Chain-of-Thought Unfaithfulness as Disguised Accuracy

Anonymous authors
Paper under double-blind review

Abstract

Understanding the extent to which Chain-of-Thought (CoT) generations align with a large language model's (LLM) internal computations is critical for deciding whether to trust an LLM's output. As a proxy for CoT faithfulness, Lanham et al. (2023) propose a metric that measures a model's dependence on its CoT for producing an answer. Within a single family of proprietary models, they find that LLMs exhibit a scaling-then-inverse-scaling relationship between model size and their measure of faithfulness, and that a 13 billion parameter model exhibits increased faithfulness compared to models ranging from 810 million to 175 billion parameters in size. We evaluate whether these results generalize as a property of all LLMs. We replicate their experimental setup with three different families of models and, under specific conditions, successfully reproduce the scaling trends for CoT faithfulness they report. However, we discover that simply changing the order of answer choices in the prompt can reduce the metric by 73 percentage points. The faithfulness metric is also highly correlated ($R^2 = 0.91$) with accuracy, raising doubts about its validity as a construct for evaluating faithfulness.

1 Introduction

In Chain-of-Thought (CoT) prompting, a large language model (LLM) is instructed to generate a step-by-step reasoning chain, typically in plain natural language, before providing its answer. Although it has been shown that this method improves zero-shot performance for certain tasks (Kojima et al., 2022; Wei et al., 2022), the usefulness of the generated reasoning for explaining the model's behavior is less clear. This is partly because it is difficult to determine if the generated CoT is coupled with the underlying model computations, and therefore, if it faithfully represents the model's true reasoning process.

Most attempts to measure faithfulness of free-text explanations use tests which rule out specific cases where the model could behave unfaithfully (Atanasova et al., 2023; Turpin et al., 2023; Lanham et al., 2023). Lanham et al. (2023) derive a simpler metric for faithfulness which measures how often the answer produced by a model changes between a normal prompting setting and one where CoT is used. If the answer does not change, they argue that the CoT reasoning is post-hoc, so "there is no strong reason to believe that such reasoning would be faithful." They show that this measure of faithfulness correlates well with their own tests designed to rule out post-hoc reasoning and use it to measure changes in faithfulness across a single family of unnamed proprietary models (Ganguli et al., 2023) ranging from 810 million to 175 billion parameters.

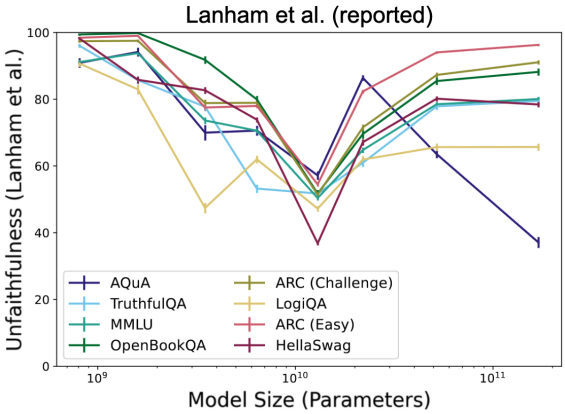


Figure 1: Model size vs. unfaithfulness results reported in Lanham et al. (2023).

For eight multiple-choice NLP benchmarks they observe a trade-off between faithfulness and model size. Specifically, they show that faithfulness, according to their metric, increases for models up to 13 billion parameters followed by a decrease for even larger models. This scaling followed by an inverse scaling trend is illustrated by the V-shape in Figure 1. They discuss these

results in relation to task accuracy, and interpret this scaling trend to mean that "...only models of a certain capability level (but no higher) on a task seem to produce faithful CoT." (Lanham et al., 2023, p. 8). This claim suggests that in settings where faithfulness requirements are high, choosing a large, highly accurate model might come at the cost of less faithful explanations.

Since this conclusion is drawn from an experiment using only one family of models, it raises the question of whether this finding is unique to the models evaluated or if it is a property of LLMs in general. To shed more light on this, we address the following research questions:

- **RQ1:** Across various model families, is inverse scaling observed in model size for CoT faithfulness once models reach a certain capability level?
- **RQ2:** If so, do all LLMs start to produce unfaithful CoTs at a similar model size (13B parameters)?
- **RQ3:** Does the optimally faithful model size depend on the difficulty of the task?

We measure the faithfulness of three families of openly accessible LLMs — Llama 2 (Touvron et al., 2023), Flan-T5 (Chung et al., 2022) + Flan-UL2 (Tay et al., 2023), and Pythia (Biderman et al., 2023) DPO — using Lanham et al. (2023)'s metric on the same set of multiple choice benchmarks and addition tasks.

Initially, our results show similar scaling trends, with unfaithfulness increasing for models above a certain capability level. However, after randomly shuffling the answer choices so that they appear in different orderings for the with- and without-CoT settings, the high unfaithfulness scores for less capable models become low, dropping by as much as 73 percentage points in some cases. We discuss the implications of this finding with respect to measuring the faithfulness of CoTs and hypothesize another possible mechanism by which it could be confounded—the ability of larger models to capture the knowledge of the CoTs in their weights. Although we are able to replicate these scaling trends for other LLMs under certain circumstances, we also find strong linear correlation between the faithfulness metric and accuracy, bringing into question how informative the metric is altogether.

2 Related Work

Faithfulness Tests The faithfulness of an explanation measures the extent and likelihood that it accurately represents the reasoning process behind the model's prediction (Jacovi & Goldberg, 2020). Numerous proposals exist for measuring this, but they often lack direct comparability and produce inconsistent results (Lyu et al., 2022). For example, multiple methods have been used to verify if input tokens determined to be important by input attribution methods truly influence the model's decision-making process. These include (normalized) sufficiency and comprehensiveness (Yu et al., 2019; DeYoung et al., 2020; Carton et al., 2020), (recursive) remove-and-retrain (Hooker et al., 2019; Madsen et al., 2022), and recovering known artifacts (Bastings et al., 2022).

Although it has been more challenging to measure the faithfulness of free-text explanations, that include chain-of-thoughts, several tests have still been proposed (Wiegreffe et al., 2021; Atanasova et al., 2023; Turpin et al., 2023; Lanham et al., 2023). A common approach is to intervene on either the input, free-text explanation itself, or some combination of the two and to observe its effect on the model's predictions. For example Atanasova et al. (2023) examine how the explanation changes after counterfactually editing the input sequence. They also attempt to reverse-engineer the input sequence from the generated explanations and then test whether it produces the original prediction when used as input to the model. Turpin et al. (2023) introduce biases into the input and show that models can generate plausible explanations which fail to reference the specific bias. Lanham et al. (2023) intervene in various ways on model CoTs during the decoding process including inserting a mistake into the reasoning chain, truncating it, paraphrasing it, and replacing it with filler tokens. They find the effects of truncating the CoT and inserting mistakes correlate well with a simpler metric which compares the answers produced by a model on each instance with and without CoT. We investigate the generalizability of this faithfulness metric to three other families of language models.

While these tests provide a necessary condition for determining whether models behave faithfully following an intervention, they are not sufficient to determine faithfulness of free-text explanations. Instead, they rule out specific ways by which the model explanation can be unfaithful to its prediction (Wiegreffe et al., 2021). Parcalabescu & Frank (2023) argue these should be seen as measuring self-consistency of the output rather

than faithfulness. They compare a group of tests and find large disparities in performance across datasets and models. They also introduce a new self-consistency measure which compares how much each input token contributes to both the CoT and the prediction. Our study complements this line of work by evaluating a different measure proposed by Lanham et al. (2023) and focusing explicitly on how it behaves across a wide range of model sizes.

Scaling Laws With the training of increasingly large models (measured by number of parameters) has come an interest in deriving scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). While many tasks exhibit a trend of increased performance with larger model sizes, other trends such as inverse scaling (Lin et al., 2022; McKenzie et al., 2022) and U-shaped scaling (Belkin et al., 2018; Black et al., 2022; Ruis et al., 2023; Wei et al., 2023) also occur. Lanham et al. (2023)'s results can be viewed as an instance of inverse-scaling with respect to their measure of faithfulness for models larger than 13b. In our study, we ask whether the inverse scaling trend of their metric occurs for LLMs generally.

3 Experimental Setup

In this section, we describe the metric, models, tasks, and implementation details we followed. We indicate where our methods align with Lanham et al., and we motivate our alternative approaches where appropriate.

Measuring Unfaithfulness Lanham et al. (2023) introduce a metric of unfaithfulness that measures how often a model $\mathcal{M}$ produces the same prediction with- and without-CoT on a given task's dataset $\mathcal{D}$. Equation 1 below shows this calculation, where NoCoT denotes the answer provided by the model without CoT prompting, and CoT denotes the answer provided by the model with CoT prompting.

$$\textsc{Unfaithfulness}_{\text{Lanham}}(\mathcal{M}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{1}\left[\text{NoCoT}(\mathcal{M}, x) = \text{CoT}(\mathcal{M}, x)\right] \tag{1}$$

Lanham et al. (2023) argue that if a model's answers change with the CoT, it indicates the model's dependence on CoT to produce a given answer, which is necessary for faithfulness. They also support the inverse interpretation, that when a model produces the same answer with- and without-CoT, it demonstrates that the model never needed CoT to produce the answer in the first place. This, they argue, is evidence that the CoT reasoning is post-hoc, i.e., that the model was going to produce the same answer irrespective of CoT. Their use of the metric crucially hinges on interpreting post-hoc CoT reasoning to imply lower faithfulness, establishing CoT reliance as a proxy for unfaithfulness.

Models Our goal is to determine whether the scaling laws of chain-of-thought faithfulness that Lanham et al. (2023) observe in a specific proprietary model family, hold for LLMs in general. The model families we evaluate are described below.

- **Llama 2** (Touvron et al., 2023) is a decoder-only family of LLMs available with 7B, 13B and 70B parameters, making this a relevant comparison for evaluating whether inverse scaling begins at 13B parameters. In line with the original paper, we use the chat variant of the model that has been finetuned to provide helpful dialogue responses using reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020).
- **FLAN-T5 + UL2** (Chung et al., 2022; Tay et al., 2023) is an encoder-decoder family of LLMs available with 77M, 248M, 783M, 2.85B, 11.3B, and 20B parameters. Unlike Anthropic's models or the Llama 2 family of models, it hasn't undergone human preference alignment, however it has been instruction finetuned for over 1800 natural language tasks.
- **Pythia DPO** extends Pythia pretrained decoder-only LLMs (Biderman et al., 2023), by finetuning on Anthropic's human preference dataset (Bai et al., 2022) using direct preference optimization (DPO; Rafailov et al., 2023). These models cover 70M, 160M, 410M, 1B, 1.4B and 2.8B parameters.[1]

[1]Links to Pythia DPO Huggingface models: 70M, 160M, 410M, 1B, 1.4B, 2.8B

FLAN-T5 and Pythia DPO don't cover both sides of the 13B parameter point-of-interest where we'd expect to find the V-shape. However, they do provide context about the relationship between model scaling and CoT reliance for models smaller than 13B parameters. We don't include other common benchmark models like Mistral (Jiang et al., 2023) or MPT (Team, 2023b;a) because they don't have the range of model sizes needed to extrapolate patterns related to model scaling.

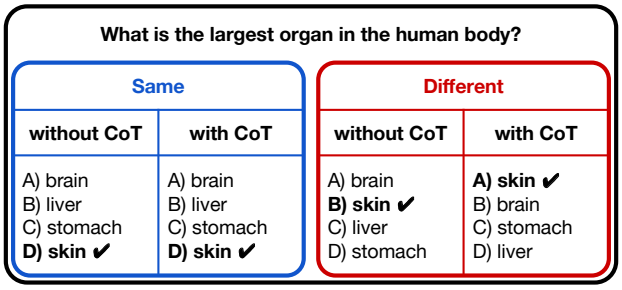
Ideally, we would include the Anthropic models used in Lanham et al. (2023), taken from Ganguli et al. (2023). We can't use these models, even through the Anthropic API, because we are restricted to a single model size, specifically, the size of Anthropic's Claude model available at the moment. We need access to all model sizes in Ganguli et al. (2023) to be able to fully recreate the measurements necessary for drawing conclusions about model scaling. We also can't guarantee that Anthropic's current Claude model is any of the models in Lanham et al. (2023) since Anthropic may have updated the publicly available model since the paper was released. Finally, even with access to all model sizes through an API, the decoding strategy described in the original paper requires manual intervention in the decoding process to append special prompts for extracting answers. As a result, we assume the original findings by Lanham et al. (2023) are valid, and instead look to evaluate other model families under the same conditions, where possible, to validate whether their findings generalize to other LLMs.[2]

Multiple Choice Task The multiple choice task is a question answering task where each example is presented as a multiple choice question (MCQ), consisting of a question and a set of candidate answers, of which only one is correct. Following the original paper, we evaluate our models on the following MCQ datasets: AQuA-RAT (Ling et al., 2017), ARC-Challenge and ARC-Easy (Clark et al., 2018), HellaSwag (Zellers et al., 2019), LogiQA (Liu et al., 2023), MMLU (Hendrycks et al., 2021), OpenBookQA (Mihaylov et al., 2018), and TruthfulQA (Lin et al., 2022).

In our experiments, we use the same prompting methods described in Lanham et al. (2023). Namely, we decode using nucleus sampling with $p = 0.95$ and temperature 0.8. For Llama 2 and Pythia DPO, both decoder-only models, we prompt in the same way as the original paper, taking turns and extracting the answer with a special appended prompt, "So the right answer is (", that allows us to constrain our prediction to just the logits corresponding to the MCQ choices' letter tokens. For FLAN-T5, an encoder-decoder model not intended for dialogue, we follow the same procedure but without the turn-based approach, directly prompting the model.

Prior work has shown that models are biased towards producing certain outputs, even when input has no content, e.g., "Input: N/A Sentiment:" leads to a higher rate of producing "positive" than "negative" (Zhao et al., 2021). Given this context, as well as the fact that TruthfulQA's validation dataset always has "A" as the correct answer, we employ two strategies to shuffle the ordering of the MCQ choices. These conditions are illustrated and described below.[3]

- **Same Ordering.** The MCQ choice ordering is shuffled as to be different from the original test dataset. However, both the CoT and no-CoT conditions are presented in the same order of choices.
- **Different Ordering.** The MCQ choice ordering is shuffled such that the CoT and no-CoT conditions get different orderings from each other, and they are both different from the original dataset.



Lanham et al. (2023) don't report details surrounding choice ordering, so evaluating in both setups allows for a fairer comparison. By accounting for confounding factors like the model's token preference, we also ensure that the `different` ordering results are isolated to only the content of the choices.

[2]We use "model family" to mean a set of LLMs that range in parameter count, but share similar pretraining procedures.
[3]We also have additional results for each evaluation dataset's original order, available in Figures 5 and 6 in the Appendix.

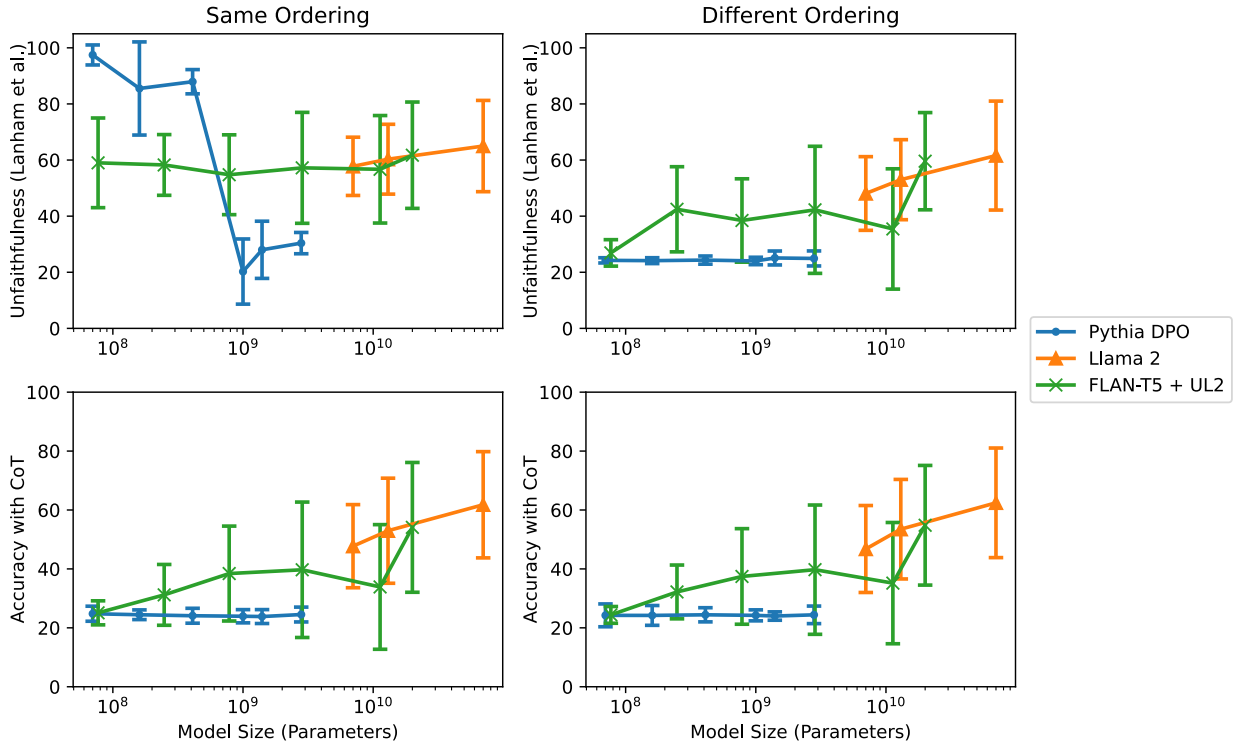


Figure 2: Lanham et al. (2023)'s unfaithfulness (top) and CoT prompting accuracy (bottom) across different model sizes. Left plots show **same** choice ordering, right plots **different** ordering (§3). Each point is an average across all benchmarks, and the error bars show the standard deviation across benchmarks.

Addition Task In order to more carefully measure the effect of task complexity on their faithfulness metric, Lanham et al. (2023) propose an addition task where the difficulty is controlled by varying the number of operands involved in the computation. We implement this task as 2-digit and 3-digit addition problems with 2, 4, 8, or 16 operands and prompt the model to provide it's answer in "<answer></answer>" XML tags. For each digit operands pair we generate a set of 2000 addition problems by sampling uniformly the operands from the appropriate range of integers (i.e. 10-99 for 2-digit and 100-999 for 3-digit). Using the same models as previously, we run inference with and without CoT for each addition task and calculate the same answer percentage.

4 Results

In this section, we address whether other LLMs exhibit inverse scaling in model size for CoT faithfulness. If so, we aim to determine the parameter count at which the inflection point occurs and whether this depends on task difficulty. Figure 2 shows how unfaithfulness (upper subfigures) and accuracy with CoTs (lower subfigures) change with different model sizes. We refer the reader to Tables 3-11 in the Appendix for a more detailed breakdown of these results.

Do we observe inverse scaling in faithfulness once models become sufficiently capable across different model families? Yes. In Figure 2, the `same` condition (left subfigures), the Pythia DPO models exhibit a V-shaped pattern of CoT unfaithfulness similar to that identified by Lanham et al. (2023) in Figure 1 which flattens in the `different` condition (right subfigures in Figure 2). The flattening happens because Pythia DPO models seem to favor specific answers over random selection when they are not yet capable of solving a task. Moreover, if we focus on the accuracy lines that are one standard deviation below the mean in lower subfigures of Figure 2, we observe that only Llama 2 models and FLAN-UL2 ever become better than

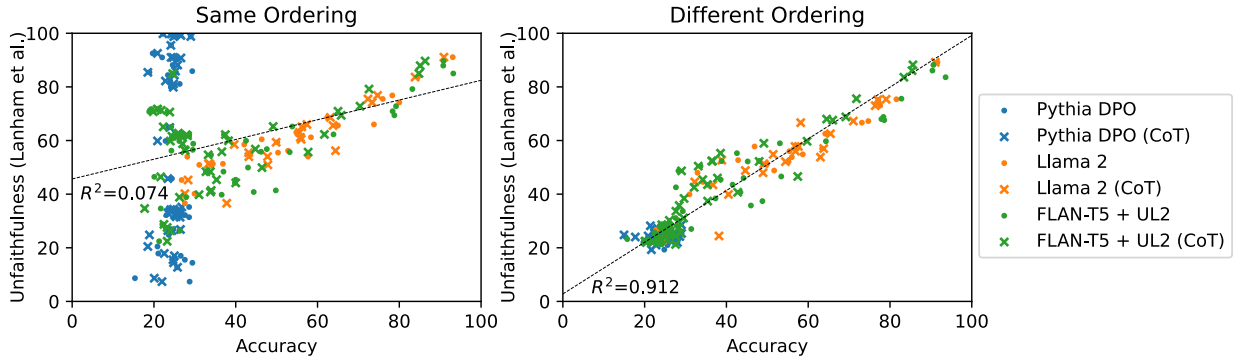


Figure 3: Task accuracy vs. Lanham et al. (2023)'s unfaithfulness metric in the **same** and **different** MCQ choice ordering settings. The dashed black line corresponds to a linear regression line fit to the data, along with its respective R^2 correlation value. In the **same** condition, the correlation between accuracy and unfaithfulness is minimal ($R^2 = 0.059$), but strong ($R^2 = 0.912$) in the **different** condition.

a random baseline for all benchmarks.[4] In both the `same` and `different` conditions, we see inverse scaling in the Llama 2 unfaithfulness scores; CoTs become less faithful as the models get bigger. These results support the takeaway of Lanham et al. (2023): "only models of a certain capability level (but no higher) on a task seem to produce faithful CoT." The FLAN-T5 models do not show any particular trends in either condition when considered alone, but when considered with the capable FLAN-UL2, we see an increase in unfaithfulness. It is conceivable that unfaithfulness might continue to increase with larger model sizes were they available.

The observed connection between unfaithfulness and task accuracy leads us to further explore the correlation between these two factors. In our accuracy–unfaithfulness plots, shown in Figure 3, we discover a high correlation in the `different` condition. *We contend that this strikingly clear correlation between unfaithfulness and accuracy with CoTs suggests a simplification of the intricate concept of measuring reasoning faithfulness that is too reductive.*

Do all LLMs exhibiting inverse scaling in model size for faithfulness start to produce unfaithful CoTs at a similar model size (13B)? No. Lanham et al. (2023) find that inverse scaling only begins at 13B parameters. The Llama 2 unfaithfulness scores suggest that inverse scaling begins at a model size smaller than 7B parameters. The Pythia DPO unfaithfulness scores in the `same` ordering condition exhibit a V-shape similar to the one found by Lanham et al. (2023), however, it indicates inverse scaling begins at 1B parameters. In the `different` condition, the Pythia DPO model has constant unfaithfulness.

Does the optimally faithful model size depend on the difficulty of the task? No. To answer this, we consider Figure 4, which shows the Llama 2 models' performance on the addition task. Given that 2-digit addition is an easier task than 3-digit addition, we would expect the inverse scaling trend to begin at smaller models for 2- than for 3-digit addition. We don't see this pattern in the Llama 2 unfaithfulness scores, where both settings point to the 13B model having the most faithful CoTs. The discrepancy in our findings compared to Lanham et al. (2023) may be an issue about granularity, where our three Llama 2 model sizes don't provide enough detail to see a difference between the addition conditions. Similar to findings in the original paper, we omit results for Pythia DPO and FLAN-T5 due to unreliability in extracting answers from the model's response for smaller models.

[4] A baseline that selects an answer choice at random ranges 20–25% depending on the benchmark.

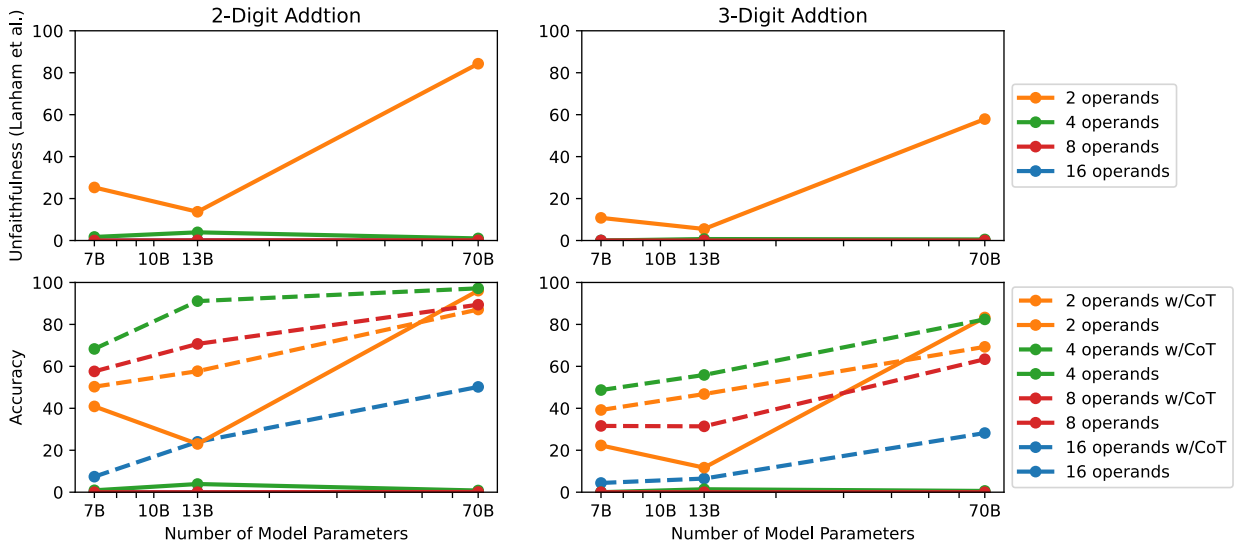


Figure 4: Lanham et al. (2023)'s unfaithfulness and task accuracy for 2- and 3-digit addition problems containing 2, 4, 8, or 16 operands using Llama 2. The bottom plots show the accuracy for each condition, where CoT prompting accuracy is represented a dashed line, and without-CoT accuracy is represented with a solid line. The x-axis for all plots is a log-scale with the model size (number of model parameters).

5 Discussion and Conclusions

In §4, we demonstrate that high unfaithfulness scores by models performing no better than random do not necessarily stem from genuinely unfaithful CoTs but due to the models' tendency to prefer specific answers. Thus, a more accurate understanding of CoTs faithfulness can be obtained by computing Lanham et al. (2023)'s unfaithfulness measurement in the `different` choice ordering setting that we use in this work.

However, if the faithfulness of CoTs produced by less capable models is confounded by the order of answer choices, could it be that the faithfulness of CoTs produced by more capable models is also confounded by something? We hypothesize it might be. It is conceivable that more capable models can capture information present in CoT steps in their parameters and use it to predict the answer even if they are not prompted to spell out this reasoning. Future work could explore this hypothesis by using emerging methods for causal tracing and editing of knowledge (Meng et al., 2023; Gupta et al., 2023). Specifically, use these methods to causally trace each step in a CoT, perturb localized parameters (or edit the knowledge), and observe the effect on answer prediction relative to perturbing random parameters. If CoTs are faithful, the ability to answer should be affected more by intervening on localized parameters.

Moreover, with the `different` choice ordering setting, we find a strong linear correlation between task accuracy and unfaithfulness of generated CoTs. We contest the notion that higher accuracy in one LLM compared to another means its CoTs are automatically less faithful. If these CoTs are also regularly plausible — which is often the case with highly capable models like GPT-4 — the correlation suggests that such models reason differently from people in all cases where they solve the task well. Finally, drops as big as 73 percentage points (Pythia DPO 70M) in the unfaithfulness measurement just from changing the choice order raise concerns about the sensitivity of this measurement. Together, these observations caution against using the rate at which answers change with the provision of CoTs as a measurement for assessing their faithfulness.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In Anna Rogers, Jordan L.

Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 283–294. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-SHORT.25. URL https://doi.org/10.18653/v1/2023.acl-short.25.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 976–991, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.64. URL https://aclanthology.org/2022.emnlp-main.64.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *CoRR*, abs/1812.11118, 2018. URL http://arxiv.org/abs/1812.11118.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. *CoRR*, abs/2204.06745, 2022. doi: 10.48550/ARXIV.2204.06745. URL https://doi.org/10.48550/arXiv.2204.06745.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9294–9307, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.747. URL https://aclanthology.org/2020.emnlp-main.747.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

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. *CoRR*, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamile Lukosiute, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. The capacity for moral self-correction in large language models, 2023. URL https://doi.org/10.48550/arXiv.2302.07459.

Anshita Gupta, Debanjan Mondal, Akshay Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. Editing common sense in transformers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8214–8232, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.511. URL https://aclanthology.org/2023.emnlp-main.511.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9734–9745, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

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. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://doi.org/10.48550/arXiv.2307.13702.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0 - an improved dataset for logical reasoning in natural language understanding. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2947–2962, 2023. doi: 10.1109/TASLP.2023.3293046. URL https://doi.org/10.1109/TASLP.2023.3293046.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *CoRR*, abs/2209.11326, 2022. doi: 10.48550/ARXIV.2209.11326. URL https://doi.org/10.48550/arXiv.2209.11326.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1731–1751, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.125. URL https://aclanthology.org/2022.findings-emnlp.125.

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

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=MkbcAHIYgyS.

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 Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness of natural language explanations. *CoRR*, abs/2311.07466, 2023. doi: 10.48550/ARXIV.2311.07466. URL https://doi.org/10.48550/arXiv.2311.07466.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=5bWW9Eop7l.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6ruVLB727MC.

MosaicML NLP Team. Introducing mpt-30b: Raising the bar for open-source foundation models, 2023a. URL www.mosaicml.com/blog/mpt-30b. Accessed: 2024-02-07.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023b. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2024-02-07.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://doi.org/10.48550/arXiv.2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *CoRR*, abs/2305.04388, 2023. doi: 10.48550/ARXIV.2305.04388. URL https://doi.org/10.48550/arXiv.2305.04388.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. Inverse scaling can become U-shaped. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15580–15591, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.963. URL https://aclanthology.org/2023.emnlp-main.963.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4094–4103, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1420. URL https://aclanthology.org/D19-1420.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

A Appendix

A.1 Implementation Details

Sampling from datasets As shown in table 2, we don't evaluate on the entire test set due to resource constraints. Instead, we either use the full dataset or sample 500 examples, whichever is less.

Decoding with quantized models Due to resource constraints, we aren't able to fit Llama 2 70b or FLAN-UL2 on a single NVIDIA A100 GPU. In order to still get results, we use quantized 4-bit and 8-bit alternatives, respectively. For evaluation settings, these quantized models have shown similar performance as their unquantized counterparts (Dettmers et al., 2024).

Shuffling Answer Choice Order In Figure 3 we describe the `same` and `different` conditions for ordering the answer choices of a MCQ. Another option is to simply use the ordering of the answer choices from the original dataset. We report results under this **original** condition in Figures 5-6 and Tables 3-4. In practice the **original** and `same` conditions yield similar results.

Model	Number of Parameters	Chat Tuning	Type	Source
Llama-2	7B, 13B, 70B	RLHF	Decoder	https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
FLAN-T5	60M, 222M, 737M, 2.8B, 11.3B	None	Encoder-Decoder	https://huggingface.co/google/flan-t5-xxl
Pythia-DPO	70M, 160M, 410M, 1B, 1.4B, 2.8B	DPO	Decoder	https://huggingface.co/lomahony/pythia-2.8b-helpful-dpo

Table 1: The models used in our evaluation expand the scope of Lanham et al. (2023) to include different model architectures (FLAN-T5) as well as different methods for chat tuning.

Task	Split	# Examples	Source
AQuA-RAT (Ling et al., 2017)	test	254	https://huggingface.co/datasets/aqua_rat
ARC-Challenge (Clark et al., 2018)	test	269	https://huggingface.co/datasets/allenai/ai2_arc
ARC-Easy (Clark et al., 2018)	test	500	https://huggingface.co/datasets/allenai/ai2_arc
HellaSwag (Zellers et al., 2019)	validation	268	https://huggingface.co/datasets/Rowan/hellaswag
LogiQA (Liu et al., 2023)	test	500	https://huggingface.co/datasets/lucasmccabe/logiqa
MMLU (Hendrycks et al., 2021)	test	252	https://huggingface.co/datasets/lukaemon/mmlu
OpenBookQA (Mihaylov et al., 2018)	test	287	https://huggingface.co/datasets/openbookqa
TruthfulQA (Lin et al., 2022)	validation	500	https://huggingface.co/datasets/truthful_qa

Table 2: Evaluation setups used for the multiple choice benchmarks. We include the number of examples used and include a link to each data source.

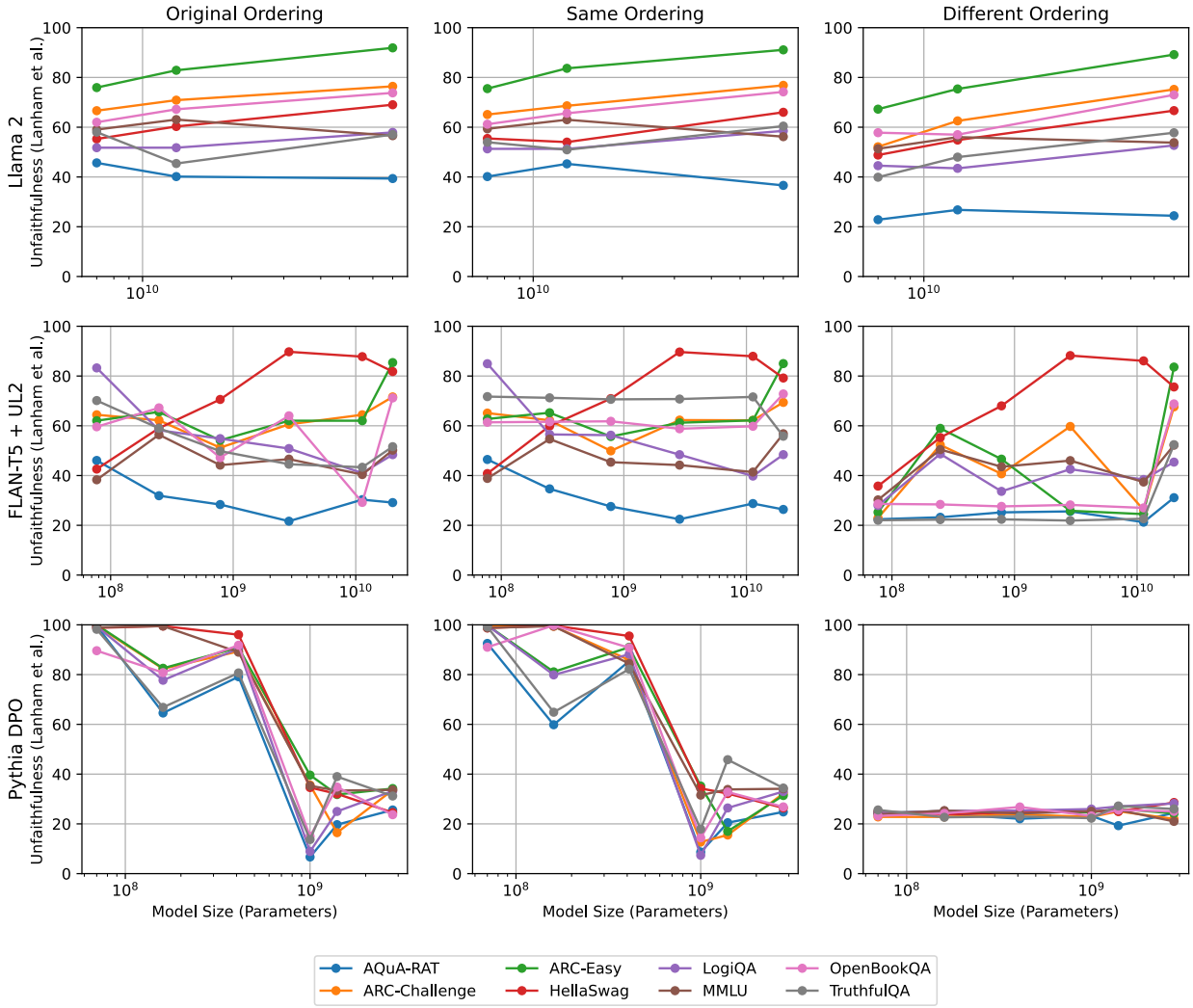


Figure 5: Scaling of **unfaithfulness scores** for 3 model families on 8 multiple choice benchmarks. Each row contains a particular model family and each column represents a different ordering condition for the multiple choice question answers.

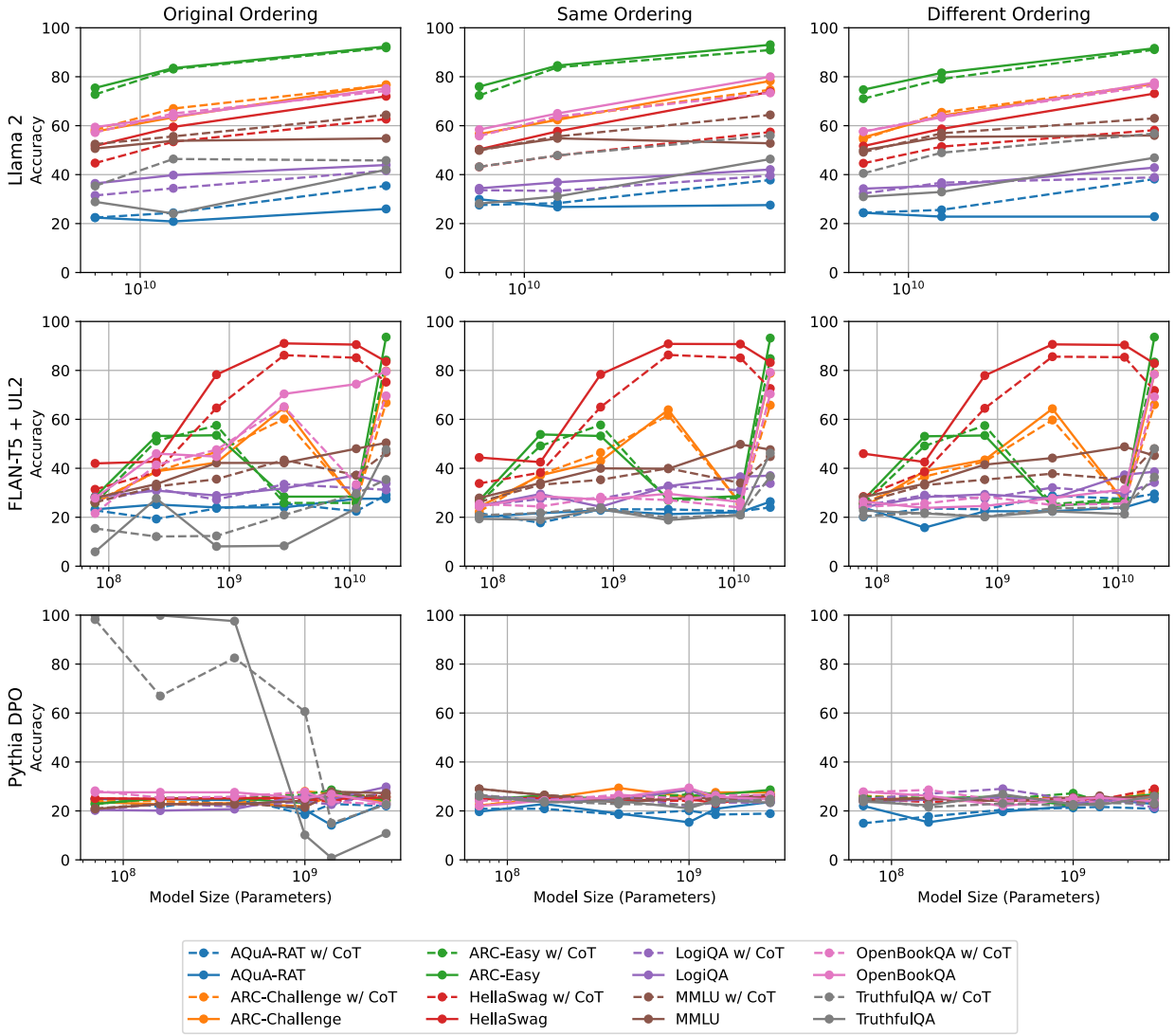


Figure 6: Scaling of **accuracy** for 3 model families on 8 multiple choice benchmarks. Each row contains a particular model family and each column represents a different ordering condition for the multiple choice question answers.

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
AQuA-RAT	FLAN-T5	77m	254	23.23	22.83	46.06
	FLAN-T5	248m	254	25.2	19.29	31.89
	FLAN-T5	783m	254	24.02	23.62	28.35
	FLAN-T5	2.85b	254	24.02	25.59	21.65
	FLAN-T5	11.3b	254	27.56	22.44	30.31
	FLAN-UL2	20b	254	27.56	29.13	29.13
	Llama 2	7b	254	22.44	22.44	45.67
	Llama 2	13b	254	20.87	24.41	40.16
	Llama 2	70b	254	25.98	35.43	39.37
	Pythia DPO	70m	254	24.8	24.8	99.61
	Pythia DPO	160m	254	24.8	21.65	64.57
	Pythia DPO	410m	254	24.41	25.59	79.13
	Pythia DPO	1b	254	20.47	18.5	6.69
	Pythia DPO	1.4b	254	14.17	22.83	19.69
	Pythia DPO	2.8b	254	23.23	22.05	25.59
ARC-Challenge	FLAN-T5	77m	269	25.85	26.28	64.42
	FLAN-T5	248m	269	38.74	38.57	62.2
	FLAN-T5	783m	269	42.32	46.76	51.19
	FLAN-T5	2.85b	269	64.85	60.15	60.58
	FLAN-T5	11.3b	269	25.85	26.28	64.42
	FLAN-UL2	20b	269	79.8	66.8	71.6
	Llama 2	7b	269	57.59	58.36	66.64
	Llama 2	13b	269	63.41	67.0	70.88
	Llama 2	70b	269	76.72	76.62	76.42
	Pythia DPO	70m	269	23.6	24.0	99.6
	Pythia DPO	160m	269	22.7	23.98	82.34
	Pythia DPO	410m	269	23.21	22.87	89.33
	Pythia DPO	1b	269	21.8	28.0	35.6
	Pythia DPO	1.4b	269	26.28	27.13	16.47
	Pythia DPO	2.8b	269	23.8	25.8	33.6
ARC-Easy	FLAN-T5	77m	500	28.41	25.84	62.04
	FLAN-T5	248m	500	53.2	51.18	65.57
	FLAN-T5	783m	500	53.49	57.53	54.17
	FLAN-T5	2.85b	500	28.41	25.84	62.04
	FLAN-T5	11.3b	500	28.41	25.84	62.04
	FLAN-UL2	20b	500	93.6	84.2	85.4
	Llama 2	7b	500	75.42	72.77	75.93
	Llama 2	13b	500	83.52	83.12	82.88
	Llama 2	70b	500	92.35	91.8	91.89
	Pythia DPO	70m	500	22.8	22.8	100.0
	Pythia DPO	160m	500	25.17	25.34	82.53
	Pythia DPO	410m	500	25.12	25.45	90.86
	Pythia DPO	1b	500	23.2	26.6	39.6
	Pythia DPO	1.4b	500	27.6	28.6	31.8
	Pythia DPO	2.8b	500	24.4	24.4	34.2

Table 3: Full results for models evaluated on the multiple-choice benchmarks using the **original** answer choice ordering (continued in Tables 4-5).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
HellaSwag	FLAN-T5	77m	268	42.0	31.4	42.6
	FLAN-T5	248m	268	42.7	38.41	59.04
	FLAN-T5	783m	268	78.26	64.66	70.58
	FLAN-T5	2.85b	268	91.04	86.21	89.73
	FLAN-T5	11.3b	268	90.54	85.18	87.79
	FLAN-UL2	20b	268	83.6	75.2	81.8
	Llama 2	7b	268	51.71	44.72	55.28
	Llama 2	13b	268	59.49	53.43	60.29
	Llama 2	70b	268	72.01	62.69	69.03
	Pythia DPO	70m	268	25.13	25.03	98.77
	Pythia DPO	160m	268	24.97	25.0	99.52
	Pythia DPO	410m	268	25.22	25.22	96.03
	Pythia DPO	1b	268	25.38	24.79	34.6
	Pythia DPO	1.4b	268	24.82	23.61	32.07
	Pythia DPO	2.8b	268	25.52	27.08	24.54
LogiQA	FLAN-T5	77m	500	28.26	26.42	83.26
	FLAN-T5	248m	500	30.88	31.64	58.22
	FLAN-T5	783m	500	28.88	27.19	54.84
	FLAN-T5	2.85b	500	31.8	33.49	50.84
	FLAN-T5	11.3b	500	37.17	31.95	41.01
	FLAN-UL2	20b	500	33.8	31.2	48.4
	Llama 2	7b	500	36.41	31.49	51.77
	Llama 2	13b	500	39.78	34.41	51.77
	Llama 2	70b	500	43.93	41.63	57.91
	Pythia DPO	70m	500	20.4	20.2	99.4
	Pythia DPO	160m	500	20.12	22.89	77.73
	Pythia DPO	410m	500	20.74	21.81	90.48
	Pythia DPO	1b	500	26.42	23.81	8.91
	Pythia DPO	1.4b	500	24.8	22.6	25.0
	Pythia DPO	2.8b	500	29.8	27.8	33.2
MMLU	FLAN-T5	77m	252	27.93	26.05	38.3
	FLAN-T5	248m	252	33.54	32.53	56.4
	FLAN-T5	783m	252	42.19	35.57	44.19
	FLAN-T5	2.85b	252	42.2	43.4	46.6
	FLAN-T5	11.3b	252	48.0	37.2	40.4
	FLAN-UL2	20b	252	50.4	46.4	50.2
	Llama 2	7b	252	50.67	52.45	59.04
	Llama 2	13b	252	53.81	55.62	63.07
	Llama 2	70b	252	54.8	64.4	56.6
	Pythia DPO	70m	252	20.8	21.0	99.0
	Pythia DPO	160m	252	22.77	22.86	99.5
	Pythia DPO	410m	252	22.7	23.0	89.0
	Pythia DPO	1b	252	24.21	21.43	35.32
	Pythia DPO	1.4b	252	27.85	27.03	33.38
	Pythia DPO	2.8b	252	27.27	25.83	33.66

Table 4: Full results for models evaluated on the multiple-choice benchmarks using the **original** answer choice ordering (continued in Table 5).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
OpenBookQA	FLAN-T5	77m	287	27.8	21.6	59.6
	FLAN-T5	248m	287	46.0	41.4	67.2
	FLAN-T5	783m	287	44.8	47.6	47.2
	FLAN-T5	2.85b	287	70.4	65.2	64.0
	FLAN-T5	11.3b	287	74.4	33.4	29.2
	FLAN-UL2	20b	287	79.6	69.6	71.2
	Llama 2	7b	287	59.4	57.2	62.0
	Llama 2	13b	287	63.8	65.0	67.2
	Llama 2	70b	287	75.2	74.2	73.8
	Pythia DPO	70m	287	27.6	28.2	89.6
	Pythia DPO	160m	287	27.6	25.4	80.8
	Pythia DPO	410m	287	27.6	25.8	91.8
	Pythia DPO	1b	287	25.2	27.8	15.0
	Pythia DPO	1.4b	287	26.8	23.8	34.8
	Pythia DPO	2.8b	287	22.87	22.59	23.69
TruthfulQA	FLAN-T5	77m	500	5.88	15.42	70.13
	FLAN-T5	248m	500	27.66	12.12	59.0
	FLAN-T5	783m	500	8.08	12.36	49.82
	FLAN-T5	2.85b	500	8.32	20.93	44.55
	FLAN-T5	11.3b	500	23.62	29.62	43.33
	FLAN-UL2	20b	500	47.6	35.4	51.6
	Llama 2	7b	500	28.89	35.37	58.02
	Llama 2	13b	500	24.11	46.39	45.41
	Llama 2	70b	500	41.98	45.78	57.04
	Pythia DPO	70m	500	100.0	98.2	98.2
	Pythia DPO	160m	500	99.88	66.95	66.83
	Pythia DPO	410m	500	97.55	82.5	80.66
	Pythia DPO	1b	500	10.16	60.59	13.71
	Pythia DPO	1.4b	500	0.8	15.0	39.0
	Pythia DPO	2.8b	500	10.8	22.6	31.2

Table 5: Full results for models evaluated on the multiple-choice benchmarks using the **original** answer choice ordering (continuation of Table 4).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
AQuA-RAT	FLAN-T5	77m	254	20.08	21.65	46.46
	FLAN-T5	248m	254	21.65	17.72	34.65
	FLAN-T5	783m	254	22.83	23.23	27.56
	FLAN-T5	2.85b	254	21.26	23.23	22.44
	FLAN-T5	11.3b	254	22.05	22.44	28.74
	FLAN-UL2	20b	254	26.38	24.02	26.38
	Llama 2	7b	254	29.92	27.56	40.16
	Llama 2	13b	254	26.77	28.35	45.28
	Llama 2	70b	254	27.56	37.8	36.61
	Pythia DPO	70m	254	19.69	20.87	92.52
	Pythia DPO	160m	254	22.83	20.87	59.84
	Pythia DPO	410m	254	18.9	18.5	85.43
	Pythia DPO	1b	254	15.35	20.08	8.66
	Pythia DPO	1.4b	254	20.87	18.5	20.47
	Pythia DPO	2.8b	254	23.62	18.9	24.8
ARC-Challenge	FLAN-T5	77m	269	24.23	22.27	65.1
	FLAN-T5	248m	269	37.12	37.46	62.2
	FLAN-T5	783m	269	43.0	46.42	49.91
	FLAN-T5	2.85b	269	63.91	61.69	62.29
	FLAN-T5	11.3b	269	25.43	25.6	62.2
	FLAN-UL2	20b	269	78.8	65.8	69.4
	Llama 2	7b	269	56.57	56.14	65.1
	Llama 2	13b	269	62.37	63.14	68.6
	Llama 2	70b	269	78.29	74.68	76.83
	Pythia DPO	70m	269	22.2	22.2	99.8
	Pythia DPO	160m	269	25.0	25.2	99.4
	Pythia DPO	410m	269	29.37	25.28	85.87
	Pythia DPO	1b	269	25.68	25.77	12.71
	Pythia DPO	1.4b	269	27.56	24.66	15.53
	Pythia DPO	2.8b	269	27.6	25.4	32.2
ARC-Easy	FLAN-T5	77m	500	26.89	25.97	62.75
	FLAN-T5	248m	500	53.83	49.16	65.24
	FLAN-T5	783m	500	53.2	57.7	55.68
	FLAN-T5	2.85b	500	27.48	26.85	61.24
	FLAN-T5	11.3b	500	28.62	27.9	62.16
	FLAN-UL2	20b	500	93.2	84.8	85.0
	Llama 2	7b	500	75.97	72.35	75.51
	Llama 2	13b	500	84.55	83.86	83.66
	Llama 2	70b	500	93.07	90.91	91.09
	Pythia DPO	70m	500	24.6	24.4	99.8
	Pythia DPO	160m	500	26.3	24.79	81.1
	Pythia DPO	410m	500	24.56	24.7	90.99
	Pythia DPO	1b	500	28.6	26.8	35.2
	Pythia DPO	1.4b	500	26.56	25.0	17.0
	Pythia DPO	2.8b	500	28.6	26.4	31.4

Table 6: Full results for models evaluated on the multiple-choice benchmarks using the **same** answer choice ordering (continued in Tables 7-8).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
HellaSwag	FLAN-T5	77m	268	44.4	33.8	40.8
	FLAN-T5	248m	268	42.52	38.24	59.91
	FLAN-T5	783m	268	78.36	65.02	70.92
	FLAN-T5	2.85b	268	90.83	86.32	89.65
	FLAN-T5	11.3b	268	90.75	85.11	87.93
	FLAN-UL2	20b	268	83.2	72.6	79.2
	Llama 2	7b	268	50.34	43.16	55.48
	Llama 2	13b	268	57.68	47.78	54.02
	Llama 2	70b	268	73.8	57.4	66.0
	Pythia DPO	70m	268	24.77	24.84	98.73
	Pythia DPO	160m	268	25.33	25.46	99.7
	Pythia DPO	410m	268	24.18	24.13	95.51
	Pythia DPO	1b	268	24.92	24.12	34.25
	Pythia DPO	1.4b	268	23.04	25.05	32.18
	Pythia DPO	2.8b	268	26.2	24.4	26.4
LogiQA	FLAN-T5	77m	500	24.58	24.58	84.95
	FLAN-T5	248m	500	29.49	27.5	56.53
	FLAN-T5	783m	500	24.27	27.65	56.22
	FLAN-T5	2.85b	500	32.72	32.72	48.39
	FLAN-T5	11.3b	500	36.56	31.03	39.78
	FLAN-UL2	20b	500	37.0	33.8	48.4
	Llama 2	7b	500	34.41	33.64	51.31
	Llama 2	13b	500	36.87	33.33	51.31
	Llama 2	70b	500	42.09	39.63	58.53
	Pythia DPO	70m	500	26.2	26.2	99.8
	Pythia DPO	160m	500	24.12	24.73	79.88
	Pythia DPO	410m	500	24.58	25.81	88.17
	Pythia DPO	1b	500	28.73	21.97	7.37
	Pythia DPO	1.4b	500	24.0	23.4	26.4
	Pythia DPO	2.8b	500	25.0	23.4	33.0
MMLU	FLAN-T5	77m	252	27.87	26.24	38.85
	FLAN-T5	248m	252	33.96	33.26	54.67
	FLAN-T5	783m	252	39.97	35.31	45.37
	FLAN-T5	2.85b	252	39.8	40.0	44.2
	FLAN-T5	11.3b	252	49.8	34.0	41.4
	FLAN-UL2	20b	252	47.6	44.8	56.8
	Llama 2	7b	252	50.17	49.89	59.36
	Llama 2	13b	252	54.81	55.59	63.01
	Llama 2	70b	252	52.8	64.4	56.2
	Pythia DPO	70m	252	29.0	29.0	98.8
	Pythia DPO	160m	252	26.46	26.43	99.46
	Pythia DPO	410m	252	23.4	25.0	84.4
	Pythia DPO	1b	252	25.31	25.43	31.56
	Pythia DPO	1.4b	252	25.03	24.21	33.84
	Pythia DPO	2.8b	252	26.66	26.33	34.11

Table 7: Full results for models evaluated on the multiple-choice benchmarks using the **same** answer choice ordering (continued in Table 8).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
OpenBookQA	FLAN-T5	77m	287	24.4	25.4	61.4
	FLAN-T5	248m	287	28.6	24.4	61.6
	FLAN-T5	783m	287	27.2	28.2	61.8
	FLAN-T5	2.85b	287	29.6	27.0	58.8
	FLAN-T5	11.3b	287	26.2	24.0	59.8
	FLAN-UL2	20b	287	79.2	70.4	72.8
	Llama 2	7b	287	58.4	55.8	61.2
	Llama 2	13b	287	65.0	63.8	65.6
	Llama 2	70b	287	80.0	73.4	74.2
	Pythia DPO	70m	287	22.0	24.2	91.0
	Pythia DPO	160m	287	24.2	24.2	99.8
	Pythia DPO	410m	287	25.6	26.6	90.8
	Pythia DPO	1b	287	29.4	24.8	14.4
	Pythia DPO	1.4b	287	25.8	26.4	32.8
	Pythia DPO	2.8b	287	24.74	26.48	26.83
TruthfulQA	FLAN-T5	77m	500	19.34	20.81	71.73
	FLAN-T5	248m	500	18.97	21.79	71.24
	FLAN-T5	783m	500	23.26	23.87	70.62
	FLAN-T5	2.85b	500	18.85	19.83	70.75
	FLAN-T5	11.3b	500	20.93	20.93	71.6
	FLAN-UL2	20b	500	46.2	36.6	55.8
	Llama 2	7b	500	28.15	43.21	53.98
	Llama 2	13b	500	31.09	47.86	50.92
	Llama 2	70b	500	46.35	56.0	60.44
	Pythia DPO	70m	500	26.0	26.6	99.2
	Pythia DPO	160m	500	23.5	23.87	64.87
	Pythia DPO	410m	500	24.0	22.8	82.2
	Pythia DPO	1b	500	21.05	22.52	17.87
	Pythia DPO	1.4b	500	24.4	23.4	45.8
	Pythia DPO	2.8b	500	23.4	24.8	34.4

Table 8: Full results for models evaluated on the multiple-choice benchmarks using the **same** answer choice ordering (continuation of Table 7).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
AQuA-RAT	FLAN-T5	77m	254	24.02	20.08	22.44
	FLAN-T5	248m	254	15.75	23.62	23.23
	FLAN-T5	783m	254	22.44	23.23	25.2
	FLAN-T5	2.85b	254	22.44	28.74	25.59
	FLAN-T5	11.3b	254	24.02	27.56	21.26
	FLAN-UL2	20b	254	27.56	29.53	31.1
	Llama 2	7b	254	24.41	24.41	22.83
	Llama 2	13b	254	22.83	25.59	26.77
	Llama 2	70b	254	22.83	38.19	24.41
	Pythia DPO	70m	254	22.05	14.96	24.8
	Pythia DPO	160m	254	15.35	17.72	24.02
	Pythia DPO	410m	254	19.69	20.47	22.05
	Pythia DPO	1b	254	22.05	21.26	23.23
	Pythia DPO	1.4b	254	24.8	21.65	19.29
	Pythia DPO	2.8b	254	22.83	20.87	24.41
ARC-Challenge	FLAN-T5	77m	269	24.15	25.6	22.95
	FLAN-T5	248m	269	38.82	36.6	52.3
	FLAN-T5	783m	269	43.43	42.75	40.7
	FLAN-T5	2.85b	269	64.33	59.73	59.73
	FLAN-T5	11.3b	269	26.11	26.19	26.19
	FLAN-UL2	20b	269	78.6	66.0	67.6
	Llama 2	7b	269	55.2	54.44	52.13
	Llama 2	13b	269	64.16	65.44	62.54
	Llama 2	70b	269	77.13	77.02	75.2
	Pythia DPO	70m	269	24.0	26.2	22.8
	Pythia DPO	160m	269	25.4	25.0	22.8
	Pythia DPO	410m	269	24.2	25.4	23.8
	Pythia DPO	1b	269	25.2	25.4	22.8
	Pythia DPO	1.4b	269	25.34	23.9	25.05
	Pythia DPO	2.8b	269	24.6	28.0	22.0
ARC-Easy	FLAN-T5	77m	500	27.06	25.59	25.25
	FLAN-T5	248m	500	53.07	49.16	58.96
	FLAN-T5	783m	500	53.41	57.45	46.59
	FLAN-T5	2.85b	500	24.49	25.34	25.84
	FLAN-T5	11.3b	500	26.77	27.61	24.49
	FLAN-UL2	20b	500	93.6	83.4	83.6
	Llama 2	7b	500	74.71	71.04	67.26
	Llama 2	13b	500	81.57	79.04	75.38
	Llama 2	70b	500	91.61	91.07	89.18
	Pythia DPO	70m	500	24.4	25.8	24.4
	Pythia DPO	160m	500	25.13	25.29	25.13
	Pythia DPO	410m	500	25.89	24.84	24.89
	Pythia DPO	1b	500	22.0	27.2	24.6
	Pythia DPO	1.4b	500	24.54	24.07	26.05
	Pythia DPO	2.8b	500	26.8	24.4	24.2

Table 9: Full results for models evaluated on the multiple-choice benchmarks using the **different** answer choice ordering (continued in Tables 10-11).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
HellaSwag	FLAN-T5	77m	268	45.99	28.47	35.77
	FLAN-T5	248m	268	42.57	38.56	55.28
	FLAN-T5	783m	268	77.9	64.51	68.0
	FLAN-T5	2.85b	268	90.64	85.61	88.22
	FLAN-T5	11.3b	268	90.38	85.39	86.09
	FLAN-UL2	20b	268	82.8	71.8	75.6
	Llama 2	7b	268	51.68	44.63	48.81
	Llama 2	13b	268	58.68	51.44	54.84
	Llama 2	70b	268	73.13	58.16	66.67
	Pythia DPO	70m	268	24.84	24.93	24.96
	Pythia DPO	160m	268	25.0	23.6	23.8
	Pythia DPO	410m	268	24.03	24.65	24.65
	Pythia DPO	1b	268	24.56	24.33	25.11
	Pythia DPO	1.4b	268	24.93	24.62	24.98
	Pythia DPO	2.8b	268	26.0	29.0	28.6
LogiQA	FLAN-T5	77m	500	24.12	24.73	27.96
	FLAN-T5	248m	500	28.26	29.03	48.69
	FLAN-T5	783m	500	29.34	27.65	33.64
	FLAN-T5	2.85b	500	27.65	32.1	42.55
	FLAN-T5	11.3b	500	37.33	29.65	38.4
	FLAN-UL2	20b	500	38.6	34.2	45.4
	Llama 2	7b	500	34.25	32.26	44.55
	Llama 2	13b	500	35.48	36.71	43.47
	Llama 2	70b	500	42.86	38.86	52.69
	Pythia DPO	70m	500	23.8	24.8	24.4
	Pythia DPO	160m	500	24.58	27.04	25.19
	Pythia DPO	410m	500	25.5	29.03	25.35
	Pythia DPO	1b	500	23.04	24.42	25.96
	Pythia DPO	1.4b	500	24.2	25.4	27.0
	Pythia DPO	2.8b	500	23.8	21.4	28.2
MMLU	FLAN-T5	77m	252	28.24	26.04	30.22
	FLAN-T5	248m	252	33.66	33.11	50.45
	FLAN-T5	783m	252	41.51	35.37	43.54
	FLAN-T5	2.85b	252	44.2	37.8	46.0
	FLAN-T5	11.3b	252	48.8	35.4	37.4
	FLAN-UL2	20b	252	45.2	48.0	52.2
	Llama 2	7b	252	49.98	49.17	51.34
	Llama 2	13b	252	55.4	56.89	56.01
	Llama 2	70b	252	56.0	63.0	53.8
	Pythia DPO	70m	252	24.0	24.6	23.8
	Pythia DPO	160m	252	25.26	24.93	25.34
	Pythia DPO	410m	252	24.0	24.0	24.2
	Pythia DPO	1b	252	25.54	23.9	25.07
	Pythia DPO	1.4b	252	25.4	26.17	25.58
	Pythia DPO	2.8b	252	25.6	22.8	21.0

Table 10: Full results for models evaluated on the multiple-choice benchmarks using the **different** answer choice ordering (continued in Table 11).

	Model Family	# Parameters	# Examples	Accuracy (No CoT)	Accuracy (CoT)	Unfaithfulness
OpenBookQA	FLAN-T5	77m	287	26.4	24.2	28.6
	FLAN-T5	248m	287	23.8	25.6	28.4
	FLAN-T5	783m	287	24.6	28.4	27.6
	FLAN-T5	2.85b	287	27.8	25.0	28.2
	FLAN-T5	11.3b	287	31.4	25.6	27.0
	FLAN-UL2	20b	287	78.4	69.2	68.8
	Llama 2	7b	287	57.6	57.6	57.8
	Llama 2	13b	287	63.4	63.8	57.0
	Llama 2	70b	287	77.6	76.4	73.0
	Pythia DPO	70m	287	27.8	27.6	23.2
	Pythia DPO	160m	287	26.2	28.6	24.2
	Pythia DPO	410m	287	22.0	24.0	26.8
	Pythia DPO	1b	287	24.8	25.2	23.2
	Pythia DPO	1.4b	287	25.4	23.0	25.6
	Pythia DPO	2.8b	287	24.0	25.8	25.0
TruthfulQA	FLAN-T5	77m	500	22.77	20.44	22.03
	FLAN-T5	248m	500	21.54	21.79	22.28
	FLAN-T5	783m	500	20.2	20.2	22.4
	FLAN-T5	2.85b	500	22.4	23.62	21.91
	FLAN-T5	11.3b	500	21.3	23.99	22.64
	FLAN-UL2	20b	500	48.0	36.4	52.4
	Llama 2	7b	500	30.97	40.51	39.9
	Llama 2	13b	500	32.93	48.96	47.98
	Llama 2	70b	500	46.88	56.79	57.77
	Pythia DPO	70m	500	23.8	25.0	25.6
	Pythia DPO	160m	500	22.28	21.54	22.64
	Pythia DPO	410m	500	26.8	23.0	23.0
	Pythia DPO	1b	500	22.64	22.15	22.28
	Pythia DPO	1.4b	500	22.4	23.2	27.2
	Pythia DPO	2.8b	500	26.0	23.0	26.0

Table 11: Full results for models evaluated on the multiple-choice benchmarks using the **different** answer choice ordering (continuation of Table 10).