Humanity's Last Exam

Organizing Team

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

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

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

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

Dataset Contributors

Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav Kazakov, 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 Gehrunger, 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, Lianghui Li, Sumeet Motwani, Christian Schroeder de Witt, Edwin Taylor, Johannes Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Justine Leon Uro, 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, Dakotah 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 Mishra, Ariel Ghislain Kemogne Kamdoum, Tobias Kreiman, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Tim Tarver, Haline Heidinger, 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, Ilia 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, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Alan Goldfarb, Sergey Ivanov, Rafał 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 Efren 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 Piccardo, Ferenc Jeanplong, Niv Cohen, Josef Tkadlec, Paul Rosu, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, John Arnold Ambay, Abhishek Shukla, Yan Carlos Leyva Labrador, Alan Givré, Hew Wolff, Vivien Rossbach, Muhammad Fayez 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, Kelin 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. Cunningham, Vladimir

[*]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.

Goryachev, 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 Sztyber-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 Raparthi, Pablo Hernández-Cámara, Freddie Martin, Dmitry Malishev, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ziye 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, Tim Santens, Khalida Meer, Harrison K Wang, Kalyan Ramakrishnan, Evan Chen, Alessandro Tomasiello, G. Bruno De Luca, 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, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Carter Harris, Jason Gross, Ilya Gusev, Asankhaya Sharma, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayezi, 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 Đuc Huy, Ruicheng Xian, Armel 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, Bita 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, Ethan Delaney, 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 Mariji, Rasoul Pouriamanesh, Wentao Wu, Gözdenur Demir, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Hsiaoyun Milliron, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Abdelkader Dendane, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Russell Campbell, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I.M.J. McInnis, Alex Hoover, Sören Möller, Song Bian, John Lai, 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 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.

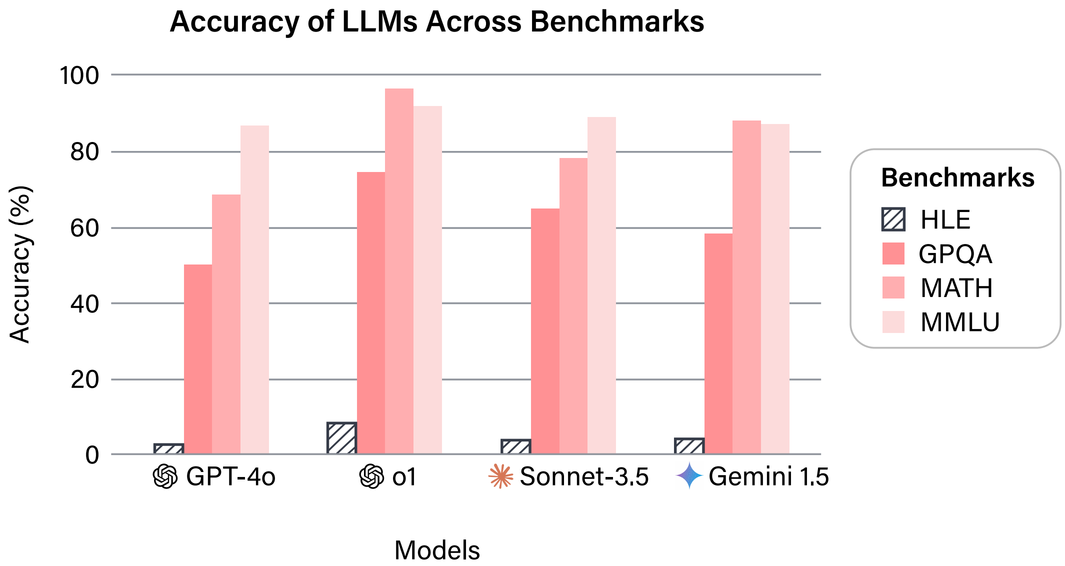


Figure 1: Compared against the saturation of some existing benchmarks, HUMANITY'S LAST EXAM accuracy remains low across several frontier models, demonstrating its effectiveness for measuring advanced, closed-ended, academic capabilities. The sources for our evaluation metrics are detailed in Appendix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

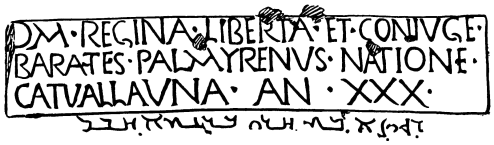
HUMANITY'S LAST EXAM (HLE) consists of 3,000 challenging questions across over a hundred subjects across. A high level summary is provided in Figure 3. We publicly release these questions, while maintaining a private test set of held out questions to assess model overfitting.

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, orginally found on a tombstone. Provide a translation for the Palmyrene script.
A transliteration of the text is provided: RGYNᵓ BT HRY BR ᶜTᵓ 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 $F, G : C \to D$ 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?

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.

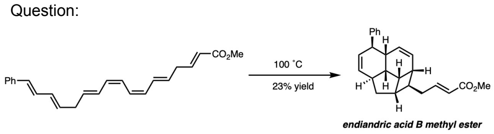
Which of the following graph classes is well-behaved?

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:



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־גַּעֲרָתְךָ֣ יְנוּס֑וּן מִן־ק֥וֹל רַ֝עַמְךָ֗ יֵחָפֵזֽוּן (Psalms 104:7) ?

Lina B
University of Cambridge

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

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 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

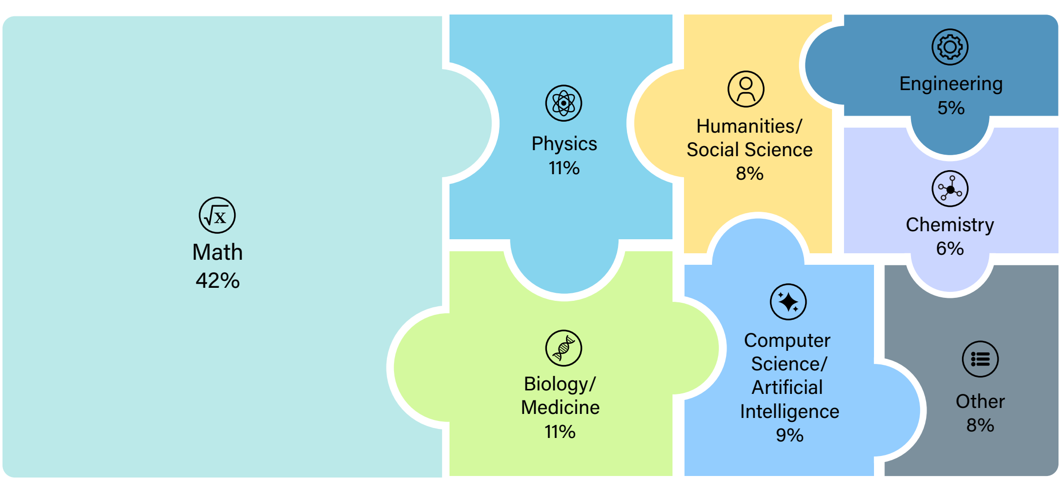


Figure 3: HLE consists of 3,000 exam questions in over a hundred subjects, grouped into high level categories here. We provide a more detailed list of subjects in Appendix B.3.

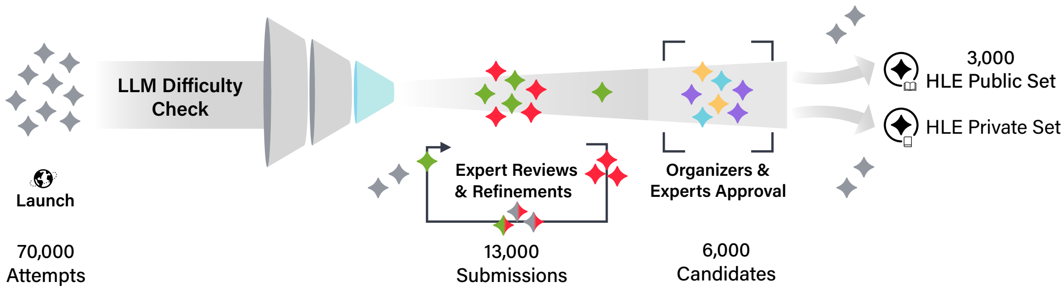


Figure 4: Dataset creation pipeline. We accept questions that make frontier LLMs fail, then iteratively refine them with the help of expert peer reviewers. Each question is then manually approved by organizers or expert reviewers trained by organizers. A private held-out set is kept in addition to the 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 (%) ↑	Calibration Error (%) ↓
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.

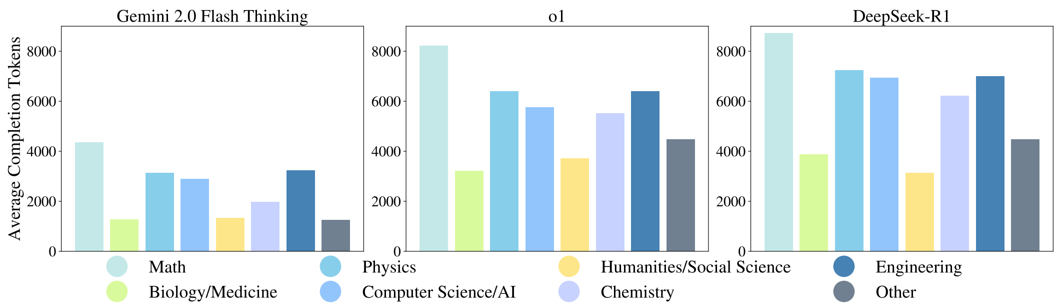


Figure 5: Average completion token counts of reasoning models tested, including both reasoning and output tokens. We also plot average token counts for non-reasoning models in Appendix C.3.

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. Mądry. 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. Omni-math: 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. Järviniemi, 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. andMadhu Gurumurthy, M. Aaron, M. Ambar, R. Fellinger, R. Wang, R. Sims, Z. Zhang, S. Goldshtein, 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. Rodriques. 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 bioscience 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. Lukošiūtė, 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=5lcPe6DqfI.

[47] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, 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. Askell, 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. Ericheva, 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 Anderson[3], Tung Nguyen[4], Mobeen Mahmood[5], Fiona Feng[6], Steven Y. Feng[7], Haoran Zhao[8], Michael Yu[3], Varun Gangal[3], Chelsea Zou[7], Zihan Wang[9], Jessica P. Wang[10], Pawan Kumar[11], Oleksandr Pokutnyi[12], Robert Gerbicz[13], Serguei Popov[14], John-Clark Levin[15], Mstyslav Kazakov[16], Johannes Schmitt[17], Geoff Galgon[18], Alvaro Sanchez[3], Yongki Lee[19], Will Yeadon[20], Scott Sauers[21], Marc Roth[22], Chidozie Agu[23], Søren Riis[22], Fabian Giska[3], Saiteja Utpala[24], Zachary Giboney[25], Gashaw M. Goshu[3], Joan of Arc Xavier[26], Sarah-Jane Crowson[27], Mohinder Maheshbhai Naiya[28], Noah Burns[7], Lennart Finke[17], Zerui Cheng[29], Hyunwoo Park[30], Francesco Fournier-Facio[15], John Wydallis[3], Mark Nandor[3], Ankit Singh[31], Tim Gehrunger[17], Jiaqi Cai[32], Ben McCarty[33], Darling Duclosel[34], Jungbae Nam[35], Jennifer Zampese[36], Ryan G. Hoerr[37], Aras Bacho[38], Gautier Abou Loume [39,40], Abdallah Galal[41], Hangrui Cao[30], Alexis C Garretson[42,43], Damien Sileo[44], Qiuyu Ren[45], Doru Cojoc[46], Pavel Arkhipov[47], Usman Qazi[48,49], Lianghui Li[50], Sumeet Motwani[51], Christian Schroeder de Witt[51], Edwin Taylor[3], Johannes Veith[52,53], Eric Singer[54], Taylor D. Hartman[55], Paolo Rissone[56], Jaehyeok Jin[46], Jack Wei Lun Shi[57], Chris G. Willcocks[20], Joshua Robinson[58], Aleksandar Mikov[50], Ameya Prabhu[59], Longke Tang[29], Xavier Alapont[26], Justine Leon Uro[3], Kevin Zhou[45], Emily de Oliveira Santos[60], Andrey Pupasov Maksimov[61], Edward Vendrow[32], Kengo Zenitani[3], Julien Guillod[62,63], Yuqi Li[64], Joshua Vendrow[32], Vladyslav Kuchkin [65], Ng Ze-An[66], Pierre Marion[50], Denis Efremov[67], Jayson Lynch[32], Kaiqu Liang[29], Andrew Gritsevskiy[68], Dakotah Martinez[3], Ben Pageler[3], Nick Crispino[69], Dimitri Zvonkine[70,71], Natanael Wildner Fraga[3], Saeed Soori[72], Ori Press[59], Henry Tang[51], Julian Salazar[73], Sean R. Green[3], Lina Brüssel[15], Moon Twayana[74], Aymeric Dieuleveut[75], T. Ryan Rogers[76], Wenjin Zhang[69], Bikun Li[77], Jinzhou Yang[78], Arun Rao[79], Gabriel Loiseau[44], Mikhail Kalinin[80], Marco Lukas[81], Ciprian Manolescu[7], Subrata Mishra[82], Ariel Ghislain Kemogne Kamdoum[83], Tobias Kreiman[45], Tad Hogg[84], Alvin Jin[32], Carlo Bosio[45], Gongbo Sun[85], Brian P Coppola[86], Tim Tarver[87], Haline Heidinger[88,89], Rafael Sayous[71], Stefan Ivanov[15], Joseph M Cavanagh[45], Jiawei Shen[69], Joseph Marvin Imperial[90,91], Philippe Schwaller[50], Shaipranesh Senthilkuma[50], Andres M Bran[50], Ali Dehghan[3], Andres Algaba[92], Brecht Verbeken[92], David Noever[93], Ragavendran P V[3], Lisa Schut[51], Ilia Sucholutsky[94], Evgenii Zheltonozhskii[95], Derek Lim[32], Richard Stanley[32,96], Shankar Sivarajan [97], Tong Yang[30], John Maar[98], Julian Wykowski[15], Martí Oller[15], Jennifer Sandlin[99], Anmol Sahu[3], Yuzheng Hu[100], Sara Fish[101], Nasser Heydari[3], Archimedes Apronti[102], Kaivalya Rawal[51], Tobias Garcia Vilchis[103], Yuexuan Zu[32], Martin Lackner[104], James Koppel[3], Jeremy Nguyen[105], Daniil S. Antonenko[106], Steffi Chern[30], Bingchen Zhao[107], Pierrot Arsene[108], Alan Goldfarb[45], Sergey Ivanov[3], Rafał Poświata[109], Chenguang Wang[69], Daofeng Li[69], Donato Crisostomi[56], Andrea Achilleos[110], Benjamin Myklebust[111], Archan Sen[45], David Perrella[112], Nurdin Kaparov[113], Mark H Inlow[114], Allen Zang[77], Elliott Thornley[115], Daniil Orel[116], Vladislav Poritski[3], Shalev Ben-David[117], Zachary Berger[32], Parker Whitfill[32], Michael Foster[3], Daniel Munro[9], Linh Ho[3], Dan Bar Hava[118], Aleksey Kuchkin[3], Robert Lauff[98], David Holmes[119], Frank Sommerhage[120], Keith Schneider[3], Zakayo Kazibwe[121], Nate Stambaugh[122], Mukhwinder Singh[123], Ilias Magoulas[124], Don Clarke[125], Dae Hyun Kim[126], Felipe Meneguitti Dias[60], Veit Elser[127], Kanu Priya Agarwal[3], Victor Efren Guadarrama Vilchis[128], Immo Klose[46], Christoph Demian[53], Ujjwala Anantheswaran[99], Adam Zweiger[32], Guglielmo Albani[129], Jeffery Li[32], Nicolas Daans[130], Maksim Radionov[131], Václav Rozhoň[132], Ziqiao Ma[86], Christian Stump[133], Mohammed Berkani[134], Jacob Platnick[135], Volodymyr Nevirkovets[136], Luke Basler[137], Marco Piccardo[138], Ferenc Jeanplong[139], Niv Cohen[94], Josef Tkadlec[140], Paul Rosu[141], Piotr Padlewski[3], Stanislaw Barzowski[3], Kyle Montgomery[69], Aline Menezes[3], Arkil Patel[5,142], Zixuan Wang[29], Jamie Tucker-Foltz[101], Jack Stade[143], Tom Goertzen[144], Fereshteh Kazemi[3], Jeremiah Milbauer[30], John Arnold Ambay[145], Abhishek Shukla[146], Yan Carlos Leyva Labrador[26], Alan Givré[147], Hew Wolff[3], Vivien Rossbach [26], Muhammad Fayez Aziz[100], Younesse Kaddar[51], Yanxu Chen[148], Robin Zhang[32], Jiayi Pan[45], Antonio Terpin[17], Niklas Muennighoff[7], Hailey Schoelkopf[3], Eric Zheng[30], Avishy Carmi[149], Adam Jones[3], Jainam Shah[150], Ethan D. L. Brown[151], Kelin Zhu[97], Max Bartolo[152], Richard Wheeler[107], Andrew Ho[153], Shaul Barkan[154], Jiaqi Wang[8], Martin Stehberger[3], Egor Kretov[155], Kaustubh Sridhar[156], Zienab EL-Wasif[157], Anji Zhang[32], Daniel Pyda[158], Joanna Tam[159], David M. Cunningham[160], Vladimir Goryachev[3], Demosthenes Patramanis[51], Michael Krause[161], Andrew Redenti[46], Daniel Bugas[3], David Aldous[45], Jesyin Lai[162], Shannon Coleman[49], Mohsen Bahaloo[163], Jiangnan Xu[164], Sangwon Lee[3], Sandy Zhao[26], Ning Tang[45], Michael K. Cohen[45], Micah Carroll[45], Orr Paradise[45], Jan Hendrik Kirchner[165], Stefan Steinerberger[8], Maksym Ovchynnikov[166], Jason O. Matos[159], Adithya Shenoy[3], Benedito Alves de Oliveira Junior[60], Michael Wang[45], Yuzhou Nie[167], Paolo Giordano[168], Philipp Petersen[168], Anna Sztyber-Betley[169], Priti Shukla[170], Jonathan Crozier[171], Antonella Pinto[172], Shreyas Verma[173], Prashant Joshi[174], Zheng-Xin Yong[175], Allison Tee[7], Jérémy Andréoletti[63], Orion Weller[176], Raghav Singhal[116], Gang Zhang[3], Alexander Ivanov[177], Seri Khoury[132], Hamid Mostaghimi[83], Kunvar Thaman[178], Qijia Chen[101], Tran Quoc Khánh[179], Jacob Loader[15], Stefano Cavalleri[180], Hannah Szlyk[69], Zachary Brown[32], Jonathan Roberts[15], William Alley[3], Kunyang Sun[45], Ryan Stendall[181], Max Lamparth[7], Anka Reuel[7], Ting Wang[69], Hanmeng Xu[106], Sreenivas Goud Raparthi[182], Pablo Hernández-Cámara[183], Freddie Martin[3], Dmitry Malishev[3], Thomas Preu[184], Tomek Korbak[185], Marcus Abramovitch[3], Dominic Williamson[144], Ziye Chen[186], Biró Bálint[3], M Saiful Bari[187], Peyman Kassani[188], Zihao Wang[77], Behzad Ansarinejad[3], Laxman Prasad Goswami[146], Yewen Sun[189], Hossam Elgnainy[190], Daniel Tordera[191], George Balabanian[156], Earth Anderson[192], Lynna Kvistad[193], Alejandro José Moyano[194], Rajat Maheshwari [195], Ahmad Sakor[81], Murat Eron[196], Isaac C. McAlister[3], Javier Gimenez[26], Innocent Enyekwe[3], Andrew Favre D.O.[197], Shailesh Shah[198], Xiaoxiang Zhou[53], Firuz Kamalov[199], Ronald Clark[51], Sherwin Abdoli[172], Tim Santens[15], Khalida Meer[26], Harrison K Wang[101], Kalyan Ramakrishnan[51], Evan Chen[32], Alessandro Tomasiello[200], G. Bruno De Luca[7], Shi-Zhuo Looi[38], Vinh-Kha Le[45], Noam Kolt[154], Niels Mündler[17], Avi Semler[51], Emma Rodman[201], Jacob Drori[3], Carl J Fossum[202], Milind Jagota[45], Ronak Pradeep[117], Honglu Fan[203], Tej Shah[204], Jonathan Eicher [205], Michael Chen[38], Kushal Thaman[7], William Merrill[94], Carter Harris[206], Jason Gross[3], Ilya Gusev[3], Asankhaya Sharma[207], Shashank Agnihotri[208], Pavel Zhelnov[72], Siranut Usawasutsakorn[209], Mohammadreza Mofayezi[72], Sergei Bogdanov[210], Alexander Piperski[211], Marc Carauleanu[212], David K. Zhang[7], Dylan Ler[3], Roman Leventov[213], Ignat Soroko[74], Thorben Jansen[214], Pascal Lauer[215,216], Joshua Duersch[217], Vage Taamazyan[218], Wiktor Morak[3], Wenjie Ma[45], William Held[7,135], Tran Đuc Huy[219], Ruicheng Xian[100], Armel Randy Zebaze[220], Mohanad Mohamed[221], Julian Noah Leser[104], Michelle X Yuan[3], Laila Yacar[222], Johannes Lengler[17], Hossein Shahrtash[223], Edson Oliveira[224], Joseph W. Jackson[225], Daniel Espinosa Gonzalez[167], Andy Zou[30,226], Muthu Chidambaram[141], Timothy Manik[3], Hector Haffenden[3], Dashiell Stander[227], Ali Dasouqi[176], Alexander Shen[228], Emilien Duc[17], Bita Golshani[3], David Stap[148], Mikalai Uzhou[229], Alina Borisovna Zhidkovskaya[230], Lukas Lewark[17], Mátyás Vincze[231,232], Dustin Wehr[3], Colin Tang[30], Zaki Hossain[233], Shaun Phillips[3], Jiang Muzhen[3], Fredrik Ekström[3], Angela Hammon[3], Oam Patel[101], Nicolas Remy[234], Faraz Farhidi[235], George Medley [3], Forough Mohammadzadeh[3], Madellene Peñaflor[236], Haile Kassahun[5], Alena Friedrich[237], Claire Sparrow[77], Taom Sakal[167], Omkar Dhamane[238], Ali Khajegili Mirabadi[49], Eric Hallman[3], Mike Battaglia[3], Mohammad Maghsoudimehrabani[239], Hieu Hoang[240], Alon Amit[241], Dave Hulbert[3], Roberto Pereira[242], Simon Weber[17], Stephen Mensah[243], Nathan Andre[244], Anton Peristyy[3], Chris Harjadi[7], Himanshu Gupta [99], Stephen Malina[245], Samuel Albanie[3], Will Cai[45], Mustafa Mehkary [72,246], Frank Reidegeld[3], Anna-Katharina Dick[59], Cary Friday[247], Jasdeep Sidhu[3], Wanyoung Kim[248], Mariana Costa[26], Hubeyb Gurdogan[79], Brian Weber[249], Harsh Kumar [250], Tong Jiang[101], Arunim Agarwal[251], Chiara Ceconello[3], Warren S. Vaz[3], Chao Zhuang[3], Haon Park[252,253], Andrew R. Tawfeek[8], Daattavya Aggarwal[15], Michael Kirchhof[59], Linjie Dai[32], Evan Kim[32], Johan Ferret[73], Yuzhou Wang[135], Minghao Yan[85], Krzysztof Burdzy[8], Lixin Zhang[26], Antonio Franca[15], Diana T. Pham[254], Kang Yong Loh[7], Joshua Robinson[255], Shreen Gul[256], Gunjan Chhablani[135], Zhehang Du[156], Adrian Cosma[257], Colin White[258], Robin Riblet[108], Prajvi Saxena[259], Jacob Votava[29], Vladimir Vinnikov[3], Ethan Delaney[260], Shiv Halasyamani[261], Syed M. Shahid[262], Jean-Christophe Mourrat[70,263], Lavr Vetoshkin[264], Renas Bacho[265], Vincent Ginis[92,101], Aleksandr Maksapetyan[26], Florencia de la Rosa[266], Xiuyu Li[45], Guillaume Malod[267], Leon Lang[148], Julien Laurendeau[50], Fatimah Adesanya [26,268], Julien Portier[15], Lawrence Hollom[15], Victor Souza[15], Yuchen Anna Zhou[269], Yiğit Yalın[270], Gbenga Daniel Obikoya[3], Luca Arnaboldi[50], Rai (Michael Pokorny)[271], Filippo Bigi[50], Kaniuar Bacho[107], Pierre Clavier[272], Gabriel Recchia[273], Mara Popescu[274], Nikita Shulga[275], Ngefor Mildred Tanwie [276], Thomas C.H. Lux[277], Ben Rank[3], Colin Ni[79], Alesia Yakimchyk[278], Huanxu (Quinn) Liu [279], Olle Häggström[280], Emil Verkama[281], Himanshu Narayan [3], Hans Gundlach[32], Leonor Brito-Santana[282], Brian Amaro[7], Vivek Vajipey[7], Rynaa Grover[135], Yiyang Fan[3], Gabriel Poesia Reis e Silva[7], Linwei Xin[77], Yosi Kratish[136], Jakub Łucki[17], Wen-Ding Li[127], Justin Xu[51], Kevin Joseph Scaria[99], Freddie Vargus[283], Farzad Habibi[284], Long (Tony) Lian[45], Emanuele Rodolà[56], Jules Robins[3], Vincent Cheng[9], Declan Grabb[7], Ida Bosio[285], Tony Fruhauff[3], Ido Akov[286], Eve J. Y. Lo[287], Hao Qi[186], Xi Jiang[77], Ben Segev[46], Jingxuan Fan[101], Sarah Martinson[101], Erik Y. Wang[101], Kaylie Hausknecht[101], Michael P. Brenner[101], Mao Mao[186], Yibo Jiang[77], Xinyu Zhang[186], David Avagian[208], Eshawn Jessica Scipio[288], Muhammad Rehan Siddiqi[289,290], Alon Ragoler[291], Justin Tan[15], Deepakkumar Patil[292], Rebeka Plecnik[3], Aaron Kirtland[175], Roselynn Grace Montecillo[293], Stephane Durand[294], Omer Faruk Bodur[3], Zahra Adoul[295], Mohamed Zekry [296], Guillaume Douville[26], Ali Karakoc[297], Tania C. B. Santos[3], Samir Shamseldeen[298], Loukmane Karim[246], Anna Liakhovitskaia[299], Nate Resman [300], Nicholas Farina[26], Juan Carlos Gonzalez[301], Gabe Maayan[186], Sarah Hoback[101], Rodrigo De Oliveira Pena[302], Glen Sherman[26], Hodjat Mariji[3], Rasoul Pouriamanesh[3], Wentao Wu[49], Gözdenur Demir[3], Sandra Mendoza[303,304], Ismail Alarab[305], Joshua Cole[306], Danyelle Ferreira[26], Bryan Johnson [307], Hsiaoyun Milliron[308], Mohammad Safdari[309], Liangti Dai[51], Siriphan Arthornthurasuk[26], Alexey Pronin[310], Jing Fan[274], Angel Ramirez-Trinidad[3], Ashley Cartwright[311], Daphiny Pottmaier[312], Omid Taheri[313], David Outevsky[314], Stanley Stepanic[315], Samuel Perry[3], Luke Askew[316], Raúl Adrián Huerta Rodríguez [3], Abdelkader Dendane[26], Sam Ali[58], Ricardo Lorena[317], Krishnamurthy Iyer[318], Sk Md Salauddin[319], Murat Islam[320], Juan Gonzalez[3], Josh Ducey[321], Russell Campbell[322], Maja Somrak[3], Vasilios Mavroudis[323], Eric Vergo[3], Juehang Qin[324], Benjámin Borbás[325], Eric Chu[73], Jack Lindsey[165], Anil Radhakrishnan[171], Antoine Jallon[3], I.M.J. McInnis[3], Alex Hoover[77], Sören Möller[326], Song Bian[85], John Lai[26], Tejal Patwardhan[271]

Affiliations

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

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

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

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

300. University of Oklahoma
301. Jala University
302. Florida Atlantic University
303. CONICET
304. Universidad Tecnológica Nacional
305. Bournemouth University
306. University of Warwick
307. University of Alabama Huntsville
308. Van Andel Institute
309. University of Hertfordshire
310. Central College
311. Sheffield Teaching Hospitals NHS Foundation Trust
312. Nottingham Trent University
313. Max Planck Institute for Intelligent Systems
314. Outevsky Bespoke Dance Education
315. University of Virginia
316. Dartmouth College
317. INESC Microsistemas e Nanotecnologias
318. University of Minnesota
319. Aligarh Muslim University
320. John Crane UK Ltd
321. James Madison University
322. University of the Fraser Valley
323. Alan Turing Institute
324. Rice University
325. HUN-REN
326. 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 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 (%) ↑	Calibration Error (%) ↓
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

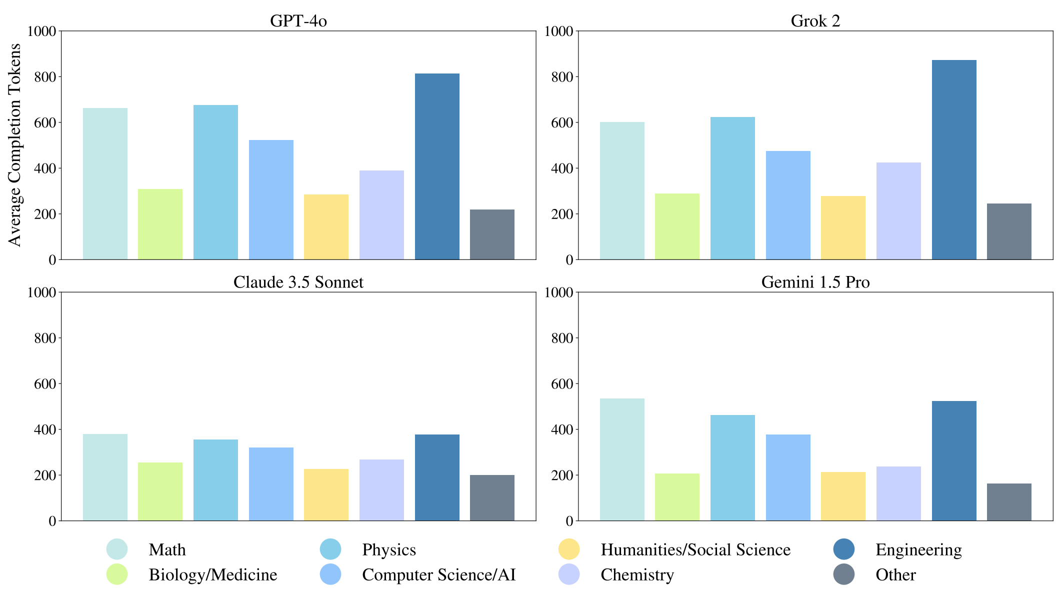


Figure 6: Average output token counts of non-reasoning models.

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].