CORRELATING AND PREDICTING HUMAN EVALUATIONS OF LANGUAGE MODELS FROM NATURAL LANGUAGE PROCESSING BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The field of natural language processing (NLP) historically evaluated language models using benchmarks with automated metrics. However, the recent advent of highly capable chat language models (LMs) has caused a tectonic shift from NLP benchmarks to human evaluations. The relationship between these two evaluation processes is unclear and underexplored for chat LMs. Broadly, to what extent are human evaluations and NLP benchmarks correlated with one another? How well can computationally inexpensive and automated benchmarks predict expensive and time-intensive human evaluations? Which benchmarks provide predictive signals for human preference for LMs? What role, if any, should benchmarks play in the era of chat LMs? To answer these questions, we conducted a large-scale study of the relationships between human evaluations and benchmarks. We show that benchmarks are broadly highly correlated with human evaluations, and we identify which benchmarks exhibit strong correlations with human evaluations and which do not. Having established that reliable correlations exist, we fit models to predict a language model's human evaluation scores from its academic evaluation scores and provide evidence that such predictive models can generalize across LM scales.

1 INTRODUCTION

For decades, the field of natural language processing (NLP) has relied on academic benchmarks and automated metrics (e.g., Accuracy, Brier Score (Brier, 1950), BLEU Papineni et al. (2002)) to evaluate the performance of language models (LMs). These NLP benchmarks provide a standardized and efficient way to measure model capabilities such as machine translation, text summarization, and question answering (Wang et al., 2018; 2019; Srivastava et al., 2022; Gao et al., 2023; Wang et al., 2023a). However, the recent emergence of highly capable chat LMs such as GPT (Ouyang et al., 2022; Achiam et al., 2023), Llama (Touvron et al., 2023a;b; Dubey et al., 2024), Gemini (Team et al., 2023; Reid et al., 2024) and Claude (Anthropic, 2023) has prompted a re-evaluation of how we assess LMs, with a growing emphasis on assessing LMs based on their ability to interact with and assist human users in real-world scenarios (Zheng et al., 2023; Reuel et al., 2024).

This shift towards human evaluations raises important questions about the relationship between NLP benchmarks and human evaluations of chat LMs. Additionally, human evaluations are not without challenges; they can be expensive, time-intensive and noisy, in contrast with computationally cheaper, faster and precise benchmarks. In this paper, we aim to explore the relationship between human evaluations and NLP benchmarks in pursuit of understanding what role, if any, benchmarks should play in the era of chat LMs. As shown in Fig. 1, we seek to answer two key research questions:

1. To what extent are human evaluations and NLP benchmarks correlated with one another?
2. How well can NLP benchmarks predict human evaluations?

To answer these questions, we conducted a large-scale study comparing human evaluations and NLP benchmarks using four Llama 2 Chat language models (LMs) (Touvron et al., 2023b). For human evaluations, we constructed a large-scale dataset of single-turn and multi-turn prompts from a diverse taxonomy (Fig. 2) and collect high quality pairwise preference data of the four Chat Llama 2 models against GPT 3.5 (Ouyang et al., 2022) from paid human annotators. For NLP benchmarks,

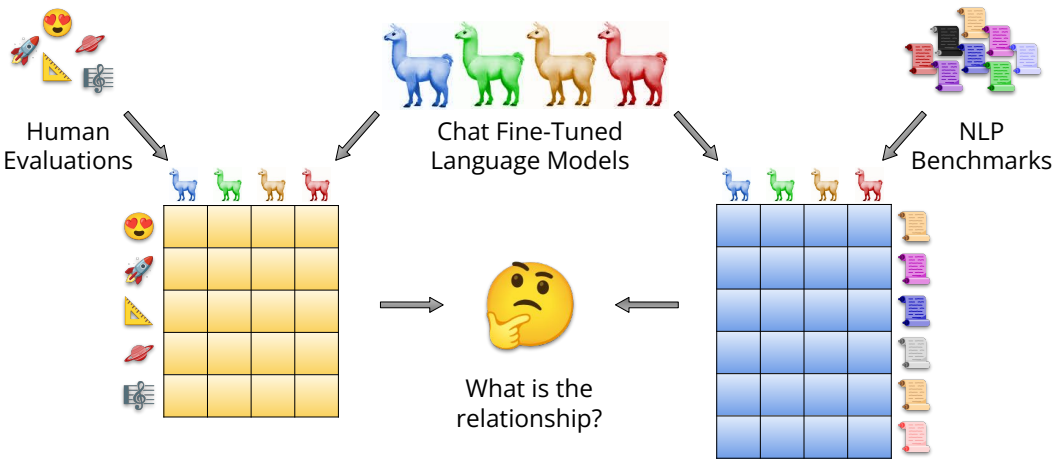

Figure 1: **Correlating and Predicting Human Evaluations of Language Models from Natural Language Processing (NLP) Benchmarks.** We evaluate chat language models on conversational tasks with human pairwise evaluations and on standard NLP benchmarks with automated metrics, then study whether scores on computationally inexpensive and fast NLP benchmarks are correlated with and predictive of expensive and time-intensive human evaluations?

we evaluate the same four Chat Llama 2 models on standard NLP benchmarks under established evaluation processes (metrics, prompting, 0-shot/few-shot, etc.). We analyze pairwise correlations between NLP benchmark and human evaluations to identify which NLP benchmarks correlate highly with human evaluations and which do not. We also aim to identify which human evaluations, if any, are uncorrelated with any NLP benchmarks. We then pivot to predicting human evaluations from NLP benchmarks using overparameterized linear regressions and leave-one-out cross-validation. We investigate the extent to which NLP benchmarks can predict human evaluations.

2 Related Work

The evaluation of language models has a rich and constantly evolving history. Human evaluations have long been considered the gold standard (Gatt & Krahmer, 2018; Van Der Lee et al., 2019; Celikyilmaz et al., 2020; Roller et al., 2020; van der Lee et al., 2021), despite serious objections raised regarding the collection, analysis, and interpretation of human evaluation scores (Novikova et al., 2018; Howcroft et al., 2020; Bowman & Dahl, 2021; Karpinska et al., 2021; Clark et al., 2021; Smith et al., 2022; Gehrmann et al., 2023; Finch et al., 2023). Many classic NLP benchmark metrics, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), ROUGE (Lin, 2004), and METEOR (Banerjee & Lavie, 2005), were introduced on the premise that they correlate with human judgments. However, subsequent studies revealed that the relationship between automated metrics and human evaluations is often complex and not straightforward (Liu et al., 2016; Novikova et al., 2017; Reiter, 2018; Karpinska et al., 2021). Another prominent class of evaluation methods are based on machine learning models, e.g., word mover distance (Kusner et al., 2015) and BERT-Score (Zhang et al., 2019) that have since evolved into using chat LMs themselves as evaluators (Wang et al., 2023b; Zheng et al., 2024; Chiang & yi Lee, 2023; Chan et al., 2023; Bavaresco et al., 2024; Fu et al., 2024), albeit with limitations, e.g., (Dorner et al., 2024; Szymanski et al., 2024; Thakur et al., 2024).

The earliest investigations into the general relationship between NLP benchmark scores and human evaluations date back to Bangalore et al. (2000), Belz & Reiter (2006), and Liu et al. (2016). In the context of natural language generation, Clinciu et al. (2021) found that embedding-based automated metrics (e.g., BERT-Score (Zhang et al., 2019) and BLEURT Sellam et al. (2020)) correlate more strongly with human judgments compared to word-overlap metrics (e.g., ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)). In the domain of natural language inference, Schuff et al. (2021) found that automated metrics do not appear to correlate with human judgment scores. However, the majority of these works predate the current era of chat LMs, which exhibit significantly more advanced capabilities compared to their predecessors. This new era motivates our work to investigate the relationship between NLP benchmarks and human evaluations when evaluating chat LMs.

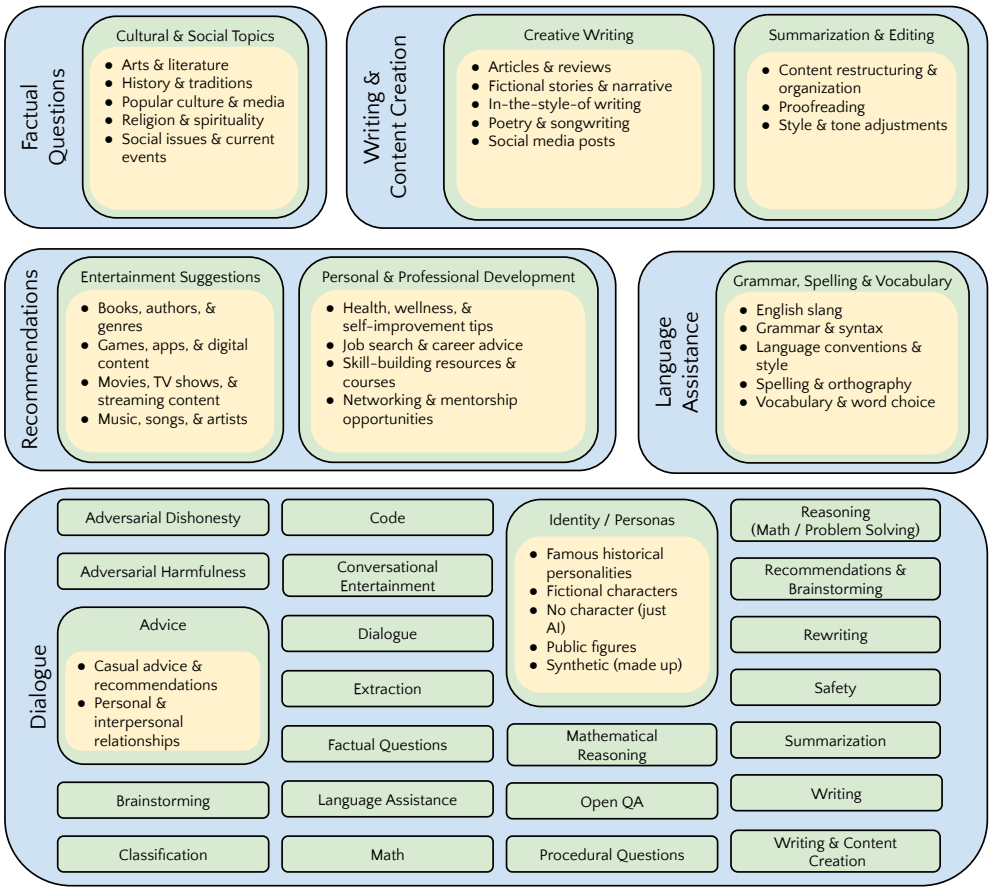

Figure 2: **Human Evaluations: Taxonomy of Single-Turn and Multi-Turn Conversations.** Single-turn and multi-turn prompts were created in a hierarchical taxonomy of 9 areas (blue), categories (green) and subcategories (yellow). Chat Llama 2 generations were then rated against ChatGPT generations by paid human annotators on a 7 point Likert scale (Likert, 1932).

3 Methods: Models, Human Evaluations and NLP Benchmarks

We briefly outline our methodology here; for additional information, please see Appendix A.

Models Our paper leverages the Llama 2 model family, consisting of four Chat LMs with 7, 13, 34, and 70 billion parameters pre-trained on 2 trillion tokens and finetuned using supervised finetuning (Sanh et al., 2021; Chung et al., 2022; Longpre et al., 2023) and reinforcement learning from human feedback (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020). We chose the Llama 2 models because at the time we collected our data, the Llama 2 family contained leading open-access chat-finetuned models spanning multiple scales with minimal variations in architecture, ensuring consistency in our analyses and a robust foundation for our investigations.

Human Evaluations: Single Turn & Multi-Turn In this work, our aim was specifically to identify which NLP benchmark scores are predictive of human preferences on open-ended prompts representative of real-world chat model usage. We chose this approach to maximize the ecological validity and generalizability of the findings to real-world use cases. For a concrete example, we may want our chat language models (LMs) to excel at providing bespoke career advice; which NLP benchmarks provide useful signals for whether models are improving at such tasks?

To answer such questions, we created a taxonomy of single-turn and multi-turn interactions (Fig. 2) between chat LMs and humans. For single-turn interactions, we generated a diverse set of prompts

spanning common areas of interest: Factual Questions, Procedural Questions, Language Assistance, Writing & Content Creation, Dialogue, Code, Reasoning, Recommendations / Brainstorming and Safety, with nested categories and subcategories. For multi-turn prompts, non-annotator humans were asked to have conversations (3 to 15 turns long) with all models on similar topics of interest: Factual Questions, Procedural Questions, Language Assistance, Writing & Content Creation, Summarization & Editing, General Dialogue, Reasoning and Recommendations / Brainstorming. This taxonomy was chosen to broadly cover common use-cases of Chat LMs. Example prompts include: "What is the tallest mountain in the world?" (Factual Question); "How do I make minestrone soup?" (Procedural Question); "Please make this sentence more friendly: I need you to stop parking in my space" (Language Assistance); "Write me a poem about getting to the weekend after a long day at work" (Writing & Content Creation). See Appendix A.2 for more information.

We then paid human annotators to evaluate each of the four Chat Llama 2 models against ChatGPT 3.5 (Ouyang et al., 2022) (gpt-3.5-0301) on a dataset of single-turn and multi-turn prompts (Fig 2). We chose gpt-3.5-0301 because, at the time this data was collected, gpt-3.5-0301 was a good balance of three desirable properties for our study: performant, cheap, and stable. For each pair of conversations (one conversation with Chat Llama responses and the other with ChatGPT responses), at least three unique human annotators independently indicated which conversation was preferred using a Likert scale (Likert, 1932) from 1 to 7, where 1 denotes the Chat Llama model was strongly preferred and 7 denotes gpt-3.5-0301 was strongly preferred. Across the 11291 single-turn samples and 2081 multi-turn samples, we had at least 3 unique human annotators per pairwise comparison, with 2104 unique human annotators overall. For our analyses, we averaged the annotators' scores for each pairwise comparison to give us an average human evaluation score per datum.

Natural Language Processing (NLP) Benchmarks We evaluated the four Chat Llama 2 models on large-scale and commonly-used NLP benchmarks: AGI Eval (Zhong et al., 2023), AI2 Reasoning Challenge (ARC; both Easy and Hard) (Clark et al., 2018), BIG Bench Hard (Srivastava et al., 2022; Suzgun et al., 2022) BoolQ (Clark et al., 2019), CommonSenseQA (Talmor et al., 2019), COPA (Roemmele et al., 2011), DROP (Dua et al., 2019), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), HumanEval (Chen et al., 2021), InverseScaling (McKenzie et al., 2022a;b; 2023), MBPP (Austin et al., 2021), MMLU (Hendrycks et al., 2020), Natural Questions (Kwiatkowski et al., 2019), OpenbookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), QuAC (Choi et al., 2018), RACE (Lai et al., 2017), SIQA (Sap et al., 2019), SQUAD (Rajpurkar et al., 2016), TLDR (Völske et al., 2017), TriviaQA (Joshi et al., 2017), WinoGrande (Sakaguchi et al., 2021) and XSum (Narayan et al., 2018). Some of these benchmarks (e.g., MMLU) contain subsets (e.g., Jurisprudence) that we do not aggregate over. These tasks cover commonsense reasoning, world knowledge, reading comprehension, coding and more. We used standard evaluation processes for all academic benchmarks including prompt formatting, metrics, 0-shot/few-shot, etc. This structured approach facilitates an exhaustive examination of model performances across varied metrics. For more information, see Appendix A.1.

Scores for Subsequent Analyses For each dataset and evaluation process (either human or NLP), we average each model's scores across all samples, yielding two matrices of scores:

$$X_{\text{NLP}} \in \mathbb{R}^{160\times 4} \qquad X_{\text{Human}} \in \mathbb{R}^{55\times 4}$$

Here, 4 is the number of models, 160 is the number of NLP benchmarks per model and 55 is the number of human evaluation area-category-subcategory scores per model. We subsequently study the correlations between X_{NLP} and X_{Human}, then test how well X_{NLP} can predict X_{Human}.

4 Correlating Human Evaluations with NLP Benchmarks

We began by computing correlations between human evaluations and NLP benchmarks, computing three standard correlations over the 4 average scores per model — Pearson (Galton, 1877), Spearman (Spearman, 1904) and Kendall (Kendall, 1938) — giving us three correlation matrices of shape 160×55 between every pair of NLP benchmark and human evaluation area-category-subcategory (Fig. 3). Pearson correlation measures the linear relationship between two continuous variables, whereas Spearman and Kendall correlations assess the monotonic relationship between two variables; Spearman correlation is based on the rank order of the data points, whereas Kendall correlation is determined by the number of concordant and discordant pairs. By using different correlation metrics, we aim to robustly characterize the relationships between human and NLP benchmarks.

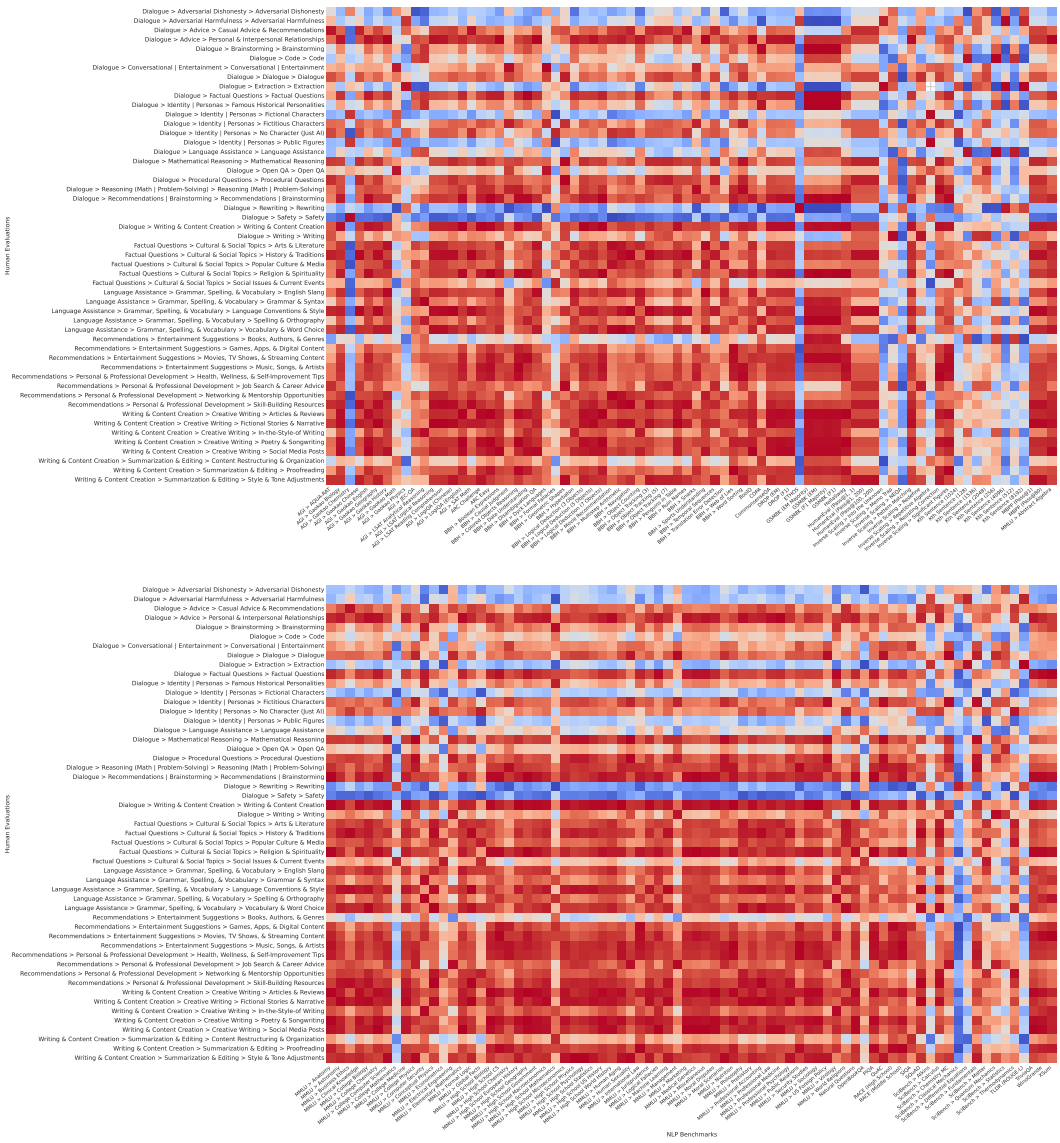

Figure 3: **Pearson Correlations Between Human Evaluations and NLP Benchmarks.** Rows: Human evaluation areas-categories-subcategories. Columns: NLP benchmarks. The heatmap is row-wrapped to fit on the page. Large positive correlations (+1) are shown in red. Large negative anticorrelations (-1) are shown in blue. Low uncorrelations ($\sim$0) are shown in light-white-gray.

Macroscopically, at the most coarse grouping of human evaluations in our taxonomy (i.e., areas) (Fig. 2), we found that average NLP benchmark scores are highly correlated with average human scores for all human evaluation areas under all three correlation metrics (Fig. 4 top). Due to the small number of models ($N = 4$), Spearman and Kendall correlations suffer discretization effects (Fig. 11), inducing an illusion of undulations. These strong correlations suggest that, at a high level, NLP benchmarks are reasonable proxies for human judgments of LM quality.

Mesoscopically, at the level of human evaluation areas and categories, we find that NLP benchmarks remain highly correlated with human evaluations, with two notable types of exceptions (Fig. 4). First, Adversarial Dishonesty, Adversarial Harmfulness, and Safety are anti-correlated with most NLP benchmarks, potentially indicating that these adversarial and safety-focused categories are more easily transgressed by more capable LMs; an alternatively hypothesis could be that safety benchmarks simply are not especially good, as demonstrated by Ren et al. (2024). Second, Language Assistance and Open Question Answering are uncorrelated with most NLP benchmarks, suggesting that these categories may require new NLP benchmarks. Open Question Answering was surprising given that

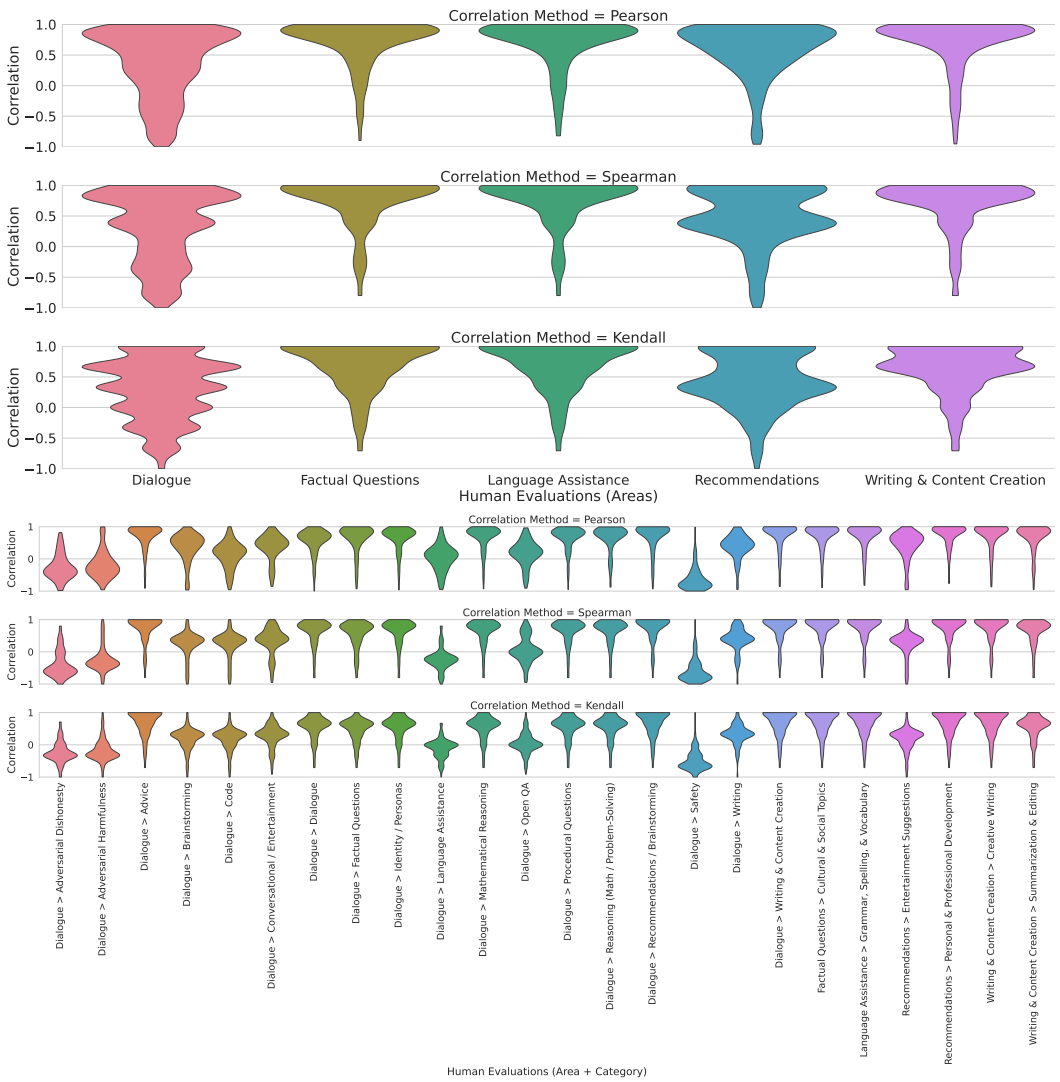

Figure 4: **Distributions of Correlations between Human Evaluations and NLP benchmarks.** Top: Macroscopically, for each human evaluation area, Chat LM scores are typically highly correlated with NLP benchmarks. Bottom: Mesoscopically, human and NLP benchmarks remain positively correlated, with notable exceptions: Adversarial Dishonesty, Adversarial Harmfulness and Safety are anticorrelated with most NLP benchmarks, and Language Assistance and Open QA are uncorrelated.

some of our NLP benchmarks are open question answering datasets, e.g., OpenBookQA (Mihaylov et al., 2018). We found the three correlations metrics visually agreed with one another and were themselves tightly coupled (App. Fig. 11), and so we present only one (Pearson) moving forward, with equivalent plots of the other two (Spearman, Kendall) deferred to the appendix.

4.1 Which human evaluations have few-to-no correlated NLP benchmarks?

To the best of our ability to discern, none. Every human evaluation seemed to have at least some NLP benchmarks that were either correlated or anticorrelated with it. This result is promising because it suggests human evaluations might be predictable from NLP benchmarks (Sec. 5).

4.2 Which NLP benchmarks exhibit high correlations with human evaluations?

To answer this question, we ordered NLP benchmarks based on their average correlation score with all of the human evaluation areas, categories and subcategories. We found many NLP benchmarks

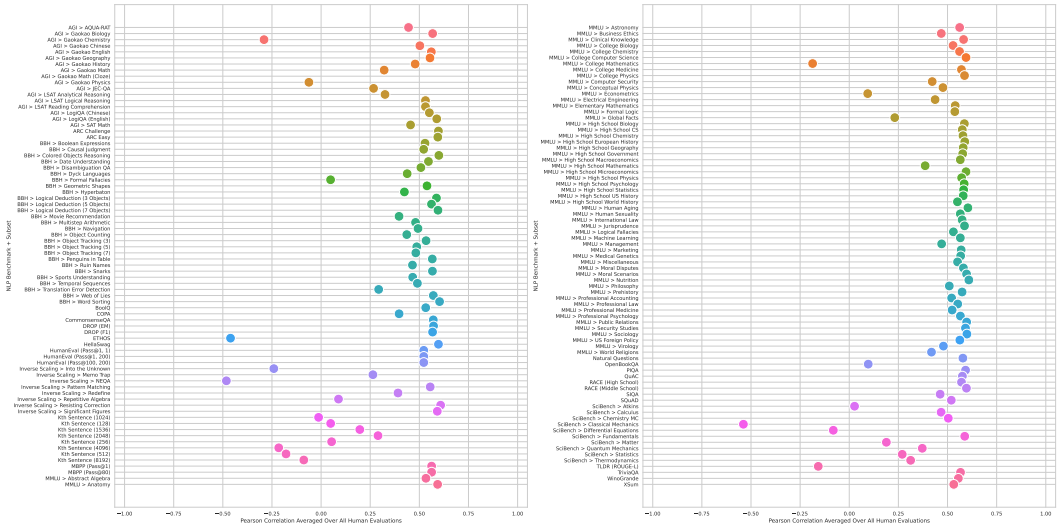

Figure 5: **NLP Benchmarks Ranked by Average Pearson Correlation over All Human Evaluations.** Certain benchmarks have higher correlations with human evaluations, including a subset of MMLU, a subset of BIG Bench Hard, HellaSwag, ARC, RACE, PIQA, NaturalQuestions, QuAC, and CommonSenseQA. Other benchmarks were weakly or uncorrelated with human evaluations: ETHOS, Kth Sentence, Inverse Scaling (with the exception of Resisting Correction Classification), OpenBookQA, COPA, SciBench (with the exception of Fundamentals of Physics) and SIQA.

have high average correlation with human evaluations (Fig. 5); the highest average correlation NLP benchmarks include a subset of MMLU (Nutrition, Human Aging, Sociology, Public Relations, Moral Scenarios, College Computer Science), a subset of BIG Bench Hard (Word Sorting, Reasoning About Colored Objects, Logical Deduction), HellaSwag, ARC, RACE, PIQA, NaturalQuestions, QuAC, CommonSenseQA, DROP and TriviaQA. Other benchmarks were less correlated or uncorrelated with human evaluations: ETHOS, Kth Sentence, Inverse Scaling (with the exception of Resisting Correction Classification), OpenBookQA, COPA, SciBench (with the exception of Fundamentals of Physics) and SIQA. Upon investigating more closely, some of the most highly correlated NLP benchmarks make sense. For instance, Inverse Scaling's Resisting Correction Classification ranked second highest for being correlated with human evaluations, and the task measures a highly desirable capability for human users: the LM's ability to follow user instructions that run counter to the LM's natural inclinations.

4.3 What communities exist between human evaluations and NLP benchmarks?

To detect what communities exist between human evaluations and NLP benchmarks, we computed the singular value decomposition of the pairwise Pearson correlation matrix between human evaluations and NLP benchmarks (Fig. 6 top). The maximum rank the correlation matrix can have is 4 because the correlations are computed over the 4 Chat Llama 2 models, but we found that the correlation matrix has only 3 non-zero singular values (App. Fig. 12). Decomposing the correlation matrix into its 3 rank-one components $\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \sigma_3 u_3 v_3^T$ revealed three levels of increasing fine-grained structure in the correlations (App. Fig. 13). We then visualized the human evaluations and NLP benchmarks in the (dimension-scaled) plane defined by the first two rank-one components of the Pearson correlation matrix (Fig. 6 bottom).

The bulk of human evaluations and NLP benchmarks live in one community; however, there are also several smaller interesting communities. Starting on the left of Fig. 6 and moving clockwise, at the top left is a loose community of Dialogue.Code, Dialogue.Language Assistant, several Kth Sentence tasks, Openbook Question Answering (OBQA), AGI.LSAT, AGI.Lawyer Qualification Test, which generally measure model capabilities at identifying and using key information within the context. On the top right, Inverse Scaling.NEQA Classification is alone; this benchmark measures whether models are tripped up by negated questions, which most humans try not to do and likely explains why

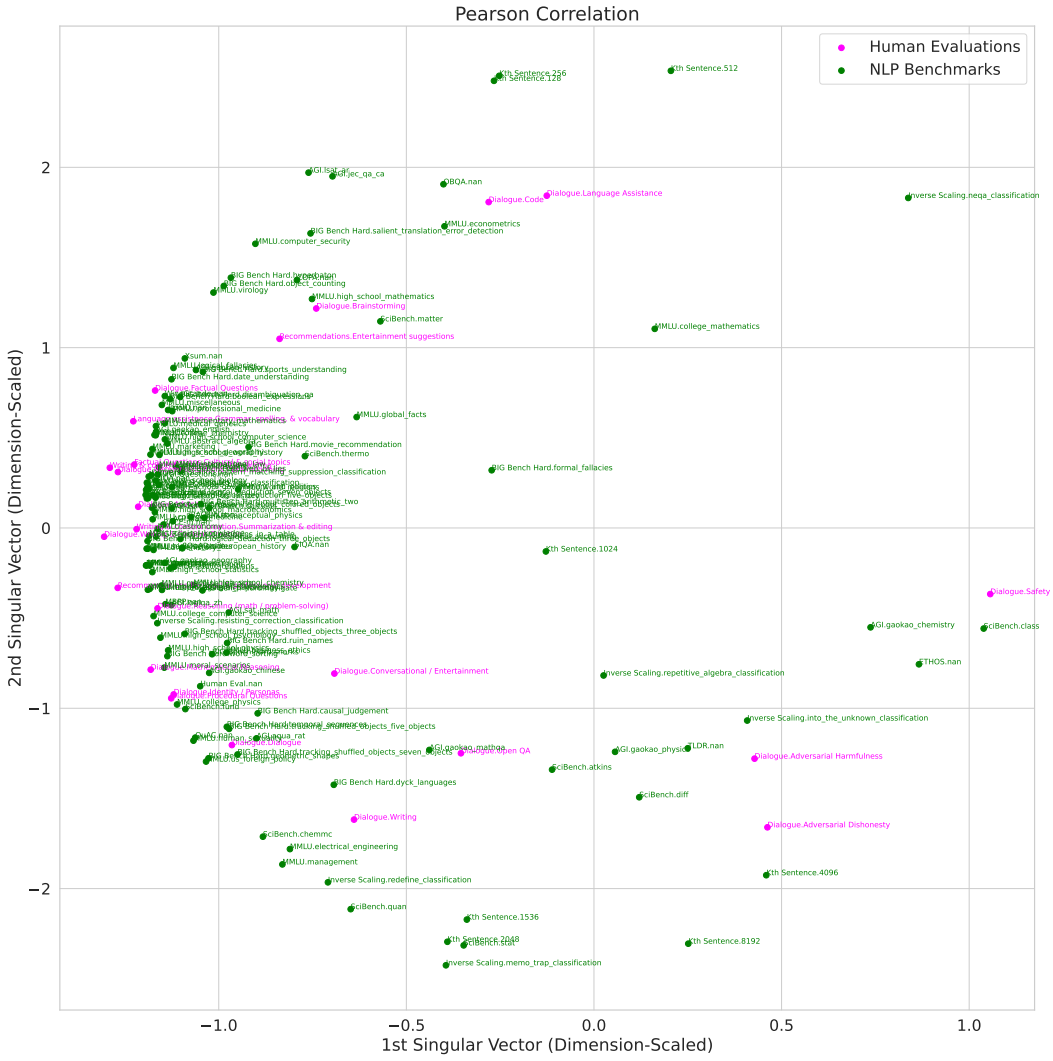

Figure 6: **Structure of Pairwise Pearson Correlations Between Human Evaluations and NLP Benchmarks.** Top: Each row is a human evaluation area and category, and each column is an NLP benchmark and task; values are Pearson correlations ranging from anticorrelated (-1) to correlated (+1). The correlation matrix has 3 non-zero singular values (App. Fig. 12). Bottom: Human evaluations and NLP benchmarks are plotted projected along the (dimension-scaled) first two singular modes of the Pearson correlation matrix. The bulk of evaluations live in one community (left), with smaller communities (top, bottom, right); for an in-depth interpretation, see Sec. 4.3.

this benchmark is isolated. On the right and lower right side, Dialog.Safety is next to ETHOS, a hate speech detection benchmark, and AGI.Gaokao Chemistry, a chemistry benchmark. This community is also close to another community in the lower right comprised of Dialogue.Adversarial Harmfulness, Dialogue.Adversarial Dishonesty, Inverse Scaling.Into the Unknown, TLDR. In the lower left, Dialogue.Open QA and Dialogue.Writing are near BIG Bench Hard's Dyck Languages, Geometric Shapes and Tracking Shuffled Objects and multiple science and factual knowledge benchmarks like MMLU's Electrical Engineering, Management,, SciBench's Quantum Chemistry (quan and chemmc). BIG Bench Hard's Formal Fallacies) and Kth Sentence (1024) lie in the center, disconnected from most other evaluations.

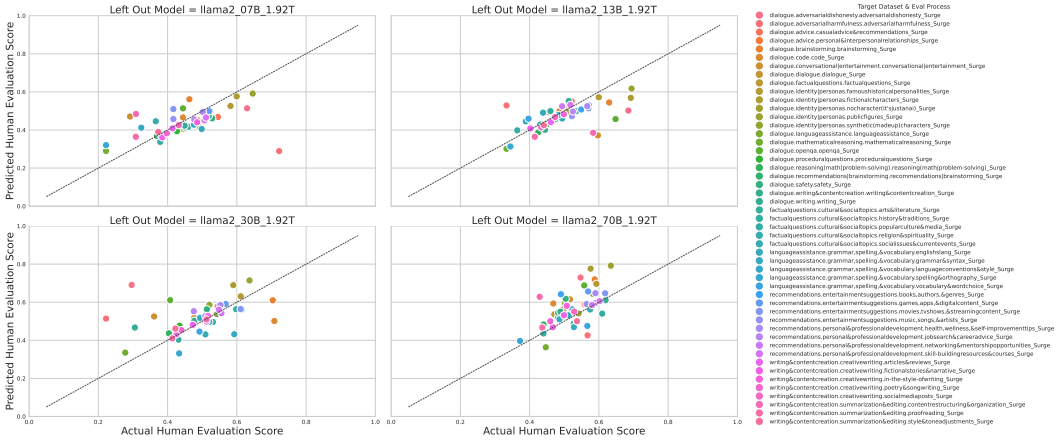

Figure 7: **Leave-one-out cross validation of overparameterized linear regressions typically accurately predict human evaluation scores from NLP benchmark scores.** Each subfigure shows predicted human evaluation scores against actual human evaluation scores on each of the four left-out Chat Llama 2 models colored by the particular area, category and subcategory of human evaluation.

5 Predicting Human Evaluations from NLP Benchmarks

Having established the existence of correlations between human evaluations and NLP benchmarks, we next investigated the feasibility of predicting human evaluations from NLP benchmarks. Our goal is to build predictive models that accurately predict a language model's average human evaluation scores per evaluation areas and categories using the model's average scores on NLP benchmarks and tasks. However, we faced a significant challenge due to the overparameterized nature of our data: for each target human evaluation area or category, there are approximately 150 covariates (NLP benchmarks and tasks) but only 4 samples (Chat Llama 2 models).

5.1 Overparameterized Linear Regressions

To predict human evaluations from NLP benchmarks, we used overparameterized linear regression. In general, overparameterized linear regression is known to be capable of generalizing (citations deferred to App. Sec. A.5), although whether linear models would generalize in this setting was an empirical question. Before fitting the models, we normalized all human evaluation scores to lie in $[0, 1]$ rather than $[-7, -1]$ (recalling that higher scores indicate the human evaluator prefers the Chat Llama 2 model compared to GPT-3.5). For each human evaluation area and category, we fit a linear model to predict a language model's average human evaluation score from its average scores on all NLP benchmarks and tasks. To assess the predictive accuracy of these overparameterized models, we employ leave-one-out cross validation: we fit four separate linear models, each time holding out one of the four chat LMs as a test sample and training on the remaining three. This approach allows us to estimate the models' performance on unseen data, albeit with limitations due to the small sample size.

Across the various human evaluation areas and categories, we found that the linear models' predicted average human evaluation scores generally align well with the actual average human evaluation scores, as evidenced by most points falling close to the identity line in the predicted score vs. actual score plane (Fig. 7). This suggests that, despite the overparameterization, the linear models can capture meaningful relationships between NLP benchmarks and human evaluations. However, we caution against over-interpreting these results, as the small sample size and the assumption of linearity may limit the generalizability of these findings to other language models or evaluation settings.

To gain insight into which NLP benchmarks are most informative for predicting human evaluation scores, we examine the learned weights of the linear models (Fig. 18). NLP benchmarks with consistently high absolute weights across different human evaluation areas and categories are likely to be more predictive of human judgments. However, due to the overparameterized nature of the models, we refrain from drawing strong conclusions about the relative importance of individual benchmarks and instead focus on the overall predictive performance. These results suggest that scaling up the

number of chat LMs and human evaluation data could unlock highly predictive models of slow, noisy and expensive but valuable human evaluations using fast, precise and cheaper NLP benchmarks.

6 DISCUSSION

In this paper, we explored the relationship between human evaluations and NLP benchmarks of chat-finetuned language models (chat LMs). Our work is motivated by the recent shift towards human evaluations as the primary means of assessing chat LM performance, and the need to understand the role that NLP benchmarks can play in this new era.

Through a large-scale study of the Chat Llama 2 model family on a diverse set of human and NLP evaluations, we demonstrated that NLP benchmarks are generally well-correlated with human judgments of chat LM quality. However, our analysis also reveals some notable exceptions to this overall trend. In particular, we find that adversarial and safety-focused evaluations, as well as language assistance and open question answering tasks, exhibit weaker or negative correlations respectively with NLP benchmarks. We also explored predicting human evaluation scores from NLP evaluation scores using overparameterized linear regression models. Our results suggest that NLP benchmarks can indeed be used to predict aggregate human preferences, although we caution that the limited sample size and the assumptions of our models may limit the generalizability of these findings. Our results suggest that NLP benchmarks can serve as fast and cheap proxies of slower and expensive human evaluations in assessing chat LMs.

Additionally, our work highlights the need for further research into NLP evaluations that can effectively capture important aspects of LM behavior, such as safety, robustness to adversarial inputs, and performance on complex, open-ended tasks. It is possible that new NLP benchmarks can provide signals on these topics, e.g., (Wang et al., 2023a). Of particular interest is developing human-interpretable and scaling-predictable evaluation processes, e.g., (Schaeffer et al., 2024a; Ruan et al., 2024; Schaeffer et al., 2024c). Developing and refining such evaluation methods (Madaan et al., 2024), as well as detecting whether evaluations scores faithfully capture models' true performance (Oren et al., 2023; Schaeffer, 2023; Roberts et al., 2023; Jiang et al., 2024; Zhang et al., 2024; Duan et al., 2024) will be crucial for ensuring that LMs are safe, reliable, and beneficial as they become increasingly integrated into real-world use cases.

In conclusion, our study provides insights into the relationship between human evaluations and NLP benchmarks of chat language models. By leveraging the complementary strengths of both human and NLP benchmarks, we can build a more complete understanding of LM capabilities and behaviors, ultimately enabling the development of models more capable, trustworthy, and beneficial to society.

REFERENCES

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.

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Anthropic. Model card and evaluations for claude models, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections, 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Srinivas Bangalore, Owen Rambow, and Steve Whittaker. Evaluation metrics for generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pp. 1–8, 2000.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, jan 2020. doi: 10.1137/20m1336072. URL https://doi.org/10.1137%2F20m1336072.

Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pp. 313–320, 2006.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding?, 2021. URL https://arxiv.org/abs/2104.02145.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL https://arxiv.org/abs/2308.07201.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Cheng-Han Chiang and Hung yi Lee. Can large language models be an alternative to human evaluations?, 2023. URL https://arxiv.org/abs/2305.01937.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://www.aclweb.org/anthology/D18-1241.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Christopher Clark, Kenton Lee, MingWei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *NAACL*, 2019.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*, 2021.

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 preprint arXiv:1803.05457*, 2018.

Miruna Clinciu, Arash Eshghi, and Helen Hastie. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2376–2387, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, 36, 2024.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pp. 138–145, 2002.

Florian E Dorner, Vivian Y Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. *arXiv preprint arXiv:2410.13341*, 2024.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. Uncovering latent memories: Assessing data leakage and memorization patterns in frontier ai models, 2024. URL https://arxiv.org/abs/2406.14549.

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. URL https://arxiv.org/abs/2407.21783.

Robert PW Duin. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pp. 1–7. IEEE, 2000.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems, 2023. URL https://arxiv.org/abs/2212.09180.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, 2024.

Francis Galton. Typical laws of heredity. *Nature*, 15(388):492–495, 1877. doi: 10.1038/015492a0. URL https://doi.org/10.1038/015492a0.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166, 2023.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pp. 169–182. Association for Computational Linguistics, 2020.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models, 2024. URL https://arxiv.org/abs/2401.06059.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation, 2021. URL https://arxiv.org/abs/2109.06835.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Anders Krogh and John A Hertz. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. The inverse scaling prize, 2022a. URL https://github.com/inverse-scaling/prize.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. Inverse scaling prize: First round winners, 2022b. URL https://irmckenzie.co.uk/round1.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. Inverse scaling prize: Second round winners, 2023. URL https://irmckenzie.co.uk/round2.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241–2252, 2017.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 72–78, 2018.

Manfred Opper. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pp. 922–925, 1995.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Tomaso Poggio, Gil Kur, and Andrzej Banburski. Double descent in the condition number. *arXiv preprint arXiv:1912.06190*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H Kim, et al. Safetywashing: Do ai safety benchmarks actually measure safety progress? *arXiv preprint arXiv:2407.21792*, 2024.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2024. URL https://arxiv.org/abs/2407.14981.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*, 2023.

Jason W Rocks and Pankaj Mehta. The geometry of over-parameterized regression and adversarial perturbations. *arXiv preprint arXiv:2103.14108*, 2021.

Jason W Rocks and Pankaj Mehta. Bias-variance decomposition of overparameterized regression with random linear features. *Physical Review E*, 106(2):025304, 2022a.

Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1):013201, 2022b.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. Open-domain conversational agents: Current progress, open problems, and future directions, 2020. URL https://arxiv.org/abs/2006.12442.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Rylan Schaeffer. Pretraining on the test set is all you need, 2023. URL https://arxiv.org/abs/2309.08632.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. *arXiv preprint arXiv:2303.14151*, 2023a.

Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Ila Fiete, Andrey Gromov, and Sanmi Koyejo. Divergence at the interpolation threshold: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023b.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024a.

Rylan Schaeffer, Zachary Robertson, Akhilan Boopathy, Mikail Khona, Kateryna Pistunova, Jason W. Rocks, Ila R. Fiete, Andrey Gromov, and Sanmi Koyejo. Double descent demystified. In *ICLR Blogposts 2024*, 2024b. URL https://d2jud02ci9yv69.cloudfront.net/2024-05-07-double-descent-demystified-54/blog/double-descent-demystified/. https://d2jud02ci9yv69.cloudfront.net/2024-05-07-double-descent-demystified-54/blog/double-descent-demystified/.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier ai models with scale remained elusive?, 2024c. URL https://arxiv.org/abs/2406.04391.

Hendrik Schuff, Hsiu-Yu Yang, Heike Adel, and Ngoc Thang Vu. Does external knowledge help explainable natural language inference? automatic evaluation vs. human ratings. In *Proceedings of the Fourth Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 26–41, 2021.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents, 2022. URL https://arxiv.org/abs/2201.04723.

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

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, 2018.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, 2024. URL https://arxiv.org/abs/2410.20266.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

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*, 2023b.

F Vallet. The hebb rule for learning linearly separable boolean functions: learning and generalization. *Europhysics Letters*, 8(8):747, 1989.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 355–368, 2019.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151, 2021.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023a.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Experimental Methodology: Data and Analyses

A.1 Data: Natural Language Processing (NLP) Benchmark Scores

We chose which NLP benchmarks to include based largely on which frontier AI models were reporting performance scores on. Llama 1 (Touvron et al., 2023a) and Llama 2 (Touvron et al., 2023b) were our primary guides, as were Gemini 1 and Gemini 1.5 (Team et al., 2023; Reid et al., 2024), Claude 3 (Anthropic, 2023), and Mistral (Jiang et al., 2023). We evaluated the 4 Llama 2 Chat models following the same evaluation processes reported in the Llama 2 paper (Touvron et al., 2023b). This included matching the prompt formatting, automated metric scoring and (when appropriate) few-shot prompting and chain-of-thought prompting. After evaluating the four Chat Llama 2 models on these NLP benchmarks and evaluation processes, we obtained 160 scores per model for our analyses.

Table 1: **Natural Language Processing Datasets and Evaluation Processes.** In this work, we used the above well-established NLP datasets and evaluation processes. "cot" means Chain-of-Thought prompting (Cobbe et al., 2021; Wei et al., 2022). "Gen" refers generations per evaluation. For more information, please see Section 3 and Appendix A.

Benchmark	Subset	Metric	Shots	Additional
AGI	aqua_rat	Acc Char	5	–
AGI	gaokao_biology	Acc Char	5	–
AGI	gaokao_chemistry	Acc Char	5	–
AGI	gaokao_chinese	Acc Char	5	–
AGI	gaokao_english	Acc Char	5	–
AGI	gaokao_geography	Acc Char	5	–
AGI	gaokao_history	Acc Char	5	–
AGI	gaokao_mathcloze	Exact Match	5	–
AGI	gaokao_mathqa	Acc Char	5	–
AGI	gaokao_physics	Acc Char	5	–
AGI	jec_qa_ca	Acc Char	5	–
AGI	jec_qa_kd	Acc Char	5	–
AGI	logiqa_en	Acc Char	3	–
AGI	logiqa_zh	Acc Char	3	–
AGI	lsat_ar	Acc Char	3	–
AGI	lsat_lr	Acc Char	3	–
AGI	lsat_rc	Acc Char	3	–
AGI	sat_math	Acc Char	5	–
ARC	Challenge	Acc Char	0	–
ARC	Easy	Acc Char	0	–
BBH	boolean_expressions	Exact Match	0	–
BBH	causal_judgement	Exact Match	0	–
BBH	date_understanding	Exact Match	0	–
BBH	disambiguation_qa	Exact Match	0	–
BBH	dyck_languages	Exact Match	0	–
BBH	formal_fallacies	Exact Match	0	–
BBH	geometric_shapes	Exact Match	0	–
BBH	hyperbaton	Exact Match	0	–
BBH	logical_deduction_3_objects	Exact Match	0	–
BBH	logical_deduction_5_objects	Exact Match	0	–
BBH	logical_deduction_7_objects	Exact Match	0	–
BBH	movie_recommendation	Exact Match	0	–
BBH	multistep_arithmetic_two	Exact Match	0	–
BBH	navigate	Exact Match	0	–
BBH	object_counting	Exact Match	0	–
BBH	penguins_in_a_table	Exact Match	0	–
BBH	reasoning_about_colored_objects	Exact Match	0	–
BBH	ruin_names	Exact Match	0	–
BBH	salient_translation_error_detection	Exact Match	0	–
BBH	snarks	Exact Match	0	–

Table 1 – continued from previous page

Benchmark	Subset	Metric	Shots	Additional
BBH	sports_understanding	Exact Match	0	–
BBH	temporal_sequences	Exact Match	0	–
BBH	tracking_shuffled_objects_3	Exact Match	0	–
BBH	tracking_shuffled_objects_5	Exact Match	0	–
BBH	tracking_shuffled_objects_7	Exact Match	0	–
BBH	web_of_lies	Exact Match	0	–
BBH	word_sorting	Exact Match	0	–
BoolQ	–	Acc Token	0	–
CommonSenseQA	–	Acc Char	7	–
COPA	–	Acc Char	?	–
DROP	–	Exact Match	0	–
DROP	–	F1	0	–
ETHOS	–	Acc Char	0	–
GSM8K	–	Exact Match	0	100 Gen
GSM8K	–	F1	0	100 Gen
GSM8K	–	Exct Mtch maj1@100	0	100 Gen
GSM8K	–	F1 maj1@100	0	100 Gen
HellaSwag	–	Acc Char	?	–
Human Eval	–	pass@1	0	1 Gen
Human Eval	–	pass@1	0	200 Gen
Human Eval	–	pass@100	0	200 Gen
Inverse Scaling	hindsight_neglect	Exact Match	0	–
Inverse Scaling	into_the_unknown	Exact Match	0	–
Inverse Scaling	memo_trap	Exact Match	0	–
Inverse Scaling	modus_tollens	Exact Match	0	–
Inverse Scaling	neqa	Exact Match	0	–
Inverse Scaling	pattern_matching_suppression	Exact Match	0	–
Inverse Scaling	redefine	Exact Match	0	–
Inverse Scaling	repetitive_algebra	Exact Match	0	–
Inverse Scaling	resisting_correction	Exact Match	0	–
Inverse Scaling	sig_figs	Exact Match	0	–
Kth Sentence	128	ROUGE-2	0	–
Kth Sentence	256	ROUGE-2	0	–
Kth Sentence	512	ROUGE-2	0	–
Kth Sentence	1024	ROUGE-2	0	–
Kth Sentence	1536	ROUGE-2	0	–
Kth Sentence	2048	ROUGE-2	0	–
Kth Sentence	4096	ROUGE-2	0	–
Kth Sentence	8192	ROUGE-2	0	–
MBPP	–	pass@1	3	80 Gen
MBPP	–	pass@80	3	80 Gen
MMLU	abstract_algebra	Acc Char	5	–
MMLU	anatomy	Acc Char	5	–
MMLU	astronomy	Acc Char	5	–
MMLU	business_ethics	Acc Char	5	–
MMLU	clinical_knowledge	Acc Char	5	–
MMLU	college_biology	Acc Char	5	–
MMLU	college_chemistry	Acc Char	5	–
MMLU	college_computer_science	Acc Char	5	–
MMLU	college_mathematics	Acc Char	5	–
MMLU	college_medicine	Acc Char	5	–
MMLU	college_physics	Acc Char	5	–
MMLU	computer_security	Acc Char	5	–
MMLU	conceptual_physics	Acc Char	5	–
MMLU	econometrics	Acc Char	5	–
MMLU	electrical_engineering	Acc Char	5	–
MMLU	elementary_mathematics	Acc Char	5	–

Table 1 – continued from previous page

Benchmark	Subset	Metric	Shots	Additional
MMLU	formal_logic	Acc Char	5	–
MMLU	global_facts	Acc Char	5	–
MMLU	high_school_biology	Acc Char	5	–
MMLU	high_school_chemistry	Acc Char	5	–
MMLU	high_school_computer_science	Acc Char	5	–
MMLU	high_school_european_history	Acc Char	5	–
MMLU	high_school_geography	Acc Char	5	–
MMLU	high_school_government_and_politics	Acc Char	5	–
MMLU	high_school_macroeconomics	Acc Char	5	–
MMLU	high_school_mathematics	Acc Char	5	–
MMLU	high_school_microeconomics	Acc Char	5	–
MMLU	high_school_physics	Acc Char	5	–
MMLU	high_school_psychology	Acc Char	5	–
MMLU	high_school_statistics	Acc Char	5	–
MMLU	high_school_us_history	Acc Char	5	–
MMLU	high_school_world_history	Acc Char	5	–
MMLU	human_aging	Acc Char	5	–
MMLU	human_sexuality	Acc Char	5	–
MMLU	international_law	Acc Char	5	–
MMLU	jurisprudence	Acc Char	5	–
MMLU	logical_fallacies	Acc Char	5	–
MMLU	machine_learning	Acc Char	5	–
MMLU	management	Acc Char	5	–
MMLU	marketing	Acc Char	5	–
MMLU	medical Genetics	Acc Char	5	–
MMLU	miscellaneous	Acc Char	5	–
MMLU	moral_disputes	Acc Char	5	–
MMLU	moral_scenarios	Acc Char	5	–
MMLU	nutrition	Acc Char	5	–
MMLU	philosophy	Acc Char	5	–
MMLU	prehistory	Acc Char	5	–
MMLU	professional_accounting	Acc Char	5	–
MMLU	professional_law	Acc Char	5	–
MMLU	professional_medicine	Acc Char	5	–
MMLU	professional_psychology	Acc Char	5	–
MMLU	public_relations	Acc Char	5	–
MMLU	security_studies	Acc Char	5	–
MMLU	sociology	Acc Char	5	–
MMLU	us_foreign_policy	Acc Char	5	–
MMLU	virology	Acc Char	5	–
MMLU	world_religions	Acc Char	5	–
NaturalQuestions	–	Exact Match	0	–
OpenBookQA	–	Acc Completion	0	–
PIQA	–	Acc Char	0	–
QuAC	–	F1	0	–
RACE	High School	Acc Char	0	–
RACE	Middle School	Acc Char	0	–
SIQA	–	Acc Char	0	–
SQuAD	–	Exact Match	0	–
SciBench	atkins	Fuzzy Match	0	–
SciBench	calculus	Fuzzy Match	0	–
SciBench	chemmc	Fuzzy Match	0	–
SciBench	class	Fuzzy Match	0	–
SciBench	diff	Fuzzy Match	0	–
SciBench	fund	Fuzzy Match	0	–
SciBench	matter	Fuzzy Match	0	–
SciBench	quan	Fuzzy Match	0	–

Table 1 – continued from previous page

Benchmark	Subset	Metric	Shots	Additional
SciBench	stat	Fuzzy Match	0	–
SciBench	thermo	Fuzzy Match	0	–
TLDR	–	ROUGE-2	0	–
TLDR	–	ROUGE-L	0	–
TriviaQA	–	Exact Match	0	–
WinoGrande	–	Acc Char	0	–
Xsum	–	ROUGE-2	1	–

A.2 DATA: HUMAN EVALUATION SCORES

Human data annotators were hired to evaluate outputs of chat language models (LMs) in single-turn and multi-turn conversations using a Likert scale (Likert, 1932) from 1 to 7. The conversations were constructed within our taxonomy of areas-categories-subcategories (Sec. 3; Fig. 2). Each conversation was evaluated by at least three unique humans for a combined total of 2104 unique human annotators. Our human annotators scored 11291 single-turn conversations and 2081 multi-turn conversations.

Table 2: **Human Evaluation Areas, Categories, and Subcategories.**

Area	**Category**	**Subcategory**
Dialogue	Adversarial Dishonesty	Adversarial Dishonesty
Dialogue	Adversarial Harmfulness	Adversarial Harmfulness
Dialogue	Advice	Casual advice & recommendations
Dialogue	Advice	Personal & interpersonal relationships
Dialogue	Brainstorming	Brainstorming
Dialogue	Classification	Classification
Dialogue	Closed QA	Closed QA
Dialogue	Code	Code
Dialogue	Conversational / Entertainment	Conversational / Entertainment
Dialogue	Conversational/Entertainment	Conversational/Entertainment
Dialogue	Dialogue	Dialogue
Dialogue	Extraction	Extraction
Dialogue	Factual Questions	Factual Questions
Dialogue	Identity / Personas	Famous historical personalities
Dialogue	Identity / Personas	Fictional characters
Dialogue	Identity / Personas	No character (it's just an AI)
Dialogue	Identity / Personas	Public figures
Dialogue	Identity / Personas	Synthetic (made up) characters
Dialogue	Language Assistance	Language Assistance
Dialogue	Math	Math
Dialogue	Mathematical Reasoning	Mathematical Reasoning
Dialogue	Open QA	Open QA
Dialogue	Procedural Questions	Procedural Questions
Dialogue	Reasoning (math / problem-solving)	Reasoning (math / problem-solving)
Dialogue	Recommendations / Brainstorming	Recommendations / Brainstorming
Dialogue	Rewriting	Rewriting
Dialogue	Safety	Safety
Dialogue	Summarization	Summarization
Dialogue	Writing	Writing
Dialogue	Writing & Content Creation	Writing & Content Creation
Factual Questions	Cultural & social topics	Arts & literature
Factual Questions	Cultural & social topics	History & traditions
Factual Questions	Cultural & social topics	Popular culture & media
Factual Questions	Cultural & social topics	Religion & spirituality
Factual Questions	Cultural & social topics	Social issues & current events
Language assistance	Grammar, spelling, & vocabulary	English slang

Table 2 – continued from previous page

Area	Category	Subcategory
Language assistance	Grammar, spelling, & vocabulary	Grammar & syntax
Language assistance	Grammar, spelling, & vocabulary	Language conventions & style
Language assistance	Grammar, spelling, & vocabulary	Spelling & orthography
Language assistance	Grammar, spelling, & vocabulary	Vocabulary & word choice
Recommendations	Entertainment suggestions	Books, authors, & genres
Recommendations	Entertainment suggestions	Games, apps, & digital content
Recommendations	Entertainment suggestions	Movies, TV shows, & streaming content
Recommendations	Entertainment suggestions	Music, songs, & artists
Recommendations	Personal & professional development	Health, wellness, & self-improvement tips
Recommendations	Personal & professional development	Job search & career advice
Recommendations	Personal & professional development	Networking & mentorship opportunities
Recommendations	Personal & professional development	Skill-building resources & courses
Writing & content creation	Creative writing	Articles & reviews
Writing & content creation	Creative writing	Fictional stories & narrative
Writing & content creation	Creative writing	In-the-style-of writing
Writing & content creation	Creative writing	Poetry & songwriting
Writing & content creation	Creative writing	Social media posts
Writing & content creation	Summarization & editing	Content restructuring & organization
Writing & content creation	Summarization & editing	Proofreading
Writing & content creation	Summarization & editing	Style & tone adjustments

A.3 Analyses: Correlations

A.4 Analyses: Community Detection

A.5 Analyses: Linear Regression

Due to space limitations in the main text, we defer citations regarding generalization of overparameterized models to here. For a nonexhaustive list, please see Vallet (1989); Krogh & Hertz (1991); Geman et al. (1992); Krogh & Hertz (1992); Opper (1995); Duin (2000); Spigler et al. (2018); Belkin et al. (2019); Bartlett et al. (2020); Belkin et al. (2020); Nakkiran et al. (2021); Poggio et al. (2019); Advani et al. (2020); Liang & Rakhlin (2020); Adlam & Pennington (2020); Rocks & Mehta (2022b; 2021; 2022a); Mei & Montanari (2022); Hastie et al. (2022); Bach (2023); Schaeffer et al. (2023a;b); Curth et al. (2024); Schaeffer et al. (2024b).

B STATISTICS OF HUMAN EVALUATIONS

As exploratory data analysis, we calculated and examined basic statistics of the human evaluations. Fig. 8 showcases how many turns (i.e., back and forth messages) are in each sample evaluated by human annotators (left) and how many human annotators evaluated each sample (right). Fig. 9 shows the empirical cumulative distributions functions of the average of human annotators' scores per datum (left) and the standard deviation of human annotators' scores per datum (right). Fig. 10 visualizes the joint distribution of means and standard deviations of human annotators' scores per datum as both a scatterplot (left) and a kernel density estimate (right).

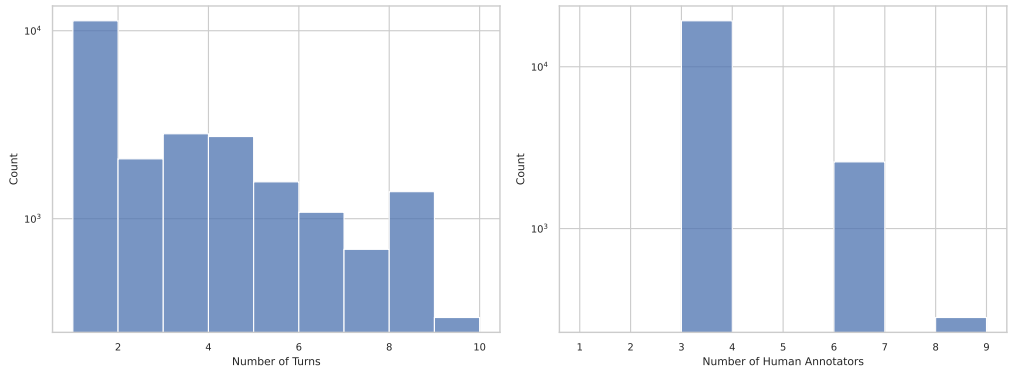

Figure 8: **Statistics of Human-Evaluated Data.** Left: Number of turns per datum. Right: Number of human annotators per human-evaluated datum.

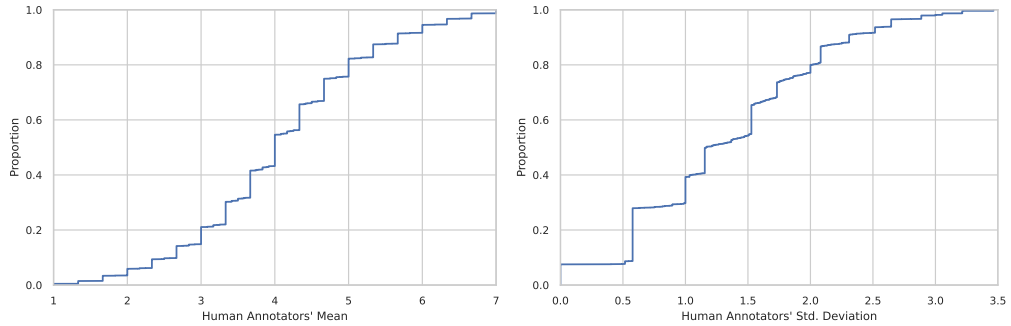

Figure 9: **Empirical Cumulative Distribution Functions of Human Annotators' Scores.** Left: Average of human annotators' score per annotated sample. Right: Human annotators' standard deviation per annotated sample.

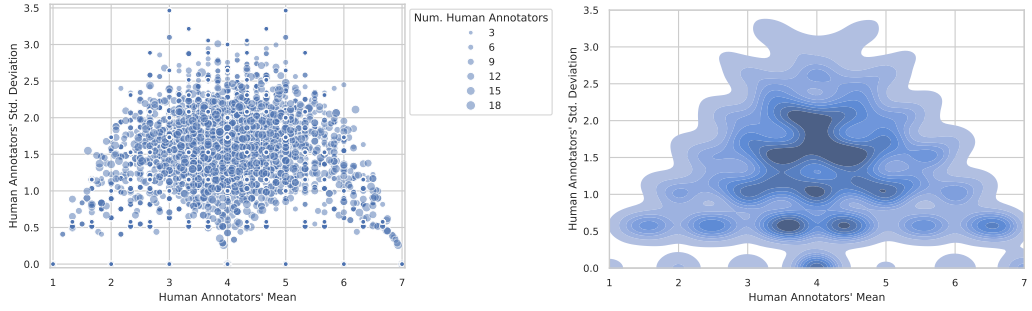

Figure 10: **Joint Distribution of Human Annotators' Average Scores per Datum vs Standard Deviation per Datum.** Left: Scatterplot. Right: Kernel Density Estimate.

C CORRELATION METRICS ARE THEMSELVES HIGHLY CORRELATED

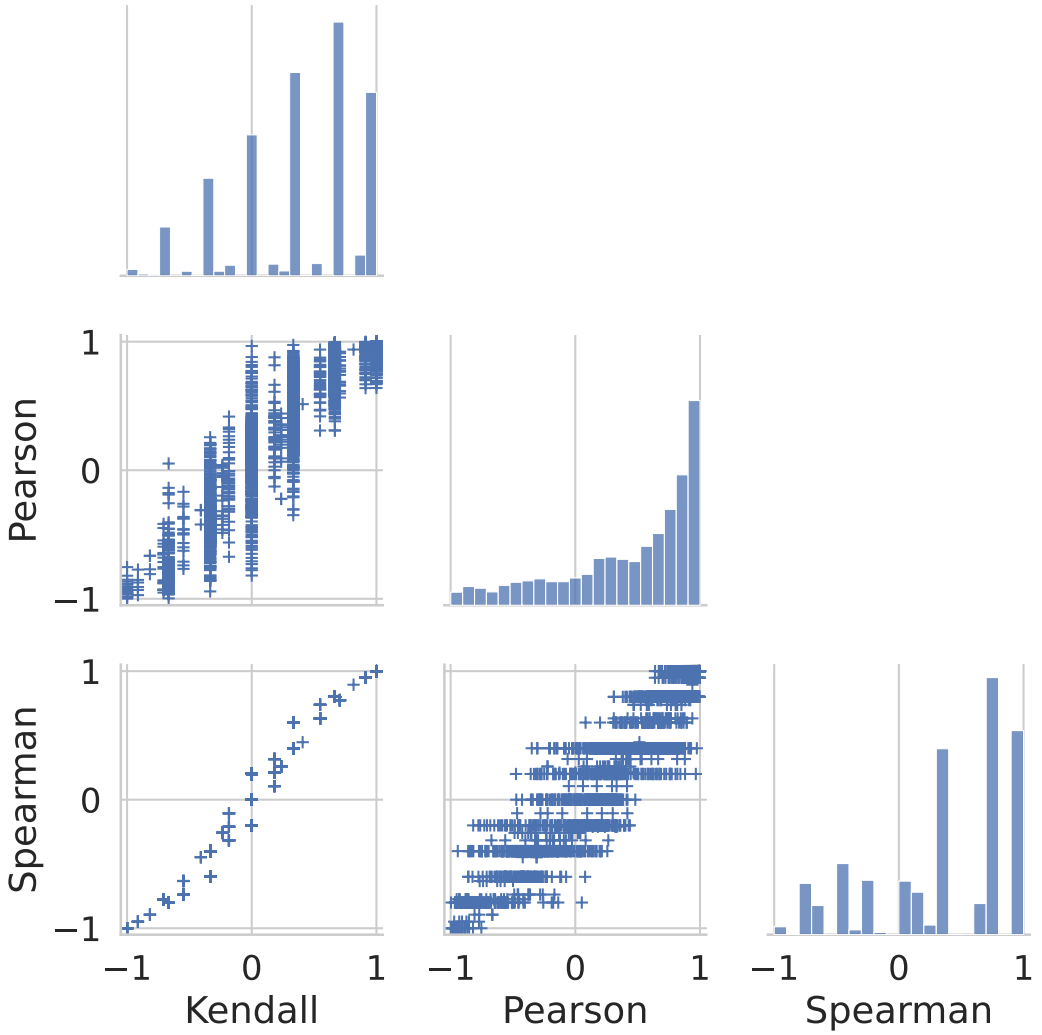

Figure 11: **Correlation of Academic-Human Evaluation Correlations Under Different Correlation Metrics**. For each pair of human evaluation (area and category) and NLP benchmark (benchmark and subset), we computed the correlation between scores under one of 3 correlation metrics: Pearson, Spearman and Kendall. We then looked at how correlated the correlation scores under the 3 correlation metrics are. In general, all 3 are correlation metrics yield correlated scores. This demonstrates that the choice of correlation metric is relatively less important.

D CORRELATION MATRICES BETWEEN HUMAN EVALUATIONS AND NLP BENCHMARKS AND THEIR SINGULAR VALUE DECOMPOSITIONS

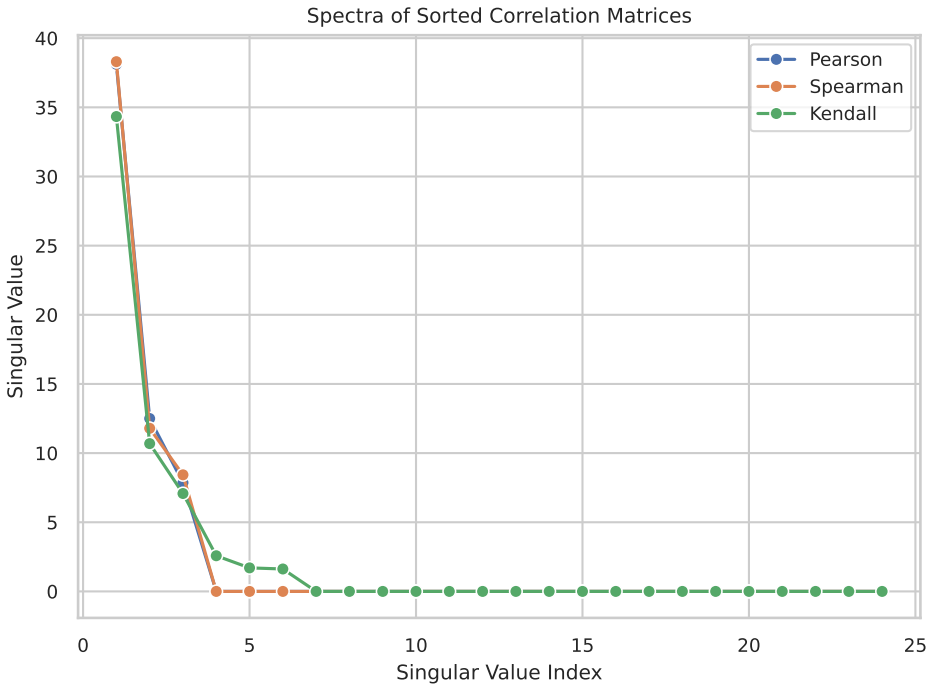

Figure 12: **Spectra of Human Evaluation-NLP Benchmark Correlation Matrices**. Because the correlation matrices are computed over four Chat Llama 2 models, the maximum matrix rank is 4. Howeer, both Pearson and Spearman correlation matrices have only 3 non-zero singular values.

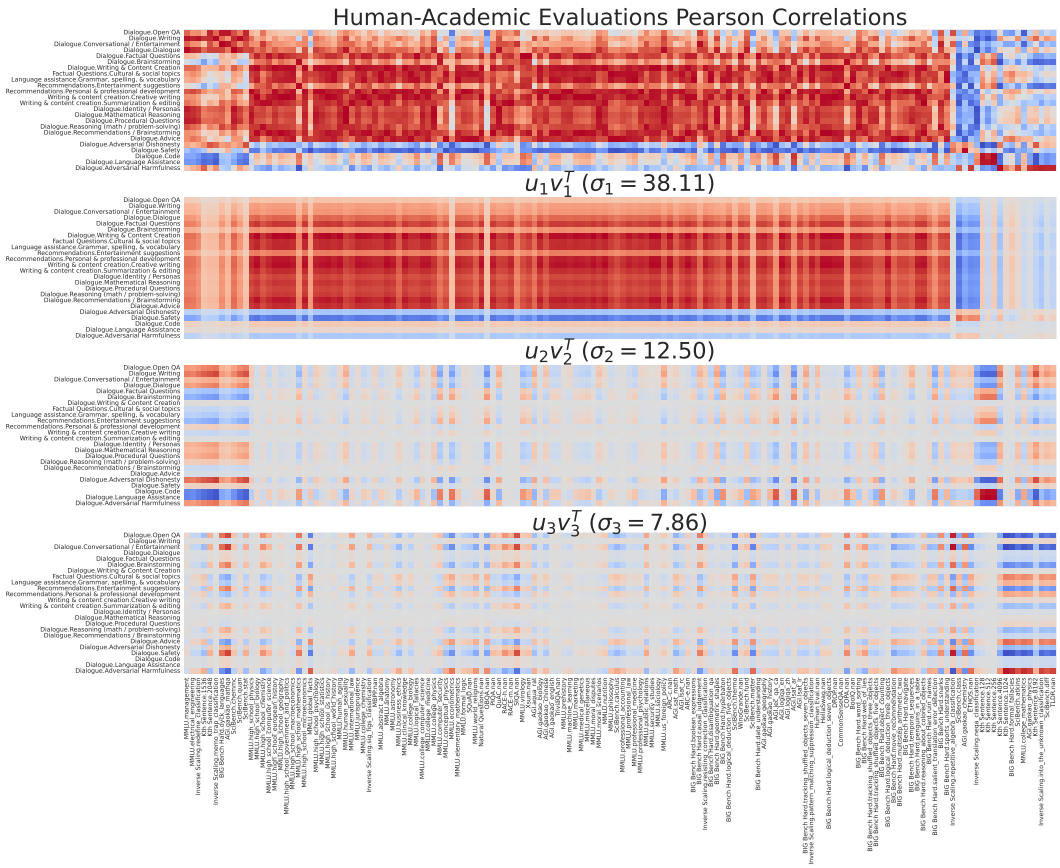

Figure 13: **Pearson Correlation Matrix between Human Evaluations and NLP Benchmarks.** The Pearson correlation matrix has 3 non-zero singular values, with corresponding modes shown in the last 3 rows.

The first component of the Pearson correlation matrix divides the human evaluations and NLP benchmarks into 3 groups (Fig. 6B): one group that is broadly uncorrelated, and two unequally-sized groups that are self-correlated and mutually anti-correlated. The *uncorrelated group* consists of human evaluations Dialogue:Code and Dialogue:Language Assistance, as well as NLP benchmarks Kth-sentence, TLDR, SciBench's Atkins and Differential Equations, MMLU's College Math and BBH's Formal Fallacies. The *smaller self-correlated group* consists of Dialogue:Adversarial Dishonesty and Safety:Harmlessness as well as ETHOS (a hate speech detection benchmark), Inverse Scaling NEQA (a negation question-answering benchmark) and AGI Gaokao Chemistry, whereas the *larger self-correlated group* consists of almost all other human evaluations and NLP benchmarks. This is more clearly visually displayed in the Spearman correlation matrix (App. Fig. 14).

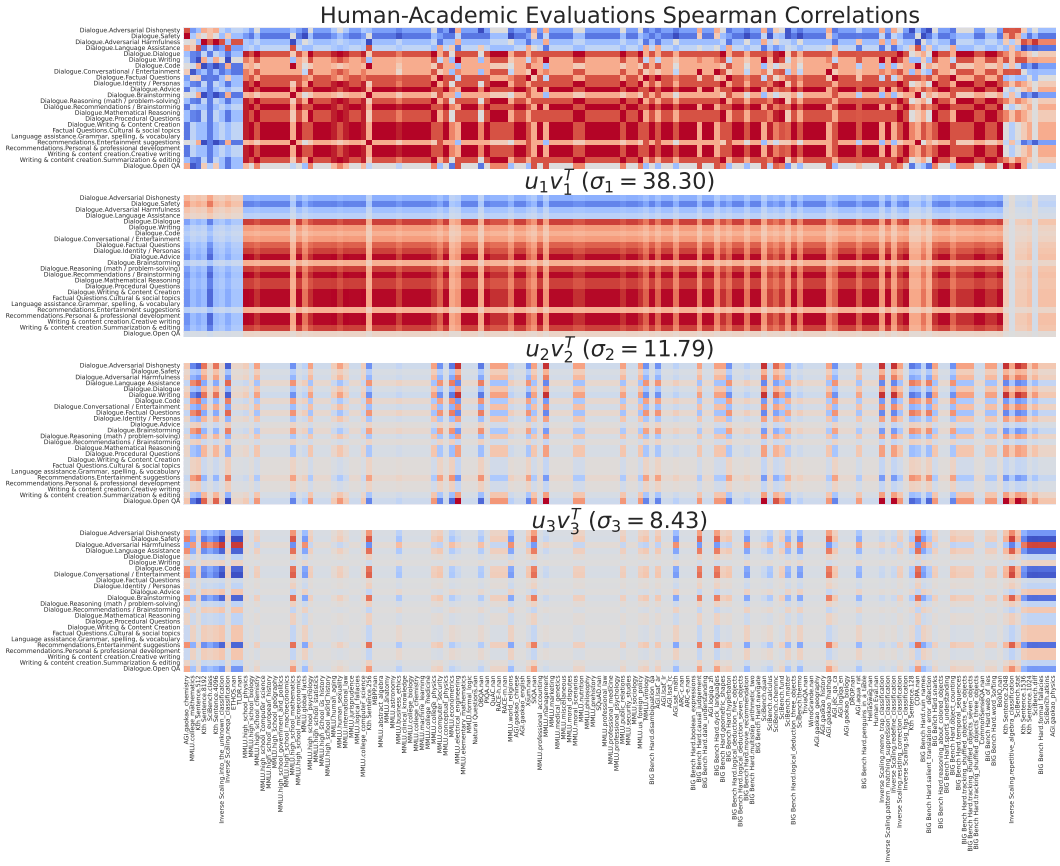

Figure 14: **Spearman Correlation Matrix between Human Evaluations and NLP Benchmarks.** The Spearman correlation matrix also has 3 non-zero singular values, with corresponding modes shown in the last 3 rows.

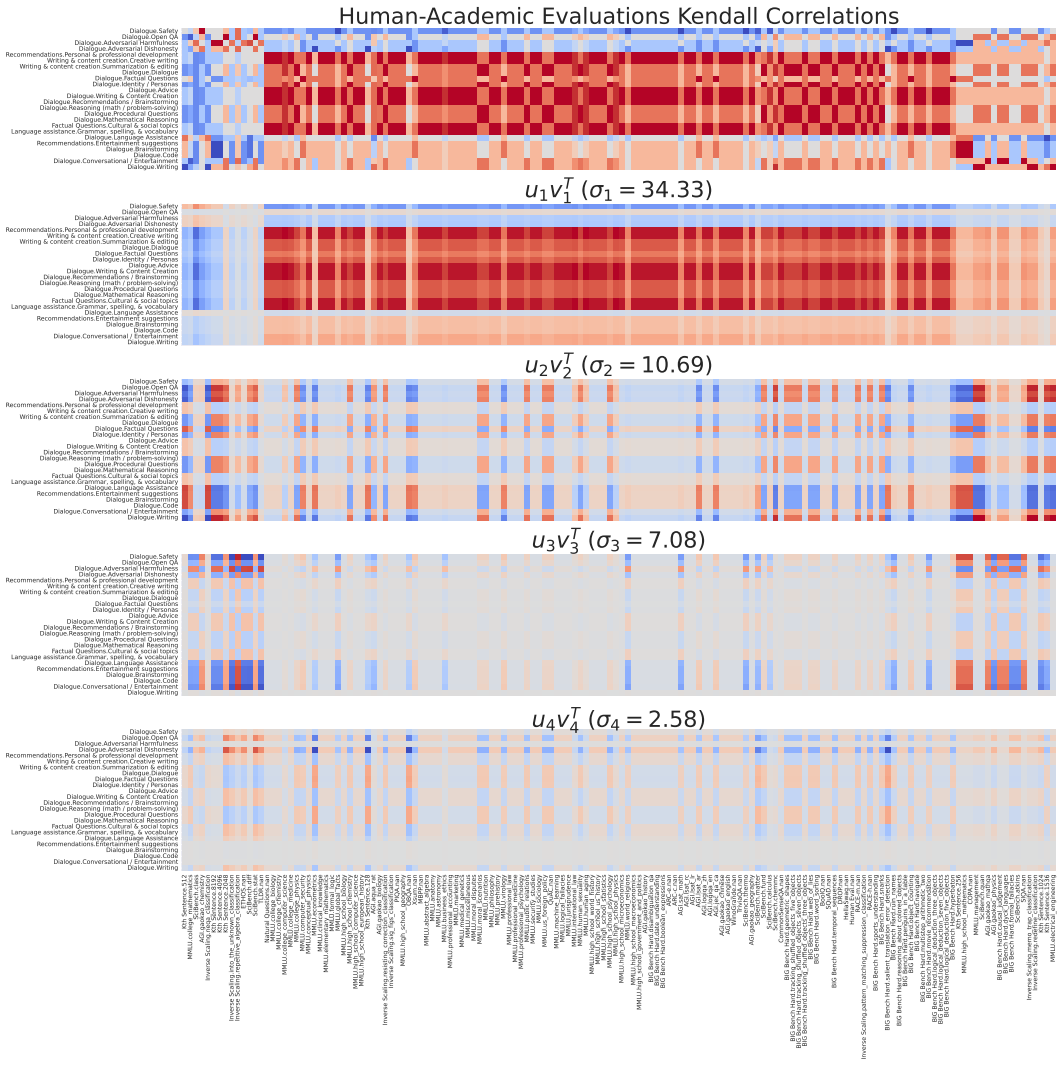

Figure 15: **Kendall Correlation Matrix between Human Evaluations and NLP Benchmarks.** The Kendall correlation matrix has four non-zero singular values, with corresponding modes shown in the last four rows.

E EMPIRICAL SCALING BEHAVIOR OF HUMAN EVALUATIONS AND NLP BENCHMARKS

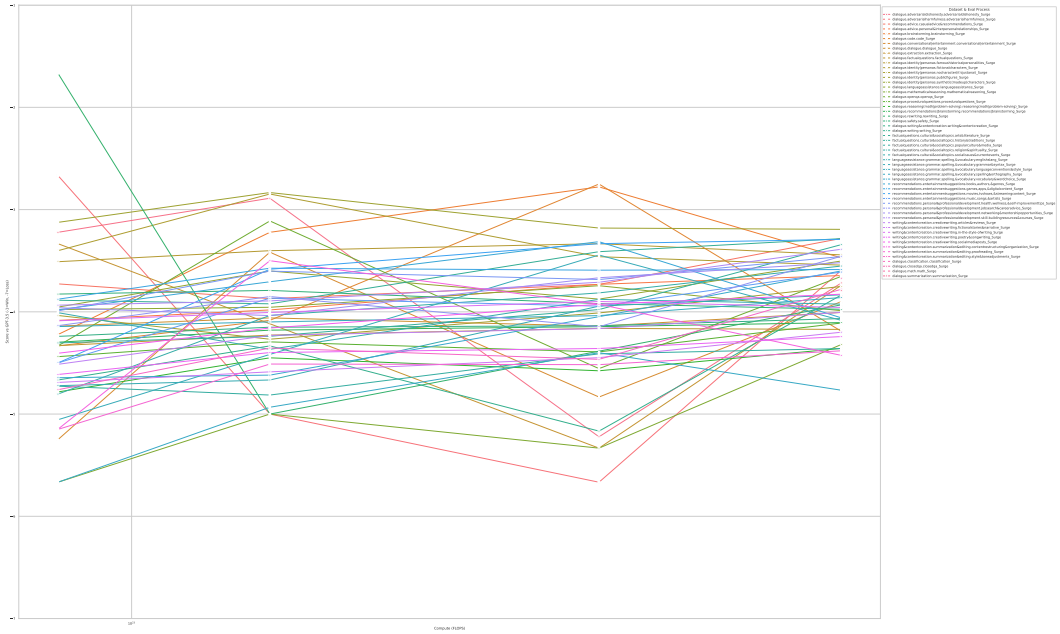

Figure 16: **Empirical Scaling Behavior of Human Evaluations with Increasing Compute.**

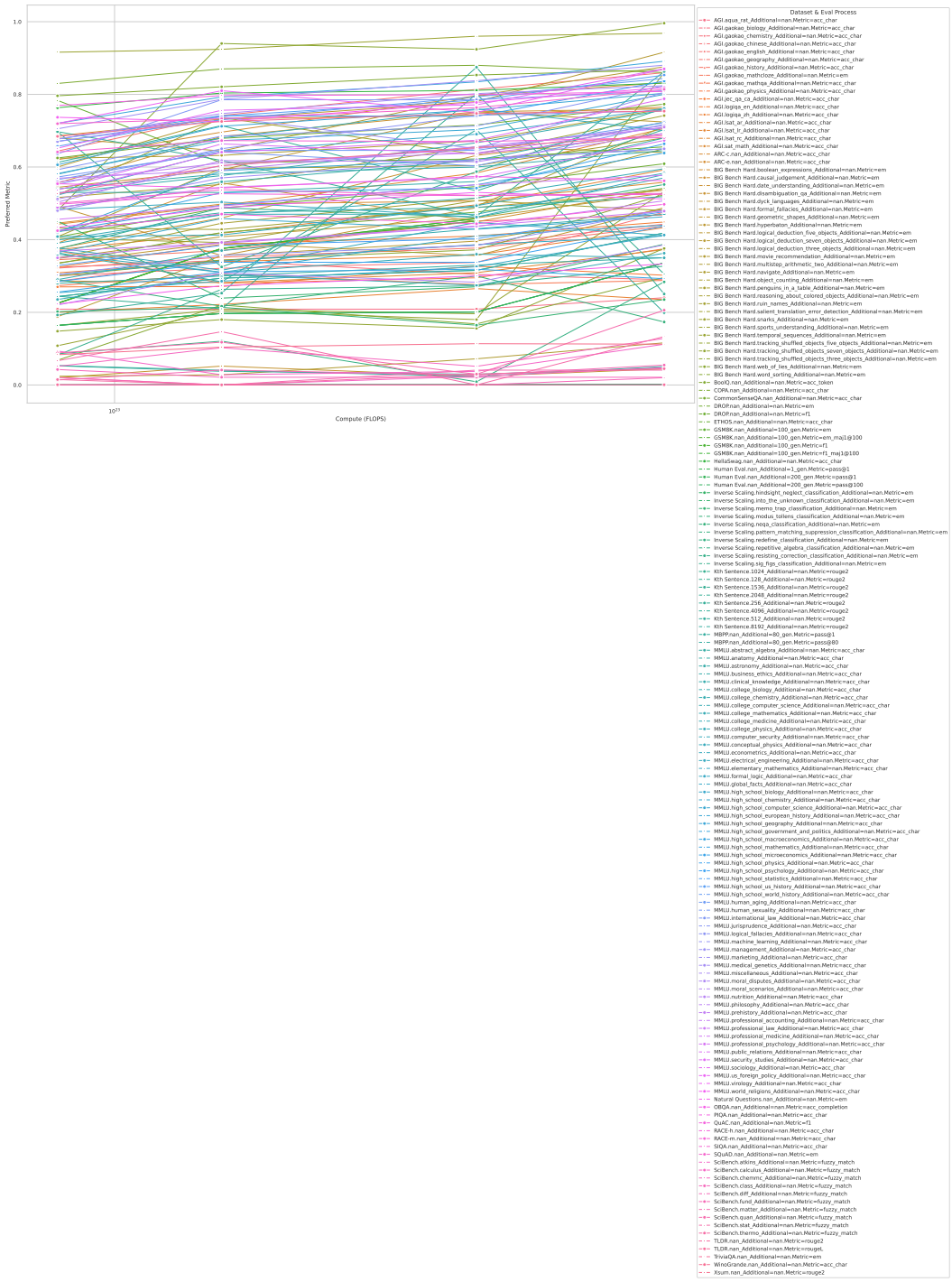

Figure 17: **Empirical Scaling Behavior of NLP Benchmarks with Increasing Compute.**

F Coefficients of Leave-One-Out Cross-Validated Linear Regressions

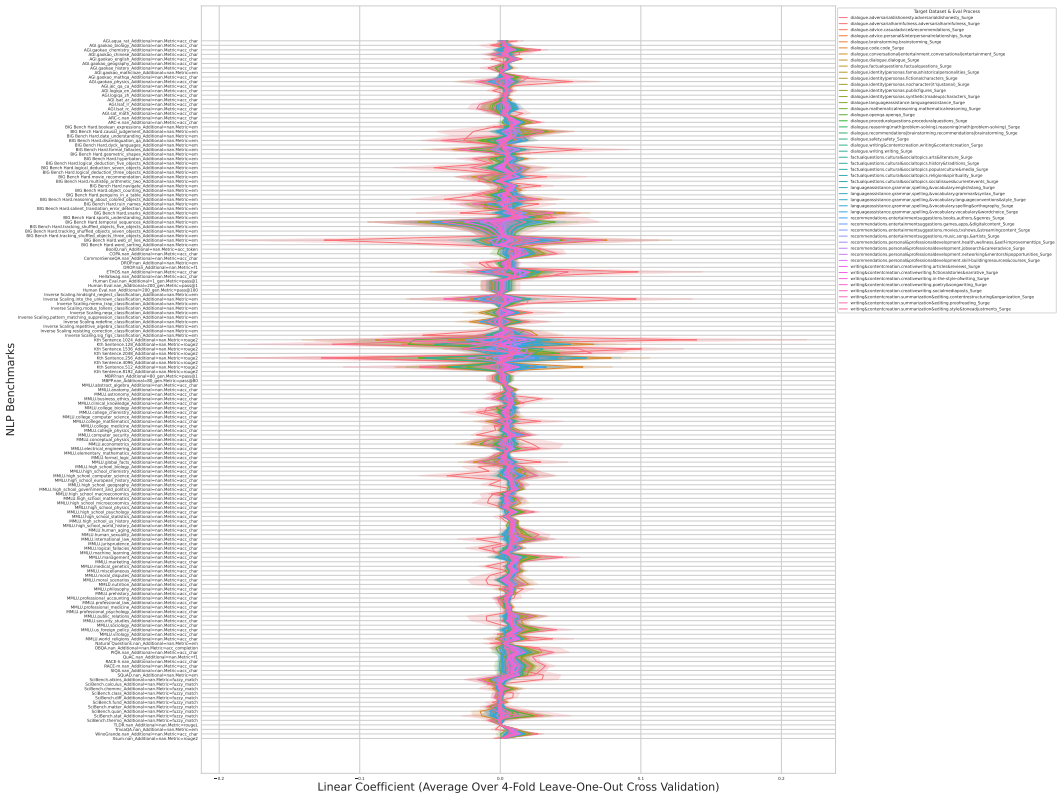

Figure 18: **Linear Coefficients of NLP Benchmarks in Predicting Human Evaluations.** For each human evaluation area, category and subcategory, we visualize the learned linear parameters per NLP benchmark averaged over the 4-fold leave-one-out cross validation process. For interpretation of results, see Sec. 5.