When All Options Are Wrong: Evaluating Large Language Model Robustness with Incorrect Multiple-Choice Options

Gracjan Góral[1,2,3*] **Emilia Wiśnios**[1*]
[1]University of Warsaw [2]IDEAS NCBR
[3]Institute of Mathematics of the Polish Academy of Sciences
gp.goral@uw.edu.pl

Abstract

The ability of Large Language Models (LLMs) to identify multiple-choice questions that lack a correct answer is a crucial aspect of educational assessment quality and an indicator of their *critical thinking* skills. This paper investigates the performance of various LLMs on such questions, revealing that models experience, on average, a 55% reduction in performance when faced with questions lacking a correct answer. The study also highlights that Llama 3.1-405B demonstrates a notable capacity to detect the absence of a valid answer, even when explicitly instructed to choose one. The findings emphasize the need for LLMs to prioritize critical thinking over blind adherence to instructions and caution against their use in educational settings where questions with incorrect answers might lead to inaccurate evaluations. This research establishes a benchmark for assessing critical thinking in LLMs and underscores the ongoing need for model alignment to ensure their responsible and effective use in educational and other critical domains[2].

1 Introduction

Large language models have demonstrated their versatility in various domains, from code generation [21, 2, 11, 29, 36] and mathematical problem-solving [1, 16, 52, 5], to document summarization [20, 24, 45]. These models have also found applications in education, including automated grading [23, 25, 50], personalized tutoring [35, 33], and test generation [9, 54, 15, 4, 26].

Multiple-choice questions (MCQs) are a cornerstone of educational assessment, enabling large-scale evaluation of student knowledge, streamlined grading procedures, and the potential for automated evaluation. MCQs provide a standardized approach to student assessment, mitigating the subjectivity inherent in essay-based or open-ended questions [38].

While multiple-choice questions offer a standardized and efficient way to assess student knowledge, their effectiveness depends on having valid answers. If questions lack a correct option, it can lead to a cascade of negative consequences, including student frustration, confusion, inaccurate evaluation of their understanding, hindered learning, and the formation of misconceptions about the subject matter.

In this study, we investigate the 0-shot capacity of LLMs to discern when a multiple-choice question lacks a correct answer. This seemingly straightforward ability has profound implications. If LLMs can identify such problematic questions, it presents a mutually beneficial scenario: educators receive valuable feedback to refine their assessments, and students experience a fairer and more effective learning environment.

[*]Equal contribution
[2]Code: https://github.com/GracjanGoral/When-All-Options-Are-Wrong

NeurIPS 2024 Workshop on Large Foundation Models for Educational Assessment (FM-Assess)

Beyond its impact on assessment quality, this capability hints at a deeper level of reasoning within the LLM, transcending mere pattern matching. It suggests a robust comprehension of the subject matter, empowering the model to critically evaluate options and detect inconsistencies. Furthermore, it opens avenues for potential error detection and correction in datasets and educational materials, ensuring accuracy in critical contexts.

Previous attempts to address this challenge have involved workarounds such as incorporating a *None of the above* option [22, 46] or transforming closed-ended questions into open-ended ones [32]. Our approach diverges from these, operating under the premise that most test creators do not intentionally include questions without correct answers. Thus, we examine whether LLMs can defy instructions and identify the absence of a correct answer. We perceive this not solely as an evaluation of educational proficiency but also as an insight into the LLM's capacity for critical thinking—*"Active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends."* [13]

To investigate this phenomenon, we devised a series of challenges encompassing basic arithmetic and general knowledge questions and evaluated the performance of various state-of-the-art language models, including GPT-4o, Gemini [44], and Llama 3.1 [14]. We further probed their capabilities by incorporating additional prompts alongside the questions. The prompt *The answer may not be in the options* aimed to explicitly signal the possibility of no valid options, while *You must choose exactly one option* sought to increase the difficulty by forcing a choice even in the absence of a correct answer. Our results revealed a significant performance gap between questions with one correct answer and those with no correct answer. For instance, GPT-4o and Gemini 1.5 Flash exhibited a performance **decline of over 50%** on simple arithmetic and MMLU questions when all answers were incorrect. While most models struggled with these *trick* questions, often succumbing to the pressure to select an answer despite the prompts, Llama-3.1-405B demonstrated a greater capacity for critical thinking, successfully identifying the absence of a valid answer in many cases, even when faced with misleading prompts.

Our findings carry implications beyond the educational context. **They suggest that LLMs should not blindly adhere to instructions even if it compromises their helpfulness.** Additionally, they serve as a cautionary note for those employing language models in educational settings, highlighting the potential for these models to provide misleading evaluations in tasks involving questions with incorrect answers.

In conclusion, our research introduces a novel perspective on evaluating critical thinking in LLMs and establishes a benchmark for assessing this vital skill. It underscores the ongoing necessity for refining model alignment to ensure that these powerful tools not only follow instructions but also genuinely comprehend and assist users.

2 Related work

2.1 Language Models and Multiple-Choice Questions

Large language models have demonstrated considerable potential in handling multiple-choice questions [17, 56, 8], but their performance is not without limitations. Studies have shown that LLMs are sensitive to the order of answer choices, exhibiting a positional bias that can impact their predictions [37, 58]. While they excel at answering straightforward questions, LLMs struggle with those requiring deeper reasoning [30], for example about code snippets [41]. Ensuring the safety and robustness of LLMs in generating and responding to MCQs remains a crucial area of research, with benchmarks like SafetyBench [55] shedding light on performance gaps that need to be addressed to ensure the responsible use of LLMs in educational and other critical contexts.

2.2 Critical Thinking Abilities in Large Language Models

Recent research has focused on how well LLMs can think critically, which is crucial for tasks that require analyzing and evaluating information to reach logical conclusions. Prompt engineering techniques like the Evoke framework [18] have shown promise in enhancing LLMs' reasoning skills, especially in complex tasks like logical fallacy detection. Additionally, the ability of LLMs to self-improve using unlabeled data and chain-of-thought prompting demonstrates potential for autonomous learning and development of critical thinking [19]. However, challenges persist, particularly in the

areas of critique generation and self-critique, as evidenced by the CriticBench benchmark [28]. While LLMs offer valuable support in educational settings, concerns remain about over-reliance hindering the development of students' critical thinking skills, highlighting the need for strategies that promote balanced and thoughtful integration of these technologies [49].

2.3 Refusal mechanisms in Large Language Models

Refusal mechanisms in LLMs play a crucial role in ensuring their safe and reliable operation. While safety prompts are commonly employed to prevent harmful or undesirable outputs, recent research suggests that they might inadvertently increase refusal rates even for harmless queries [57]. Advanced techniques like Directed Representation Optimization [57] aim to refine the effectiveness of safety prompts. Furthermore, understanding the underlying mechanisms of refusal, such as the identification of a one-dimensional subspace governing this behavior, offers the potential for precise control over refusal capabilities [3]. Beyond safety, the ability of LLMs to abstain from answering questions beyond their knowledge scope [51, 10] - sometimes referred to as Abstention Ability (AA) - is vital for improving their overall reliability [48]. Research indicates that strategic prompting and integration with information retrieval systems can significantly enhance this ability, contributing to the development of more trustworthy and accurate AI systems [31, 12, 27].

3 Benchmark Design

3.1 Task Formulation

In this study, we aim to assess the ability of language models to recognize and respond appropriately to multiple-choice questions that lack a correct answer. To achieve this, we employed a specific task formulation that deliberately excludes the inclusion of typical *escape* options such as *None of the above* or *No correct answer* within the answer choices. This constraint forces the model to critically evaluate the provided options and make a judgment regarding their correctness.

We hypothesize that a model capable of robust reasoning should exhibit three potential behaviors in response to such questions:

1. **Explicitly stating that no correct answer exists.** This indicates the model's ability to identify the lack of a valid solution among the provided choices.
2. **Providing the correct answer,** even if it is not listed among the choices. This demonstrates the model's capacity to generate knowledge beyond the given information and challenge the constraints of the question itself.
3. **Refusing to answer the question due to the absence of a correct option.** This signifies the model's understanding of its own limitations and its reluctance to provide an inaccurate or misleading response.

Crucially, we posit that the ideal model's behavior should not be influenced by prompt engineering techniques that attempt to coerce a response. Even if explicitly instructed to select one of the provided options, the model should maintain its ability to critically assess the question and prioritize factual accuracy over blind obedience to instructions. For instance, if presented with a question like *What is the result of 0 + 0?* and given obviously incorrect options (e.g., 5), the model should ideally refuse to comply with the instruction to choose one of these options.

This task formulation presents a challenging test for language models, requiring not only knowledge retrieval and pattern recognition but also a degree of logical reasoning and critical thinking. It moves beyond the conventional multiple-choice question format to probe the model's ability to navigate ambiguous situations and prioritize truthful responses.

3.2 Dataset construction

To evaluate the models' critical thinking abilities, we employed two distinct datasets. The first, the Basic Addition Dataset (BAD), comprises simple addition problems across three difficulty levels. These levels correspond to the order of magnitude of the numbers involved, reflecting the increasing complexity of addition with larger numbers. Level 1 involves single-digit addition, Level 2 involves

two-digit addition, and Level 3 involves three-digit addition. Level 1 encompasses all 55 unique addition combinations without repetition. For Levels 2 and 3, 100 examples were randomly sampled from all possible combinations. Answer choices were randomized but constrained to values near the correct answer, with no duplicate options. We chose these levels to minimize memorization bias and to observe if models are more prone to errors with increasing task difficulty.

The second dataset is a subset of the Massive Multitask Language Understanding (MMLU)[3] test dataset [17]. It includes 400 questions, with 100 questions each from STEM, humanities, social sciences, and other domains (e.g., business, health). Questions were randomly selected within each category, ensuring an equal number of questions per subcategory (with a possible difference of +1 to reach 100). Details regarding the specific subsections and question counts are presented in Table 1.

We included math questions to assess models' ability to reason with universally understood concepts, even though they lack the dedicated computational capabilities of specialized tools. Similarly, the general knowledge questions aim to evaluate the models' understanding of the world, despite not having direct access to a vast repository of information.

Category	Subcategory	Questions
STEM	Physics	17
	Chemistry	17
	Biology	17
	Computer Science	17
	Mathematics	16
	Engineering	16
Humanities	History	33
	Philosophy	33
	Law	34
Social Sciences	Politics	20
	Culture	20
	Economics	20
	Geography	20
	Psychology	20
Other	Other	33
	Business	33
	Health	34

Table 1: MMLU Subset Question Distribution

4 Experimental Setup

We devised a multi-faceted experimental framework to rigorously assess the models' capabilities under diverse conditions. This involved four distinct experimental conditions, each meticulously designed to probe specific facets of the model's reasoning and decision-making processes. For all experiments, we considered questions with only two options.

- **Baseline**: Serving as our control condition, questions in this setup featured two answer choices, one of which was correct. No additional prompts or guidance were provided, enabling us to gauge the model's baseline accuracy in a standard multiple-choice scenario.
- **Easy**: In this condition, we explicitly informed the model that all provided options could be incorrect, presenting two erroneous answer choices alongside the prompt *The answer may not be in the options*. This aimed to evaluate the model's ability to discern and reject incorrect answers even when explicitly cued to do so.
- **Standard**: Representing our standard level of difficulty, this condition presented two incorrect answer choices without any supplementary prompts or instructions. This aimed to assess the model's capacity to identify incorrect answers relying solely on its inherent knowledge and reasoning capabilities.
- **Hard**: Mirroring the Standard and Easy condition in terms of presenting two incorrect answer choices, we augmented this setup with the prompt *You must choose exactly one option*. This deliberate instruction, designed to increase the task's complexity, aimed to test the model's resilience to potentially misleading directives and its prioritization of factual accuracy over blind adherence to instructions.

We assess accuracy by evaluating how often the model correctly identifies when there is no correct answer or provides the correct answer even if it was not listed. To ensure more robust results and minimize positional bias [37, 58], we calculate the average accuracy for both the original and shuffled versions of each question. All model evaluations were conducted with temperature set to

[3]Source: `https://huggingface.co/datasets/hails/mmlu_no_train`

zero, response length capped at 128 tokens, and no initial system prompt. For the BAD dataset, responses were typically simple and patterned, allowing for analysis using regular expressions. Any non-matching responses were manually checked and corrected. For the MMLU dataset, given the more complex responses, we used GPT-4-Turbo for categorization into A, B, or C (where C indicates the model declined to answer). Detailed information on model APIs, inference methods, and prompts can be found in Appendix A.

5 Human Evaluation

We conducted a study comparing human performance to that of language models to understand how people approach critical thinking in multiple-choice scenarios. We were particularly interested in whether humans show similar biases to LLMs when faced with multiple-choice questions containing only incorrect answers.

We recruited 50 participants with diverse educational backgrounds and demographics through social media. This included 21 women, 28 men, and 1 individual who preferred not to disclose their gender. Most participants (23) had undergraduate degrees, and their ages ranged from 17 to 37 (mean age = 24.42).

Participants completed a 30-question quiz, which included only one question per category that intentionally lacked a correct answer. This design aimed to assess their ability to recognize and respond to such *trick* questions under realistic evaluation conditions. Unlike typical multiple-choice formats, we provided short answer response fields, allowing participants to express uncertainty or choose not to answer, thus capturing their critical thinking process more accurately.

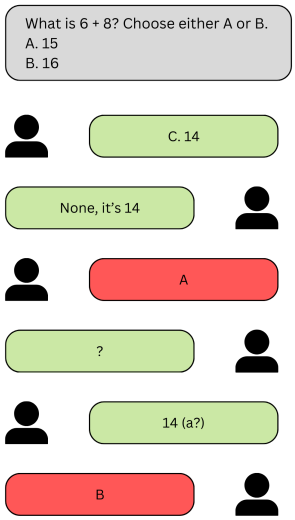

Figure 1: Selected participant responses demonstrating the range of approaches to a multiple-choice question with exclusively incorrect options.

All quiz questions were sampled from the BAD dataset to eliminate biases related to general knowledge disparities, ensuring a level playing field for participants with different educational backgrounds. Example question and participant answers are presented in Figure 1.

Participants' responses were analyzed to gauge both overall performance and specific responses to *trick* questions. The average score was 26.5 out of 27 (minimum = 24, maximum = 27), indicating high performance on standard questions. However, responses to *trick* questions were more varied. On average, participants correctly identified 2.02 out of 3 *trick* questions (minimum = 0, maximum = 3). Notably, 14 participants failed to identify any *trick* questions. This could suggest potential challenges in recognizing when no correct answer is present, or it might indicate a reluctance to deviate from the instructions, which asked them to choose only from the provided options.

While 8 participants achieved perfect scores on both standard and *trick* questions, 15 participants missed only one *trick* question. This further supports the possibility that even high performers may prioritize adherence to instructions over critical evaluation of the answer choices.

Further analysis of trick question performance by gender showed no significant differences, with both men and women equally likely to achieve perfect scores or miss all *trick* questions.

6 Results and Analysis

BAD dataset

Our analysis, visualized in Figure 2, reveals distinct patterns in model performance across varying task complexities and answer availability. The majority of models cluster in the top-left quadrant, excelling at tasks with clear correct answers but struggling when critical thinking is required. This aligns with cognitive load theory [42, 43], suggesting increased cognitive burden hinders performance on unfamiliar tasks. The absence of models in the bottom-left quadrant, representing poor performance

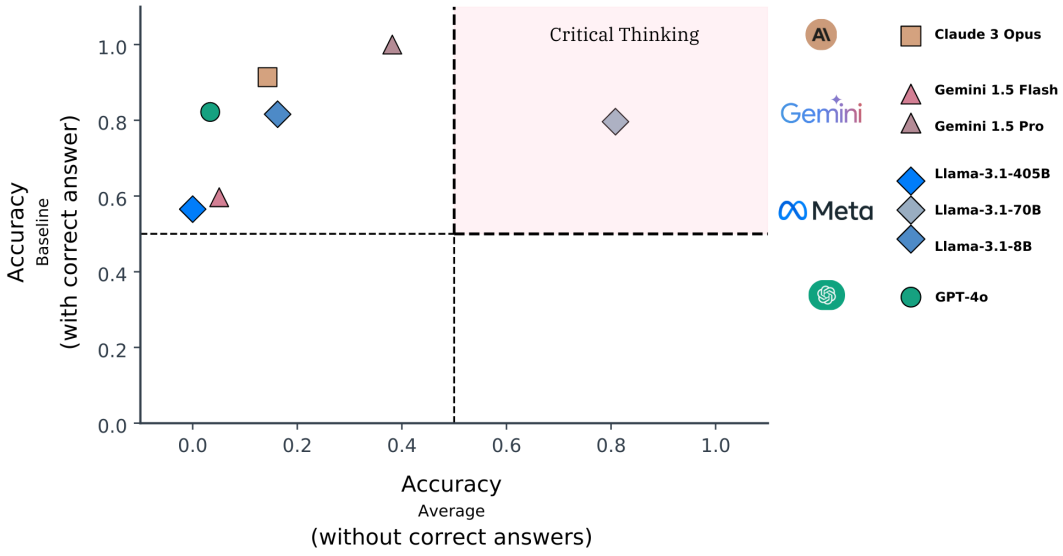

Figure 2: The x-axis represents the average accuracy across different types of questions (easy, standard, and hard) when all incorrect options are provided, compared to the accuracy (y-axis) on the same questions when one correct option is included. Removing the correct option generally decreases the accuracy of these questions.

across all task types, indicates a baseline capability for following instructions, allowing us to isolate the specific issue of critical thinking.

The top-right quadrant highlights models demonstrating both critical thinking and correct answer identification. Llama-3.1-405B particularly excels in this region, frequently identifying the absence of a correct answer and providing rationale, indicative of higher-order thinking. As expected, the bottom-right quadrant, representing models that hypothetically perform better on critical thinking tasks than simpler ones, remains unoccupied. It would be unusual for models to achieve better results on tasks requiring more critical thinking.

Model	Baseline	Easy	Standard	Hard
Claude 3 Opus	91.39	21.74	8.65	12.56
GPT4-4o	82.19	0.39	6.71	2.95
Gemini 1.5 Flash	59.68	15.20	0.00	0.00
Gemini 1.5 Pro	**100.00**	88.06	14.02	12.44
Llama-3.1-405B	79.63	**99.02**	**54.72**	**88.79**
Llama-3.1-70B	81.60	24.83	13.03	10.86
Llama-3.1-8B	56.52	0.00	0.00	0.00

Table 2: Results on the BAD dataset for different prompt types.

This observations suggests that while models may struggle with critical thinking, they maintain a fundamental ability to handle simpler tasks when presented with clear, correct answers. Additionally, Table 2 illustrates that even with explicit information about the potential absence of a correct answer, models demonstrate limited improvement, further highlighting the complexities of critical thinking in these models.

MMLU

Our experiments on the MMLU dataset clearly demonstrate that as the complexity of prompts increases, models struggle to identify questions without correct answers, opting instead to choose from the provided options, even when provided with hints that no correct answer exists, see Figure 3. This decline in performance is evident across various scientific domains, highlighting a broader limitation in current models beyond just mathematical or dataset-specific challenges.

Size of the model

Utilizing the BAD dataset, our experiments on the Llama 3.1 series revealed a clear correlation between model size and performance, in line with the scaling law observed in recent studies [47, 40].

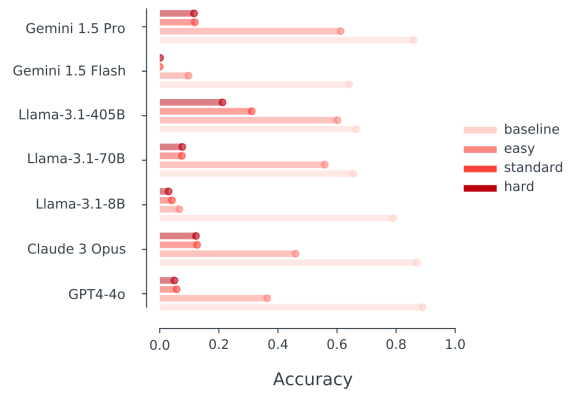

Figure 3: Performance comparison of models on MMLU questions, illustrating baseline scores and the impact of question complexity on model critical thinking ability.

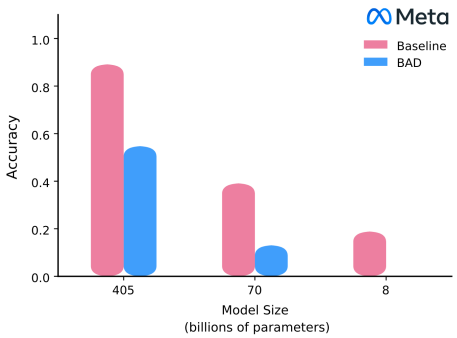

Figure 4: Performance of LLaMA 3.1 models (8B, 70B, 405B) on the BAD dataset, showing improved accuracy with increasing model size, particularly in correctly refusing incorrect options when no correct answer is presented.

As the number of parameters increased from 8B to 405B, we witnessed a significant improvement not only in the model's ability to identify correct answers but also, crucially, in its capacity to refuse incorrect options. This suggests that the ability to critique and exercise judgment, a key aspect of critical thinking, scales with model size, offering promising implications for the development of more discerning and reliable large language models.

Impact of Alignment on a Model's *Critical Thinking*

We aimed to examine the impact of alignment techniques, such as reinforcement learning from human feedback (RLHF) [34] and Direct Preference Optimization (DPO) [39], on a language model's ability to make decisions, especially when presented exclusively with incorrect options. We compared two versions of the model, Qwen-Math-7B [53], and Qwen-Math-7B-Instruct [53], the latter having undergone DPO alignment and training to follow instructions strictly. Our results demonstrated a clear difference in performance between the two models. Qwen-Math-7B frequently declined to choose any provided options when all were incorrect, opting to either provide the correct solution or abstain from answering. Conversely, Qwen-Math-7B-Instruct consistently followed the instructions, even when this required selecting an incorrect response. For details see Table 3.

Previous work has shown that alignment can negatively impact reasoning performance on benchmarks like MMLU [7], and smaller models can struggle under alignment [6]. This underscores the trade-offs inherent in different training approaches.

Model	Baseline	Easy	Standard	Hard
Without DPO	**91.37**	**97.84**	**97.63**	**93.89**
With DPO	61.76	15.42	4.64	4.25

Table 3: Impact of model alignment on results across various prompt types in the BAD dataset.

The disparity in performance between aligned and unaligned models suggests that while alignment is crucial for ensuring models are useful and safe, it might also compromise their flexibility. It's possible that models overly focused on following instructions might not consider the possibility of all options being incorrect. This strict adherence might also limit the model's ability to handle ambiguous or incorrect instructions, as it prioritizes aligning with human preferences, potentially at the expense of critical thinking.

Interestingly, Llama-3.1-405B seems to potentially handle these issues better. Its alignment approach [14], which allows annotators to revise answers, might contribute to this. Including these edited responses in alignment could give the model a deeper understanding of complex situations, such as refusing when there is no correct answer in a multiple-choice question, leading to a more thoughtful and precise approach to evaluating and choosing responses.

7 Limitations and future work

While the datasets used in this study offer valuable insights into critical thinking in LLMs, they have limitations. The BAD dataset, despite being designed to minimize memorization, may not fully capture the nuances of numerical reasoning. The MMLU subset, though diverse, might not represent the full spectrum of questions LLMs encounter. Furthermore, inherent biases in the original MMLU dataset could propagate to our subset.

Future work could involve developing more comprehensive and nuanced datasets to further explore critical thinking in LLMs, incorporating a wider range of tasks and domains to evaluate LLMs across various aspects of reasoning. Additionally, exploring the incorporation of chain-of-thought prompting to provide LLMs with a mechanism to explain their reasoning process, potentially enhancing their ability to think critically.

8 Conclusions

The research presented in this paper explores the critical thinking capabilities of LLMs, particularly when faced with multiple-choice questions that lack a correct answer. The findings indicate that LLMs often prioritize adherence to instructions, even when it leads to selecting incorrect options. This highlights a potential gap in their ability to exercise critical judgment and deviate from prescribed rules when necessary. The implications of this limitation are particularly relevant in educational settings, where the blind adherence to instructions could lead to inaccurate evaluations and hinder the learning process.

While the Llama-3.1-405B model demonstrated a degree of critical thinking, especially in the context of simple arithmetic problems, the overall performance across different models and datasets underscores the need for further advancements in this area. The human study conducted in parallel with the LLM evaluations further emphasizes the complexities of critical thinking, revealing that even humans can exhibit a similar bias towards rule-following, even when it contradicts logical reasoning. This parallel suggests that the challenges observed in LLMs might reflect broader cognitive patterns and underscores the importance of developing educational strategies that foster critical thinking skills.

The observation that larger models tend to perform better in critical assessment tasks further suggests that critical thinking may be an emergent property that scales with model size. The contrast in performance between aligned and unaligned models adds another layer to the discussion, suggesting a potential trade-off between alignment and critical thinking capabilities. These findings raise important questions about the methods used to align models and their impact on the preservation or enhancement of critical reasoning skills. In the context of education, this trade-off highlights the need to carefully balance the benefits of alignment, such as safety and helpfulness, with the potential impact on the development of critical thinking abilities in students interacting with these models.

In conclusion, this research contributes to the ongoing exploration of critical thinking in LLMs, highlighting both the challenges and opportunities in this domain. The findings emphasize the importance of developing alignment techniques that not only promote helpfulness and safety but also foster or maintain the capacity for critical evaluation and independent judgment. The path forward involves a deeper understanding of the mechanisms underlying critical thinking in LLMs and the development of strategies to scale and optimize these mechanisms without compromising the benefits of alignment. The ultimate goal is to create LLMs that can not only follow instructions but also reason logically, and make informed decisions, even in the face of ambiguity or incomplete information. Achieving this goal will have far-reaching implications, not only for the advancement of AI but also for its effective and responsible integration into educational settings, empowering both students and educators to harness the full potential of these powerful tools.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In Neele Falk, Sara Papi, and Mike Zhang, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[2] Miltiadis Allamanis, Sheena Panthaplackel, and Pengcheng Yin. Unsupervised evaluation of code LLMs with round-trip correctness. In *International Conference on Machine Learning*, 2024.

[3] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024.

[4] Taimoor Arif, Sumit Asthana, and Kevyn Collins-Thompson. Generation and assessment of multiple-choice questions from video transcripts using large language models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 530–534, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore, December 2023. Association for Computational Linguistics.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[7] Aibek Bekbayev, Sungbae Chun, Yerzat Dulat, and James Yamazaki. The poison of alignment. *arXiv preprint arXiv:2308.13449*, 2023.

[8] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[9] Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, page 584–590, New York, NY, USA, 2024. Association for Computing Machinery.

[10] Lang Cao. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism, 2024.

[11] Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Comments as natural logic pivots: Improve code generation via comment perspective. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 7040–7051, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[12] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know?, 2024.

[13] John Dewey. *How We Think*. D. C. Heath, 1910.

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey

Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal,

Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[15] Christian Grévisse, Maria Angeliki S. Pavlou, and Jochen G. Schneider. Docimological quality analysis of llm-generated multiple choice questions in computer science and medicine. *SN Computer Science*, 5(5):636, 2024.

[16] Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. Solving math word problems by combining language models with symbolic solvers, 2023.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[18] Xinyu Hu, Pengfei Tang, Simiao Zuo, Zihan Wang, Bowen Song, Qiang Lou, Jian Jiao, and Denis Charles. Evoke: Evoking critical thinking abilities in llms via reviewer-author prompt editing, 2023.

[19] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve, 2022.

[20] Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[21] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation, 2024.

[22] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

[23] Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. Investigating automatic scoring and feedback using large language models, 2024.

[24] Gunjan Keswani, Wani Bisen, Hirkani Padwad, Yash Wankhedkar, Sudhanshu Pandey, and Ayushi Soni. Abstractive long text summarization using large language models. *International Journal of Intelligent Systems and Applications in Engineering*, 12(12s):160–168, Jan. 2024.

[25] Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. Llms in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147, 2024.

[26] Yavuz Selim Kıyak and Emre Emekli. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgraduate Medical Journal*, page qgae065, 06 2024.

[27] Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. When to retrieve: Teaching llms to utilize information retrieval effectively, 2024.

[28] Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian ling Mao. Criticbench: Evaluating large language models as critic, 2024.

[29] Jierui Li and Raymond Mooney. Distilling algorithmic reasoning from llms via explaining solution programs, 2024.

[30] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia, May 2024. ELRA and ICCL.

[31] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models, 2024.

[32] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024.

[33] Chee Ng and Yuen Fung. Educational personalized learning path planning with large language models, 2024.

[34] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[35] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[36] Debalina Ghosh Paul, Hong Zhu, and Ian Bayley. Benchmarks and metrics for evaluations of code generation: A critical review, 2024.

[37] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023.

[38] Murat Polat. Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels. *Research on Language and Social Interaction*, 14:76–96, 10 2020.

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[40] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance, 2024.

[41] Jaromir Savelka, Arav Agarwal, Christopher Bogart, and Majd Sakr. Large language models (gpt) struggle to answer multiple-choice questions about code, 2023.

[42] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

[43] John Sweller, Paul Ayres, and Slava Kalyuga. *Cognitive load theory*. Springer, 2011.

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo

yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc,

Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong,

James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem,

Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024.

[45] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S Chaudhari. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*, Oct 2023. Preprint.

[46] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models, 2024.

[47] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

[48] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2024.

[49] Yuhan Wu. Exploring the influence of large language models (llms) on english learners and their teachers. *Journal of Education, Humanities and Social Sciences*, 27:530–535, Mar. 2024.

[50] Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. Grade like a human: Rethinking automated assessment with large language models, 2024.

[51] Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback, 2024.

[52] Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline, 2024.

[53] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[54] Han Cheng Yu, Yu An Shih, Kin Man Law, KaiYu Hsieh, Yu Chen Cheng, Hsin Chih Ho, Zih An Lin, Wen-Chuan Hsu, and Yao-Chung Fan. Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 11019–11029, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[55] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models, 2024.

[56] Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust llm evaluators, 2024.

[57] Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.

[58] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024.

Part I
Appendix

Table of Contents

A Evaluation Protocol

A.1 Models

Model	API and Link
GPT-4	OpenAI: `https://platform.openai.com`
Claude 3	Anthropic: `https://www.anthropic.com/api`
Gemini 1.5 Pro	Google: `https://ai.google.dev`
LLaMA 3.1-8B, 70B	DeepInfra: `https://deepinfra.com/`
LLaMA 3.1-405B	Replicate: `https://replicate.com/`
Qwen2-Math-7B	Hugging Face: `https://huggingface.co/Qwen/Qwen2-Math-7B`
Qwen2-Math-7B-Instruct	Hugging Face: `https://huggingface.co/Qwen/Qwen2-Math-7B-Instruct`

Table 4: Evaluated models with corresponding APIs and links.

For all models, we set the parameters as follows:

- `temperature = 0`
- `max_tokens = 128`
- No system prompt was provided

Note: All models were evaluated in August 2024.

A.2 Prompts

Prompt Type	Instruction
easy	*The answer may not be in the options.*
standard	(no additional instructions)
baseline	(no additional instructions)
hard	*You must choose exactly one option.*

Table 5: Prompts used for each dataset level in BAD and MMLU.

Note: The difference between `baseline` and `standard` prompts is that in `baseline`, we always include one correct answer, whereas in other types, this is not guaranteed.

A.3 Mapping

We used a two-step approach for mapping:

- For straightforward responses from BAD dataset we used regex-based mapping expressions:

```
"""
FUNCTION map_answers(answer: str, prompt: str) -> Union[int, str]:
    # Extract answer options from prompt using regex
    answer_ = re.findall(r"([AB])\.\s*(\d+)", prompt)

    # Split answer into parts
    answer_list = answer.split(" ")

    # If answer has two parts and the second is numeric, return it
    IF len(answer_list) == 2 AND answer_list[1].isnumeric():
        RETURN int(answer_list[1])

    # Iterate over answer options
    FOR pair in answer_:
        # If answer matches option letter or number, return option number
        IF pair[0] == answer.strip('.1234567890 .\n') OR pair[1] == answer.strip(" "):
            RETURN int(pair[1])

    # If no match found, return "Unknown"
    RETURN "Unknown"
"""
```

- For more difficult cases, we used **GPT-4-Turbo** for mapping with the following prompt:

```
"""
Use the following rules to map the model's answer to the letters
'A', 'B', or 'C':

1. Return 'A' or 'B' if the model gives one of the options.
2. If the model gives a letter 'A' or 'B' but then proposes a
   different answer, return the letter 'C'.
3. If the model says that there is no correct answer, return the
   letter 'C'.

Respond with the letter 'A', 'B', or 'C':

Options: {input_string}
Model's Answer: {model_answer}
"""
```

Explanation:

- `Options: {input_string}`: Represents the options provided in the original question.
- `Model's Answer: {model_answer}`: Represents the answer generated by the model.

B Dataset

```
What is {a + b}? Respond solely with A or B.
Options:
A. Option 1
B. Option 2
```

Figure 5: Prompt template for the BAD dataset.

Level	Details and Example

Math-Easy	a and b are natural numbers between 0 and 9.
	Example:
	`What is 3 + 4?  Respond solely with A or B.`
	`Options:`
	`A. 7`
	`B. 14`

Math-Medium	a and b are natural numbers between 10 and 99.
	Example:
	`What is 59 + 63?  Respond solely with A or B.`
	`Options:`
	`A. 117`
	`B. 122`

Math-Hard	a and b are natural numbers between 100 and 999.
	Example:
	`What is 341 + 410?  Respond solely with A or B.`
	`Options:`
	`A. 658`
	`B. 751`

Table 6: Details and examples for each level in the BAD dataset.

Note: For the **MMLU dataset**, we add *Respond solely with A or B* for the baseline. For each level, the appropriate prompts are applied as described in Table 5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims presented in the introduction and abstract accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the *Limitations and future work section*.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There are no theoretical results in our work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details regarding evaluation are presented in the main text or in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have shared the link to the public repository with code and data used.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details regarding hyperparameters, prompts, model's accessed are available either in the main text or in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This will be changed in the future.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All details regarding API and computational resources used are presented in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We respect the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any model. The only data we have used was the subset of already published MMLU dataset and dataset of addition on simple numbers which does not pose any danger.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have only used MMLU dataset which os on the open license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The BAD dataset we have created is described in detail in the paper and available in our repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We have described full annotation process. There was no compensation for answering 30 question regarding number addition, all participants were volunteers.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our study only involved asking 30 questions about number addition on the primary school level difficulty. All participants were volunteers.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.