Quantifying the Bitter Lesson: How Safety Benchmarks Measure Capabilities Instead of Safety

Richard Ren[*1,2], **Steven Basart**[*1], **Adam Khoja**[1,3], **Alexander Pan**[3],

Alice Gatti[1], **Long Phan**[1], **Xuwang Yin**[1], **Mantas Mazeika**[1],

Gabriel Mukobi[1], **Ryan Hwang Kim**[1,4], **Stephen Fitz**[5], **Dan Hendrycks**[1]

[1]Center for AI Safety

[2]University of Pennsylvania

[3]University of California, Berkeley

[4]Yale University

[5]Keio University

Abstract

Performance on popular ML benchmarks is highly correlated with model scale, suggesting that most benchmarks tend to measure a similar underlying factor of general model capabilities. However, substantial research effort remains devoted to designing new benchmarks, many of which claim to measure novel phenomena. In the spirit of the Bitter Lesson, we ask whether such effort is wasteful. To quantify this question, we leverage spectral analysis to measure an underlying capabilities component, the direction in benchmark-performance-space which explains most variation in model performance. In an extensive analysis of existing safety benchmarks, we find that variance in model performance on many safety benchmarks is largely explained by the capabilities component. In response, we argue that safety research should prioritize metrics which are not highly correlated with scale. Our work provides a lens to analyze both novel safety benchmarks and novel safety methods, which we hope will enable future work to make differential progress on safety.

1 Introduction

Benchmarks serve as crucial standards, providing metrics by which models and techniques are evaluated. The AI safety community has invested extensively in creating benchmarks aimed at measuring distinct safety-relevant properties [71, 46, 38, 9, 61, 24, 65]. While these benchmarks have driven significant advancements, there is a critical oversight: the performance on safety benchmarks intended to measure bias, ethics, adversarial robustness, or fairness is often strongly correlated with general capabilities benchmarks such as MMLU [25], MATH [26], and GSM8K [14]. This correlation means that simply enhancing the general capabilities of models, such as by scaling parameters and increasing training data, often boosts performance across all benchmarks indiscriminately [33, 39].

This oversight is problematic because safety benchmarks have seldom been scrutinized for this correlation [84]. Consequently, this lack of scrutiny obscures the development of techniques that

[*]Equal Contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

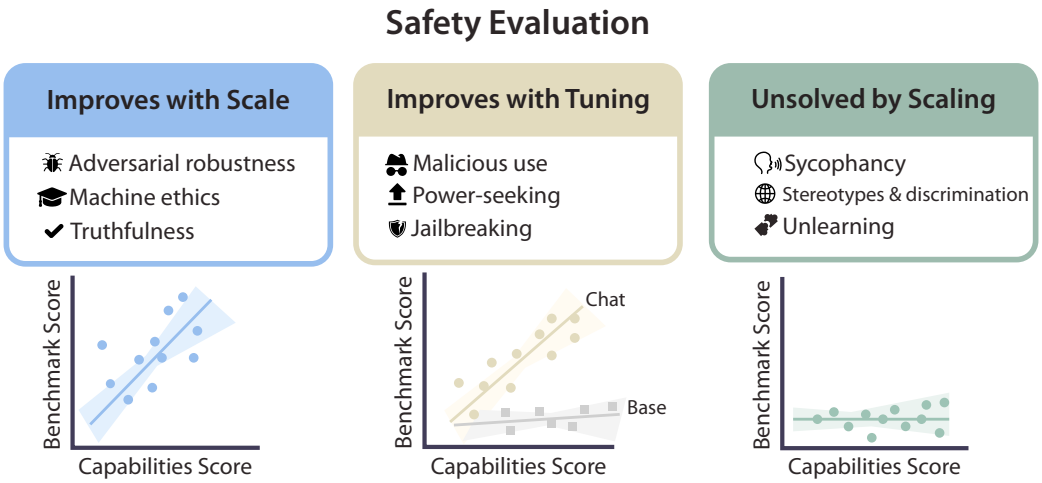

Figure 1: Our analysis identifies three classes of safety tasks according to the correlation between their scores and the capabilities scores. Tasks whose scores improve with scale have a positive correlation between benchmark scores and capabilities score. Tasks whose scores improve with tuning show a safer correlation on specific model classes, e.g., chat/instruct-tuned models. Finally, tasks whose scores do not improve naturally with model scale show no correlation between benchmark scores and capabilities scores.

specifically and differentially improve safety. Without clear and distinct metrics and goals, efforts to advance AI safety are hindered [79, 56]. The conflation of general capability improvements with safety-specific advancements not only misleads progress assessments but also undermines the incentive to develop targeted safety solutions [2]. To address this issue effectively, it is crucial to distinguish and prioritize safety-specific goals within the broader context of AI development.

Given this context, a pivotal question arises: how should the AI safety community allocate its efforts to differentially improve model safety? We can derive some insight from the "Bitter Lesson" [64], which observes that compute is becoming exponentially more available over time, and that AI research methodologies which optimize performance at a constant level of compute are subsumed by new paradigms that effectively leverage greater compute. Rather than over-indexing on the strengths and weaknesses of present-day models, this framework suggests that effective safety research should anticipate and address the flaws that will emerge or remain in future generations of models, and deemphasize issues likely to be resolved through general model scaling.

Similarly, success in new safety methods should be measured not only by improvements in safety benchmark scores, but also by how much these methods make desired safety properties more correlated with scale. For example, Reinforcement Learning from Human Feedback (RLHF) [5, 44] has successfully associated toxicity reduction with model scale, an achievement that basic pretraining and instruction fine-tuning struggled to attain. By concentrating on properties and methods that specifically enhance safety independently of capabilities advancements, the safety community can make more effective use of its resources and significantly contribute to the development of safer AI systems.

2 Related Work

Safety vs. capabilities. One paradigm of measuring AI progress is a decomposition into datasets that measure "safety" vs datasets that measure "capabilities" [23]. While the distinction between safety and capabilities is sometimes blurred, safety research tends to study empirical phenomena that are negative side effects of model deployment [78, 52, 54, 59, 47], are malicious use of models [75, 86, 35], or do not improve with scale [8, 39]. In particular, a popular debate (e.g., between McKenzie et al. [39] and Wei et al. [76]) is whether a given safety dataset is in fact tightly correlated with scale. Our work addresses this debate through a meta-analysis of safety datasets, quantifying the degree to which safety datasets are entangled with capabilities.

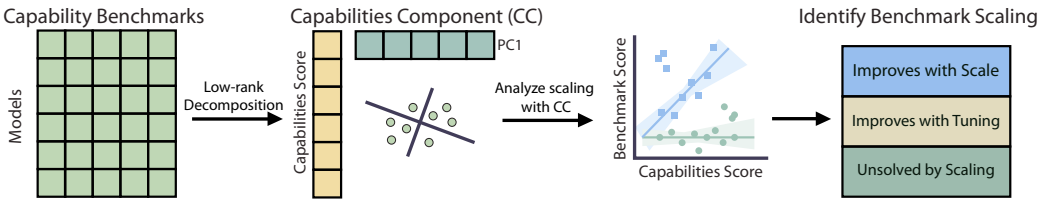

Figure 2: Illustration of the safety task identification pipeline. We first produce a matrix of scores for a set of language models evaluated on a set of capabilities benchmarks (first step). We extract the first principal component and use it to compute a capabilities score for each model (second step). We perform analysis of base and chat/instruct-tuned models on a variety of tasks representing major areas of AI safety (third step). Finally, we identify tasks whose scores are correlated with scale, tasks whose scores improve with scale only with chat/instruct-tuning, and tasks which are uncorrelated with scale (fourth step).

Scaling laws and the Bitter Lesson. The Bitter Lesson [64] argues that the main technique to improving ML models has consistently been scale. NLP has seen the biggest embrace of this trend, with developers focused on scaling the Transformer [70] with more data and compute [69, 67, 11, 1, 12, 66, 3, 4]. To aid in such engineering effort, there has been an extensive body of literature on quantitatively modeling scaling laws for loss, mapping out model performance as a function of compute and data [27, 33, 42, 28] or even hyperparameter choice [81]. Similar trends have taken hold in vision [20, 83, 22, 51] and robotics [68].

Scaling tends to improve not only training loss, but also downstream task performance [72]. A common finding is that models with lower pretraining loss also have higher accuracy on downstream tasks [77, 80, 30, 21, 19, 18, 15, 34]. Importantly, most prior work examines scaling laws from a model perspective (i.e., how does performance improve with scale), whereas our work examines scaling laws from a dataset perspective (i.e., how do benchmarks saturate with scale).

Metrics and capabilities correlations. Recent advances in observational scaling laws have provided a methodology that allows researchers to gain a deeper understanding of the underlying capabilities of machine learning models [33, 42]. Previous studies, such as those by [27, 12], have demonstrated that these scaling laws can predict model performance across various tasks. This body of work has established a foundation for using observational scaling laws as a powerful method for enhancing model training and evaluation processes by predicting performance trends based on scaling behavior. Recent research further supports the utility of these scaling laws, for extrapolating model performance, enabling a more nuanced assessment of model capabilities [60, 31].

However, while significant progress has been made in identifying and leveraging these underlying factors for general capabilities, there has been a noticeable gap in exploring how these scaling laws correlate with non-capability properties of networks, such as safety attributes. Although the identification of fundamental scaling relationships has been beneficial, there is a lack of research focusing on the implications of these relationships for safety datasets. Understanding how scaling impacts safety properties is crucial for developing datasets and benchmarks that can properly measure the intended effects and not by a "third variable" (i.e. capabilities). This paper aims to bridge this gap by examining the correlation between scaling laws and safety-specific characteristics, thereby providing insights that can guide the development of safer AI systems and future AI Safety datasets.

3 Capabilities Correlations for Evaluating Differential Progress on Safety

Estimating capabilities using benchmark scores. Inspired by prior work which applied factor analysis to matrices of model-benchmark scores [31], and concurrent to Ruan et al. [60], we apply spectral analysis of benchmark scores to identify a unified underlying *capabilities score* for models in terms of their performance on a range of benchmarks. Given a set of n models and a suite of m capabilities benchmarks (e.g. MMLU [25], Winogrande [62], GSM8K [14], etc.) we construct a matrix of scores $A \in \mathbb{R}^{n \times m}$, such that A_{ij} is the score of the i-th model on the j-th benchmark, normalized so that columns have mean 0 and variance 1.

Spectral analysis of capabilities scores. Naive composite benchmarks usually weight their component tests equally, averaging test scores. A more principled approach can involve weighting component benchmarks according to the strength of their association with each other, with higher weight placed on benchmarks that account for greater variance in model performance across benchmarks. To achieve this, we compute a correlation matrix $C \in \mathbb{R}^{m \times m}$ associated with A, such that C_{ab} is the correlation between task a and task b performance across all models. We extract the largest eigenvalue λ of C and its associated unit eigenvector v. The components of v act as the weights of the composite benchmark, and $Av \in \mathbb{R}^n$ gives the capabilities scores of each model.

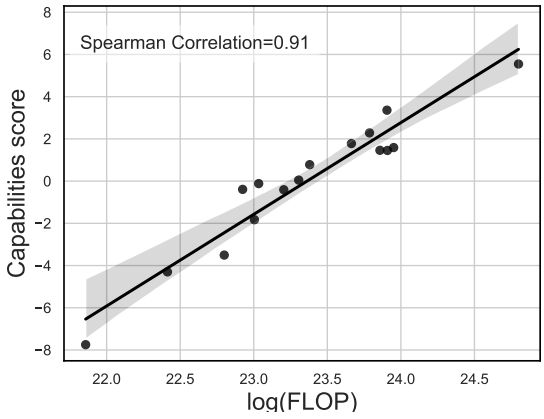

Figure 3: We observe a strong correlation between training FLOPs and relative capabilities score.

When C is the Pearson correlation matrix $A^T A$ [50], $\sqrt{\lambda}$ is the largest singular value of A, and v is its associated top principal component [17]. λ/m then represents the proportion of total variance in normalized model scores explained by the principal component vector. Additionally, the outer product of the capabilities scores and benchmark weights $(Av)v^T$ is the best rank-1 approximation of A [16]. However, using Pearson correlation can be sensitive to outliers, which becomes relevant when dealing with large model sets and a heterogeneous collection of benchmarks. For that reason, our analysis takes C to be the Spearman correlation matrix [63], in which case λ/m represents the explained variance in rank scores.

Using capabilities scores to measure capabilities correlations. To evaluate the relationship between a new benchmark and general capabilities, which we call the *capabilities correlation* of the benchmark, we can evaluate a set of models with known capability scores on the new benchmark and measure the correlation between capability scores and benchmark scores (we use Spearman correlation for these calculations as well). These general ability components allow for quantitative, intuitive, and principled evaluations of task relationship to general model abilities. Ultimately, however, these correlations depend on the set of models used, as well as the benchmarks chosen to produce their capabilities scores. In the Appendix, we perform a sensitivity analysis to explore the robustness of this methodology to different choices of models and benchmarks.

Safety techniques can alter capabilities correlations. In our analysis, we categorize models into distinct classes—base models and instruct (including chat) models—to better understand how different training paradigms impact performance on safety tasks. Base models, RLHF'ed models, light adversarial training, and future safety techniques could all be considered different model classes, with different profiles of capabilities correlations. Ideally, we should develop training regimens which produce high capabilities correlations with all relevant safety properties. By running separate analyses for each model class, we can identify the relative strengths of these techniques as models scale.

4 Results

We come to our central question: which tasks or datasets are correlated with capabilities? Towards answering this question, we analyze the overall capabilities scores as captured by tasks in 4.1, the tasks of Adversarial Robustness in 4.2, Bias and Toxicity in 4.3, Machine Ethics in 4.4, Malicious Use in 4.5, and Rogue AI Risk in 4.6.

To provide ease of understanding, we define a positive capabilities coefficient as yielding a safer system with scale, while a negative capabilities coefficient indicates less safe systems with scale.

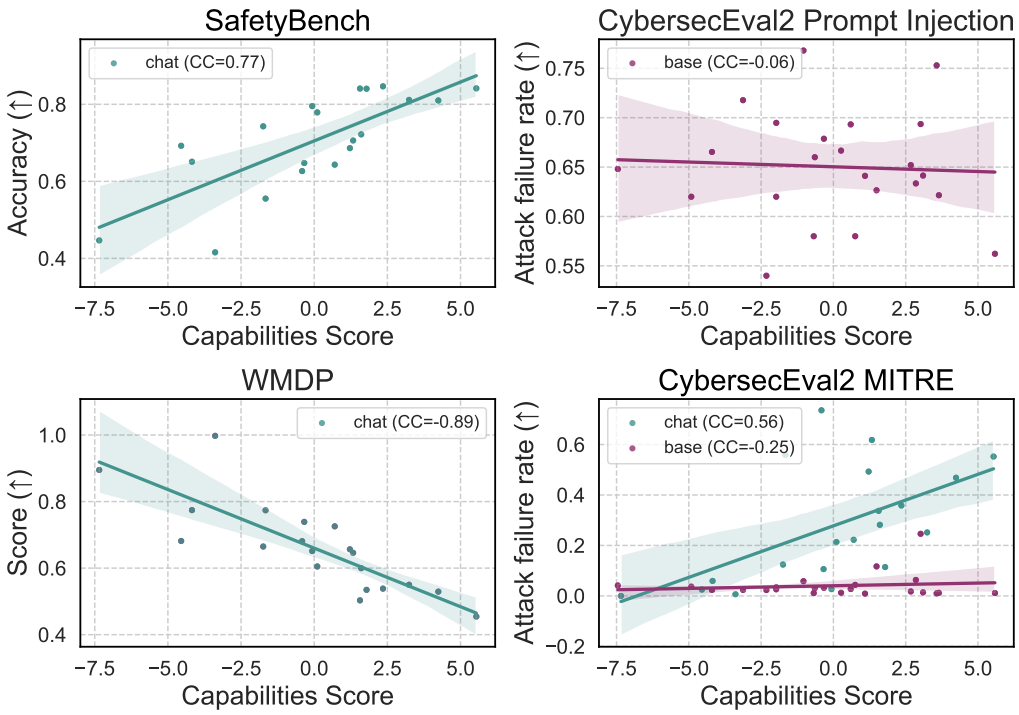

Figure 4: Observed correlations between capabilities scores and models' performance. Top left: safety task positively correlated with capabilities score. Top right: safety task not correlated with capabilities score. Bottom left: safety task negatively correlated with capabilities score. Bottom right: safety task where chat models are positively correlated with capabilities score while base models are not. CC indicates the capabilities component of the corresponding benchmark.

4.1 General Capabilities and Overview of Model Class Correlations

Most variance in capabilities datasets is explained by a capabilities component. We run analyses for base and chat models, finding that 72% and 71% of variance is captured by the capabilities component respectively. We calculate the capabilities component from the following benchmarks: LogiQA [36], PIQA [10], Hellaswag [82], Winogrande [62], COPA [58], MedQA [32], ARC Challenge [13], MMLU [25], MATH [26], LAMBADA [48], Wikitext [40], GSM8K [14], GPQA [57], and BBH [6]. We also use a diverse set of model classes and derivatives to ensure robustness in results, as results can be skewed if they come from a derivative of one model (e.g. Llama-2 [69]); we list the 24 base models and 22 instruct/chat used for our analysis in the Appendix.

The capabilities component is strongly correlated with scale. We quantify the correlation of model capabilities scores with log FLOP for base (r=0.96) and chat (r=0.96) models, and plot chat models n in Figure 5. We calculate training FLOP via the approximation of 6 * params * train_tokens described in [33].

Observed properties of capabilities correlations. In our experiments, we observe three high-level categories of result:

a.) Some safety benchmarks [84] are already highly correlated with capabilities, obeying "scaling laws" (top left).

b.) Some safety benchmarks are not aligned with scale (top right) or are negatively correlated with scale (bottom left), obtaining worse safety properties as capabilities increase. At times, these problems are not solved by any type of model class.

c.) Some correlations are strengthened or weakened through safety techniques; for example, chat models exhibit on a higher correlation on CybersecEval2 MITRE [9], a task for measuring refusal to assist in malicious cyberattacks, than base models (bottom right).

In Figure 4, we examples of these scenarios. In the following sections, we continue to explore how instruction tuning affects models and highlight the need for alternative directions to be pursued across safety areas.

4.2 Adversarial Robustness

Adversarial robustness evaluates models' ability to maintain performance when faced with adversarial examples. In the vision domain, adversarial robustness is known to have different properties from general capabilities [74, 85]. However, the relation between general capabilities and adversarial robustness is less clear for LLMs. Many different adversarial robustness benchmarks have been developed to assess different aspects of their robustness [73, 62]. We now analyze whether these benchmarks measure novel properties or are highly correlated with general capabilities.

We compute the correlation between the capability score and safety scores on the following benchmarks: AdvGLUE [73], AdvGLUE++ [29], AdvDemonstration [29], and HarmBench [38]. Full results on all datasets and model classes are in the Appendix.

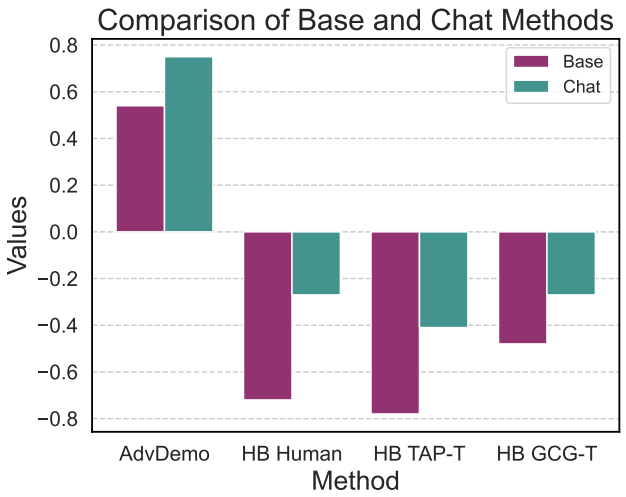

Figure 5: For many of the benchmarks we evaluate, capabilities correlations are higher (or less negative) among Chat models. This demonstrates that evaluating correlations for multiple model classes is crucial for understanding whether a benchmark will be solved as general capabilities improve.

Some robustness benchmarks are correlated with general capabilities. We find that some adversarial robustness benchmarks are moderately correlated with general capabilities, while others have low or even negative correlation. For example, AdvGLUE, AdvGLUE++, and AdvDemonstration have respective capabilities correlations of 0.68, 0.58, and 0.75 for the instruct/chat model class. On the other hand, general capabilities are anti-correlated with robustness on HarmBench. These results may be due to the different nature of the tasks in these datasets. In other words, some robustness properties are likely to be solved as general capabilities improve, while others are not yet strongly correlated with capabilities.

Different model classes have different scaling properties. Just as adversarial training significantly alters the robustness properties of vision models, different classes of general-purpose AI models can yield different scaling properties for safety benchmarks. In Figure 5, we show how some LLM adversarial robustness benchmarks have higher capabilities correlations when using instruct/chat models. This demonstrates that improving the capabilities correlation of a safety benchmark is possible. Once the correlation reaches a high enough value, additional work on the benchmark is unnecessary, as it will be solved automatically as general capabilities improve.

4.3 Bias

We investigate bias datasets aimed at quantifying language models' propagation of social stereotypes and harmful preconceptions. It is well-known that pretraining on internet data introduces bias, and one might expect that training larger models on more data would increase the amount of bias present. We test this hypothesis by measuring the capabilities coefficient of different LLM bias benchmarks.

Bias is often weakly correlated with capabilities, but not always. Our findings reveal that for some bias measures, the capabilities correlation is weak as expected. For example, in Figure 6 (left) we show that BBQ Disambiguated [49], Anthropic Discrimination Evaluation [65], and CrowS-Pairs English [43] display this pattern across both base and instruct/chat models.

Bias Evaluation	Capabilities Correlation	
	Base	Instruct/Chat
CrowS-Pairs	-0.32	0.18
Anthropic Discrim.	0.36	0.40
BBQ Disambig.	0.25	0.29

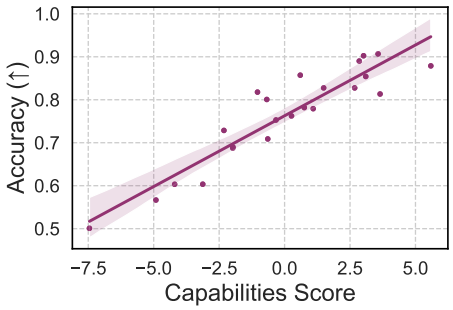

Figure 6: Left: We find that for several common bias benchmarks, bias is not reduced by general capabilities improvements, indicated by low capabilities correlations. Right: However, on BBQ Ambiguated capabilities score is strongly correlated with reducing bias.

However, for other measures, improvements to general capabilities can actually reduce bias. In Figure 6 (right), we plot the capabilities score against accuracy on BBQ Ambiguated [49] and find that bias reduction is highly correlated with general capabilities. This observation contrasts with conventional wisdom, which suggests that scaling up models exacerbates bias due to associations in the training data [7].

4.4 Machine Ethics

Machine ethics benchmarks probe models' understanding of moral concepts. There are several benchmarks that analyze machine ethics, such as ETHICS [24] and STEER Rationality [55]. We report the capabilities correlation of these benchmarks in Table 1.

Table 1: Capabilities correlations for various machine ethics datasets. For brevity, we show instruct/chat models only, although correlations are also high for base models.

Ethics Evaluation	Capabilities Correlation
ETHICS (Average)	0.80
ETHICS Commonsense	0.72
ETHICS Deontology	0.41
ETHICS Justice	0.49
ETHICS Utilitarianism	0.74
ETHICS Virtue	0.77
STEER Rationality	0.54

High capabilities correlation. We find that machine ethics benchmarks tend to be highly correlated with general capabilities. Many subsets of ETHICS have an extremely high capabilities coefficient for both base and instruct/chat models. These findings corroborate isolated observations of scale improving performance on machine ethics benchmarks [33, 37], indicating that internet-scale pretraining imbues LLMs with an understanding of ethics and morality. However, our results also show that this correlation is not identical across all areas of machine ethics. Some topics improve much more slowly with general capabilities, suggesting a need to ensure a balanced understanding of different ethical perspectives is present in models.

4.5 Malicious Use

Malicious use evaluations test whether models can resist being exploited for harmful ends, including spreading misinformation or enabling cybercrime. Benchmarks like HarmBench [38], CyberSecEval2 [9], and WMDP [35] are used to assess the susceptibility of models to malicious use. To bypass refusal training, many of these evaluations also employ adversarial prompts. We analyze the capabilities coefficients of resistance to malicious use benchmarks under current models and present results in Table 2.

General capabilities exacerbate malicious use. Many base models cause more harmful responses as their capabilities increase, as indicated by negative capabilities correlations. This includes many splits of HarmBench and CyberSecEval2, as well as WMDP (an unlearning dataset that penalizes high performance).

We find that instruction tuning weakens many capabilities correlations, indicating that models no longer become less safe with scale. In the MITRE task of CyberSecEval2, which measures refusal to

Rogue AI Evaluation	Capabilities Correlation
MACHIAVELLI Power	0.46
MACHIAVELLI Utility	0.48
MACHIAVELLI Violations	0.55
Sycophancy	-0.73
TruthfulQA MC1	0.83

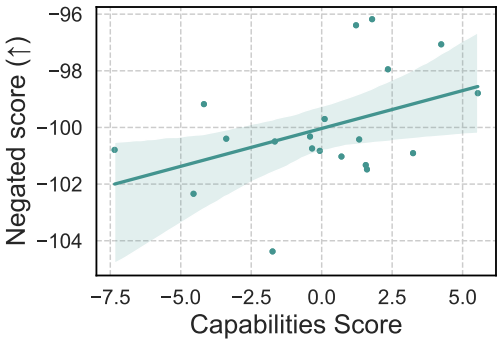

Figure 7: Left: Capabilities correlations on instruct/chat models on Rogue AI evaluations. Right: Capabilities correlations on instruct/chat models with accuracy on MACHIAVELLI Power.

participate in cyberattacks, the effect is even stronger, with the capabilities correlation changing from negative to positive.

These results demonstrate that instruct/chat models have improved over base models in their ability to leverage general capabilities to reduce malicious use risk. However, in most cases the correlations remain negative or weak, suggesting there is still considerable work to be done on this problem.

4.6 Rogue AI

Rogue AI risk evaluations probe risks related to deceptive model behavior, dishonesty, and power-seeking tendencies. Previously, it was unknown whether models become more power-seeking as they scale. We report the capabilities correlations of these benchmarks in Figure 7 (left).

Power-seeking tendencies decrease with scale, but sycophancy does not. On the MACHIAVELLI dataset [46], we find that measures of power-seeking tendencies and ethical violations decrease as general capabilities improve, with moderate capabilities correlations ranging from 0.46 to 0.55. On the other hand, sycophancy [53] becomes worse as models become more capable, with a capabilities correlation of -0.73. This highlights how different aspects of rogue AI risk are correlated with general capabilities to different extents.

Unlike power-seeking and sycophancy, we find that TruthfulQA MC1 variance is strongly correlated with general capabilities. This could be explained by training leakages or may indicate that models are able to discern fact from human falsehood as capabilities advance. Regardless, we find that TruthfulQA does not seem to measure a meaningfully different metric from capabilities benchmarks.

Table 2: Malicious Use Evaluations and Metrics

Malicious Use Evaluation	Capabilities Correlation	
	Base	Instruct/Chat
HarmBench DR		
Biochemical	-0.54	-0.04
Cybercrime	-0.50	-0.07
Harassment	-0.45	-0.16
Harmful	-0.42	0.24
Illegal	-0.41	0.09
Misinfo	-0.44	-0.37
WMDP		
WMDP Bio	-0.91	-0.87
WMDP Chem	-0.88	-0.86
WMDP Cyber	-0.86	-0.87
CybersecEval2		
Autocomplete	-0.74	-0.77
Exploit	-0.31	-0.49
Instruct	-0.43	-0.90
MITRE	-0.25	0.55
Prompt Injection	-0.02	-0.17

5 Discussion

As our experiments show, safety and capabilities metrics can be intertwined, with some safety metrics improving naturally as a consequence of general capabilities advancements. We discuss several implications of this finding below.

The importance of low capabilities correlation. While many datasets measure interesting aspects of safety, these aspects are often not unique and instead are highly correlated with general capabilities. We argue researcher time for developing methods should be allocated toward solving benchmarks that won't be solved with scale and general capabilities advancements. Thus, capabilities correlation can be used as a metric for identifying which problems to spend research effort making progress on. If capabilities correlation for a benchmark is low, this means it will likely require additional algorithmic effort to make progress on.

Measuring properties that improve with scale is still valuable. Even if an evaluation is expected to be solved with scale, it is still useful to measure it. For instance, knowing the dangerous capabilities that emerge with scale is crucial. Reporting how evaluations scale with model size can help predict future risks, particularly with dangerous capabilities. This information is highly relevant, as it can indicate which problems may worsen with scale. The goal is not to measure the usefulness of a safety dataset, but to understand how to allocate research efforts efficiently. Evaluations showing strong correlation or anti-correlation with capabilities are valuable for tracking the evolution of dangerous capabilities and ensuring precise measurement of safety metrics, even if they are eventually solved with scale.

Improving capabilities correlation as a goal for safety research. If a meaningful safety metric is strongly correlated with general capabilities, this is a good outcome, because it means the problem will likely be solved by scaling even if present-day models struggle. As a corollary, safety research should seek to develop new methods and model classes that cause safety metrics to correlate more strongly with capabilities. However, this should not be taken too far. Past a certain threshold, further efforts to align safety metrics with general capabilities are unnecessary. Once this is achieved, research efforts can be re-allocated elsewhere.

New safety evaluations should report their correlation with capabilities. This practice can ensure that evaluations initially measure a meaningful safety property that requires research effort to improve, rather than simply increased scale. This is notably done in some past papers, such as RuLES [41] and EQ Bench [45]. For example, it can be useful to know if a malicious use benchmark worsens with scale. A low correlation does not necessarily imply that the safety metric is irrelevant; it may indicate flaws in the dataset, such as insufficient data to detect small changes in progress, or that the dataset measures a different aspect entirely.

Broad application of correlation analysis. This analysis can be applied in a broad range of scenarios when determining whether an evaluation measures a meaningfully different property. In a broad range of scenarios, it may be the case that a confounding variable that better explains performance, rather than what a benchmark claims it measures. Future investigations can also investigate correlations of safety with various types of capabilities; while previous research already shows that performance on different categories of capabilities (such as reasoning, knowledge, coding, or mathematics) seem to be tied to scale, other papers have found that reasoning and knowledge can represent different components of a PCA analysis [17].

6 Conclusion

We have shown that a wide variety of safety benchmarks are tightly correlated with general model capabilities, calling their importance into question. By considering the Bitter Lesson and the continued scaling of deep learning, we argued that research in AI safety should anticipate the possibility of safety metrics being correlated with general capabilities, such that they are naturally solved with scale. To quantify this, we developed a methodology to measure the correlation of safety metrics with general capabilities via spectral analysis of accuracy on capabilities datasets. In experiments, we measured the capabilities correlation of a wide variety of safety benchmarks and datasets, finding that many prior datasets are strongly correlated with general capabilities. We make two specific recommendations: future safety benchmarks should aim for low correlation with general capabilities, while future safety methods should aim to increase correlation between relevant safety metrics and general capabilities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016.

[3] Anthropic. Introducing Claude. https://www.anthropic.com/index/claude-2, 2023.

[4] Anthropic. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family, 2024.

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[6] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023. ISSN 2835-8856.

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922.

[8] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". arXiv preprint arXiv:2309.12288, 2023.

[9] Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models, 2024.

[10] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei,

Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113, 2023.

[13] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv:1803.05457v1, 2018.

[14] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

[15] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. arXiv preprint arXiv:2403.15796, 2024.

[16] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. Psychometrika, 1:211–218, 1936. doi: 10.1007/BF02288367.

[17] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

[18] Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. arXiv preprint arXiv:2109.07740, 2021.

[19] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-supervised pretraining of visual features in the wild, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[21] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4917–4926, 2019. doi: 10.1109/ICCV.2019.00502.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.

[23] Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research, 2022.

[24] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[26] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. NeurIPS, 2021.

[27] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409, 2017.

[28] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[29] Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, et al. Decoding compressed trust: Scrutinizing the trustworthiness of efficient llms under compression. arXiv preprint arXiv:2403.15447, 2024.

[30] Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. arXiv preprint arXiv:2404.09937, 2024.

[31] David Ilić. Unveiling the general intelligence factor in language models: A psychometric approach, 2023.

[32] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14):6421, 2021.

[33] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019.

[35] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024.

[36] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020.

[37] Nestor Maslej, Luciano Fattorini, Rowan Perrault, Valerio Parli, Anna Reuel, Erik Brynjolfsson, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Michael Waldrop, Raymond Li, Bart Selman, and Eric Horvitz. Artificial intelligence index report 2023, 2023.

[38] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. 2024.

[39] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. arXiv preprint arXiv:2306.09479, 2023.

[40] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

[41] Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. Can llms follow simple rules? arXiv, 2023.

[42] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. Advances in Neural Information Processing Systems, 36, 2024.

[43] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154.

[44] 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.

[45] Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2023.

[46] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org, 2023.

[47] Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. arXiv preprint arXiv:2402.06627, 2024.

[48] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1144.

[49] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165.

[50] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, 1901.

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4195–4205, 2023.

[52] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

[53] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela

Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847.

[54] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv preprint arXiv:2310.03693, 2023.

[55] Narun Raman, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. Steer: Assessing the economic rationality of large language models, 2024.

[56] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, and Deepak John Reji. Safe and responsible large language model development, 2024.

[57] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

[58] Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. 01 2011.

[59] Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of lm agents with an lm-emulated sandbox. arXiv preprint arXiv:2309.15817, 2023.

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

[61] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024.

[62] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. arXiv preprint arXiv:1907.10641, 2019.

[63] C. Spearman. The proof and measurement of association between two things. The American Journal of Psychology, 15(1):72–101, 1904.

[64] Rich Sutton. The bitter lesson, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[65] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023.

[66] Gemini Team. Introducing gemini: our largest and most capable ai model. Google Blog, 2023. URL https://blog.google/technology/ai/google-gemini-ai/.

[67] Llama Team. Llama 3: An open large language model. Meta AI Blog, 2023. URL https://ai.meta.com/blog/meta-llama-3/.

[68] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213, 2024.

[69] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[71] Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, Leon Derczynski, Debojyoti Dutta, Ian Eisenberg, James Ezick, Heather Frase, Brian Fuller, Ram Gandikota, Agasthya Gangavarapu, Ananya Gangavarapu, James Gealy, Rajat Ghosh, James Goel, Usman Gohar, Sujata Goswami, Scott A. Hale, Wiebke Hutiri, Joseph Marvin Imperial, Surgan Jandial, Nick Judd, Felix Juefei-Xu, Foutse Khomh, Bhavya Kailkhura, Hannah Rose Kirk, Kevin Klyman, Chris Knotz, Michael Kuchnik, Shachi H. Kumar, Srijan Kumar, Chris Lengerich, Bo Li, Zeyi Liao, Eileen Peters Long, Victor Lu, Sarah Luger, Yifan Mai, Priyanka Mary Mammen, Kelvin Manyeki, Sean McGregor, Virendra Mehta, Shafee Mohammed, Emanuel Moss, Lama Nachman, Dinesh Jinenhally Naganna, Amin Nikanjam, Besmira Nushi, Luis Oala, Iftach Orr, Alicia Parrish, Cigdem Patlak, William Pietri, Forough Poursabzi-Sangdeh, Eleonora Presani, Fabrizio Puletti, Paul Röttger, Saurav Sahay, Tim Santos, Nino Scherrer, Alice Schoenauer Sebag, Patrick Schramowski, Abolfazl Shahbazi, Vin Sharma, Xudong Shen, Vamsi Sistla, Leonard Tang, Davide Testuggine, Vithursan Thangarasa, Elizabeth Anne Watkins, Rebecca Weiss, Chris Welty, Tyler Wilbers, Adina Williams, Carole-Jean Wu, Poonam Yadav, Xianjun Yang, Yi Zeng, Wenhui Zhang, Fedor Zhdanov, Jiacheng Zhu, Percy Liang, Peter Mattson, and Joaquin Vanschoren. Introducing v0.5 of the ai safety benchmark from mlcommons, 2024.

[72] Pablo Villalobos. Scaling laws literature review. Epoch AI, 2023. URL https://epochai.org/blog/scaling-laws-literature-review.

[73] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In Advances in Neural Information Processing Systems, 2021.

[74] Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In CVPR, 2024.

[75] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems, 36, 2024.

[76] Jason Wei, Najoung Kim, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. arXiv preprint arXiv:2211.02011, 2022.

[77] 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. arXiv preprint arXiv:2206.07682, 2022.

[78] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359, 2021.

[79] Laura Weidinger, Joslyn Barnhart, Jenny Brennan, Christina Butterfield, Susie Young, Will Hawkins, Lisa Anne Hendricks, Ramona Comanescu, Oscar Chang, Mikel Rodriguez, Jennifer Beroshi, Dawn Bloxwich, Lev Proleev, Jilin Chen, Sebastian Farquhar, Lewis Ho, Iason Gabriel, Allan Dafoe, and William Isaac. Holistic safety and responsibility evaluations of advanced ai models, 2024.

[80] Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. arXiv preprint arXiv:2212.09803, 2022.

[81] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022.

[82] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[83] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12104–12113, 2022.

[84] 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 with multiple choice questions. arXiv preprint arXiv:2309.07045, 2023.

[85] Wanqi Zhou, Shuanghao Bai, Qibin Zhao, and Badong Chen. Revisiting the adversarial robustness of vision language models: a multimodal perspective, 2024.

[86] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043, 2023.

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 are presented and discussed thoroughly in the results and appendix.

   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: We cover this in the discussion section but also in the appendix.

   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: No theoretical results presented.

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: The method as described in 3 and the datasets as listed in 4 should be sufficient to reproduce the work. We will also be releasing all relevant code.

   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 will release all of our code. All of the data sources are public. We will work to provide an anonymized evaluation script that evaluates on the data.

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 of the evaluations used the default parameters from their respective repositories such as lm-eval-harness or from RULES etc.

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: [Yes]

Justification: The plots report the error bars all of the other data is derived from primary sources and therefore have their own internal noise.

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: We include this in the appendix but all experiments were run on A100-80 GB DGX boxes.

   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: The paper does not cause any harms directly or indirectly to people. It also has proper attribution.

   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: The paper broadly discusses the impacts that this work has on evaluation of AI Safety benchmarks and datasets. It also discusses the monetary implications briefly.

    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 new models or datasets along with this paper.

    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: All of the datasets we use are open-source and cited. We also create forks of the repositories used to be able to recreate the work.

    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: [NA]

    Justification: We do not present any new datasets or assets and instead repurpose existing datasets into a benchmark.

    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: [NA]

    Justification: We do not use human subjects.

    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: [NA]

    Justification: We do not use human subjects.

    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.