Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations

Swarnadeep Saha[1] **Peter Hase**[1] **Nazneen Rajani**[2] **Mohit Bansal**[1]
[1]UNC Chapel Hill [2]Hugging Face
{swarna, peter, mbansal}@cs.unc.edu, nazneen@huggingface.co

Abstract

Recent work on explainable NLP has shown that few-shot prompting can enable large pre-trained language models (LLMs) to generate grammatical and factual natural language explanations for data labels. In this work, we study the connection between explainability and sample hardness by investigating the following research question – "Are LLMs and humans equally good at explaining data labels for both easy and hard samples?" We answer this question by first collecting human-written explanations in the form of generalizable commonsense rules on the task of Winograd Schema Challenge (Winogrande dataset). We compare these explanations with those generated by GPT-3 while varying the hardness of the test samples as well as the in-context samples. We observe that (1) GPT-3 explanations are as grammatical as human explanations regardless of the hardness of the test samples, (2) for easy examples, GPT-3 generates highly supportive explanations but human explanations are more generalizable, and (3) for hard examples, human explanations are significantly better than GPT-3 explanations both in terms of label-supportiveness and generalizability judgements. We also find that hardness of the in-context examples impacts the quality of GPT-3 explanations. Finally, we show that the supportiveness and generalizability aspects of human explanations are also impacted by sample hardness, although by a much smaller margin than models.[1]

1 Introduction

Prior work on explainable NLP (Wiegreffe and Marasovic, 2021) has explored different forms of explanations ranging from extractive rationales (Zaidan et al., 2007; DeYoung et al., 2020), semi-structured, and structured explanations (Jansen et al., 2019; Mostafazadeh et al., 2020; Saha et al., 2021) to free-text explanations (Camburu et al., 2018). Due to the flexibility of free-text explanations, they have emerged as a popular form of explanations with multiple benchmarks developed around them, as well as models that generate such explanations using seq2seq language models (Ehsan et al., 2018; Camburu et al., 2018; Rajani et al., 2019; Narang et al., 2020). Few-shot prompting (Radford et al., 2019; Schick and Schütze, 2021) with Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) has been shown to produce highly fluent and factual natural language explanations that are often preferred over crowdsourced explanations in existing datasets (Wiegreffe et al., 2022). However, past work has not yet explored a critical dimension of datapoint-level explanations, which is how *hard* the data point is to classify correctly. Given recent work on measuring *hardness* of individual data points (Swayamdipta et al., 2020), we study how sample hardness influences both LLMs' and humans' ability to explain data labels. In summary, we are interested in investigating the following three research questions:

1. **RQ1.** *Do LLMs explain data labels as well as humans for both easy and hard examples?*
2. **RQ2.** *How much do LLM explanations vary based on the size and the hardness of the retrieval pool for choosing in-context samples?*
3. **RQ3.** *Are humans equally good at explaining easy and hard examples?*

As a case study, we investigate these questions for a classical commonsense reasoning task, Winograd Schema Challenge (Levesque et al., 2012) on a large-scale dataset, Winogrande (Sakaguchi et al., 2020) (examples in Fig. 1). We first collect generalizable rule-based explanations from humans like "If X is larger than Y, then X does not fit in Y.". To measure data *hardness*, we use *Data Maps* (Swayamdipta et al., 2020), an approach based on the training dynamics of a classification model.

[1]Supporting code and data are available at https://github.com/swarnaHub/ExplanationHardness.

Sentence	Options (Answer)	Human Explanation	GPT-3 Explanation
Katrina had the financial means to afford a new car while Monica did not, since _ had a high paying job.	Katrina/Monica	If X has a high paying job, then X is able to afford expensive purchases	If X has a higher paying job than Y, then X would have more money to afford a new car.
Angela tried to calm Carrie's nerves at the airport because _ was scared of flying in airplanes.	Angela/Carrie	If X is afraid and Y is not, then Y can help calm X down.	If X is scared of flying, then Y will try to calm X down.
Donald was always reading while Kenneth did not, as _ liked to read books all the time.	Donald/Kenneth	If X enjoys reading more than Y, X will read more books than Y	If X reads more than Y, then X likes to read books all the time.
Justin spent less on dog food than Eric because _ had a puppy dog at their house.	Justin/Eric	If X has a puppy and Y has a dog, then a puppy eats less so X spends less on dog food.	If X has a Y at their house, then X will spend less on Y.
I grabbed a pile of clothes from the basket that contained a T-shirt and a Sweater, since it was hot outside I put on the _ to wear.	Sweater/T-Shirt	If X is cooler than Y, then X will be chosen when it is hot outside.	If X is hot, then Y should be worn.
Helen was curious about how the ball broke the window, because the _ was hard.	ball/window	If it is surprising that X is broken then X must have been hard, otherwise it wouldn't be surprising that Y broke X.	If X is hard and Y isn't, then Y is more likely to break when hit by X.

Figure 1: Representative examples of explanations for Winograd Schema written by humans and generated by GPT-3 for easy (first 3 rows) and hard examples (last 3 rows). For easy examples, GPT-3 explanations are almost as good as humans, although less generalizable. For example, humans can generalize 'cars' to 'expensive purchases' while the model does not. For hard examples, GPT-3 explanations are often much worse than human ones.

Similar to Wiegreffe et al. (2022), we generate post-hoc explanations by conditioning on the answer leveraging GPT-3 with in-context learning. We perform human evaluation of the crowdsourced and model-generated explanations and compare them on the basis of 'grammaticality', 'supportiveness' and 'generalizability'. In summary, we report the following findings:

- LLM-generated explanations match the grammaticality/fluency of human-written explanations regardless of the hardness of test samples.
- For easy examples, both models and humans write 'supportive' explanations, but humans write more 'generalizable' explanations that can explain multiple similar data points. For hard examples, humans write explanations that are not only more 'generalizable' but also significantly more 'supportive' of the label.
- While choosing in-context examples, factors like size and hardness of the retrieval pool affect the quality of model-generated explanations.
- Humans, while much better than models in explaining hard examples, also struggle with writing generalizable explanations for these points, succeeding only about 2/3rd of the time.

2 Method and Experimental Setup

Our method first estimates hardness of the samples using Data Maps (Swayamdipta et al., 2020) and then chooses a subset of easy, medium, and hard examples, for which we collect human-written explanations and generate explanations from a state-of-the-art model. Next, we answer our research questions by comparing the explanations against multiple granular evaluation axes.

Data Maps. We estimate sample hardness via a model-based approach[2] called Data maps (Swayamdipta et al., 2020). Data Maps characterize points x_i in a dataset along two dimensions according to a classifier's behavior during training: (1) confidence $\hat{\mu}_i$ which measures the mean model probability of the true label y_i^* across E epochs, and (2) variability $\hat{\sigma}_i$ which measures the standard deviation of the model probability of the true label across epochs.

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^*|x_i)$$

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta^{(e)}}(y_i^*|x_i) - \hat{\mu}_i)^2}{E}}$$

where $p_{\theta^{(e)}}$ denotes the model's probability with parameters $\theta^{(e)}$ at the end of the e^{th} epoch. These two metrics give rise to different portions in the dataset including *easy-to-learn* examples where the model consistently predicts the sample correctly across epochs (high confidence, low variability), *hard-to-learn* examples where the model rarely predicts the sample correctly (low confidence, low variability) and *ambiguous* examples where the model is indecisive about its predictions (high variability). We

[2]We do not rely on human annotations for hardness quantification because of its subjectivity. Data Maps also provide a hardness ranking of the samples, which might be difficult to obtain from humans.

fine-tune RoBERTa-large (Liu et al., 2019) on the Winogrande dataset to compute the confidence and variability of each training sample in the dataset. The two metrics are then used to rank the samples from *easy* to *hard* (most confident to least confident) and *least-ambiguous* to *most-ambiguous* (least variable to most variable). As discussed later, we choose a subset of these examples to compare human and model-generated explanations.

Explanations for Winograd Schema. Next, we define the structure of explanations for the Winograd Schema Challenge (Levesque et al., 2012). Specifically, these are semi-structured if-then commonsense rules as shown in Fig. 1. This characterization of explanations allows us to (1) capture generalizable commonsense knowledge via placeholders X (and Y) capable of explaining a number of similar data points, (2) enforce the common structural form of an if-then rule for all data points in this task, while still maintaining the flexibility of free-text explanations (see Fig. 1 for some examples), (3) ensure non-trivial explanations that do not leak the label (Hase et al., 2020), with the aim of avoiding explanations that only repeat the label without providing generalizable background knowledge (a common issue in past explanation datasets), (4) evaluate explanation properties with reduced human subjectivity due to their semi-structural form.

Human Explanation Collection. Using the above criteria for constructing explanations (see detailed instructions in Fig. 6), we collect human-written explanations on Amazon Mechanical Turk. In order to ensure that the explanations do not explicitly leak the label, the annotators are asked to write explanations in the form of generalizable commonsense rules consisting of placeholders X (and Y) without mentioning the actual options. We collect explanations for 500 easiest and 500 hardest samples, along with 100 examples with medium hardness (around the median confidence). We do not collect explanations separately for least and most ambiguous samples because ambiguity correlates strongly with hardness, i.e., the least ambiguous examples are often the easiest while the most ambiguous examples are also typically the hardest.

Explanation Generation via GPT-3. Next, we select GPT-3 (Brown et al., 2020) as a representative candidate of today's NLP model landscape to generate explanations from. For each set of 500 easy and hard samples, we randomly split them into 400 samples for retrieving in-context samples and 100 samples for testing. We generate explanations for the test samples using the largest (175B) "text-davinci-002" InstructGPT model of GPT-3 by conditioning on the context and the gold label (as shown in Fig. 8). The in-context samples are chosen by computing the embeddings of the test sample and the retrieval samples using Sentence BERT (Reimers and Gurevych, 2019) and selecting the top-k samples (see Appendix C for examples). We set k to 5 in our experiments. Further details of our prompting method are in Appendix B.

Explanation Evaluation. Having obtained human and model explanations, we now describe their evaluation process. Due to the limitations of automatic metrics for evaluating explanation quality (Clinciu et al., 2021), we follow Wiegreffe et al. (2022) to conduct human evaluation of both crowdsourced and GPT-3 explanations on MTurk based on three attributes – *grammaticality*, *supportiveness*, and *generalizability*. When evaluating explanations for *grammaticality*, we evaluate their syntax and fluency while ignoring spelling mistakes and typos (which also hardly ever appear in model explanations). Given the semi-structured nature of our explanations, we evaluate *supportiveness* as whether, when appropriately replacing X and Y with the two options, the explanation answers the question "Why does this point receive the label it does?" (Miller, 2019). Lastly, we evaluate *generalizability* as how applicable the explanation is for other samples with different X and Y. We maintain a trained pool of annotators for both explanation authoring and verification while ensuring that they do not verify their own data. Each explanation is evaluated by 3 different annotators and the final results are obtained by majority voting. We report moderate inter-annotator agreement scores of Krippendorff's α (Krippendorff, 2011) between 0.4-0.6, details of which are discussed in Appendix A.

3 Results

3.1 RQ1: Do LLMs explain data labels as well as humans for both easy and hard examples?

In Fig. 2, we compare the human and GPT-3 explanations for easy, medium, and hard[3] examples

[3]Some hard examples can have incorrect labels (Swayamdipta et al., 2020). When collecting explanations from humans, we ask if they agree with the label (see Fig. 6). If they do not, we discard such examples (about

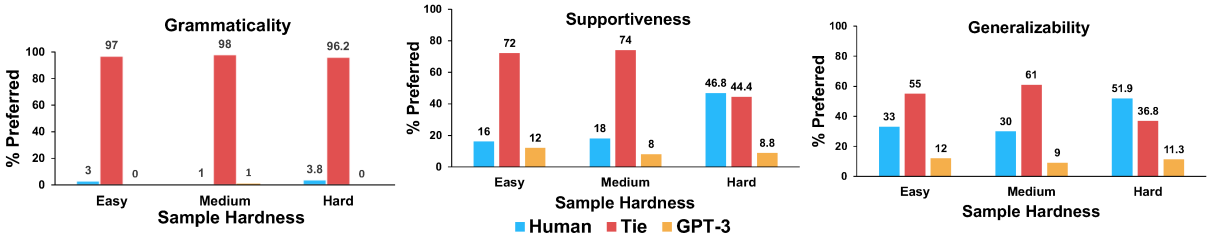


Figure 2: Head-to-head comparison of human and GPT-3 explanations for easy, medium and hard examples along the axes of grammaticality, supportiveness and generalizability.

along the three axes. We observe that GPT-3 is not only able to learn the if-then structure of the explanations but also matches humans in terms of generating grammatically fluent explanations, regardless of the sample hardness. For easy examples, GPT-3 explanations are almost as supportive of the label as human explanations, sometimes even outperforming humans. However, humans are typically better at writing more generalizable explanations that apply to broader contexts (see examples 1-3 in Fig. 1). For hard examples, GPT-3 often fails to generate sufficiently supportive explanations and hence significantly underperforms humans in more than 46% of the cases (see examples 4-6 in Fig. 1 and Appendix D for some common errors). This, in turn, also hurts the generalizability aspect of the model-generated explanations. Medium-hard examples show a trend similar to easy examples because their confidence values are much closer to the easy examples than the hard ones.

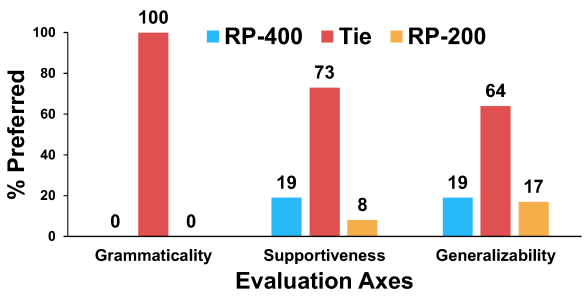
Significance Testing. Pertaining to the above results, we further use a non-parametric bootstrap test (Efron and Tibshirani, 1994) to evaluate whether the human win-rate differs significantly from the model win-rate, while treating ties as neutral. We encode human wins as 1, model wins as -1, and ties as 0 and test whether the average score is not equal to 0 (meaning that the win-rate differs between human and model). In summary, for easy and medium samples, humans' generalizability is significantly better than the model's (difference in win rate is 20 points with $p<0.001$), while for hard samples, both humans' generalizability and supportiveness are better than the model's (differences in win rates are 0.38 and 0.4 respectively, with $p<$1e-4). Next, for grammaticality, we test if GPT-3 explanations matches human explanations within a win-rate threshold of $\pm r$ points. For a threshold of $r=0.1$ (testing that grammaticality win-rates are within 10 percentage points of each other), we obtain $p<0.05$, and a threshold of $r=0.15$ yields $p<$1e-4. This suggests that the model's grammaticality significantly matches human's for *threshold* values around 0.1.

3.2 RQ2: How much do model explanations vary based on the size and the hardness of the retrieval pool for choosing in-context samples?

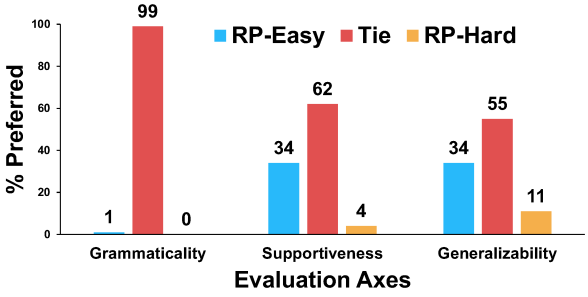
We investigate RQ2 by conducting two experiments in which we compare the explanations generated by GPT-3 for 100 easy examples. In the first, we vary the size of the retrieval pool (RP) for selecting in-context examples from 400 to 200 while keeping the average hardness constant, and in the second, we vary the hardness of the retrieval pool from easy to hard examples with the size of the pool (400) remaining constant. As shown in Fig. 3, the grammaticality of the explanations is unaffected in both experiments. However, supportiveness drops when the in-context samples are retrieved from a smaller pool. A larger pool increases the likelihood of having more similar in-context examples to the test sample, and we conclude that similar in-context examples improve the supportiveness of the explanation. We also find that when explaining easy examples, having a retrieval pool of similar easy examples helps the model generate better explanations, possibly because of more similar in-context examples. Combining with RQ1, we conclude that hardness of both in-context and test samples can affect the quality of model explanations.

We also conduct a similar study for comparing the explanation quality of 100 hard test examples by varying the hardness of the retrieval pool. In contrast to easy test examples, we do not observe statistically significant differences in explanation quality for hard examples when the retrieval pool's hardness is varied. In particular, with respect to supportiveness, the win percentages for hard and easy pool are 20% and 18% respectively, with remaining 62% being ties, while for generalizability, they are 33% and 25% respectively, with remain-

14%) from evaluation.



(a) Size of Retrieval Pool



(b) Hardness of Retrieval Pool

Figure 3: Head-to-head comparison of GPT-3 explanations for easy examples by varying the size (400/200) and hardness (easy/hard) of the Retrieval Pool (RP) for choosing in-context examples.

ing 42% being ties. We believe that the quality of hard examples may not be sensitive to changes in the in-context examples simply because the corresponding explanations are not very good to begin with.

3.3 RQ3: Are humans equally good at explaining easy vs. hard examples?

In RQ1, we compared the relative performance of model and humans in explaining easy, medium, and hard examples. RQ3 now evaluates the absolute quality of human-written explanations. In particular, we ask the annotators to rate whether the explanations demonstrate *acceptable* grammaticality, supportiveness, and generalizability. Fig. 4 shows the fraction of *acceptable* human explanations along these three axes for easy and hard examples. We observe that humans also find it hard to write generalizable explanations for some hard examples. Overall, the quality of human explanations is also impacted by the hardness of the samples, although to a lesser extent than GPT-3 since human explanations become clearly preferable to model explanations as hardness increases (RQ1).

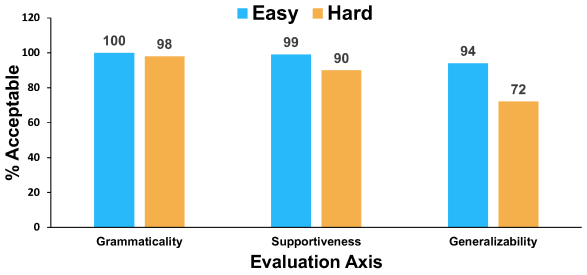


Figure 4: Percentage of acceptable human explanations for easy and hard examples across evaluation axes.

4 Related Work

There has been significant progress made in recent years on both curating natural language explanation datasets (Camburu et al., 2018; Rajani et al., 2019; Brahman et al., 2021; Aggarwal et al., 2021, *inter alia*) as well as generating them (Rajani et al., 2019; Shwartz et al., 2020). Related to the Winograd Schema Challenge, WinoWhy (Zhang et al., 2020) contains explanations only for the WSC273 test set (Levesque et al., 2012) and does not follow the structure of our commonsense rule-based explanations, thereby leading to label leakage. Label leakage makes evaluation of explanations harder because supportiveness can become trivial. Our study builds on top of prior works that also generate free-text explanations using in-context learning with GPT-3 (Marasović et al., 2022; Wiegreffe et al., 2022). However, our novelty lies in investigating the connection between explainability and sample hardness. A number of concurrent works have also explored free-text explanations for in-context learning in various reasoning tasks (Nye et al., 2021; Chowdhery et al., 2022; Wei et al., 2022; Lampinen et al., 2022; Wang et al., 2022; Ye and Durrett, 2022), primarily focusing on improving model performance with explanations and not evaluating explanation properties or factors that might influence them.

5 Conclusion

We studied the effect of sample hardness on the quality of post-hoc explanations generated by LLMs for data labels. We concluded that while LLM explanations are as fluent as human explanations regardless of the sample hardness, humans are typically better at writing more generalizable explanations and specifically, for hard examples, human explanations are also more supportive of the label. Factors like the hardness and size of the retrieval pool for choosing in-context examples can further impact the explanation quality. We also observe that the generalizability aspect of human explanations drops for harder examples, although by a smaller margin than models.

Limitations

The goal of our study is to evaluate how well models explain the data labels and not their own answers for the data points. Hence, both humans and models write or generate post-hoc explanations by conditioning on the gold labels. This also leads us to evaluate the explanations for how acceptable they are to the humans rather than their faithfulness to the model decisions (Wiegreffe and Pinter, 2020; Jacovi and Goldberg, 2020). The notion of data maps-driven instance difficulty (Swayamdipta et al., 2020) is primarily model dependent, and it is conceivable that different choices of models (or model-families) would yield different rank-ordering of data points by hardness. However, we measure the relative hardness of the data points and it is very unlikely that the k-easiest samples for RoBERTa (which is used to estimate sample hardness) will be the k-hardest samples for GPT-3 (which is used to generate explanations) or vice versa. In addition, we find that humans also struggle to explain our estimated 'hard' examples. These factors make our results fairly generalizable and future work can explore this direction further. It would also be interesting to see how our results generalize to other forms of explanations in NLP like rationales or structured explanations.

Acknowledgements

We thank the reviewers for their helpful feedback and the annotators for their time and effort. This work was supported by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, DARPA MCS Grant N66001-19-2-4031, ONR Grant N00014-18-1-2871, and Google PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592–12601.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. 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*, pages 2376–2387.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4351–4367.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Peter A Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T Morrison. 2019. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 2732–2740. European Language Resources Association (ELRA).

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. WANLI: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 410–424. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. Glucose: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In *NAACL*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe and Yuval Pinter. 2020. Attention is not not explanation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 11–20. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot in-context learning. *arXiv preprint arXiv:2205.03401*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745.

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

A Crowdsourcing Details

All our crowdsourcing studies are done on Amazon Mechanical Turk. We select crowdworkers who are located in the US with a HIT approval rate higher than 96% and at least 1000 HITs approved. We conduct qualification tests before crowdworkers are allowed to write and verify explanations. As shown in Figure 5, it tests the annotator's understanding of the Winograd Schema Challenge by asking to choose the correct option given the sentence and get all questions correct. In Figure 6, we show the instructions and interface for collecting human-written explanations. Finally, in Figure 7, we show the interface for explanation verification. We pay annotators \$0.10 for each HIT of explanation construction and \$0.15 for each HIT of explanation verification at an hourly wage of \$12-15.

	Easy	Hard
Grammaticality	0.63	0.61
Supportiveness	0.51	0.43
Generalizability	0.45	0.37

Table 1: Inter-annotator agreement scores (Krippendorff's α (Krippendorff, 2011)) for human evaluation of explanations for easy and hard examples along three evaluation axes.

Sentence	Options
I wanted to buy small tweezer to fit in my wristlet, but they still didn't fit. The _ were too small.	tweezer / wristlet
The documents contained in the files could not fit properly. The _ were too large.	documents / files
I measured the area in my kitchen, but the stove didn't fit because the _ was too small.	kitchen / stove

Table 2: Examples from the Winogrande dataset requiring the same commonsense knowledge that "If X is larger than Y, then X does not fit in Y".

Inter-annotator Agreement. Each explanation is evaluated by three annotators. We report inter-annotator agreement using Krippendorff's α (Krippendorff, 2011). Despite the subjective nature of our task, we observe moderate agreement scores among annotators, as reported in Table 1. Perhaps unsuprisingly, we find the agreement score for grammaticality to be the highest and that of generalizability to be the lowest. For supportiveness, we observe an α in the range of 0.4–0.5. Between easy and hard examples, the agreement scores for hard examples are lower, which also shows that these examples are harder for humans to agree on.

B Prompting Details

We avoid prompt tuning by largely following Wiegreffe et al. (2022) for prompt construction and choosing a layout that resembles Wiegreffe et al. (2022)'s CommonsenseQA prompt. Following Liu et al. (2022), we order the in-context samples in increasing order of similarity to the test sample such that the most similar sample is last in the context. All our generated explanations are obtained using the largest "text-davinci-002" model of GPT-3[4] with greedy decoding and maximum token limit of 50. While prior works (Zhao et al., 2021; Lu et al., 2022) have shown that in-context learning methods have high variance based on the hyperparameters chosen or the order of examples, we find that our generated explanations are fairly ro-

[4] https://beta.openai.com/docs/models/gpt-3

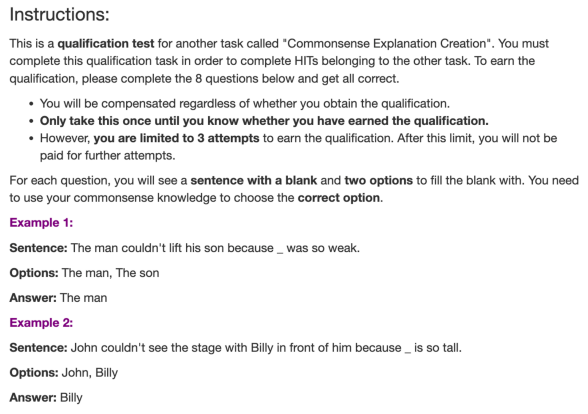


Figure 5: Instructions for the qualification test for writing and verifying explanations for the task of Winograd Schema Challenge.

bust to such variations due to their semi-structured form. We also note that finding the most optimal prompt is not the main focus of our work. Instead, we are interested in understanding the connection between explanation quality and sample hardness when other factors like hyperparameters, decoding strategy, etc are kept unaltered.

C Examples of Similar Examples Retrieved for In-Context Learning

We find that our simple method of using sentence embeddings to retrieve top-k similar examples for in-context learning works well in practice. In Table 3, we show some representative examples, demonstrating the presence of similar commonsense knowledge between the test sample and the top-2 similar samples.

D Analysis of GPT-3-generated Explanations for Hard Samples

In Table 4, we show more examples of bad explanations generated by GPT-3 for some of the hard examples. While the model is able to learn the semi-structured nature of the explanations, it often makes mistakes in identifying what X and Y are (first example), misses the core reasoning concept (second and third examples) or are non-contextual (last example), thereby either not properly supporting the label or completing refuting the label (fourth example). Consequently, the 'generalizability' aspect of these explanations also suffer.

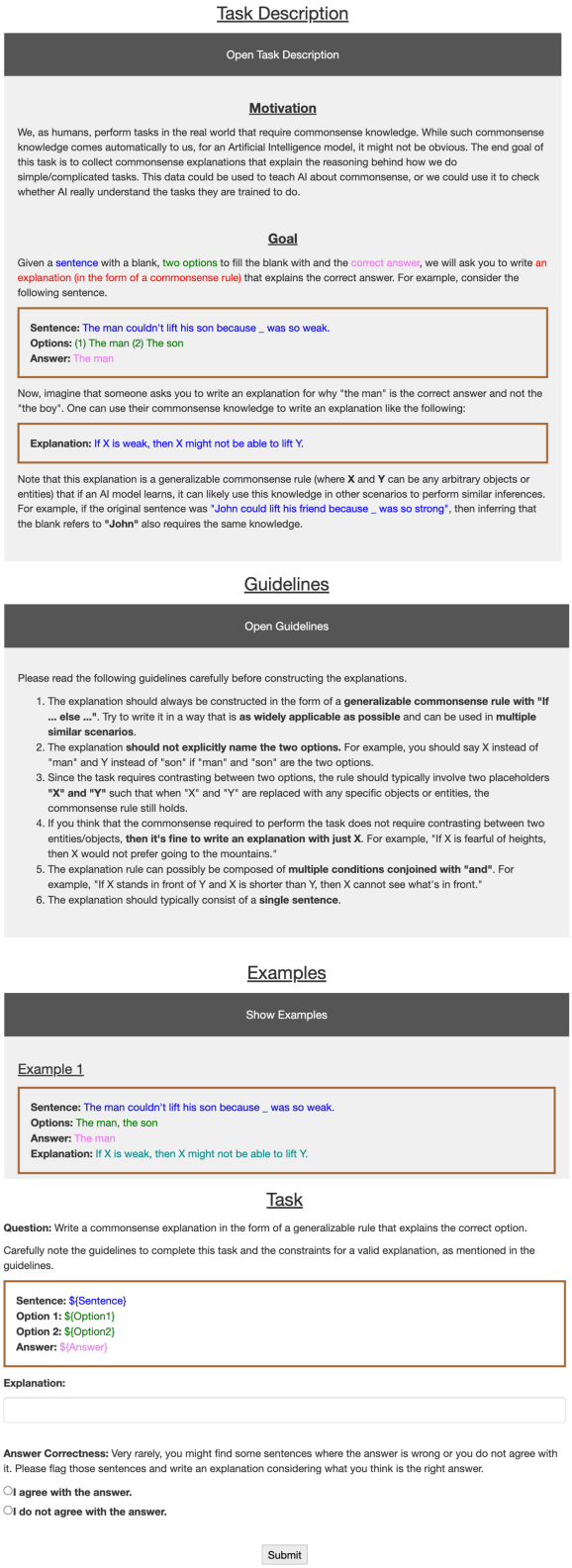


Figure 6: Explanation Creation Guidelines and Interface on Amazon Mechanical Turk. We ask crowdworkers to follow the guidelines when constructing explanations. We also perform in-browser checks to ensure that the options are not explicitly mentioned in the explanations.

Sample	1st Similar Example	2nd Similar Example
Katrina had the financial means to afford a new car while Monica did not, since _ had a high paying job.	Leslie was able to buy new paint for his house this weekend unlike Nelson, because _ was wealthy.	Kyle was not able to have a lavish lifestyle but Lawrence could because _ had lots of money.
Bill's new houseboat he purchased would not fit in his garage, the _ was too small.	I tried to set the plant in the pot, but it didn't work because the _ was too large.	The bottles supplied is not enough to collect the water. The _ is too much.
She had a cold and decided to ditch the vitamins and use medicines, because the _ were less effective.	My cousin preferred the treatments over the procedures because the _ were better for your health.	I removed beef from my diet and added pork, as the _ turned out to be unhealthy.

Table 3: Examples of similar samples retrieved for in-context learning. The similar examples require similar commonsense reasoning for inferring the correct answer.

Sentence	Options	Answer	GPT-3 Explanation	Reason
Matthew is trying to make William's wiener dog happy, because _ is left alone with it.	Matthew/William	Matthew	If X is left with Y then X is responsible for Y.	Unable to understand what X and Y are in the context.
The couple couldn't decide whether to get a new refrigerator or new laundry machine. They went with the former because their old _ was dilapidated.	refrigerator/laundry machine	refrigerator	If X is old and Y is new, you will go with X.	Doesn't explain the core concept that old things must be replaced first.
Jean preferred to use the scrub over the body soap to clean off dead skin because the _ was harsher.	scrub/body soap	scrub	If X is harsher than Y, then X is preferred.	Misses the important part of "preferred for cleaning" without which the explanation is incomplete and not generalizable.
We tried to get the bookcase through the door, but the _ was not narrow enough.	door/bookcase	bookcase	If X is not narrow enough, then Y can't get through.	Wrong explanation as it does not support the label.
Because Lindsey wanted to make a mess and Victoria did not, _ cracked her eggs and opened them over the bowl.	Lindsey/Victoria	Victoria	If X wanted to make a mess, and Y didn't, then Y would do the opposite of X.	Explanation not relevant to the context.

Table 4: A few representative examples of bad explanations generated by GPT-3 for hard examples.

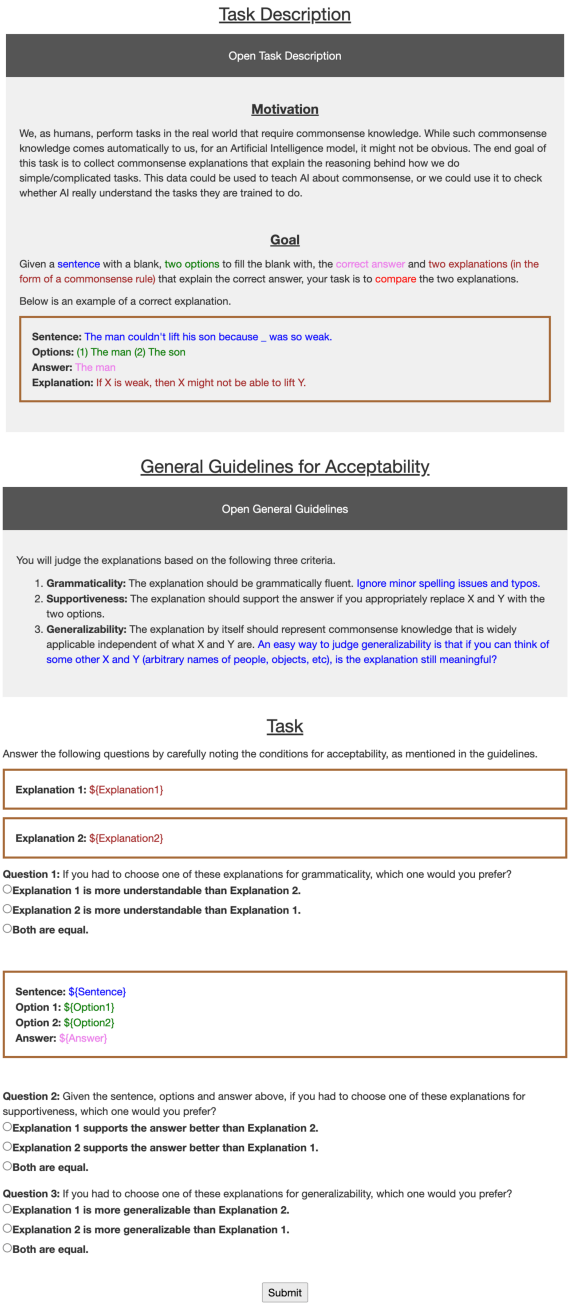


Figure 7: Explanation Verification Guidelines and Interface on Amazon Mechanical Turk.

Let's explain commonsense questions.
question: Katrina gave a more dynamic speech than Angela during the presentation since _ was creative. What does the "_" refer to?
Katrina, Angela
Katrina
why? If X is more creative than Y, X would make the more dynamic speech.
###
question: Lindsey decided to buy a nice piece of ink wash artwork from Katrina because _ was a great artist. What does the "_" refer to?
Lindsey, Katrina
Katrina
why? If X decided to buy an artwork from Y, then Y was a great artist. ###
question: Using Rose water as a tan remover worked for Amy but not Katrina because _ disliked the smell. What does the "_" refer to?
Amy, Katrina
Katrina
why? If X doesn't like the smell of rose water, then Y is more likely to use the rose water tan remover. ###
question: Kyle was not able to have a lavish lifestyle but Lawrence could because _ had lots of money. What does the "_" refer to?
Kyle, Lawrence
Lawrence
why? If X has more money than Y, then X would know more about lavish lifestyles.
###
question: Leslie was able to buy new paint for his house this weekend unlike Nelson, because _ was wealthy. What does the "_" refer to?
Leslie, Nelson
Leslie
why? If X is wealthy, then X will be able to buy new paint, and Y will not.
###
question: Katrina had the financial means to afford a new car while Monica did not, since _ had a high paying job.
Katrina, Monica
Katrina
why?

Figure 8: Example of a prompt with five in-context samples for Winogrande. Each sample consists of the question, two options, the correct answer and an explanation in the form of a generalized commonsense rule. The in-context samples are arranged in increasing order of similarity to the test sample. GPT-3 generates a free-text explanation for the current sample: *If X has a higher paying job than Y, then X would have more money to afford a new car.*