Language Models are Unreliable Spatial Reasoners

Joseph Cappadona, Yimin Chen, Jiaxuan Lan, Cara Leong
{jcappadona, yc3294, jl13072, caraleong}@nyu.edu

Abstract

We construct a natural language inference (NLI) benchmark[1] to test the spatial reasoning capabilities of language models (LMs). Many commonsense benchmarks test spatial reasoning abilities simultaneously with other reasoning abilities, and a few benchmarks test certain specific spatial reasoning abilities rigorously, but no prior works have sought to evaluate LMs on a comprehensive suite of spatial reasoning tasks. To do this, we use templates to procedurally generate a large quantity of NLI questions relating to four categories of spatial reasoning – motion, distance, orientation, containment – and evaluate state-of-the-art NLI and QA models on these questions. We find that all families of LMs perform poorly on many of the spatial reasoning categories, demonstrating clear gaps in their understanding of space and its relationship to people and objects.

1 Introduction

Benchmarking language models' (LMs) abilities to perform commonsense reasoning is an essential step toward building models that can be deployed in the real world. LMs have shown promise on many commonsense reasoning benchmarks (Zellers et al., 2018; Singh et al., 2021; Zhang et al., 2018; Talmor et al., 2019). However, most of these benchmarks aim for broad coverage over several reasoning domains, often testing multiple types of reasoning simultaneously. Doing so makes it difficult to provide comprehensive coverage of a single domain, which in turn reduces a benchmark's ability to diagnose where the weaknesses of models lie.

To this end, many physical commonsense reasoning benchmarks have emerged, seeking to narrow in on testing the *physical* reasoning abilities of LMs. For example, PIQA (Bisk et al., 2019) is a multiple-choice task that requires LMs to reason about the physical consequences of actions on an environment (see Appendix B.1). PROST (Aroca-Ouellette et al., 2021) is a cloze-style multiple-choice benchmark that requires reasoning about concepts like stackability, breakability, and rollability of objects (see Appendix B.2).

Within the domain of physical commonsense reasoning, spatial commonsense reasoning has also received some attention. For example, Weston et al. (2015) introduce BABI, a set of 20 toy tasks which aim to test a model's ability to reason about the properties and relationships of agents and objects in a variety of everyday contexts. Each task tests a unique reasoning capability and is designed so that most adult humans could potentially score 100%. Several of the tasks proposed in BABI relate to spatial reasoning, for example, Task 17 tests positional reasoning, Task 18 tests size reasoning, and Task 19 tests path finding (see Appendix B.3). The dataset is generated using a simulation of agents and objects in an environment where agents randomly interact with the environment and the corresponding actions and states are collected and used to construct task examples.

SPARTQA (Mirzaee et al., 2021) builds off BABI's positional reasoning task (Task 17) by building a spatial reasoning benchmark that rigorously tests concepts like containment and orientation (see Appendix B.4). The benchmark is constructed by applying hand-designed context free grammars and context-sensitive rules to images from the Natural Language for Visual Reasoning dataset (Suhr et al., 2017).

Our test suite covers a superset of the spatial relations covered in BABI and SPARTQA, aiming to provide a more comprehensive benchmark for spatial reasoning. We diverge from the literature in that we are not interested in seeing how LMs do when fine-tuned on our benchmark. Instead, we test pre-trained LMs out-of-the-box, aiming to see how well they generalize from their training data.

[1]Code available for download at https://github.com/josephcappadona/spatialQA

	Input	Expected	#
Motion	P: {Tom} is {swimming, running}. H: {Tom} is {stationary, not in motion}.	contradiction	72
Orientation	P: {The park} is {north} of {the theater}. H: {The theater} is {north, east, west} of {the park}.	neutral or contradiction	72
Distance	P: {The ball} is {touching} {the water}. H: {The water} is {touching, in contact with} {the ball}.	entailment	120
Containment	P: {A bag} {contains} {apples}. H: {Apples} are {inside, within} {a bag}.	entailment	18
Metaphor	P: {John} got out of {doing chores}. H: {John} was {inside, contained by} {doing chores}.	neutral or contradiction	12

Table 1: Example premises (P) and hypotheses (H) for each type of spatial relation. Curly braces indicate lexical items that can be substituted programmatically using templates, with some non-exhaustive example words and phrases. The final column is the number of examples generated for the corresponding template.

The need for this type of generalization testing is demonstrated well by Ribeiro et al.'s (2020) CHECKLIST, an application for efficiently producing suites of tests for comprehensive behavioral testing of LMs. The approach draws inspiration from software engineering research, which emphasizes rigorous black-box testing paradigms to validate software systems and check for critical failures.

A variety of techniques have been used in the literature to generate benchmark data, like crowdsourcing (Singh et al., 2021), simulations (Weston et al., 2015), and grammars (Mirzaee et al., 2021). However, some recent publications like PROST opt for a CHECKLIST-like approach and use templates where lexical items are substituted in. This is the approach we employ due to the ability to generate a large amount of test data very quickly.

To summarize our contributions, we expand upon the existing work on spatial reasoning benchmarks by introducing a natural language inference (NLI) benchmark which tests four categories of spatial reasoning: motion, orientation, distance, and containment. An additional category, metaphor, is used to test whether models can distinguish between literal and figurative uses of spatial reasoning concepts. We then test state-of-the-art QA and NLI models out-of-the-box (without fine-tuning on our benchmark) and report the results.

2 Test Suite

Our test suite consists of 20,480 premise-hypothesis pairs created from 92 templates[2] that test a model's ability to make inferences about spatial relations. Each example contains a single-sentence premise containing a spatial cue (e.g. in, above, near), and a hypothesis that tests the model's ability to identify and reason about that spatial relation. Drawing heavily from psycholinguistic research of human spatial typology (Levinson et al., 2003) and the tasks and relations covered by BABI and SPARTQA, we identify four distinct categories of spatial reasoning worth testing: motion, orientation, distance, and containment. Examples of premise-hypothesis pairs from these categories are given in Table 1.

Test examples are generated through the use of templates that test one particular spatial relation type. For instance, one template to test the symmetric property of proximity is given below:

Premise: {A} is {near} to {B}.
Hypothesis: {B} is {near} to {A}.

Each template is then populated with relevant lexical items. Varying the lexical inputs of each template tests the robustness of a model's spatial inferences with regard to irrelevant changes in the input — all else being equal, a model should predict the same labels for all premise-hypothesis pairs in the template. In some cases, we allow multiple entailment relationships as shown in Table 1.

For each spatial category, we also include metaphorical premise-hypothesis pairs that contain the same syntactic triggers as physical spatial relations. Metaphorical template structures were adapted from MetaNet (Dodge et al., 2015)[3]. We include these metaphorical templates to test whether a model can identify when spatial relationship cues

[2]See Table 3 in Appendix A for a numerical breakdown of the number of tests and examples for each category

do not actually trigger a spatial relationship. For instance, both "The school is **near** the theater" and "John is **near** death" contain the proximity cue "near". However, the first sentence entails "The school is in close physical proximity to the theater", while the second sentence does not entail "John is in close physical proximity to death".

3 Testing

We test our spatial reasoning benchmark on models from multiple families that are at or near state-of-the-art[4] on related natural language inference tasks like MultiNLI (Multi-Genre NLI) (Williams et al., 2018), QNLI (Question-answering NLI) (Wang et al., 2018), WNLI (Winograd NLI) (Wang et al., 2018), and ReCOrD (Reading Comprehension with Commonsense Reasoning Dataset) (Zhang et al., 2018).

In particular, we test the following models: GPT-3[5], UnifiedQAv2, T5, DeBERTa fine-tuned on MultiNLI, RoBERTa fine-tuned on MultiNLI, ALBERTv2 fine-tuned on MultiNLI, and XLNet fine-tuned on MultiNLI. GPT-3 was accessed using the OpenAI API[6] and all other models were downloaded from HuggingFace (Wolf et al., 2020). Model sizes are reported in Table 4 of Appendix A.

4 Results

Results from testing each model on our test suite are shown in Table 2. Acc_{wopc} represents the average accuracy for each model across test templates *without partial credit*. In this paradigm, a model must correctly answer all entailments within a particular test template to receive any credit for that template. Acc_{pc} represents the average accuracy *with partial credit*, where tests are no longer all-or-nothing. For example, a model that correctly predicts the labels of 99% of a template's examples obtains an Acc_{pc} score of 99% and an Acc_{wopc} score of 0% for that template.

Despite our benchmark's relative simplicity, all of the models fail on parts of our test suite with no model passing all tests in any category.

[3] https://metaphor.icsi.berkeley.edu/pub/en/index.php/

[4] As reported on https://paperswithcode.com/ leaderboards as of April 2022

[5] Since GPT-3, UnifiedQAv2, and T5 are not NLI models by design, we slightly modify the task to fit their respective input formats. See Appendix C for details.

[6] https://openai.com/api/

Model	Acc_{wopc}	Acc_{pc} (st.dev)
GPT-3	26.5	**61.1 (6.8)**
UnifiedQAv2	**32.7**	33.3 (15.7)
T5	29.5	51.9 (23.3)
DeBERTa	30.2	56.5 (21.0)
RoBERTa	30.9	54.7 (18.7)
ALBERT	29.7	43.8 (22.2)
XLNet	30.5	50.9 (17.6)

Table 2: Mean model accuracy with/without partial credit, averaged over all reasoning categories, for the largest model in each family.

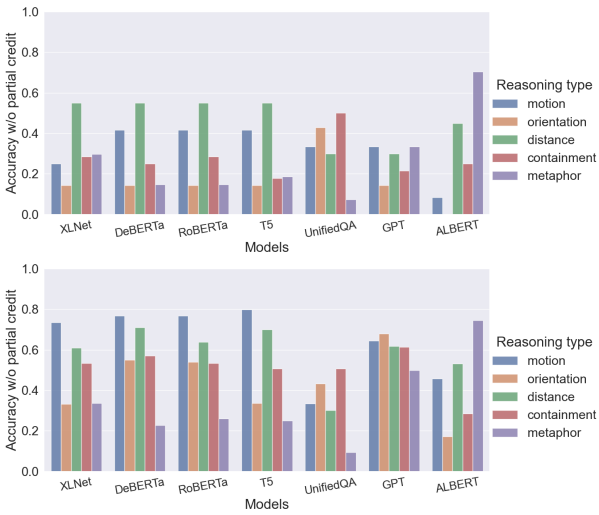

Figure 1: Mean model accuracy with (bottom) and without (top) partial credit, for each reasoning category, for the largest model in each model family. Precise numerical results for both graphs are present in Tables 5 and 6 of Appendix A.

Table 2 shows that there is not a clear relationship between Acc_{wopc} and Acc_{pc}, since GPT-3 had the lowest Acc_{wopc} score and the highest Acc_{pc} score and UnifiedQAv2 had the highest Acc_{wopc} score and the lowest Acc_{pc} score. Additionally, as evidenced by the generally large standard deviations, model performance varied widely across categories for most models, with GPT-3's performance across categories being the most stable by a fair margin with a standard deviation of 6.8.

4.1 Analysis

Figure 1 breaks down model performance by reasoning category. One trend across models is that 5 out of 7 models were able to make inferences related to motion and distance better than they were

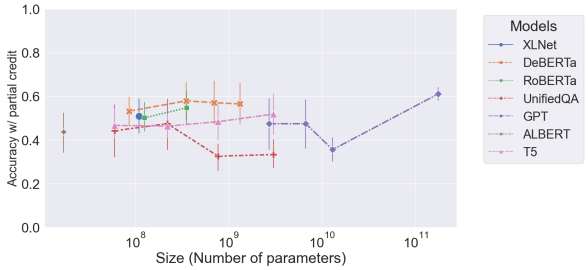

Figure 2: Mean model accuracy *with* partial credit, averaged over all reasoning categories, plotted with respect to the size of each model

able to make inferences related to containment and orientation. Why this is the case is unclear, but one hypothesis is that distance and motion relations may be more frequently represented in the training corpora for these models. We leave exploration of this and related questions for future work.

Considering each spatial relation in isolation, most models performed worse on metaphorical templates in that category than on literal templates. One exception to this rule was ALBERT, which was the only model whose best category was metaphor. However, ALBERT's improved performance on metaphorical templates comes with concomitant poor performance on literal templates. We hypothesize that models only generalize one entailment pattern for a given spatial relation, and since two patterns of reasoning are required to deal with literal and metaphorical meanings, better generalization on literal readings directly results in worse generalization on metaphorical readings, and vice versa.

Additionally, model performance was often highly sensitive to lexical changes. For instance, GPT-3, UnifiedQAv2, DeBERTa and RoBERTa were able to correctly identify the correct entailment for between 40% and 100% of the examples generated from the following template: "A is in B and B is in C" $\rightarrow$ "A is in C". However, none of these models were able to identify the correct entailment for "A fits in B and B fits in C" $\nrightarrow$ "A is in C" in any of the 792 examples for that template. More examples of heavily failed templates are visible in Table 9 of Appendix A.

4.2 Model size

Figure 2 shows that there is not a clear relationship between model size and overall performance on our test suite. While some models like T5 and RoBERTa appear to perform slightly better as model size increases, others like UnifiedQAv2 appear to perform worse. The relationship between model size and performance is broken down by category in Figures 7 and 8 in Appendix A. There we can see that for the metaphor category in particular, there is a clear decrease in performance for all model families as model size increases. However, the lack of a clear relationship between model size and performance in general suggests that spatial reasoning abilities might not emerge simply from increasing model size; rather, it is possible that either the training data or the language modeling task (or both) may not provide a suitable signal for such reasoning to be learned.

5 Future Work

Future work should involve:

- testing models fine-tuned on different datasets to see how they affect generalization on our task
- taking the test suite beyond synthetic data, e.g., by mining examples from real-world corpora, or by crowdsourcing templates, lexical items, and/or labels
- verifying the performance of GPT and UnifiedQAv2 when the task is more explicitly designed as a QA task rather than an NLI task
- analyzing the frequency of spatial relations in common training corpora and attempting to relate it to relative category performance

6 Conclusion

We constructed an NLI benchmark to evaluate the spatial reasoning abilities of language models. Our test suite was constructed using templates where sets of lexical items are substituted in to produce a large number of examples from relatively few templates. We tested four categories of spatial reasoning – motion, orientation, distance, and containment – which were chosen to provide a broad coverage of the types of spatial relations used in English. We also test inference over metaphors which use similar types of spatial relations. We found that even the largest models which perform state-of-the-art on a variety of NLI and QA benchmarks did poorly on all categories, with no model passing all tests in any category, demonstrating that there are clear gaps in the commonsense reasoning abilities of language models which are unlikely to be fixed simply by scaling up and training on more data.

7 Ethical Considerations

Given that our test suite is designed to evaluate real-world models, it is important that the templates that we have written accurately reflect the language and spatial reasoning used by a wide variety of English speakers. However, our test suite has only been looked over by the four authors, and so we cannot make any broad claims about the coverage of our test cases. It is possible that we have neglected certain types of spatial relations when constructing our templates. Additionally, the labels for our templates were also written by the authors, and there might be disagreements over these labels when outsiders are consulted. Therefore, before the results of our benchmark can be fully trusted, a more thorough vetting and analysis of our test suite must be done, ideally through a crowdsourcing platform where workers can validate the labels. Until such work is done, we caution researchers using our task for training or evaluation.

8 Collaboration Statement

Joe and Cara were primarily responsible for constructing the suite of tests. Yimin and Jiaxuan were primarily responsible for testing the models. All group members helped with analyzing the results and writing.

References

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

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.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system.

Stephen Levinson, Sérgio Meira, The Language, and Cognition Group. 2003. 'natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3):485–516.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the*

2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, abs/1810.12885.

A Appendix A: Additional Figures

Category	# templates	# examples
Motion	12	552
Orientation	7	2192
Distance	20	2260
Containment	28	14550
Metaphor	27	926
	94	20480

Table 3: A numerical breakdown of the number of templates and total number of examples generated for each category

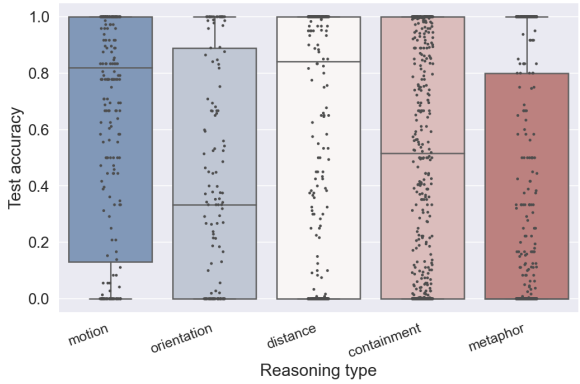

Figure 3: Box plot of Acc_{pc} for all tests across all models, broken down by category

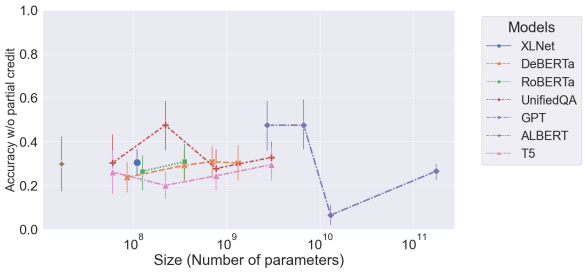

Figure 4: Mean model accuracy *without* partial credit, averaged over all reasoning categories, plotted with respect to the size of each model

Model		# parameters
GPT-3	B	2.7B
	M	6.7B
	L	13B
	XL	175B
UnifiedQAv2	S	60M
	B	220M
	L	770M
	XL	3B
T5	S	60M
	B	220M
	L	770M
	XL	3B
DeBERTa	B	86M
	L	350M
	XL	700M
	XXL	1.32B
RoBERTa	B	125M
	L	355M
ALBERT	L	17M
XLNet	B	110M

Table 4: The number of parameters for each model

Model	Motion	Orient.	Dist.	Contain.	Metaph.	Avg
GPT-3	33.3	14.3	30.0	21.4	33.3	26.5
UnifiedQAv2	33.3	42.9	30.0	50.0	7.4	**32.7**
T5	41.7	14.3	55.0	17.9	18.5	29.5
DeBERTa	41.7	14.3	55.0	25.0	14.8	30.2
RoBERTa	41.7	14.3	55.0	28.6	14.8	30.9
ALBERT	8.3	0.0	45.0	25.0	70.4	29.7
XLNet	25.0	14.3	55.0	28.6	29.6	30.5

Figure 5: Mean model accuracy *without* partial credit, broken down by category, for the largest model in each family

Model	Motion	Orient.	Dist.	Contain.	Metaph.	Avg (st.dev)
GPT-3	64.4	68.0	61.7	61.4	49.8	**61.1 (6.8)**
UnifiedQAv2	33.3	43.3	30.0	50.7	9.3	33.3 (15.7)
T5	79.9	33.7	70.0	50.7	25.0	51.9 (23.3)
DeBERTa	76.7	55.0	71.0	57.1	22.8	56.5 (21.0)
RoBERTa	76.7	53.9	63.9	53.4	26.0	54.7 (18.7)
ALBERT	45.7	17.2	53.0	28.4	74.5	43.8 (22.2)
XLNet	73.5	33.2	61.0	53.4	33.6	50.9 (17.6)

Figure 6: Mean model accuracy *with* partial credit, broken down by category, for the largest model in each family

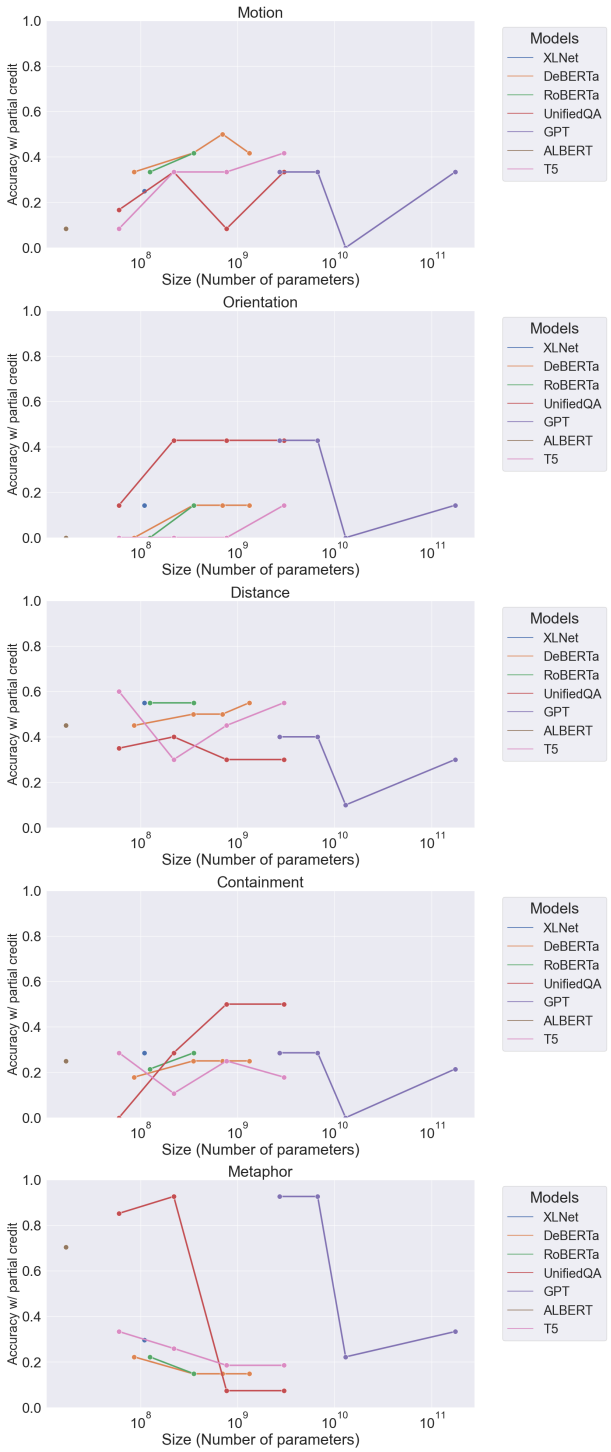

Figure 7: Mean accuracy *without* partial credit for each category, plotted with respect to model size

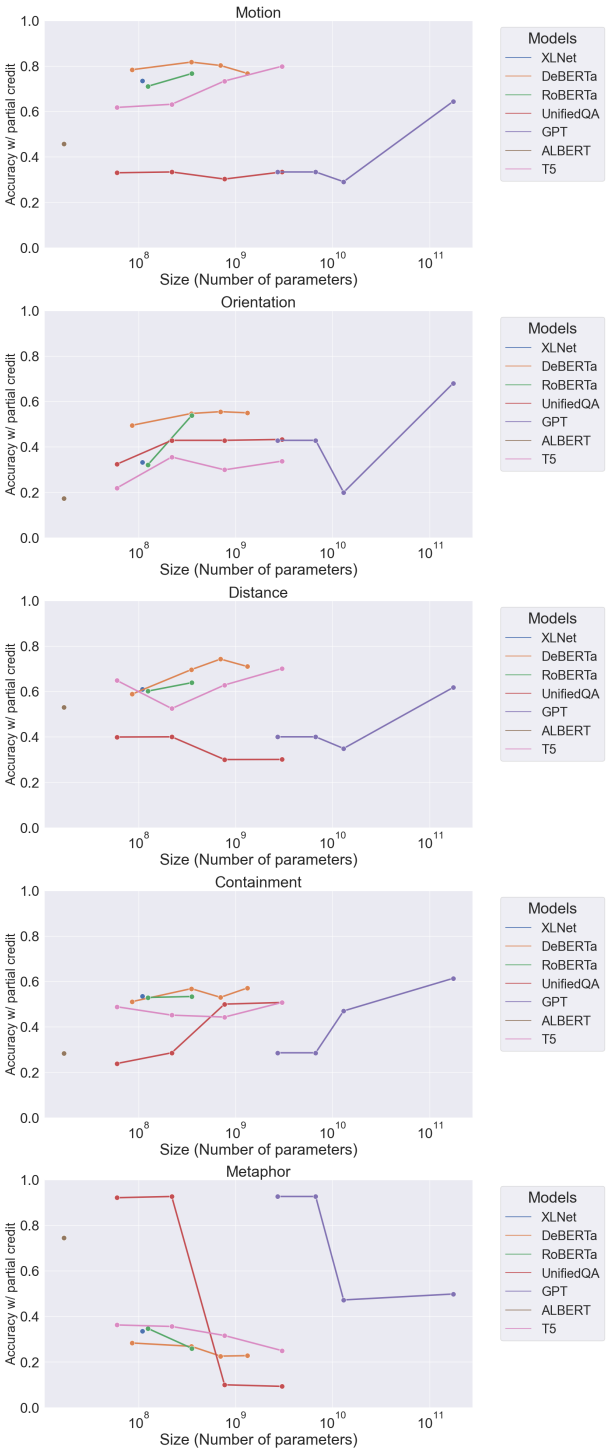

Figure 8: Mean accuracy *with* partial credit for each category, plotted with respect to model size

Category	Template	Expected	Avg. Acc_{pc}
Motion	P: {John} is {running}. H: {John} is {thinking}.	neutral	14.3
Motion	P: {John} is {in motion}. H: {John} is {thinking}.	neutral	37.0
Motion	P: {John} is {laying down}. H: {John} is {thinking}.	neutral	49.7
Orientation	P: The {theater} is {north} of the {park}. H: The {theater} is {north} of the {school}.	neutral	2.7
Orientation	P: The {block} is {above} the {book}. H: The {book} is {above} the {block}.	contradiction	39.8
Orientation	P: The {cup} is {above} of the {block}. H: The {block} is {below} of the {the cup}.	entailment	41.2
Distance	P: The {ball} is not {touching} the {cup}. H: The {cup} is {near} the {ball}.	neutral	11.1
Distance	P: The {ball} is not {touching} the {cup}. H: The {ball} is not {near} the {cup}.	neutral	14.3
Distance	P: The {ball} is not {touching} the {cup}. H: The {cup} is not {near} the {ball}.	neutral	14.3
Containment	P: The {block} {fits in} the {cup}, and the {cup} {fits in} the {cabinet}. H: The {cabinet} {contains} the {cup}.	neutral	14.3
Containment	P: The {block} {fits in} the {cup}, and the {cup} {fits in} the {cabinet}. H: The {block} {is in} the {cabinet}.	neutral	14.3
Containment	P: The {block} {fits in} the {cup}. The {cup} {fits in} the {cabinet}. H: The {cabinet} {contains} the {block}.	neutral	14.3
Metaphor	P: {John} is jumping {to conclusions}. H: {John} is {in motion}.	neutral or contradiction	2.7
Metaphor	P: The {girl} is {forced} into {movement}. H: The {girl} is {in} {movement}.	neutral or contradiction	12.7
Metaphor	P: {John} is skipping the {show}. H: {John} is {in motion}.	neutral or contradiction	13.4

Figure 9: A sample of three test cases for each category with low (relative to other test cases) Acc_{pc} scores averaged across the largest model in each model family

Category	Template	Expected	Avg. Acc_{pc}
Motion	P: {John} is {moving}. H: {John} is {in motion}.	entailment	85.7
Motion	P: {John} is {not moving}. H: {John} is {stationary}.	entailment	83.3
Motion	P: {John} is {sitting}. H: {John} is {stationary}.	entailment	71.7
Orientation	P: The {theater} is {north} of the {library}. H: The {library} is {north, east, west} of the {theater}.	neutral or contradiction	62.1
Orientation	P: The {post office} is {north} of the {library}. H: The {library} is {south} of the {post office}.	entailment	57.1
Orientation	P: The {ball} is {left} of the {block}. H: The {block} is {right} of the {ball}.	entailment	51.0
Distance	P: The {block} is {touching} the {book}. H: The {book} is {close to} the {block}.	entailment	85.7
Distance	P: The {block} is {far from} the {book}. H: The {book} is not {close to} the {block}.	entailment	85.7
Distance	P: The {block} is {far from} the {book}. H: The {block} is {touching} the {book}.	contradiction	85.0
Containment	P: The {ball} {cannot fit in} the {cup}. H: The {cup} {can contain} the {ball}.	contradiction	88.5
Containment	P: The {apples} are {placed} the {bucket}. H: The {apples} are {in} the {bucket}.	entailment	84.1
Containment	P: The {ball} is {in} the {cup}. H: The {cup} is {contains} the {ball}.	entailment	81.8
Metaphor	P: The {boy} put the {fiasco} behind him. H: The {boy} is in front of the {fiasco}.	neutral or contradiction	85.7
Metaphor	P: {John} is in {disguise}. H: {John} is physically contained in {disguise}.	neutral or contradiction	71.4
Metaphor	P: The {boy} is looking forward to the {party}. H: The {boy} is behind the {party}.	neutral or contradiction	57.1

Figure 10: A sample of three test cases for each category with high (relative to other test cases) Acc_{pc} scores averaged across the largest model in each model family

B Appendix B: Dataset Examples

B.1 PIQA Dataset Examples

```
To separate egg whites from the yolk
using a water bottle, you should...

(a) Squeeze the water bottle and
    press it against the yolk.
    Release, which creates suction
    and lifts the yolk.

(b) Place the water bottle and press
    it against the yolk. Keep pushing,
     which creates suction and lifts
    the yolk
```

B.2 PROST Dataset Examples

```
A person drops a glass, a pillow, a
    coin, and a pen from a balcony.
The [MASK] is most likely to break

A) glass B) pillow C) coin D) pen
```

B.3 bAbI Dataset Examples

Task 17: Positional Reasoning

```
The triangle is to the right of the
    blue square.
The red square is on top of the blue
    square.
The red sphere is to the right of the
     blue square.
Is the red sphere to the right of the
     blue square? A:yes
Is the red square to the left of the
    triangle? A:yes
```

Task 18: Size Reasoning

```
The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase? A:
    yes
Will the cupboard fit in the box? A:
    no
```

Task 19: Path Finding Task

```
The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A:
     west, north
How do you go from office to bathroom
    ? A: north, west
```

B.4 SpartQA Dataset Examples

```
STORY:
We have three blocks, A, B and C.
    Block B is to the right of block
    C and it is below block A. Block
    A has two black medium squares.
    Medium black square number one is
     below medium black square number
     two and a medium blue square. It
     is touching the bottom edge of
    this block. The medium blue
    square is below medium black
    square number two. Block B
    contains one medium black square.
     Block C contains one medium blue
     square and one medium black
    square. The medium blue square is
     below the medium black square.

QUESTIONS:
FB: Which block(s) has a medium thing
     that is below a black square? A,
     B, C
FB: Which block(s) doesn't have any
    blue square that is to the left
    of a medium square? A, B
FR: What is the relation between the
    medium black square which is in
    block C and the medium square
    that is below a medium black
    square that is touching the
    bottom edge of a block? Left
CO: Which object is above a medium
    black square? the medium black
    square which is in block C or
    medium black square number two?
    medium black square number two
YN: Is there a square that is below
    medium square number two above
    all medium black squares that are
     touching the bottom edge of a
    block? Yes
```

C Appendix C: Text-to-Text Prompts

C.1 GPT-3 Prompt

In the original paper, GPT-3 was evaluated on ANLI. Therefore, we use a prompt format adapted from Brown et al. (2020) (Figure G.7). We used few-shot prompting due to unreliability of outputs when prompted in a zero-shot and one-shot setting. Few-shot examples were drawn from SNLI.

```
A soccer game with multiple males
    playing.
Question: Some men are playing a sport.
    True, false, or neither?
Answer: True

A statue at a museum that no seems to be
     looking at.
Question: Tons of people are gathered
    around the statue. True, false, or
    neither?
Answer: False

A woman with a green headscarf, blue
    shirt and a very big grin.
Question: The woman is young. True,
    false, or neither?
Answer: Neither

{premise}
Question: {hypothesis} True, false, or
    neither?
Answer:
```

C.2 UnifiedQA Prompt

Since UnifiedQA was not trained or evaluated on any sort of NLI task, we reframe our task as a multiple-choice task. We use a similar framing to the GPT-3 prompt described in Appendix C.1 where the model must choose between "true", "false", and "neither". We follow the prompt setup used for the MCTest task as described in Khashabi et al. (2020) (Table 1). Similar to with GPT-3, we used few-shot prompting due to unreliability of outputs when prompted in a zero-shot and one-shot setting. Few-shot examples were drawn from SNLI.

```
Some men are playing a sport.
(A) True (B) False (C) Neither
A soccer game with multiple males
    playing.
True

Tons of people are gathered around the
    statue.
(A) True (B) False (C) Neither
A statue at a museum that no seems to be
     looking at.
False

The woman is young.
(A) True (B) False (C) Neither
A woman with a green headscarf, blue
    shirt and a very big grin.
Neither

{hypothesis}
(A) True (B) False (C) Neither
{premise}
```

C.3 T5 Prompt

T5 was originally trained on MultiNLI. Therefore, we use the prompt format used for training described in Raffel et al. (2019) (Appendix D.5).

```
mnli hypothesis: {hypothesis} premise: {
    premise}
```