Probing Physical Reasoning with Counter-Commonsense Context

Kazushi Kondo,[1] **Saku Sugawara,**[2] **Akiko Aizawa**[2]
[1]The University of Tokyo, [2]National Institute of Informatics
kkkazu@g.ecc.u-tokyo.ac.jp, {saku,aizawa}@nii.ac.jp

Abstract

In this study, we create a CConS (Counter-commonsense Contextual Size comparison) dataset to investigate how physical commonsense affects the contextualized size comparison task; the proposed dataset consists of both contexts that fit physical commonsense and those that do not. This dataset tests the ability of language models to predict the size relationship between objects under various contexts generated from our curated noun list and templates. We measure the ability of several masked language models and generative models. The results show that while large language models can use prepositions such as "in" and "into" in the provided context to infer size relationships, they fail to use verbs and thus make incorrect judgments led by their prior physical commonsense.

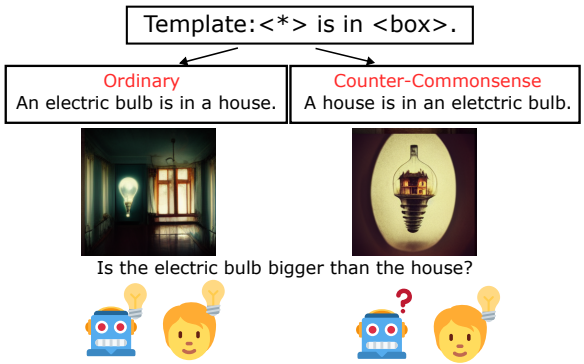

Figure 1: Examples of contexts that do or do not accord with ordinary commonsense. Humans can imagine the situation and make correct inferences, but language models are drawn to commonsense and make incorrect judgments. The example images are generated by Midjourney (https://midjourney.com).

1 Introduction

Humans possess physical commonsense regarding the behavior of everyday objects. Physical commonsense knowledge is relevant to their physical properties, affordances, and how they can be manipulated (Bisk et al., 2020). While a significant amount of physical commonsense can be expressed in language (Forbes and Choi, 2017; Bisk et al., 2020), direct sentences describing facts such as "people are smaller than houses" rarely appear because of reporting bias (Gordon and Van Durme, 2013; Ilievski et al., 2021). Recent language models have succeeded in tasks that do not require contextual reasoning, such as size comparison and prediction of event frequency (Talmor et al., 2020).

However, what about inferences that are context-dependent? Whether a language model can make correct inferences in various contexts is important because physical reasoning is highly context-dependent (Ogborn, 2011). Several studies on contextual physical reasoning (Forbes et al., 2019; Bisk et al., 2020; Aroca-Ouellette et al., 2021; Zellers et al., 2021) have been conducted to produce datasets that assess the ability to recognize physical situations described in writing. Without context, however, these datasets may be answered by commonsense.

Humans also can reason in ways that differ from simply using commonsense. For instance, if the context "there is a house inside a light bulb." is provided, humans can still imagine the situation and reason that the bulb must be larger than the house. In other words, commonsense is just a sweeping generalization, and reasoning about context must be independent of commonsense. This reasoning with defeasibility, which reflects the ability to reason logically without relying only on commonsense, seems to have been overlooked in the study of language models compared to the acquisition of commonsense. Previous investigations of contextual physical reasoning (Aroca-Ouellette et al., 2021; Yu et al., 2022) failed to distinguish physical reasoning from the simple use of physical commonsense. To appropriately measure physical reasoning ability, we must use contexts that go

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 603–612

against commonsense to rule out the possibility that the model is overconfident in physical commonsense.

In this study, we investigate the behavior of the language model concerning physical commonsense given the context of a situation that contradicts commonsense. We choose the size comparison task despite various possible domains of physical commonsense (Ilievski et al., 2021). The task is one of the easiest physical commonsense reasoning tasks for language models (Forbes and Choi, 2017; Goel et al., 2019), and it is also easy to add a context to change the relationship between sizes. For example, in this study, the context is a sentence that implies a size relationship, such as "<obj1> contains <obj2>."

For this purpose, we created a new dataset, CConS (Counter-commonsense Contextual Size comparison)[1]. This dataset contains 1,112 sentences generated from 139 templates and tests the ability of language models to infer the size relationship between objects using a cloze-style prompt. Figure 1 shows the size comparison examples with or without contexts that (do not) agree with ordinary commonsense. Our experiments using recent language models show that GPT-3(text-davinci-003) (Brown et al., 2020) correctly reasons in context when it is consistent with commonsense, yielding 85% accuracy. In contrast, even GPT-3 can only show poor performance (41 % accuracy) for examples that contradict commonsense. This suggests that the models may not effectively distinguish between physical commonsense and inferences based on contexts, leading to incorrect predictions. Nevertheless, when prepositions hint at the relationships, the accuracy rate exceeded 55%, even for counter-commonsense examples. In summary, our counter-commonsense examples reveal the difference in influence between prepositions and verbs in contextualized physical reasoning.

The contributions of this study are as follows:

1. We create a dataset that assesses size comparison ability more precisely by contrasting examples that conform to physical commonsense with ones that do not.

2. We show that physical commonsense prevents measuring the language models' ability of contextual physical reasoning.

3. We demonstrate that even large models perform poorly when making inferences that violate physical commonsense. Specifically, they struggle to infer size relations implied by verbs and can infer only when prepositions indicate.

[1] https://github.com/cfkazu/Counter-Commonsense-Context

2 Related Works

Size Comparison Task The size comparison task, which previous studies (Yang et al., 2018; Goel et al., 2019) investigated since the earlier linguistic representations, such as GloVe (Pennington et al., 2014) or ELMo (Peters et al., 2018), is one of the easiest physical common-sense inference tasks for language models (Forbes and Choi, 2017; Goel et al., 2019). While there are many prior studies (Elazar et al., 2019; Zhang et al., 2020) on this topic, VerbPhysics (Forbes and Choi, 2017) is the most similar to this study in that it focuses on the relationship between sizes and verbs. There are also some other approaches, such as methods that extract external knowledge (Elazar et al., 2019), filling-masks (Talmor et al., 2020), or generate images (Liu et al., 2022). These results suggest that the commonsense of comparing object size is encoded in recent language models. However, these studies do not consider the context that might influence the results of size comparisons.

Defeasible Reasoning According to Koons (2022), defeasible reasoning is an argument that is rationally persuasive but not completely valid as a deduction. This defeasible reasoning is similar to the subject of this study in that it involves the recognition that commonsense and assumptions in a given context are not entirely correct propositions. Therefore, this study can be seen as an investigation into whether a language model can capture commonsense as defeasible reasoning. The creation of a dataset dealing with defeasible reasoning has been discussed by Rudinger et al. (2020) and Allaway et al. (2022). Our study is similar to Allaway et al. (2022) in that it generates sentences that violate the context by fitting words to a template. However, this study differs in that we also generate examples contrary to commonsense for measuring the actual performance of the language model as well as the differences from the ordinary case.

3 Dataset Creation

In this study, we create 139 templates and automatically generate 1,112 examples. Table 1 lists

Template	Generated: Ordinary Examples	Generated: Counter-Commonsense Examples
He found <portable> in <box>.	He found a key in a key box.	He found a monitor in a key box.
<box> contains <portable>.	A key box contains a key.	A key box contains a monitor.
<*> fills <box>.	A marble fills a bin.	A refrigerator fills a bin.
<*> is covered by <flat>.	A pen is covered by a newspaper.	A desk is covered by a handkerchief.

Table 1: Examples of the templates. <tag> constrains possible nouns to be filled. For example, <box> means that the noun entering there must have the attribute "box," that is, it must be able to hold things. <*> indicates that any words in the noun list (only material nouns) can be inserted.

examples of these templates.

Designing Template We focus on the comprehensiveness of verb phrases while designing templates to ensure that the choice of verbs is not arbitrary. Therefore, we extract 139 verb phrases that indicate size relationships from the Oxford 5000 dictionary [2] and manually assemble simple sentences. For example, the statement "<obj1> beats <obj2>" is not included in this template because this statement is not informative enough to determine a size relation.

Moreover, in comparing sizes, we also notice not only verbs but the usage of prepositions such as "in" or "into" may provide clear clues about the size relationships. Therefore, we select templates that contain only examples with these prepositions and distinguish them as easy templates from those that do not as hard templates. In subsequent experiments, we also investigate the effect of this difference on the behavior of the language model.

Restriction on Noun If nouns are arbitrarily inserted, the resulting sentences may be nonsensical or impossible for a human to imagine. For example, we choose not to include the sentence "the stone threw the dog" because it is beyond imagination.

We place restrictions on the nouns used in the sentence templates by defining tags to avoid this nonsense. A single placeholder can have constraints (multiple tags). There are 18 types of tags, including "have_hands," "box," and "portable." Tags are manually determined to abstract the properties of verb phrases. We also use the Oxford 5000 dictionary to obtain a list of nouns referring to physical objects. One of the nouns that satisfy all constraints is randomly selected from a list of 195 nouns and inserted.

Generating Sentences The template tags are replaced with the corresponding nouns to generate the context, and the questions asking for size comparisons are combined. For example, the contextualized question text provided to the masked language models is as follows:

"`«context» In this situation, the size of <obj1> is probably much [MASK] than the size of <obj2>.`"

Contexts and questions are used to generate input for each of the masked language models and generative models. We classify generated sentences to the Ordinary or Counter-Commonsense (CCommon) subset based on whether the size relationship between objects indicated by the template accords commonsense.

4 Experiment

Task Definition We measure the ability of masked language models and generative models to recognize size relationships by providing sentences for each architecture. These sentences are generated from templates (Section 3). We also see how the language model's behavior changes when context sentences follow or do not follow a general common-size relationship.

Comparison Aspects We investigate how language models create physical reasoning without being biased by their prior physical commonsense.

1. How do the physical reasoning results of the language model change when contexts are consistent or inconsistent with commonsense?

2. How does the performance of a language model change when comparing an easy dataset that contains certain prepositions that hint at size relationships with a hard dataset that does not?

Model Settings In this study, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020) are used to assess the performance of the masked language models. We also investigate how the size of the model affects

[2] https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000

physical reasoning. We choose T0 (Sanh et al., 2022) and GPT-3(text-davinci-003) to evaluate the performance of the generative model.

According to Talmor et al. (2020), RoBERTa-Large outperforms BERTs and RoBERTa-Base in a no-context size comparison task. Proceeding from this analysis we attempt to detect whether commonsense influences physical reasoning by giving examples contrary to commonsense as context.

Tasks Format Details The tasks are performed by inputting sentences according to the format defined for each of the models, as follows.

Format for Masked Language Models
WithContext: «context» In this situation, the size of <obj1> is probably much [MASK] than the size of <obj2>.
WithoutContext: The size of <obj1> is probably much [MASK] than the size of <obj2>.

The candidates for [MASK] are "larger," "bigger," "smaller," and "shorter." If the sum of the probabilities of the first two options exceeds 0.5, language models predict that obj1 is larger than obj2. Therefore, the language model always makes binary decisions.

Format for Generative Models
WithContext: «context» Which is bigger in this situation, <obj1> or <obj2>?
WithoutContext: Which is bigger in general, <obj1> or <obj2>?
«context» is a sentence generated from templates.

Human Evaluation We ask crowdworkers to perform the same size comparison task to measure the accuracy of humans in this task. Thus, we can test the validity of the automatically generated questions. The crowdworkers are given the same context and make a choice that is larger. (See Appendix B for details.) Five crowdworkers are assigned to each question. We use some intuitive examples, such as "<obj1> contains <obj2>," which are provided for qualification, and exclude those who get such examples wrong or choose the same answer for all examples.

Model	Ordinary	CCommon	NoCon
BERT-B	0.483	0.515	0.495
BERT-L	0.500	0.521	0.494
RoBERTa-B	0.554	0.443	0.507
RoBERTa-L	0.692	0.413	0.639
ALBERT-B	0.500	0.521	0.494
ALBERT-XXL	0.720	0.346	0.701
T0++	0.682	**0.530**	0.589
T0	0.684	0.443	0.574
GPT-3	**0.856**	0.415	**0.764**
Human	0.814	0.798	0.791

Table 2: The inference results of the language model for data sets where the context follows and does not follow commonsense and context is removed.

5 Result and Analysis

Tables 2 and 3 exhibit the performance of the language model on our datasets. GPT-3 outperforms other models in Ordinary and NoCon setups. RoBERTa-Large and ALBERT-XXLarge show better reasoning ability than the other masked language models in the Ordinary dataset. However, for the CCommon dataset, the performance of the pretrained language model decreases, particularly in ALBERT-XXLarge. This result suggests that commonsense built into the model hinders its ability to make accurate judgments. Other models struggle to capture size relationships. These results without context (NoCon) are generally consistent with the findings of a previous investigation of the no-context size comparison task conducted by Talmor et al. (2020).

In some CCommon examples, BERT performs better than RoBERTa. This may be because BERT is less equipped with commonsense, allowing it to make simpler judgments without being influenced.

Impact of Prepositions Prepositions did not significantly impact the prediction for the masked language models in the Ordinary dataset. However, there is a significant difference in the correct response rates in the CCommon dataset. RoBERTa-Large performs well in easy data, regardless of whether the context defies commonsense. This result indicates that RoBERTa-Large recognizes the connection between the prepositions and size relationships. The ALBERT-XXLarge model does not perform well for the CCommon dataset, even if the setting is easy; therefore, we consider that it merely answers according to commonsense rather than making inferences. In short, context is not useful for ALBERT when the prepositions do not

Model	Ordinary		CCommon	
	Easy	Hard	Easy	Hard
BERT-B	0.506	0.471	0.460	0.557
BERT-L	0.527	0.479	0.480	0.553
RoBERTa-B	0.557	0.550	0.473	0.419
RoBERTa-L	0.711	0.671	0.467	0.369
ALBERT-B	0.527	0.479	0.480	**0.553**
ALBERT-XXL	0.744	0.693	0.353	0.346
T0++	0.762	0.607	**0.593**	0.480
T0	0.726	0.638	0.473	0.424
GPT-3	**0.940**	**0.788**	0.567	0.296
Human	0.835	0.796	0.829	0.769

Table 3: Comparison results of reasoning ability of the language model for datasets that follow the commonsense and those that do not. Sentences with prepositions "in" or "into" are included in the easy dataset and otherwise in the hard.

provide direct hints.

GPT-3 uses prepositions more effectively than other models and performs better on the Easy dataset, while the model struggles to answer the CCommon dataset in the hard setting. This result means GPT-3 learns commonsense well but cannot make physical logical inferences.

6 Conclusion

We develop a method providing a counter-commonsense context to measure physical reasoning ability. Our proposed contextualized physical commonsense inference dataset reveals that current language models can partially predict size relations but do not perform as well as humans in contexts that contradict commonsense. These judgments are possible to a limited extent in the presence of certain prepositions such as "in" and "into." While we focused on size comparison tasks in this study, the importance of context in physical reasoning is not limited to this task. Increasing the size and scope of the datasets for contextual commonsense inference is necessary to build language models that more closely resemble humans and differentiate between general commonsense and the facts at hand.

Limitations

The main limitation of our method is that it requires human effort to increase the variety of templates, which makes it difficult to create large datasets. Using templates to generate data reduces the time required to create data manually, but the need for human labor remains an obstacle. To resolve this, the templates themselves need to be generated automatically, although the tags that constrain the nouns also need to be generated automatically, which is a difficult problem.

Acknowledgment

We would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported by JST PRESTO Grant Number JPMJPR20C4 and JSPS KAKENHI Grant Number 21H03502.

References

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2022. Penguins Don't Fly: Reasoning about Generics through Instantiations and Exceptions. ArXiv:2205.11658 [cs].

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical Reasoning about Objects through Space and Time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439. Number: 05.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How Large Are Lions? Inducing Distributions over Quantitative Attributes. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Maxwell Forbes and Yejin Choi. 2017. Verb Physics: Relative Physical Knowledge of Actions and Objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do Neural Language Representations Learn Physical Commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*, page 7.

Pranav Goel, Shi Feng, and Jordan Boyd-Graber. 2019. How Pre-trained Word Representations Capture Commonsense Physical Comparisons. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 130–135, Hong Kong, China. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, AKBC '13, pages 25–30, New York, NY, USA. Association for Computing Machinery.

Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. 2021. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347.

Robert Koons. 2022. Defeasible Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2022 edition. Metaphysics Research Lab, Stanford University.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

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:1907.11692 [cs].

Jon Ogborn. 2011. Science and commonsense. *Revista Brasileira de Pesquisa em Educação em Ciências*, 6.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking Like a Skeptic: Defeasible Inference in Natural Language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yiben Yang, Larry Birnbaum, Ji-Ping Wang, and Doug Downey. 2018. Extracting Commonsense Properties from Embeddings with Limited Human Guidance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 644–649, Melbourne, Australia. Association for Computational Linguistics.

Model	Model-FullName
BERT-B	bert-base-uncased
BERT-L	bert-large-uncased
RoBERTa-B	roberta-base
RoBERTa-L	roberta-large
ALBERT-B	albert-base-v2
ALBERT-XXL	albert-xxlarge-v2
T0++	bigscience/T0pp
T0	bigscience/T0

Table 4: Paths for using the Hugging Face models used in this study. These models were used without modification.

Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. PACS: A Dataset for Physical Audiovisual CommonSense Reasoning. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 292–309, Cham. Springer Nature Switzerland.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World. 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 2040–2050, Online. Association for Computational Linguistics.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do Language Embeddings capture Scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 292–299, Online. Association for Computational Linguistics.

A Experiment Details

We used a language model published on hugging face Transformers (Wolf et al., 2020) except GPT-3 under MIT (RoBERTa) or Apache-2.0 (BERT, ALBERT, T0, T0++) license. For GPT-3, the OpenAI API (text-davinci-003[3]) is used. All of these models are designed to solve downstream natural language tasks. Table 4 lists the paths for accessing the models via hugging face.

We use a GPU Tesla V100-PCIE-32GB. The total computation time was 1 hour for the masked language models and 2 hours for the generative models.

B Human Evaluation Details

We evaluate human accuracy in a size comparison task using Amazon Mechanical Turk. We provide the following instructions and let the crowdworkers choose their answers: We calculate the reward as $15 per hour. Figure 2 shows the instructions for the contextualized size comparison task. The choices are virtually two-option questions, except "I can't imagine the situation," etc. Figure 3 shows the instructions for the non-contextualized size comparison task. The choices are "obj1","obj2," and "N/A (cannot determine)."

No personal information is obtained. Crowdworkers live in the United Kingdom, the United States, and Canada. By accepting Amazon Mechanical Turk's participation agreement [4], crowdworkers consent to the collection and use of non-personal data for research purposes.

[3] https://platform.openai.com/docs/models/gpt-3-5

[4] https://www.mturk.com/participation-agreement

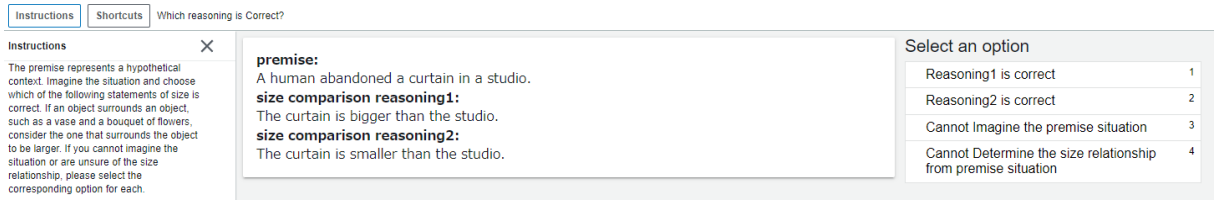

Figure 2: An instruction and options given to Amazon Mechanical Turk crowdworkers for contextualized size comparison task. Annotators are asked to read a context and determine which object is larger in the situation.

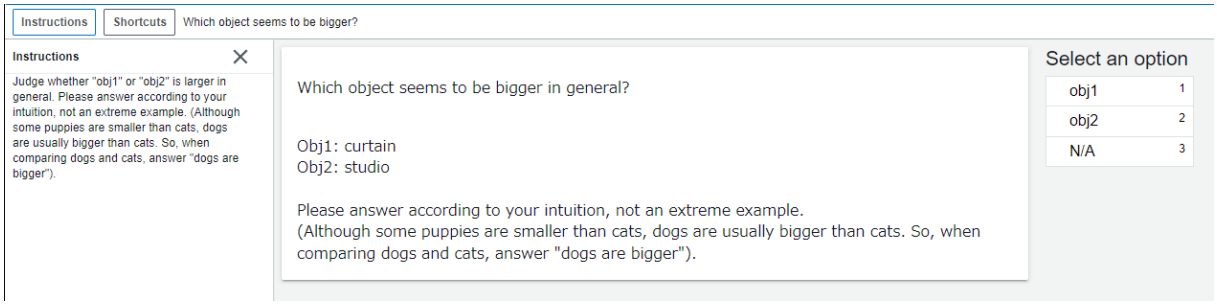

Figure 3: An instruction and options given to Amazon Mechanical Turk crowdworkers for contextualized size comparison task. Annotators are asked to judge which object is generally larger.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
The paper is about the simple task of comparing the sizes of two objects, and we believe there is no such risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1 and Abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3,4, Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Section 1,4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3, Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

4, Appendix A, B

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, Appendix A

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

4, Appendix C

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix C

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix C

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The task is a simple task of comparing the sizes of two objects and obviously does not pose any problems with user safety, health, or personal information.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C