nature portfolio

Peer Review File

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

Corresponding Author: Center for AI Safety .

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
This paper presents a new benchmark for evaluating AI systems, "Humanity's Last Exam" (HLE for short). The benchmark consists of crowdsourced multiple choice and short answer question across a variety of domains, with a particular emphasis on mathematics. Several current state-of-the-art machine learning systems are evaluated on this benchmark, both for overall performance and calibration (ie. the extent to which their stated confidence aligns with the probability of success in answering a question). These models are able to solve only a fraction of the problems, suggesting that it will serve as an effective benchmark for at least the next generation of models.

This paper is generally a well-executed piece of work designing an impressive benchmark for AI models. It follows in a tradition of collaborative creation of benchmarks, so the approach is not particularly innovative, but the design of the collaboration was done with care and the results appear to be of high quality. The paper is likely to be highly cited, as this is a field in which many papers are published every day.

That said, I'm not sure this work is appropriate for publication in Nature, and it may make more sense for it to be published in a more specialized venue. My major reason for making this argument is that while the benchmark is impressive, it does not provide any scientific insight into the operation of these systems. To provide an analogy, if a psychologist came up with a new kind of intelligence test, that probably wouldn't be appropriate for publication in Nature either. However, scientific contributions that related the results of people taking that new test to other aspects of human behavior or life outcomes might be, as might work that connected different dimensions of intelligence assessed by the test to different parts of the human brain, or that designed interventions based on the test that could help people succeed in performing tasks they otherwise couldn't perform. In the AI setting, the benchmark itself doesn't answer any scientific questions. However, it seems like it sets up a lot of opportunities to answer questions that the authors haven't pursued, and the answers to those questions might be something that it makes sense to publish in Nature.

Specifically, I think the scientific questions here are things like:

1. Can we understand why some models outperform others? Can we trace that to aspects of the underlying architecture?

2. Can we understand why some questions are easier than others? Can we find mechanisms within the AI models that allow them to answer these questions?

3. Are the questions that are easy for one model easy for other models? Is there something in the training data that could explain this? Are specific domains easier than others? How does this compare to human performance across these domains?

4. For the reasoning models, are the number of tokens generated correlated with model success on a per-question basis?

5. Are there any interventions that improve the performance of the models? Since reasoning models generally seem to produce higher performance, can similar levels be reached by other models using appropriate prompts (e.g. chain of thought prompts)?

The current dataset provides an opportunity to engage with these questions at scale, and the results would provide us insight into the models -- and how to improve upon them -- rather than just a measure of how they are doing on what is honestly a fairly arbitrary set of questions. The current paper feels to me like a missed opportunity to do some interesting science.

There are several other minor issues:

The current format is more consistent with an AI conference submission than a Nature paper (in fact it looks like it was intended to be submitted to the NeurIPS conference).

MMLU should be spelled out in the abstract to increase accessibility to a general audience.

Figure 1 omits the reasoning models which actually show the highest performance on this task. They should be included. If there are non-negligible error bars on the accuracy scores they should also be shown.

The non-zero accuracy scores summarized in Section 4.2 are interesting and might be worth digging into further. Research in education and psychometrics has developed a variety of ways for correcting for guessing that could potentially be used to clean up these results.

Error bars or equivalent should be reported for Table 1 and Figure 5.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
This paper introduces a new benchmark called Humanity's Last Exam (HLE), which is designed to evaluate the capabilities of language models at the frontier of human academic knowledge. It comprises 2,500 challenging, multi-modal questions across numerous academic subjects, created by global experts. The benchmark aims to provide a more accurate measure of frontier LLM capabilities by presenting questions that cannot be easily answered via internet search or simple reasoning, requiring deep knowledge and complex problem-solving skills. Evaluations show that even state-of-the-art models perform poorly and are poorly calibrated on HLE, which indicates a significant gap between current AI performance and expert human knowledge in these domains. HLE offers a high-fidelity signal for assessing model capabilities and is publicly released to support more informed research and policy decisions.

Strengths:

- HLE addresses the problem of benchmark saturation. Many popular benchmarks have become inadequate for evaluating frontier LLMs. HLE is designed to be significantly harder, targeting the frontier of human knowledge to better assess advanced model capabilities.
- The benchmark is developed by subject-matter experts globally. Specifically, HLE questions are created by academics and domain experts from over 500 institutions across 50 countries. Each question undergoes rigorous multi-stage review by graduate-level and expert reviewers to ensure high difficulty, precision, and relevance.
- The benchmark includes 2,500 original, unambiguous questions across a broad range of academic disciplines, including mathematics, humanities, and natural sciences. Questions are designed to be resistant to simple internet lookup or database retrieval.
- HLE provides a clear measure of the gap between current LLM capabilities and expert-level academic performance.

Weaknesses:

- The framing of the benchmark as "humanity's last exam" feels a bit strong to me. While the benchmark is pitched as "multi-modal", only around 14% of questions (around 350 questions) require comprehending both text and an image. Additionally, it is limited to closed-ended, verifiable academic questions, which restricts its ability to assess broader academic capabilities such as open-ended reasoning, synthesis, or multi-step problem solving across richer modalities.

- The paper would benefit from deeper analysis beyond headline accuracy and calibration error. Currently, there are no ablation studies or breakdowns that help explain model failure modes. For example, it is unclear how performance varies across question types, difficulty levels, or subject domains, or whether models perform worse on problems ranked as harder by the organizers. Additional analysis, such as the effect of test-time scaling, prompt variants, or different prompting strategies would strengthen the paper's empirical value and help the community better understand what makes HLE challenging.

- Reviewers of the questions were instructed not to verify full solution details if doing so would take more than five minutes. As a result, organizers launched a post-release bug bounty and community audit to catch major errors in the dataset. These after-the-fact corrections indicate that, despite multi-layer vetting, non-trivial noise may still remain and confound benchmarking results.

Minor comment(s):

In Table 1's caption, it is unclear to me why low accuracy combined with high calibration error is interpreted as evidence of hallucination ("indicative of hallucination"). A model selecting an incorrect answer with high confidence may reflect limitations in problem-solving or reasoning ability, rather than hallucination per se.

(Remarks on code availability)
The GitHub repository includes a link to the dataset, along with simple evaluation scripts to generate model responses and judge results.

Referee #3

(Remarks to the Author)
Benchmarks are critical to understand AI system capabilities. However, benchmarks targeted at assessing academic knowledge have been largely saturated, rendering their utility for measuring AI systems futile. The authors aim to address this issue by curating and releasing a new, large-scale dataset to test whether models' capabilities on expert level knowledge across a range of domains.

I applaud the authors on their mammoth data collection effort. It is clear the authors developed an extensive evaluation procedure for question verification. The collected dataset ("Humanity's Last Exam", HLE) is no doubt a significant advance for the AI community – and broader academic communities touched by AI that want to understand current and future model capabilities.

However, I think there are some substantive issues with the framing of this benchmark and several key properties of HLE and its curation process that are absent from the current manuscript.

First, the authors frame HLE as "the final" academic benchmark ("HLE may be the last academic exam we need to give models"). I think this is highly problematic language. Researchers are continually discovering new knowledge – don't we want to probe whether models (and people) understand new knowledge over time? Moreover, the topic distribution that the authors collect is highly skewed towards STEM. Only 9% of the questions are in the humanities. If this is meant to be "humanity's last exam," I would think we want a far higher proportion of humanities-related questions (given the many centuries of the humanities, and ability for such questions to also probe critical thinking). I assume this is in part from the recruited pool (it is likely easier for AI researchers to be in touch with other STEM researchers!). However, pool bias was not substantively discussed – and the broad generality of framing the benchmark as "humanity's last exam" I think is a touch too hype-ey. That is likely hard to change at this stage (as the dataset has received much discussion in the broader AI community); however, I would highly encourage the authors to soften the text around this exam being "final."

The benchmark itself consists of both multiple choice and exact match short answer questions. This makes sense for easy checking, but at least in my academic career, I often encountered many longer form exams! Again, this need not be included (the strict format makes sense, in this first version, for easy checking – but it begs the need for future versions that do go beyond such limited answer forms). Making it not "the final" academic benchmark.

Relatedly — the authors regularly report averaged statistics over the dataset and over topics. However, there is no breakdown of performance by multiple choice (MC) vs. short-answer (SA) questions. The authors note how it is possible for models to happen upon the answer randomly. Breaking up the performance into MC vs. SA would be helpful to better understand how serious the "risk" of a model just happening upon the right answer is. How many categories did most MC questions have? These kind of details would be valuable to include somewhere; at minimum in the Supplement.

Each plot showing average would also be well-served to include error bars — what's the variance of the accuracy, over the individual questions?

There are a few other miscellaneous details that would be valuable to include about dataset construction:
How many people, on average, reviewed each question in the second phase? If 1-3 people reviewed questions in the first phase, what is the distribution of individual reviewers (did some reviewers account for the majority of the reviews?)
How were the prizes determined, and is there a list of the distribution (did the same submitters make the most quality submissions?)
For the cases where the temperature was non-zero, were models sampled multiple times? This is helpful to know wrt the cases where models attain non-zero accuracy.
The authors prompt for "rationales" as well – were these inspected systematically in any form, as part of the evaluation? If not, can they be? (alongside accuracy and calibration). The authors do count tokens, but that is different from soundness of the rationales.
While each of these questions is relatively minor – the large absence of them together is worth addressing in a revision, as the dataset itself is the primary contribution of the paper.

There is also some discussion of a public/private split. The authors minimally discuss the likely over-optimization that will come from the release of their first public dataset – rendering the public HLE likely the same fate as MMLU. If permissible to

share (though I could see reasons not to), what is the size (number of questions) in the private dataset, and what is the authors' release plan (timeline)? At minimum, it is worth including a discussion of saturation on the public set.

Overall, I think HLE is a tremendous and substantive data collection effort and admire the authors' work to do so. However, I do find the framing throughout the paper too strong on the hype, and think that the framing of the paper as "the final" exam is misleading. There are also several important details around the data collection effort that are worth including (and hopefully relatively light to include).

Minor: Why are DeepSeek-r1 and o3 not included in Table 3/full?

(Remarks on code availability)
I have reviewed the code. I appreciate that the authors put the public dataset on HuggingFace and have a nice README. This is great for the broader community. Performance also includes error bars here; hence, I do wish that was included in the main paper - as noted in my review!

Referee #4

(Remarks to the Author)
* Summary of the key results
This paper is about a new challenging benchmark, named Humanity's Last Exam (HLE), consisting of 2500 challenging problems for LLM spanning across many subjects such as Math (41%), Biology/Medicine (11%), CS/AI (10%), Physics (9%), Humanities/Social science (9%), Chemistry (7%), Engineering (4%) and Other (9%). HLE is a multi-modal benchmark (14% includes images), in which answers to questions are easy to verify (24% multiple-choice and 76% exact-match). What sets HLE apart is its difficulty, in which the maximum performance of existing LLMs was only 13.4% max at the time of the paper writing.

* Originality and significance: if not novel, please include reference
This is no doubt an impressive benchmark with 1000 subject expert contributors affiliated with over 500 institutions across 50 countries. The organization and execution of the benchmark creation process is excellent. The work is novel in terms of scale, diversity, and difficulty of the benchmark thanks to the incentive structure and the 3-stage review process (LLM, 2 rounds of expert reviews and organizers' approvals). There exists a Frontier Math benchmark, which is similar in spirit to HLE in terms of difficulty, but it's not as diverse as HLE.

* Data & methodology: validity of approach, quality of data, quality of presentation
The process of curating and validating examples to be included in the benchmark, is sound. The results in Table 1 with accuracy on HLE and model calibration error are solid, demonstrating the difficulty of the benchmark.
I found the comparison in terms of token count between reasoning and non-reasoning models not convincing. It's obvious that "Models with reasoning require substantially more inference time compute". We should use reasoning models but compare the token counts used for HLE vs other (easier) benchmarks to show that HLE requires more reasoning compute.

* Conclusions: robustness, validity, reliability
This is solid work. The benchmark is of high caliber and very interesting. The evaluation section is a little bit short. I would love to see more analyses and insights before publishing at Nature.

* Suggested improvements: experiments, data for possible revision
Here are a few suggestions:
a. It would be great to add results of newer models.
b. It would be more convincing to show token counts of reasoning models between HLE vs other datasets (instead of btw reasoning and non-reasoning models on HLE)
c. I'd love to see more analyses and insights from testing modes. For example, more discussions on the results of individual categories (currently reported in section C.3 Categorical Results; perhaps these can be brought over to the main text) + some concrete examples of model failures.

Thang Luong

(Remarks on code availability)
The code have 2 main scripts:
run_judge_results.py: to automatically extract answers, confidence, etc from LLM's outputs.
run_model_predictions.py: to generate LLM's outputs.

The purpose of the code is relatively simple, so I think the results should be reproducible. The README has sufficient details & I verified that the system prompts match what were described in the paper.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
I appreciate the revisions to the paper, which have certainly improved it, but my major concern remains: there just isn't a lot of science in this paper. I appreciate that it was a significant effort and will be a useful benchmark for AI models, but to me that's not sufficient motivation for publishing it in Nature. The emphasis of the paper is on the scores that AI models receive on this benchmark. That means that its primary utility will be to a small number of large AI companies that have the resources to develop new models. There might be other value in the dataset to the broader community of researchers working in AI and related areas, through the insight that it offers into how and why those models are failing, but the current paper doesn't provide any evidence that this would be the case. That's what I was asking for in my previous review.

As a useful comparison, it is worth taking a look at the paper reporting the initial sequence of the human genome, which was published in Nature: https://www.nature.com/articles/35057062

In addition to describing the huge effort that went into this process and the ways that this effort was coordinated (analogous to the present paper), the paper provides a list of 11 insights into the nature of the human genome that came out of this process (beginning after "many points are already clear."). These were part of the scientific contribution associated with this paper, and these insights illustrate the potential for further insights in the future. An important disanalogy to the current paper is that the sequenced genome was useful to a broad range of scientists, not just a metric to be used to score whether a small number of companies had done a better job of predicting the contents of the genome. In my opinion, the current asymmetries of resources in AI research make it even more important to demonstrate that there are scientific insights (not just metrics) that come from a project like this.

The note that frontier labs would need to do this science isn't convincing -- rather, it plays into the critique that these results will only be useful to a small number of companies. There are open weights models that can be used to do versions of these analyses, and some of my suggested scientific questions are about model behavior rather than internals.

As a further note about the concern about scientific contributions, I think it is essential that every result in the main paper (not just the supplement or on the leaderboard) have either error bars or associated statistical tests, otherwise it is difficult to assess any claims about e.g. differences between models.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
I believe this benchmark could be a valuable contribution to the community. Most of my comments on the weaknesses of the paper in the previous round have been adequately addressed. More specifically:

- The authors revised the title, abstract, introduction, and discussion to soften the "Humanity's Last Exam" framing and clarified that HLE targets expert-level questions rather than being a literal final exam.

- The authors explained that the benchmark is closed-ended by design to support automation and reproducibility, and they updated Section 5 to acknowledge the limitations in assessing broader academic capabilities.

- The authors added Appendix C Table 2 (performance by subject), a new Table 3 (performance by question type), and Figure 5 (test-time scaling) to provide a deeper analysis beyond headline accuracy and calibration error.

- The authors clarified that review signals are meant to guide contributors in designing high quality, close-ended questions, and they introduced a community audit and bug bounty program to mitigate remaining quality issues.

- The authors emphasized that calibration error measures overconfidence independently of accuracy and can be interpreted as a form of hallucination, which justifies its use alongside task performance metrics.

Overall, the revised manuscript is more clear and more rigorous.

(Remarks on code availability)
The GitHub repository includes a link to the dataset, along with simple evaluation scripts to generate model responses and judge results.

Referee #3

(Remarks to the Author)
Thank you to the authors for their openness on addressing my and other reviewers' concerns in this revision. I am personally very happy to see the title now revised.

The authors note in their Response letter that they de-emphasize "last"/"final" language; however, I found it still quite prominent in the text. At least in having the benchmark named "Humanity's Last Exam"! If possible, I would recommend the

authors at minimum add a sentence in the introduction (when defining HLE) to make clearer how they define "last." The comments from my first review still stand --- knowledge will fundamentally continue to evolve! And this benchmark only contains a small subset of all knowledge to date (far weighed towards STEM). It is great that the authors added a line in the Limitations on the STEM focus! However, I think ", albeit with a stronger representation on STEM disciplines" could be fleshed out a tiny bit more (e.g., re-add the % that are STEM vs. non-STEM, perhaps).

On error bars: I think it is important to show, even if they are low. That's good to report! I would encourage adding error bars regardless.

Thank you as well for clarifying many of my questions on data collection and evaluation. I did not fully understand "Our dataset size allows for low error accuracy estimates even without resampling" in the context of non-zero temperature. What about the data makes it "okay" (relative to other datasets?) to not repeatedly sample from the models? I think this is worth adding a bit more clarity on in the text.

Thank you as well for adding the multiple-choice table. That is helpful and interesting to see!

Overall, I think the text is much improved --- but could benefit from a few further tweaks (namely: the "last" part of HLE warrants clarification; ideally add error bars; clarify resampling). Each of which is hopefully minor.

Minor: broken reference on page 31 next to "chain-of-thought prompts".

(Remarks on code availability)
I did not go back through the code again during this revision, but had looked at the code during my first review.

Response to Referee #1

We thank the reviewer for their thoughtful and constructive feedback.

***Regarding [1], [2], [3], [5]**,* we believe this is beyond the scope of our contribution, it should be work that frontier scaling labs should explore (model architecture ablations, training data influence, finetuning methods etc.), mainly due to computational and model weight access limitations. We agree these are interesting scientific questions and encourage frontier labs to work on these, which is why we fully open-sourced the dataset to encourage research.

***Regarding [4]**,* we added a new Figure 5 on inference time compute in the main text. As many previous works on reasoning models have also shown, increased test-time compute will improve the performance of models on many tasks in a log-linear fashion.

***Regarding the result table**,* we added major recent models' results.

Regarding the error bars, our rationale for omitting error bars from the main table was their small magnitude (typically 1-2%), given the large size of our dataset. However, we still report the error bars for research reference in our live update of the leaderboard at **https://scale.com/leaderboard/humanitys_last_exam**.

***Regarding non-zero accuracy scores**,* there is a performance floor for random guessing within the dataset. This exists primarily because a portion of the questions are multiple-choice and easier to guess correctly (Appendix C Table 3). Furthermore, even exact-match questions can occasionally be answered correctly by chance if a model happens to guess a key premise in the question's design.

Response to Referee #2

We thank the reviewer for their thoughtful and constructive feedback. Below we respond to the main points raised in the Weaknesses and Minor comments sections, and describe the revisions we made accordingly.

“The framing of the benchmark as "Humanity's Last Exam" feels a bit strong to me.”

We agree with the reviewer that the initial wording was overly strong and could be misinterpreted. In response to this and similar feedback from the editors and other reviewers, we have thoroughly revised the manuscript to soften this claim. Specifically, we have amended the title, abstract, introduction, and discussion to ensure our framing is more measured, focusing on HLE as a benchmark with expert-level questions rather than as a literal "final" benchmark.

“It is limited to closed-ended, verifiable academic questions, which restricts its ability to assess broader academic capabilities such as open-ended reasoning, synthesis, or multi-step problem solving across richer modalities.”

Our motivation was to make this a lightweight, automatic benchmark for the open-source community to reproduce, in the spirit of prior capability benchmarks (eg. MMLU). This necessitates making the benchmark closed-ended by design, as open-ended questions can better assess creativity and reasoning, but they typically require manual evaluation, introduce significant subjectivity, and are challenging to scale. Nevertheless, we have revised the Limitations section (Section 5) to highlight these limitations. We also now explain that the choice of the “final” or “last” wording is intended to highlight the transition towards a new class of benchmarks focused on more dynamic and open-ended AI capabilities, and that HLE is intended to be the last closed-ended benchmark.

“The paper would benefit from deeper analysis beyond headline accuracy and calibration error.”

We report performance across different subjects in the Appendix C Table 2 and added a new Table 3 breaking down performance by question types. We added a new Figure 5 on test-time scaling.

Regarding “Reviews”: We want to highlight that the review signal is mainly intended as a sanity check and feedback to help the question contributors (who are mainly academics and researchers from diverse fields, not just CS or AI) better design questions that are close-ended**,** robust, and of high quality for benchmarking AI. We have further clarified this point in the main manuscript. To mitigate remaining issues even further, we launched a post-release community audit and bug bounty program. While some residual noise is inevitable in a benchmark of this scale and scope, we believe it is unlikely to significantly affect the floor of model performance. Instead, any errors are more likely to slightly limit ceiling accuracy when HLE is saturated. This limit on ceiling accuracy is also seen in established benchmarks like MMLU.

Regarding "Calibration Error as indicative of hallucination": We consider calibration error and accuracy to be independent metrics, as a model with high accuracy can still be poorly calibrated. Therefore, we use calibration error specifically as a measure of a model's overconfidence, independently of its task performance, which can be interpreted as a form of hallucination.

Response to Referee #3

We sincerely thank the reviewer for their thoughtful and constructive feedback. We appreciate the recognition of the significance of our data collection effort and the usefulness of HLE for the AI and academic communities. We have carefully revised the manuscript in light of the reviewer's comments and addressed the concerns as follows.

> ***"The authors frame HLE as "the final" academic benchmark ("HLE may be the last academic exam we need to give models"). I think this is highly problematic language."***

We agree with the reviewer that the initial wording was overly strong and could be misinterpreted. In response to this and similar feedback from the editors and other reviewers, we have thoroughly revised the manuscript to soften this claim. Specifically, we have amended the title, abstract, introduction, and discussion to ensure our framing is more measured, focusing on HLE as a benchmark with expert-level questions rather than as a literal "final" benchmark.

> ***"The topic distribution that the authors collect is highly skewed towards STEM. Only 9% of the questions are in the humanities."***

We have now explicitly acknowledged this in the Limitations (Section 5) of the revised manuscript. We clarify that HLE has a strong representation of STEM disciplines.

> **"*I would highly encourage the authors to soften the text around this exam being "final."***

> ***"The strict format makes sense, in this first version, for easy checking – but it begs the need for future versions that do go beyond such limited answer forms). Making it not "the final" academic benchmark."***

In response to these points, we have revised the Limitations section (Section 5). We now explain that the choice of the "final" or "last" wording is intended to highlight the transition towards a new class of benchmarks focused on more dynamic and open-ended AI capabilities. We thank the reviewer for highlighting where our initial phrasing was unclear and strong.

The authors prompt for "rationales" as well – were these inspected systematically in any form, as part of the evaluation? If not, can they be? (alongside accuracy and calibration).

To disambiguate:

(1) The question author's written rationales are only meant to provide auditing during the review process, to sanity check the right answers.
(2) Model generated rationales at evaluation time (aka reasoning or CoT) are not systematically inspected, as (1) several model providers hide it and (2) it's difficult to scale up such an evaluation automatically and still be accurate. We prompt for them because it's standard to have a model reason for a bit before answering to get good performance. Instead, focusing on the final answer makes this a lightweight benchmark for the community to run in a similar vein to prior benchmarks of its kind.

Regarding *"Error Bars"*

Thank you for this suggestion. Our rationale for omitting error bars from the main table was their small magnitude (typically 1-2%), given the large size of our dataset. However, we still report the error bars for research reference in our live update of the leaderboard at **https://scale.com/leaderboard/humanitys_last_exam.** If adding error bars is still a strong recommendation for the main table in the manuscript, we are happy to include them.

Regarding *"Reviewers":* During the first round of reviews, each question was reviewed by 2-4 expert annotators. In the second round, each question and its associated reviews were assessed again by at least one selected expert in that subject to further minimize bias. As a final step before open-sourcing the dataset (although not directly discussed), our organizing team manually conducted a final sanity check of all questions and reviews. We want to highlight that the review signal is mainly intended as a sanity check and feedback to help the question contributors (who are mainly academics and researchers from diverse fields, not just CS or AI) better design questions that are close-ended**,** robust, and of high quality for benchmarking AI. We have further clarified this point in the main manuscript.

Regarding *"Prizes":* The prize-selection process involved a manual review by the organizers of all questions that received high scores during the review phase. This final selection prioritized overall quality and diversity, both at the individual question level and the broader domain level.

Regarding *"Temperature":* Most reasoning models were run with a non-zero temperature as the optimal generation config. Due to computational constraints, we did not sample multiple times (this is a practice for datasets of a smaller order of magnitude such as GPQA). Our dataset size allows for low error accuracy estimates even without resampling.

Regarding *"rationales in prompt":* As HLE was developed concurrently with the emergence of reasoning models, we prompted for rationales to provide "chain-of-thought" space, enabling non-reasoning or hybrid models to perform robustly on the dataset. During the dataset submission and design process, we worked with question contributors (through review and feedback) to ensure each final, succinct answer was easily verifiable without the need for a rationale check.

Regarding multiple choice (MC) vs. short-answer (SA) questions

Thank you for your suggestion, we included a new Table 3 in the Appendix for these statistics. We indeed find evidence models are better at guessing on the multiple choice questions (which are much higher accuracy) and updated our “Accuracy” section with this finding.

Regarding *“Private Split”*

Thanks for understanding our reasons for keeping the private set confidential, HLE is a high demand benchmark and we want to avoid any sort of cheating given there is incentive for model builders. The private set acts as a deterrence to training on the public test set. We’re refraining from disclosing more details about the private set, though we confirm that it is smaller than the public set but will nonetheless give good error bars. We will evaluate on the private set once models exceed 50% accuracy or if we suspect overfitting. At the present time, we do not suspect overfitting on any model given the relatively low accuracies, and will not publish any analysis on it.

Why are DeepSeek-r1 and o3 not included in Table 3/full?

We have now updated the result table to include newer models. Although the paper is static, we will continue to host live leaderboards of new model releases at https://scale.com/leaderboard/humanitys_last_exam.

Response to Referee #4

We thank the reviewer for their thoughtful and constructive feedback. Below, we respond to the main points raised in the "Suggested improvements" section:

[a]. ***It would be great to add results of newer models.***

In the revised manuscript, we have included results for several newer models. These additions are reflected in both the main results table (Table 1) and the extended analysis in Appendix C. Since the paper is static, we also now point the paper to our live leaderboards at https://scale.com/leaderboard/humanitys_last_exam

[b]. It would be more convincing to show token counts of reasoning models between HLE vs other datasets (instead of btw reasoning and non-reasoning models on HLE)

We agree that comparing inference token usage across benchmarks can sometimes offer useful insights. However, we chose not to include such a comparison in this work, as the prior capability benchmarks of this kind (e.g., GSM8k, MMLU, GPQA) vary significantly in structure and expected answer length, making direct comparisons of token counts difficult to interpret meaningfully. For example, GSM8k triggers very short reasoning lengths due to the simplicity of the problems, while MMLU and GPQA prompt short responses due to being multiple choice questions. Instead, we focused our analysis on token usage within HLE to highlight how reasoning impacts model performance in our benchmark specifically by updating Figure 5, and we found similar log-linear scaling trends as reported in other reasoning papers.

[c]. I'd love to see more analyses and insights from testing modes. For example, more discussions on the results of individual categories (currently reported in section C.3 Categorical Results; perhaps these can be brought over to the main text) + some concrete examples of model failures.

We thank you for these helpful suggestions. In response, our revised Figure 5 now presents a log2 scale plot of output tokens versus accuracy for reasoning models on HLE.

In this analysis, we note our observation that scaling reasoning tokens beyond a certain threshold can lead to diminishing returns on accuracy. This suggests that future models will need to improve both accuracy and computational efficiency simultaneously. We also find that models tend to reason more on mathematical categories than biology or humanities categories. Regarding the categorical results, we will consider moving more of this analysis to the main text if space permits in the camera-ready version, although we are currently at the 6 figure limit.

Regarding the concrete examples of model failures, we believe that the current Figure 2 also serves the same purpose. We manually picked challenging and interesting held-out questions for the readers' purpose of understanding the distribution in HLE.

Response to Referee #1

We thank the reviewer for the thoughtful and constructive feedback, especially on the clarification request about the scientific contributions of the work.

We share the reviewer's concern about resource asymmetry in AI research. The primary goal of this project, and the core contribution of the paper, was to help address this imbalance. We did so by creating a foundational, open-source dataset of uniquely diverse, expert-level knowledge, making the study of frontier AI more accessible to the entire scientific community.

While only a few organizations can build new AI models, a vast community of researchers studies their technical capabilities, failure modes, and broader social, economic, and ethical implications. HLE is designed to serve as a common ground for evaluation and study by scientists in both industry and academia. This collaborative role is underscored by the project's community-driven nature, with contributions from more than 1000 researchers across more than 100 institutions worldwide.

Beyond benchmarking model improvements, HLE's value lies in its broader applications as a multipurpose scientific tool:

- **For Science and Policy:** The dataset provides a shared, evidence-based view of AI's current capabilities. This allows natural and physical scientists to gauge AI's reliability on complex research tasks, while enabling social scientists and policymakers to understand its implications for critical societal issues. For example, we were informed recently that the upcoming *International AI Safety Report (previous version) will possibly highlight HLE's progress over time as an important dimension of the covered issues.* This cross-disciplinary clarity is important for responsible innovation.
- **For Enabling New Research:** The open-source dataset allows any researcher to conduct deep, qualitative analyses of model failure modes or probe for specific reasoning pathways compared to human experts. As an example, our own analysis introduces a calibration error metric, showing that even top models remain poorly calibrated on expert-level questions despite claimed improvements in reducing hallucinations. Another example is that HLE is becoming a common evaluation framework for the emerging field of AI research agents. This direction evaluates how well AI systems can tackle HLE questions by independently using tools (like coding and literature search) or employing advanced methods like multi-agent reasoning, where systems collaborate on potential solutions (for example, DeepMind's Gemini DeepThink system that recently achieved the gold-medal standard at this year IMO). Progress on HLE in this context helps illustrate the capabilities of agent-based AI on larger-scale research-level tasks.

We also added the error bars in the main results tables as advised by the reviewer.

Response to Referee #2

We thank the reviewer for the thoughtful and constructive feedback on the revision of the paper.

Response to Referee #3

We thank the reviewer for the thoughtful and iterative feedback on what can still be improved.

HLE and Humanity's Last Exam name: Thank you for your iterative suggestion on clarifying the naming of the benchmark. We understand that the current full name of the benchmark may not be perfectly precise. Upon discussion with the Editor of Nature Journal, we will now simplify the main name to **"***HLE (originally known as Humanity's Last Exam, …)***"** in the introduction and will continue working with the editors in the process to refine the precise wording of the description of **"last exam"** for the general audiences. We greatly appreciate your concern about the importance of precision in this matter.

"..., albeit with a stronger representation on STEM disciplines". : We now add a direct reference to Figure 3 and Appendix B.4 for a better illustration of stronger representation on Math and STEM disciplines mentioned. We changed the sentence too: "..., albeit with a stronger representation on Math and STEM disciplines as shown in Figure 3 and Appendix B.4"

On error bars: We now add the error bars to the main result tables.

Regarding sampling and temperature: Our evaluations are run on either temperature 0 or the default/recommended temperature by the model builders. We saw that the size of 2,500 questions in HLE is large enough for a low error bar between different temperature sampling. We are also able to reproduce different official reported results from different organizations' model releases without the need to tune the generation parameters.

We also fixed the broken reference on page 31