clark Levin¹⁵, Johannes Schmitt¹⁶, Geoff Galgon¹⁷, Alvaro Sanchez³, Yongki Lee¹⁸, Will Yeadon¹⁹, Scott Sauers²⁰, Marc Roth²¹, Chidozie Agu²², Søren Riis²¹, Fabian Giska³, Saiteja Utpala²³, Zachary Giboney²⁴, Gashaw M. Goshu³, Joan of Arc Xavier²⁵, Sarah-Jane Crowson²⁶, Mohinder Maheshbhai Naiya²⁷, Noah Burns⁷, Lennart Finke¹⁶, Zerui Cheng²⁸, Hyunwoo Park²⁹, Francesco Fournier-Facio¹⁵, John Wydallis³, Mark Nandor³, Ankit Singh³⁰, Tim Gehringer¹⁶, Jiaqi Cai³¹, Ben McCarty³², Darling Duclosel³³, Jungbae Nam³⁴, Jennifer Zampese³⁵, Ryan G. Hoerr³⁶, Aras Bacho³⁷, Gautier Abou Loume^{38,39}, Abdallah Galal⁴⁰, Hangrui Cao²⁹, Alexis C Garretson^{41,42}, Damien Sileo⁴³, Qiuyu Ren⁴⁴, Doru Cojoc⁴⁵, Pavel Arkhipov⁴⁶, Usman Qazi^{47,48}, Lianghui Li⁴⁹, Sumeet Motwani⁵⁰, Christian Schroeder de Witt⁵⁰, Edwin Taylor³, Johannes Veith^{51,52}, Taylor D. Hartman⁵³, Paolo Rissone⁵⁴, Jaehyeok Jin⁴⁵, Jack Wei Lun Shi⁵⁵, Chris G. Willcocks¹⁹, Joshua Robinson⁵⁶, Aleksandar Mikov⁴⁹, Ameya Prabhu⁵⁷, Longke Tang²⁸, Xavier Alapont²⁵, Kevin Zhou⁴⁴, Emily de Oliveira Santos⁵⁸, Andrey Pupasov Maksimov⁵⁹, Edward Vendrow³¹, Kengo Zenitani³, Julien Guillod^{60,61}, Yuqi Li⁶², Joshua Vendrow³¹, Vladyslav Kuchkin⁶³, Ng Ze-An⁶⁴, Pierre Marion⁴⁹, Denis Efremov⁶⁵, Jayson Lynch³¹, Kaiqu Liang²⁸, Andrew Gritsevskiy⁶⁶, Dakotah Martinez³, Ben Pageler³, Nick Crispino⁶⁷, Dimitri Zvonkine^{68,69}, Natanael Wildner Fraga³, Saeed Soori⁷⁰, Ori Press⁵⁷, Henry Tang⁵⁰, Julian Salazar⁷¹, Sean R. Green³, Lina Brüssel¹⁵, Moon Twayana⁷², Aymeric Dieuleveut⁷³, T. Ryan Rogers⁷⁴, Wenjin Zhang⁶⁷, Bikun Li⁷⁵, Jinzhou Yang⁷⁶, Arun Rao⁷⁷, Gabriel Loiseau⁴³, Mikhail Kalinin⁷⁸, Marco Lukas⁷⁹, Ciprian Manolescu⁷, Subrata⁸⁰, Ariel Ghislain Kemogne Kamdoun⁸¹, Tobias Kreiman⁴⁴, Tad Hogg⁸², Alvin Jin³¹, Carlo Bosio⁴⁴, Gongbo Sun⁸³, Brian P Coppola⁸⁴, Tim Tarver⁸⁵, Haline Heidinger^{86,87}, Rafael Sayous⁶⁹, Stefan Ivanov¹⁵, Joseph M Cavanagh⁴⁴, Jiawei Shen⁶⁷, Joseph Marvin Imperial^{88,89}, Philippe Schwaller⁴⁹, Shaipranesh Senthilkuma⁴⁹, Andres M Bran⁴⁹, Ali Dehghan³, Andres Algaba⁹⁰, Brecht Verbeken⁹⁰, David Noever⁹¹, Ragavendran P V³, Lisa Schut⁵⁰, Ilia Sucholutsky⁹², Evgenii Zheltonozhskii⁹³, Derek Lim³¹, Richard Stanley^{31,94}, Shankar Sivaranjan⁹⁵, Tong Yang²⁹, John Maar⁹⁶, Julian Wykowski¹⁵, Martí Oller¹⁵, Jennifer Sandlin⁹⁷, Anmol Sahu³, Yuzheng Hu⁹⁸, Sara Fish⁹⁹, Nasser Heydari³, Archimedes Apronti¹⁰⁰, Kaivalya Rawal⁵⁰, Tobias Garcia Vilchis¹⁰¹, Yuexuan Zu³¹, Martin Lackner¹⁰², James Koppel³, Jeremy Nguyen¹⁰³, Daniil S. Antonenko¹⁰⁴, Steffi Chern²⁹, Bingchen Zhao¹⁰⁵, Pierrot Arsene¹⁰⁶, Alan Goldfarb⁴⁴, Sergey Ivanov³, Rafał Poświata¹⁰⁷, Chenguang Wang⁶⁷, Daofeng Li⁶⁷, Donato Crisostomi⁵⁴, Andrea Achilleos¹⁰⁸, Benjamin Myklebust¹⁰⁹, Archan Sen⁴⁴, David Perrella¹¹⁰, Nurdin Kaparov¹¹¹, Mark H Inlow¹¹², Allen Zang⁷⁵, Elliott Thornley¹¹³, Daniil Orel¹¹⁴, Vladislav Poritski³, Shalev Ben-David¹¹⁵, Zachary Berger³¹, Parker Whitfill³¹, Michael Foster³, Daniel Munro⁹, Linh Ho³, Dan Bar Hava¹¹⁶, Aleksey Kuchkin³, Robert Lauff⁹⁶, David Holmes¹¹⁷, Frank Sommerhage¹¹⁸, Keith Schneider³, Zakayo Kazibwe¹¹⁹, Nate Stambaugh¹²⁰, Mukhwinder Singh¹²¹, Ilias Magoulas¹²², Don Clarke¹²³, Dae Hyun Kim¹²⁴, Felipe Meneguitti Dias⁵⁸, Veit Elser¹²⁵, Kanu Priya Agarwal³, Victor Efren Guadarrama Vilchis¹²⁶, Immo Klose⁴⁵, Christoph Demian⁵², Ujjwala Ananthaswaran⁹⁷, Adam Zweiger³¹, Guglielmo Albani¹²⁷, Jeffery Li³¹, Nicolas Daans¹²⁸, Maksim Radionov¹²⁹, Václav Rozhoň¹³⁰, Ziqiao Ma⁸⁴, Christian Stump¹³¹, Mohammed Berkani¹³², Jacob Platnick¹³³, Volodymyr Nevirkovets¹³⁴, Luke Basler¹³⁵, Marco Piccardo¹³⁶, Ferenc Jeanplong¹³⁷, Niv Cohen⁹², Varun Gangal³, Josef Tkadlec¹³⁸, Paul Rosu¹³⁹, Piotr Padlewski³, Stanislaw Barzowski³, Kyle Montgomery⁶⁷, Aline Menezes³, Arkil Patel^{5,140}, Zixuan Wang²⁸, Jamie Tucker-Foltz⁹⁹, Jack Stade¹⁴¹, Tom Goertzen¹⁴², Fereshteh Kazemi³, Jeremiah Milbauer²⁹, John Arnold Ambay¹⁴³, Abhishek Shukla¹⁴⁴, Yan Carlos Leyva Labrador²⁵, Alan Givré¹⁴⁵, Hew Wolff³, Vivien Ross-

bach²⁵, Muhammad Fayez Aziz⁹⁸, Younesse Kaddar⁵⁰, Yanxu Chen¹⁴⁶, Robin Zhang³¹, Jiayi Pan⁴⁴, Antonio Terpin¹⁶, Niklas Muennighoff⁷, Hailey Schoelkopf³, Eric Zheng²⁹, Avishy Carmi¹⁴⁷, Adam Jones³, Jainam Shah¹⁴⁸, Ethan D. L. Brown¹⁴⁹, Kelin Zhu⁹⁵, Max Bartolo¹⁵⁰, Richard Wheeler¹⁰⁵, Andrew Ho¹⁵¹, Shaul Barkan¹⁵², Jiaqi Wang⁸, Martin Stehberger³, Egor Kretov¹⁵³, Kaustubh Sridhar¹⁵⁴, Zienab EL-Wasif¹⁵⁵, Anji Zhang³¹, Daniel Pyda¹⁵⁶, Joanna Tam¹⁵⁷, David M. Cunningham¹⁵⁸, Demosthenes Patramanis⁵⁰, Michael Krause¹⁵⁹, Andrew Redenti⁴⁵, Daniel Bugas³, David Aldous⁴⁴, Jesyin Lai¹⁶⁰, Shannon Coleman⁴⁸, Mohsen Bahaloo¹⁶¹, Jiangnan Xu¹⁶², Sangwon Lee³, Sandy Zhao²⁵, Ning Tang⁴⁴, Michael K. Cohen⁴⁴, Micah Carroll⁴⁴, Orr Paradise⁴⁴, Jan Hendrik Kirchner¹⁶³, Stefan Steinerberger⁸, Maksym Ovchinnikov¹⁶⁴, Jason O. Matos¹⁵⁷, Adithya Shenoy³, Benedito Alves de Oliveira Junior⁵⁸, Michael Wang⁴⁴, Yuzhou Nie¹⁶⁵, Paolo Giordano¹⁶⁶, Philipp Petersen¹⁶⁶, Anna Sztzyber-Betley¹⁶⁷, Priti Shukla¹⁶⁸, Jonathan Crozier¹⁶⁹, Antonella Pinto¹⁷⁰, Shreyas Verma¹⁷¹, Prashant Joshi¹⁷², Zheng-Xin Yong¹⁷³, Allison Tee⁷, Jérémy Andréoletti⁶¹, Orion Weller¹⁷⁴, Raghav Singhal¹¹⁴, Gang Zhang³, Alexander Ivanov¹⁷⁵, Seri Khoury¹³⁰, Hamid Mostaghimi⁸¹, Kunvar Thaman¹⁷⁶, Qijia Chen⁹⁹, Tran Quoc Khanh¹⁷⁷, Jacob Loader¹⁵, Stefano Cavalleri¹⁷⁸, Hannah Szlyk⁶⁷, Zachary Brown³¹, Jonathan Roberts¹⁵, William Alley³, Kunyang Sun⁴⁴, Ryan Stendall¹⁷⁹, Max Lamparth⁷, Anka Reuel⁷, Ting Wang⁶⁷, Hanmeng Xu¹⁰⁴, Sreenivas Goud Raparathi¹⁸⁰, Pablo Hernández-Cámara¹⁸¹, Freddie Martin³, Dmitry Malishev³, Thomas Preu¹⁸², Tomek Korbak¹⁸³, Marcus Abramovitch³, Dominic Williamson¹⁴², Ziye Chen¹⁸⁴, Biró Bálint³, M Saiful Bari¹⁸⁵, Peyman Kassani¹⁸⁶, Zihao Wang⁷⁵, Behzad Ansarinejad³, Laxman Prasad Goswami¹⁴⁴, Yewen Sun¹⁸⁷, Hossam Elgarnay¹⁸⁸, Daniel Tordera¹⁸⁹, George Balabanian¹⁵⁴, Earth Anderson¹⁹⁰, Lynna Kvistad¹⁹¹, Alejandro José Moyano¹⁹², Rajat Maheshwari¹⁹³, Ahmad Sakor⁷⁹, Murat Eron¹⁹⁴, Isaac C. McAlister³, Javier Gimenez²⁵, Innocent Enyekwe³, Andrew Favre D.O.¹⁹⁵, Shailesh Shah¹⁹⁶, Xiaoxiang Zhou⁵², Firuz Kamalov¹⁹⁷, Ronald Clark⁵⁰, Sherwin Abdoli¹⁷⁰, Khalida Meer²⁵, Harrison K Wang⁹⁹, Evan Chen³¹, Alessandro Tomasiello¹⁹⁸, Shi-Zhuo Looi³⁷, Vinh-Kha Le⁴⁴, Noam Kolt¹⁵², Niels Mündler¹⁶, Avi Semler⁵⁰, Emma Rodman¹⁹⁹, Jacob Drori³, Carl J Fossum²⁰⁰, Milind Jagota⁴⁴, Ronak Pradeep¹¹⁵, Honglu Fan²⁰¹, Tej Shah²⁰², Tej Shah²⁰³, Jonathan Eicher²⁰⁴, Michael Chen³⁷, Kushal Thaman⁷, William Merrill⁹², Carter Harris²⁰⁵, Jason Gross³, Ilya Gusev³, Asankhaya Sharma²⁰⁶, Shashank Agnihotri²⁰⁷, Pavel Zhelnov⁷⁰, Siranüt Usawasutsakorn²⁰⁸, Mohammadreza Mofayez⁷⁰, Sergei Bogdanov²⁰⁹, Alexander Piperski²¹⁰, Marc Caraleanu²¹¹, David K. Zhang⁷, Dylan Ler³, Roman Leventov²¹², Ignat Soroko⁷², Thorben Jansen²¹³, Pascal Lauer^{214,215}, Joshua Duersch²¹⁶, Vage Taamazyan²¹⁷, Wiktor Morak³, Wenjie Ma⁴⁴, William Held^{7,133}, Tran Duc Huy²¹⁸, Ruicheng Xian⁹⁸, Armel Randy Zebaze²¹⁹, Mohanad Mohamed²²⁰, Julian Noah Leser¹⁰², Michelle X Yuan³, Laila Yacar²²¹, Johannes Lengler¹⁶, Hossein Shahrtash²²², Edson Oliveira²²³, Joseph W. Jackson²²⁴, Daniel Espinosa Gonzalez¹⁶⁵, Andy Zou^{29,225}, Muthu Chidambaram¹³⁹, Timothy Manik³, Hector Haffenden³, Dashiell Stander²²⁶, Ali Dasouqi¹⁷⁴, Alexander Shen²²⁷, Emilien Duc¹⁶, Bitu Golshani³, David Stap¹⁴⁶, Mikalai Uzhou²²⁸, Alina Borisovna Zhidkovskaya²²⁹, Lukas Lewark¹⁶, Mátyás Vincze^{230,231}, Dustin Wehr³, Colin Tang²⁹, Zaki Hossain²³², Shaun Phillips³, Jiang Muzhen³, Fredrik Ekström³, Angela Hammon³, Oam Patel⁹⁹, Nicolas Remy²³³, Faraz Farhidi²³⁴, George Medley³, Forough Mohammadzadeh³, Madellene Peñaflor²³⁵, Haile Kassahun⁵, Alena Friedrich²³⁶, Claire Sparrow⁷⁵, Taom Sakal¹⁶⁵, Omkar Dhamane²³⁷, Ali Khajegili Mirabadi⁴⁸, Eric Hallman³, Mike Battaglia³, Mohammad Maghsoudimehrabani²³⁸, Hieu Hoang²³⁹, Alon Amit²⁴⁰, Dave Hulbert³, Roberto Pereira²⁴¹, Simon Weber¹⁶, Stephen Mensah²⁴², Nathan Andre²⁴³, Anton Peristyy³, Chris Harjadi⁷, Himanshu Gupta⁹⁷, Stephen Malina²⁴⁴, Samuel Albanie³, Will Cai⁴⁴, Mustafa Mehkary^{70,245}, Frank Reidegeld³, Anna-Katharina Dick⁵⁷, Cary Friday²⁴⁶, Jasdeep Sidhu³, Wanyoung Kim²⁴⁷, Mariana Costa²⁵, Hubeyb Gurdogan⁷⁷, Brian Weber²⁴⁸, Harsh Kumar²⁴⁹, Tong Jiang⁹⁹, Arunim Agarwal²⁵⁰, Chiara Ceconello³, Warren S. Vaz³, Chao Zhuang³, Haon Park^{251,252}, Andrew R. Tawfeek⁸, Daattavya Aggarwal¹⁵, Michael Kirchhof⁵⁷, Linjie Dai³¹, Evan Kim³¹, Johan Ferret⁷¹, Yuzhou Wang¹³³, Minghao Yan⁸³, Krzysztof Burdzy⁸, Lixin Zhang²⁵, Antonio Franca¹⁵, Diana T. Pham²⁵³, Kang Yong Loh⁷, Joshua Robinson²⁵⁴, Shreen Gul²⁵⁵, Gunjan Chhablani¹³³, Zhehang Du¹⁵⁴, Adrian Cosma²⁵⁶, Colin White²⁵⁷, Robin Riblet¹⁰⁶, Prajvi Saxena²⁵⁸, Jacob Votava²⁸, Vladimir Vinnikov³, Shiv Halasyamani²⁵⁹, Syed M. Shahid²⁶⁰, Jean-Christophe Mourrat^{68,261}, Lavr Vetoshkin²⁶², Renas Bacho²⁶³, Vincent Ginis^{90,99}, Aleksandr Maksapetyan²⁵, Florencia de la Rosa²⁶⁴, Xiuyu Li⁴⁴, Guillaume Malod²⁶⁵, Leon Lang¹⁴⁶, Julien Laurendeau⁴⁹, Fatimah Adesanya^{25,266}, Julien Portier¹⁵, Lawrence Hollom¹⁵, Victor Souza¹⁵, Yuchen Anna Zhou²⁶⁷, Yiğit Yalın²⁶⁸, Gbenga Daniel Obikoya³, Luca Arnaboldi⁴⁹, Rai (Michael Pokorny)²⁶⁹, Filippo Bigi⁴⁹, Kaniuar Bacho¹⁰⁵, Pierre Clavier²⁷⁰, Gabriel Recchia²⁷¹, Mara Popescu²⁷², Nikita Shulga²⁷³, Ngefor Mildred Tanwie²⁷⁴, Thomas C.H. Lux²⁷⁵, Ben Rank³, Colin Ni⁷⁷, Alesia Yakimchyk²⁷⁶, Huanxu (Quinn) Liu²⁷⁷, Olle Häggström²⁷⁸, Emil Verkama²⁷⁹, Himanshu Narayan³, Hans Gundlach³¹, Leonor Brito-Santana²⁸⁰, Brian Amaro⁷,

Vivek Vajipey⁷, Rynaa Grover¹³³, Yiyang Fan³, Gabriel Poesia Reis e Silva⁷, Linwei Xin⁷⁵, Yosi Kratish¹³⁴, Jakub Łucki¹⁶, Wen-Ding Li¹²⁵, Justin Xu⁵⁰, Kevin Joseph Scaria⁹⁷, Freddie Vargus²⁸¹, Farzad Habibi²⁸², Long (Tony) Lian⁴⁴, Emanuele Rodolà⁵⁴, Jules Robins³, Vincent Cheng⁹, Declan Grabb⁷, Ida Bosio²⁸³, Tony Fruhauff³, Ido Akov²⁸⁴, Eve J. Y. Lo²⁸⁵, Hao Qi¹⁸⁴, Xi Jiang⁷⁵, Ben Segev⁴⁵, Jingxuan Fan⁹⁹, Sarah Martinson⁹⁹, Erik Y. Wang⁹⁹, Kaylie Hausknecht⁹⁹, Michael P. Brenner⁹⁹, Mao Mao¹⁸⁴, Yibo Jiang⁷⁵, Xinyu Zhang¹⁸⁴, David Avagian²⁰⁷, Eshawn Jessica Scipio²⁸⁶, Muhammad Rehan Siddiqi^{287,288}, Alon Ragoler²⁸⁹, Justin Tan¹⁵, Deepakkumar Patil²⁹⁰, Rebeka Plecnik³, Aaron Kirtland¹⁷³, Roselynn Grace Montecillo²⁹¹, Stephane Durand²⁹², Omer Faruk Bodur³, Zahra Adoul²⁹³, Mohamed Zekry²⁹⁴, Guillaume Douville²⁵, Ali Karakoc²⁹⁵, Tania C. B. Santos³, Samir Shamseldeen²⁹⁶, Loukmane Karim²⁴⁵, Anna Liakhovitskaia²⁹⁷, Nate Resman²⁹⁸, Nicholas Farina²⁵, Juan Carlos Gonzalez²⁹⁹, Gabe Maayan¹⁸⁴, Sarah Hoback⁹⁹, Rodrigo De Oliveira Pena³⁰⁰, Glen Sherman²⁵, Hodjat Mariji³, Rasoul Pouriamanesh³, Wentao Wu⁴⁸, Gözdenur Demir³, Sandra Mendoza^{301,302}, Ismail Alarab³⁰³, Joshua Cole³⁰⁴, Danyelle Ferreira²⁵, Bryan Johnson³⁰⁵, Hsiaoyun Milliron³⁰⁶, Mohammad Safdari³⁰⁷, Liangti Dai⁵⁰, Siriphan Arthornthurasuk²⁵, Alexey Pronin³⁰⁸, Angel Ramirez-Trinidad³, Ashley Cartwright³⁰⁹, Daphiny Pottmaier³¹⁰, Omid Taheri³¹¹, David Outevsky³¹², Stanley Stepanic³¹³, Samuel Perry³, Luke Askew³¹⁴, Raúl Adrián Huerta Rodríguez³, Abdelkader Dendane²⁵, Ricardo Lorena³¹⁵, Krishnamurthy Iyer³¹⁶, Sk Md Salauddin³¹⁷, Murat Islam³¹⁸, Juan Gonzalez³, Josh Ducey³¹⁹, Russell Campbell³²⁰, Maja Somrak³, Vasilios Mavroudis³²¹, Eric Vergo³, Juehang Qin³²², Benjámín Borbás³²³, Eric Chu⁷¹, Jack Lindsey¹⁶³, Anil Radhakrishnan¹⁶⁹, Antoine Jallon³, I.M.J. McInnis³, Alex Hoover⁷⁵, Sören Möller³²⁴, Tejal Patwardhan²⁶⁹

Affiliations

- | | |
|--|--|
| 3. Independent Researcher | 29. Carnegie Mellon University |
| 4. Texas A&M University | 30. Hemwati Nandan Bahuguna Garhwal University |
| 5. McGill University | 31. Massachusetts Institute of Technology |
| 6. Queen's University | 32. Accenture Labs |
| 7. Stanford University | 33. Escuela Superior de Medicina- Instituto Politécnico Nacional |
| 8. University of Washington | 34. CICMA |
| 9. University of California, San Diego | 35. University of Canterbury |
| 10. RWTH Aachen University | 36. Metropolitan State University of Denver |
| 11. Pondicherry Engineering College | 37. California Institute of Technology |
| 12. Institute of Mathematics of NAS of Ukraine | 38. Université de Yaoundé I |
| 13. ELTE | 39. Ecole Nationale Supérieure Polytechnique de Yaoundé |
| 14. University of Porto | 40. Tanta University |
| 15. University of Cambridge | 41. Tufts University |
| 16. ETH Zürich | 42. The Jackson Laboratory |
| 17. Nimbus AI | 43. Inria |
| 18. Georgia Southern University | 44. University of California, Berkeley |
| 19. Durham University | 45. Columbia University |
| 20. University of Minnesota Twin Cities | 46. Institute of Science and Technology Austria |
| 21. Queen Mary University of London | 47. RUSM |
| 22. Alberta Health Services | 48. University of British Columbia |
| 23. Microsoft Research | 49. École Polytechnique Fédérale de Lausanne |
| 24. ZG Law | 50. University of Oxford |
| 25. Outlier | 51. Charité – Universitätsmedizin |
| 26. Hereford College of Arts | 52. Humboldt-Universität zu Berlin |
| 27. Auckland University of Technology | |
| 28. Princeton University | |

53. Northern Illinois University
54. Sapienza University of Rome
55. National University of Singapore
56. University of Southern California
57. University of Tübingen
58. University of Sao Paulo
59. Universidade Federal de Juiz de Fora
60. Sorbonne Université
61. École Normale Supérieure
62. C. N. Yang institute for Theoretical Physics
63. University of Luxembourg
64. University of Malaya
65. Rockwell Automation
66. Contramont Research
67. Washington University
68. CNRS
69. Université Paris-Saclay
70. University of Toronto
71. Google DeepMind
72. University of North Texas
73. Institut Polytechnique de Paris
74. TRR Designs
75. University of Chicago
76. Maastricht University
77. University of California, Los Angeles
78. Martin-Luther-University Halle-Wittenberg
79. Leibniz University Hannover
80. Indian Institute of Technology Bombay
81. University of Calgary
82. Institute for Molecular Manufacturing
83. University of Wisconsin-Madison
84. University of Michigan
85. Bethune-Cookman University
86. St. Petersburg College
87. La Molina National Agrarian University
88. University of Bath
89. National University Philippines
90. Vrije Universiteit Brussel
91. PeopleTec, Inc.
92. New York University
93. Technion – Israel Institute of Technology
94. University of Miami
95. University of Maryland
96. Technische Universität Berlin
97. Arizona State University
98. University of Illinois Urbana-Champaign
99. Harvard University
100. Royal Holloway, University of London
101. Universidad Iberoamericana
102. TU Wien
103. Swinburne University of Technology
104. Yale University
105. University of Edinburgh
106. École Normale Supérieure Paris-Saclay
107. National Information Processing Institute
108. University College London
109. Ecco IT
110. University of Western Australia
111. Snorkel AI
112. Indiana State University
113. Oxford University
114. Mohamed bin Zayed University of Artificial Intelligence
115. University of Waterloo
116. Manhattan School of Music
117. Universiteit Leiden
118. Synbionix
119. Corteva Agriscience
120. Diverging Mathematics
121. Saint Mary's University
122. Emory University
123. Sanford Burnham Preybs
124. Yonsei University
125. Cornell University
126. University of Leeds
127. Politecnico di Milano
128. KU Leuven
129. Brandenburg University of Technology
130. INSAIT
131. Ruhr University Bochum
132. University Mohammed I
133. Georgia Institute of Technology
134. Northwestern University
135. University of Arizona
136. Universidade de Lisboa,

137. Mānuka Honey and Beekeeping Consultancy Ltd
138. Charles University
139. Duke University
140. Mila
141. University of Copenhagen
142. The University of Sydney
143. University of Technology Sydney
144. Indian Institute of Technology Delhi
145. University of Buenos Aires
146. University of Amsterdam
147. Ben-Gurion University
148. blurrylogic
149. Donald and Barbara Zucker School of Medicine
150. Cohere
151. Ivy Natal
152. Hebrew University
153. Fraunhofer IMTE
154. University of Pennsylvania
155. National Institute of Laser Enhanced Sciences
156. Drexel University
157. Northeastern University
158. EHC Investments LLC
159. University of Windsor
160. St. Jude Children's Research Hospital
161. GC
162. Rochester Institute of Technology
163. Anthropic
164. CERN
165. University of California, Santa Barbara
166. University of Vienna
167. Warsaw University of Technology
168. EF Polymers Pvt Ltd
169. North Carolina State University
170. Independent researcher
171. Simplr AI, Asurion
172. All India Institute of Medical Sciences
173. Brown University
174. Johns Hopkins University
175. Ruhr-Universität Bochum
176. Standard Intelligence
177. Posts and Telecommunications Institute of Technology
178. Clearhorse Ltd
179. Cranfield University
180. JNTU
181. Image Processing Lab, Universitat de Valencia
182. Universität Zürich
183. UK AI Safety Institute
184. Boston University
185. SDAIA
186. Children's Hospital of Orange County
187. The Ohio State University
188. Cairo University Specialized Pediatric Hospital
189. Universidad de Valencia
190. University of Arkansas
191. Monash University
192. OncoPrecision
193. Genomia Diagnostics Research Pvt Ltd
194. IEEE Life Member
195. Larkin Community Hospital
196. The University of Texas at Dallas
197. Canadian University Dubai
198. Università di Milano-Bicocca
199. University of Massachusetts Lowell
200. Virginia Tech
201. University of Geneva
202. Tej Shah
203. Rutgers University
204. MolMind
205. Cal Poly San Luis Obispo
206. Patched Codes, Inc
207. University of Mannheim
208. Chulalongkorn University
209. Ecole polytechnique
210. Stockholm University
211. AE Studio
212. Gaia Lab
213. Leibniz Institute for Science and Mathematics Education
214. Australian National University
215. Saarland University
216. College of Eastern Idaho
217. Intrinsic Innovation LLC
218. HUTECH
219. INRIA
220. King Saud University

221. Universidad de Buenos Aires
222. Pennsylvania College of Technology
223. CERo Therapeutics Holdings, Inc.
224. The Univeirsty of Tennessee
225. Gray Swan AI
226. EleutherAI
227. University of Montpellier
228. HomeEquity Bank
229. Materials Platform for Data Science LLC
230. University of Trento
231. Fondazione Bruno Kessler
232. Cambridge University
233. LGM
234. Georgia State University
235. Polytechnic University of the Philippines
236. University of Oregon
237. University of Mumbai
238. University of Guelph
239. Case Wester Reserve University
240. Intuit
241. CTTC / CERCA
242. National University
243. Talishar
244. Dyno Therapeutics
245. The Hospital for Sick Children
246. Lewis Katz School of Medicine
247. Fyaora Labs
248. Intelligent Geometries
249. Indian Institute of Technology (BHU)
250. Center for AI Safety
251. AIM Intelligence
252. Seoul National University
253. The University of Texas at Arlington
254. The Hartree Centre
255. Missouri University of Science and Technology
256. POLITEHNICA Bucharest National University of Science and Technology
257. Abacus.AI
258. German Research Center for Artificial Intelligence
259. University of Houston
260. Eastern Institute of Technology (EIT)
261. ENS Lyon
262. Czech Technical University in Prague
263. CISPA Helmholtz Center for Information Security
264. Universidad de Morón
265. Université Paris Cité and Sorbonne Université
266. Sheffield Hallam University
267. The New School
268. Max Planck Institute for Software Systems
269. OpenAI
270. École Polytechnique
271. Modulo Research
272. Heidelberg University
273. La Trobe University
274. University of Yaoundé I
275. Lux Labs
276. University of Innsbruck
277. Nabu Technologies Inc
278. Chalmers University of Technology
279. KTH Royal Institute of Technology
280. Unidade Local de Saúde de Lisboa Ocidental
281. Quotient AI
282. University of California, Irvine
283. University of Padua
284. Aalto University
285. Royal Veterinary College
286. The Future Paralegals of America
287. RMIT University
288. Universal Higher Education
289. Eastlake High School
290. CSMSS Chh. Shahu College of Engineering
291. Central Mindanao University
292. University of Montreal
293. University of Bradford
294. Beni Suf University
295. Bogazici University
296. Mansoura University
297. Univerisity of Bristol
298. University of Oklahoma
299. Jala University
300. Florida Atlantic University
301. CONICET
302. Universidad Tecnológica Nacional

303. Bournemouth University	314. Dartmouth College
304. University of Warwick	315. INESC Microsistemas e Nanotecnologias
305. University of Alabama Huntsville	316. University of Minnesota
306. Van Andel Institute	317. Aligarh Muslim University
307. University of Hertfordshire	318. John Crane UK Ltd
308. Central College	319. James Madison University
309. Sheffield Teaching Hospitals NHS Foundation Trust	320. University of the Fraser Valley
310. Nottingham Trent University	321. Alan Turing Institute
311. Max Planck Institute for Intelligent Systems	322. Rice University
312. Outevsky Bespoke Dance Education	323. HUN-REN
313. University of Virginia	324. Forschungszentrum Jülich

B Dataset

B.1 Submission Process

To ensure question difficulty, we automatically check the accuracy of frontier LLMs on each question prior to submission. Our testing process uses multi-modal LLMs for text-and-image questions (GPT-4O, GEMINI 1.5 PRO, CLAUDE 3.5 SONNET, O1) and adds two non-multi-modal models (O1-MINI, O1-PREVIEW) for text-only questions. We use different submission criteria by question type: exact-match questions must stump all models, while multiple-choice questions must stump all but one model to account for potential lucky guesses. Users are instructed to only submit questions that meet this criteria. We note 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.

We use a standardized system prompt (Appendix C.1.1) to structure model responses into “Reasoning” and “Final Answer” formatting, and employ an automated GPT-4O judge to evaluate response correctness against the provided answers.

B.2 Human Review Instructions

Questions which merely stump models are not necessarily high quality – they could simply be adversarial to models without testing advanced knowledge. To resolve this, we employ two rounds of human review to ensure our dataset is thorough and sufficiently challenging as determined by human experts in their respective domains.

B.2.1 Review Round 1

We recruit human subject expert reviewers to score, provide feedback, and iteratively refine all user submitted questions. This is similar to the peer review process in academic research, where reviewers give feedback to help question submitters create better questions. We train all reviewers on the instructions and rubric below.

Reviewer Instructions

- Questions should usually (but do not always need to) be at a graduate / PhD level or above. (Score 0 if the question is not complex enough and AI models can answer it correctly.)
 - If the model is not able to answer correctly and the question is below a graduate level, the question can be acceptable.
- Questions can be any field across STEM, law, history, psychology, philosophy, trivia, etc. as long as they are tough and interesting questions.
 - For fields like psychology, philosophy, etc. we usually check if the rationale contains some reference to a book, paper or standard theories.

- For fields like law, the question text can be adjusted with “as of 2024”. Make sure questions about law are time-bounded.
- Questions do not always need to be academic. A handful of movie, TV trivia, classics, history, art, or riddle questions in the dataset are OK.
- Trivia or complicated game strategy about chess, go, etc. are okay as long as they are difficult.
- We generally want things that require a high level of human intelligence to figure out.
- Questions should ask for something precise and have an objectively correct, univocal answer.
 - If there is some non-standard jargon for the topic/field, it needs to be explained.
 - Questions must have answers that are known or solvable.
 - Questions should not be subjective or have personal interpretation.
 - Questions like “Give a proof of...”; “Explain why...”; “Provide a theory that explains...” are usually bad because they are not closed-ended and we cannot evaluate them properly. (Score 0)
 - No questions about morality or what is ethical/unethical. (Score 0)
- Questions should be original and not derived from textbooks or Google. (Score 0 if searchable on web)
- Questions need to be in English. (Score 1 and ask for translation in the review if the question is written in a different language)
- Questions should be formatted properly. (Score 1-3 depending on degree of revisions needed)
 - Question with numerical answers should have results approximated to max 2-3 decimals.
 - Fix LaTeX formatting if possible. Models often get questions right after LaTeX formatting is added or improved.
 - Questions that can be converted to text should be (converting images to text often helps models get them right).

Other Tips

- Please write detailed justifications and feedback. This is going out to the question submitter so please use proper language and be respectful.
 - Explanations should include at least some details or reference. If the rationale is unclear or not detailed, ask in the review to expand a bit.
 - Please check if the answer makes sense as a possible response to the question, but if you do not have knowledge/context, or if it would take more than 5 minutes to solve, that is okay.
- Please prioritize questions with no reviews and skip all questions with more than 3 reviews.
- Please double check that the model did actually answer the question wrong.
 - Sometimes the exact match feature does not work well enough, and there are false negatives. We have to discard any exact match questions that a model got right.
- On the HLE dashboard, look at least 10 examples reviewed by the organizers before starting to review, and review the examples from training.
- The average time estimated to review a question 3-5 minutes.
- Use a “-1 Unsure” review if the person submitting seems suspicious or if you’re not convinced their answer is right.

Score	Scoring Guideline	Description
0	Discard	The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.
1	Major Revisions Needed	Major revisions are needed for this question or the question is too easy and simple.
2	Some Revisions Needed	Difficulty and expertise required to answer the question is borderline. Some revisions are needed for this question.
3	Okay	The question is sufficiently challenging but the knowledge required is not graduate-level nor complex. Minor revisions may be needed for this question.
4	Great	The knowledge required is at the graduate level or the question is sufficiently challenging.
5	Top-Notch	Question is top-notch and perfect.
Unsure	-	Reviewer is unsure if the question fits the HLE guidelines, or unsure if the answer is right.

B.2.2 Review Round 2

To thoroughly refine our dataset, we train a set of reviewers along with organizers to pick the best questions. These reviewers are identified by organizers from round 1 reviews as particularly high quality and thorough in their feedback. Different than the first round of reviews, reviewers are asked to grade both the question and look at feedback from round 1 reviewers. Organizers then approve questions based on reviewer feedback in this round. We employ a new rubric for this round below.

Score	Scoring Guideline	Description
0	Discard	The question is out of scope, not original, spam, or otherwise not good enough to be included in the HLE set and should be discarded.
1	Not sure	Major revisions are needed for this question or you're just unsure about the question. Please put your thoughts in the comment box and an organizer will evaluate this.
2	Pending	You believe there are still minor revisions that are needed on this question. Please put your thoughts in the comment box and an organizer will evaluate this.
3	Easy questions models got wrong	These are very basic questions that models got correct or the question was easily found online. Any questions which are artificially difficult (large calculations needing a calculator, requires running/rendering code, etc.) should also belong in this category. The models we evaluate cannot access these tools, hence it creates an artificial difficulty bar. Important: "Found online" means via a simple search online. Research papers/journals/books are fine
4	Borderline	The question is not interesting OR The question is sufficiently challenging, but 1 or more of the models got the answer correct.
5	Okay to include in HLE benchmark	Very good questions (usually has score of 3 in the previous review round). You believe it should be included in the HLE Benchmark.
6	Top question in its category	Great question (usually has a score of 4-5 in the previous review round), at a graduate or research level. Please note that "graduate level" is less strict for Non-STEM questions. For Non-STEM questions and Trivia, they are fine as long as they are challenging and interesting.

B.3 Subject List

We allow question contributors to choose or declare a subject the author felt best suited their question. We present the top fifty most popular subjects in HLE below, although we note there are over a hundred subjects in the overall dataset.

Mathematics, Physics, Computer Science, Chemistry, Applied Mathematics, Trivia, Electrical Engineering, Biology, Linguistics, Medicine, Genetics, History, Economics, Ecology, Artificial Intelligence, Musicology, Philosophy, Neuroscience, Law, Art History, Biochemistry, Astronomy, Classics, Chess, Chemical Engineering, Microbiology, Classical Ballet, Materials Science, Poetry, Quantum Mechanics, Aerospace Engineering, Civil Engineering, Mechanical Engineering, Geography, Robotics, Data Science, Molecular Biology, Statistics, Immunology, Education, Logic, Computational Biology, Psychology, English Literature, Machine Learning, Puzzle, Cultural Studies, Marine Biology, Archaeology, and Biophysics.

C Evaluation

C.1 Prompts

C.1.1 Evaluation

We use the following system prompt for evaluating LLMs on multiple-choice questions:

```
Your response should be in the following format:  
Explanation: {your explanation for your answer choice}  
Answer: {your chosen answer}  
Confidence: {your confidence score between 0% and 100% for your answer}
```

We use the following system prompt for evaluating LLMs on exact-match questions:

```
Your response should be in the following format:  
Explanation: {your explanation for your final answer}  
Exact Answer: {your succinct, final answer}  
Confidence: {your confidence score between 0% 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 gpt-4o-2024-08-06 with structured decoding enabled to get an extracted_final_answer, reasoning, correct, confidence extraction for each output.

```
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 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.

C.2 Text-Only Results

Model	Accuracy (%) \uparrow	Calibration Error (%) \downarrow
GPT-4o	2.9	90.4
GROK 2	3.9	92.5
CLAUDE 3.5 SONNET	4.2	87.0
GEMINI 1.5 PRO	4.8	91.1
GEMINI 2.0 FLASH THINKING	5.9	92.1
o1	8.9	92.0
DEEPSEEK-R1	9.4	81.8

Table 2: Accuracy and RMS calibration error of models from Table 1 on the text-only questions of HLE, representing 90% of the public set.

C.3 Non-Reasoning Model Token Counts

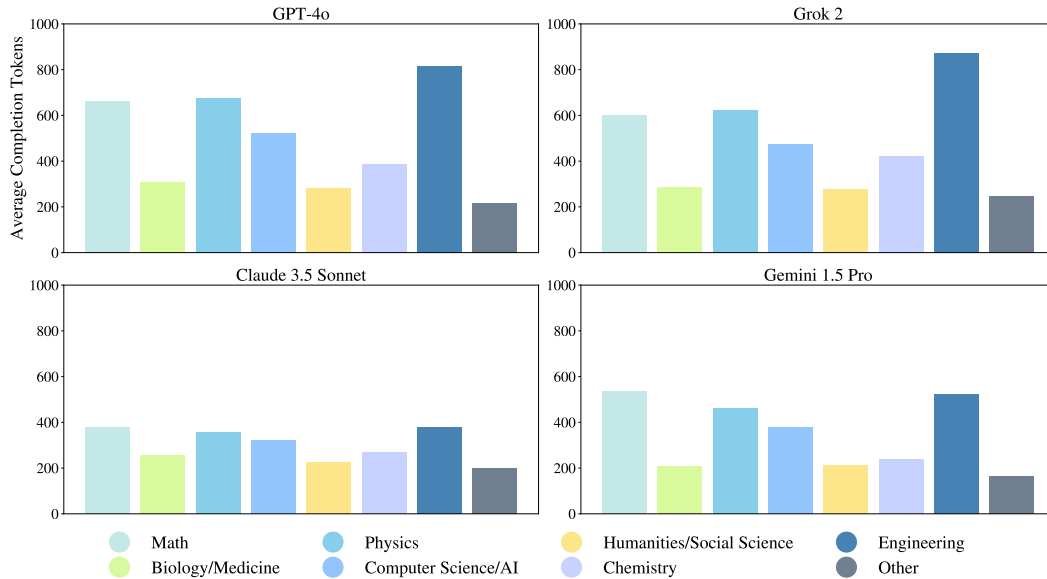


Figure 6: Average output token counts of non-reasoning models.

C.4 Model Versions

Model	Version
GPT-4o	gpt-4o-2024-11-20
GROK 2	grok-2-latest
CLAUDE 3.5 SONNET	claude-3-5-sonnet-20241022
GEMINI 1.5 PRO	gemini-1.5-pro-002
GEMINI 2.0 FLASH THINKING	gemini-2.0-flash-thinking-exp-1219
o1	o1-2024-12-17
DEEPSEEK-R1	January 20, 2025 release

Table 3: Evaluated model versions. All models use temperature 0 when configurable.

C.5 Benchmark Difficulty Comparison

In Figure 1, we evaluate the accuracy of all models on HLE using our zero-shot chain-of-thought prompts (Appendix C.1.1). On prior benchmarks, we list our sources here.

For GPT-4o and O1-PREVIEW, we report zero-shot, chain-of-thought results from OpenAI found at <https://github.com/openai/simple-evals>.

For GEMINI 1.5 PRO, we report 5-shot MMLU Team et al. [49] and other results from [Google’s reported results here](#).

For CLAUDE 3.5 SONNET, we report 0-shot chain-of-thought results from Anthropic [4].