Do language models have coherent mental models of everyday things?

Yuling Gu and **Bhavana Dalvi Mishra** and **Peter Clark**
Allen Institute for AI, Seattle, WA
{yulingg,bhavanad,peterc}@allenai.org

Abstract

When people think of everyday things like an egg, they typically have a mental image associated with it. This allows them to correctly judge, for example, that "the yolk surrounds the shell" is a false statement. Do language models similarly have a coherent picture of such everyday things? To investigate this, we propose a benchmark dataset consisting of 100 everyday things, their parts, and the relationships between these parts, expressed as 11,720 "X relation Y?" true/false questions. Using these questions as probes, we observe that state-of-the-art pre-trained language models (LMs) like GPT-3 and Macaw have fragments of knowledge about these everyday things, but do not have fully coherent "parts mental models" (54-59% accurate, 19-43% conditional constraint violation). We propose an extension where we add a constraint satisfaction layer on top of the LM's raw predictions to apply commonsense constraints. As well as removing inconsistencies, we find that this also significantly improves accuracy (by 16-20%), suggesting how the incoherence of the LM's pictures of everyday things can be significantly reduced.[1]

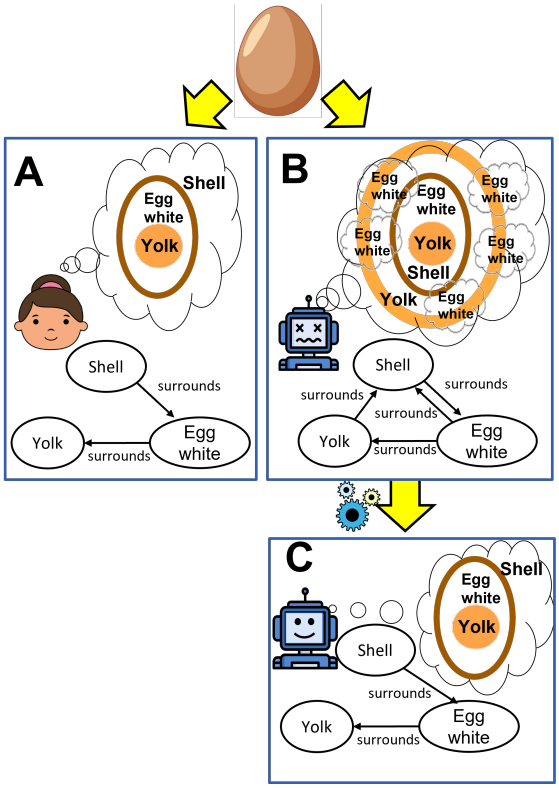


Figure 1: While humans appear to have coherent mental pictures of everyday things (e.g., an egg, **A**), our question-asking probes suggest that LMs do not (e.g., one LM answered that the egg white both surrounds and is surrounded by the shell, **B**). This model incoherence can be reduced by applying commonsense constraints (e.g., surrounds is asymmetric), resulting in a more coherent parts model (**C**).

1 Introduction

Psychologists and cognitive scientists hypothesize that humans develop mental models of the world, namely internal, conceptual representations of the environment which we base our decisions and actions on (Ha and Schmidhuber, 2018; Jonassen and Henning, 1996). Hespos and Spelke (2004) observed that 5-month-old human infants exhibit understanding of mechanical properties of objects in terms of arrangements and motions of surfaces, well before they can understand language. Drawing loosely on this idea, but without making any claims about how LMs reason internally (Shanahan, 2022; Andreas, 2022), we investigate if pre-trained language models show evidence of coherent internal representations of everyday things, analogous to human mental models, via probing. We focus on mental models in the context of ordinary objects that we encounter in our everyday lives. Such commonsense knowledge helps us understand how these everyday things work and how to interact with them. For example, when someone tries to make a fried egg, they know that it has a shell and

[1] We make our data and code publicly available at https://github.com/allenai/everyday-things.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1892–1913
July 9-14, 2023

that it can be cracked open to reveal the egg white and yolk inside. However, if a system does not have a coherent picture of such everyday things, thinking that the egg yolk surrounds the shell, then it might have to resort to ridiculous approaches such as trying to scrape the egg yolk off the shell into the pan.

We explore a first version of this, in which we consider only knowledge about an object's parts and their relationships. We refer to this knowledge as a parts mental model. We first create a benchmark dataset of 100 everyday things, by asking human annotators to draw a graph representing their parts mental model (e.g., Figure 2) depicting the parts of an everyday thing, spatial relationships, connections between its parts and functional dependencies (if any). Then we probe two representative state-of-the-art LMs with questions about these everyday things. We find that the LMs' parts mental models are generally of poor quality. Further, model predictions can violate basic consistency constraints e.g. transitivity. To alleviate this, we apply constraint reasoning to derive more accurate and consistent mental models of everyday things, correcting some of the LMs' original inconsistencies. This is illustrated in Figure 1.

Our contributions are:

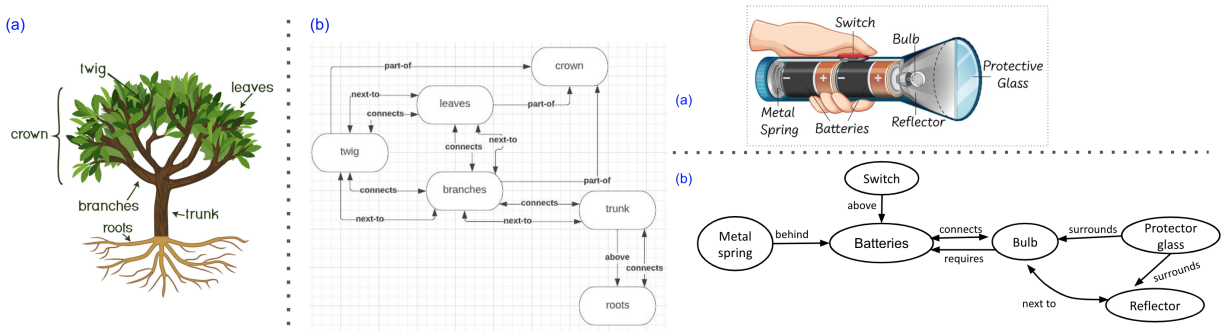
1. We present a benchmark dataset of parts mental models consisting of 100 everyday things, 2.2K parts and 11.7K relationships.
2. We show that SOTA LMs like GPT-3 and Macaw are poor at answering relationship queries between parts of everyday things. The parts mental models derived using their predictions are only 54-59% accurate, and significantly inconsistent (19-43% conditional violation τ).
3. We propose a neuro-symbolic method that applies constraint reasoning on top of raw LM predictions as a way of obtaining more consistent (0% conditional violation τ) and more accurate mental models (16-20% improvement). This suggests a broader cognitive architecture (LM + reasoner) for future systems, to better construct mental models than the LM alone.

2 Related work

Mental models: The idea of mental models (Johnson-Laird, 1983) is not new. Many years ago, Craik (1943) proposed that thinking itself is the manipulation of internal representations of the world. Craik (1943) described mental models as a 'small-scale model' of external reality and of its own possible actions within someone's head. Such a mental model is useful in many ways, including allowing one to try out various alternatives, make conclusions, react to future situations, learn from past events, and in general, improve competency. Years later, when Johnson-Laird (2006) outlined the mental processes that underlie human reasoning, he based his discussion on the fundamental assumption that human beings can construct internal representations of spatial layouts, and specified mental models to be iconic. In his words, a mental model's "parts and the relations among them correspond to the parts of the layout and the relations among them." While coherent internal representations of spatial layouts are crucial for human reasoning, their role, coherence, and even existence in LMs have not been systematically explored. In this work, we try to bridge this gap by proposing a benchmark dataset and methodology to compare human internal representations of spatial layouts of everyday things with those of LMs.

Prior datasets: Prior works on reasoning about object/body parts include Li et al. (2019b) which focused on human body parts and human interaction with other objects. The PTR benchmark (Hong et al., 2021) is a QA dataset about objects and their parts, combining 5 everyday things: chair, table, bed, refrigerator, and cart, to create questions across 70K different scenes. Ji et al. (2022) used tangram puzzles to analyze shape naming, part naming and segmentation divergence across participants when they see a certain shape. Contributing to this existing body of datasets, the dataset we introduce serves as a resource for researchers to study canonical parts mental models for a wide variety of everyday things, focusing on relationships between parts of objects, which is fundamental to how humans think and interact with these things.

Large language models: Despite recent advances in LMs, studies suggest that they still struggle at reasoning with real-world entities and concepts. Bisk et al. (2020) found that when LMs answer questions involving physical commonsense reasoning, their performance at that time was near chance level for questions involving spatial relations like "top" and "bottom." Sahu et al. (2022) demonstrated the lack of conceptual consistency in LMs by correlating models' answers on commonsense reasoning questions (CSQA dataset) and their



Tree Flashlight

Figure 2: Our everyday things dataset, ParRoT, covers different entities, both natural (e.g. tree) and man-made (e.g. flashlight). Above are two examples of such everyday things. In each case, we show a (a) diagram of the entity, and (b) parts graph of the everyday thing drawn by crowdworkers. The parts graphs illustrate how our dataset contains a variety of relations between parts.

answers on associated conceptual questions from ConceptNet knowledge base. To improve existing systems, progress has been made such as by imposing constraints with neuro-symbolic approaches (Nye et al., 2021; Mitchell et al., 2022) and incorporating both textual and visual information (Dan et al., 2020). Inspired by recent progress, we propose a constraint reasoning method that applies hard commonsense constraints (e.g., if 'A above B' is *True* then 'A below B' cannot be *True*) on top of raw LM predictions to produce more accurate and consistent mental models of everyday things.

3 Parts mental models and Task

We define "parts mental model" for everyday things in this section. Then in the rest of the paper, we describe how we collect a dataset for them, measure LMs' coherence on them, and finally apply external reasoning to improve the accuracy and consistency of LMs' parts mental model.

Here, we use parts mental model to mean a parts-focused subset of a complete mental model of an entity. We represent a parts mental model as a directed graph where parts of the everyday thing form the nodes of this graph and these nodes are connected with edges indicating how these parts are related to each other. Based on prior works such as Renz (2002) and Gunning et al. (2010), we selected 11 spatial orientation relations to focus on. In addition, we augmented these with relations describing connectivity and functional dependency. In total, we consider 14 relationships (across these 3 categories) between parts, listed in Table 2.

Note that the notion of a single "parts mental model" for an everyday thing is somewhat unconstrained (e.g., which parts to pick? what version of the entity are we talking about?). To make this task more well-defined, we also provide a predefined list of parts as a guide (details in Section 4.1), and the task for annotators or a model is to specify relationships between them as they see appropriate, using our ontology of relationships. This is important so that we can do meaningful comparisons between language models and humans' notion of parts mental models of everyday things.

Figure 2 shows two examples of parts mental models in our dataset, where edges encode relationships between parts. E.g., in a tree, "trunk is above the roots"; in a flashlight, "bulb requires the batteries," etc. Inspired by previous literature, we envision that such parts mental models would play a key role when one carries out daily activities involving these everyday things.

Task

Here we define our task: "Construct a parts mental model for everyday things" with the following input/output specifications:

- Input: Everyday thing, Parts list, Relation vocabulary (14 relations).
- Output: List of tuples (x, r, y) where relation r holds between parts x and y.

In Section 4 we describe how we acquire a benchmark dataset by asking human annotators to carry out this task. Once we have collected gold-standard parts mental models for everyday things based on the human annotations, we prompt LMs for their

[2] A requires B denotes A cannot perform its primary function without B.

	Given as seed (unique)	Annotated mental models	Avg. annotated per mental model	Annotated + enriched (*) (Total)	Total avg. per mental model (Total / # mental models)
# everyday things	100	100	-	100	-
# mental models	-	300	-	300	-
# parts	716	2191	7.30	2191	7.30
# relations (p1, rln, p2)	8	2752	9.17	11720	39.07
# spatial relations	6	1858	6.19	9956	33.19
# connectivity relation(s)	1	818	2.73	1612	5.37
# functional relation(s)	1	76	0.25	152	0.51

Table 1: Statistics of ParRoT, our Everyday Things Dataset. *Enriched refers to implied relations, see Section 4.3

Type	Relations
Spatial orientation	part of, has part, inside, contains, in front of, behind, above, below, surrounds, surrounded by, next to*
Connectivity	directly connected to*
Functional dependency	requires[2], required by

Table 2: Relationships encoded in "parts mental models" of everyday things. Among these relations, 'next to' and 'directly connected to' relations are bi-directional, whereas the other 12 relations are uni-directional.

parts mental models and evaluate how well they do on this task. Our proposed method to measure this is described in Section 5. In particular, we are interested in (1) how accurate are LM-generated parts mental models when compared to gold-standard models in our dataset and (2) ignoring accuracy, how consistent are these generated parts mental models with respect to basic commonsense constraints? I.e., Do they at least conform to the 4 types of commonsense constraints laid out in Section 5.2 e.g., '*above*' and '*below*' are inverse relations, so if the LM predicts that in a tree, (trunk is *above* the roots) then it should also predict (roots are *below* the trunk).

4 Everyday Things Dataset: ParRoT (Parts and Relations of Things)

We created a dataset of common entities that one would encounter in their daily life. For each everyday thing, our dataset (ParRoT) contains a "parts mental model" in the form of a graph, which depicts parts of the entity and relational information about the parts. Such a graph encodes a parts-focused mental model of that everyday thing, potentially useful for reasoning about how the entity works and how to interact with it.

4.1 Everyday entities

We first compiled a list of entities from children's books, vocabulary lists (Grades 1-8), and online web search.[3] For the unique entities in this list, the authors manually filtered out those entities that are not common in everyday setting or have too few (i.e. only 1 or 2 parts) or too many parts (composite scenes). Specifically, we kept 100 entities that are common everyday things that a child would be familiar with, with a mix of natural and man-made things. This annotation task involves answering the following question for each item in the list: "Do you imagine this is something that most people would have seen in their everyday lives?"

We recognize there could be many variants of a single everyday entity e.g. different types of coffee makers. To narrow down the possibilities, the authors picked a diagram for each everyday thing via web search and carefully annotated a parts list for each of them to guide the level of granularity we are looking for. In some cases, the entity name was qualified to disambiguate further e.g. "digital clinical thermometer" instead of just "thermometer."

4.2 Mental model annotations

We ask crowdworkers to draw sketches of everyday things covering spatial relations, connectivity, and functional dependencies between parts (Table 2). To encourage the format of the mental model graphs to be more standardized across annotators, we ask that the nodes (in circles) mainly contain labels from the "Parts list" provided. However, to collect mental models that are most natural to the workers, they were also told that they can ignore parts in the "Parts list" if they seem unimportant, or add extra parts that seem important. We also specified for edges to be labeled with the relations

[3]Appendix A provides more details on the source of the list of everyday things.

shown in Table 2.[4]

Given the name of an everyday thing, list of parts, and example diagram, 3 crowdworkers were recruited to sketch mental models for each everyday thing.[5] Figure 2 shows examples of such sketches. According to Norman (2013), mapping that takes advantage of spatial analogies leads to immediate understanding and is more natural. Sketching out such a graph allows workers more flexibility in taking advantage of spatial analogies between the actual entity and the sketch (see flashlight example in Figure 2). Therefore, we hypothesize that drawing a graph would be easier or more natural for crowdworkers than typing a list of relations.[6]

4.3 Statistics

ParRoT consists of 100 everyday things ranging from devices like coffee maker, space heater to natural entities like tree and butterfly with number of parts (provided as a seed list to crowdworkers) ranging from 3-14. We collected 3 mental models per everyday thing. We take the parts mental models annotated by crowdworkers to be correct but not complete. I.e., they may include only those relations that they think are salient for the everyday thing, and also omit the ones that can be easily inferred from what they have annotated e.g., when (trunk is $above$ the roots) is annotated, (roots are $below$ the trunk) can be omitted (Figure 2, tree example). For each everyday thing's mental model annotation, with the relation tuples annotated, we automatically add relations that are implied via enrichment based on 4 types of constraints (symmetric, asymmetric, inverse, and transitive). The inferred relations include both relations that are labeled True (e.g. A above B being True implies that B below A is True) and relations that are labeled False (e.g. A above B being True implies B above A is False). This gives a total of 11.7K gold relation tuples (6894 with "True" as gold labels and 4826 with "False" as gold labels). Table 1 provides additional dataset statistics. Appendix C discusses the unanimity and diversity of mental models for these everyday things.

[4]For ease of annotation, they do not need to repeat annotations that mean the same thing. e.g. if they annotated (x, above, y), they do not need to annotate (y, below, x) again. We automatically generate these in our data post-processing.

[5]More details can be found in Appendix B.

[6]Later these sketches are transcribed into (x, r, y) tuples.

5 Measuring and Improving Parts Mental Models

Our proposed approach, ParRoT-Con,[7] comprises two main components.[8] The first component "Probing a Pre-trained Language Model" sends an exhaustive list of relation queries to a LM querying for every relation between each pair of parts (e.g. all relationships between egg white, yolk, shell, shell membrane and air cell). This gives us a large set of candidate relation tuples along with the model's confidence in each of them. Incorrect relation predictions can result in inconsistencies in the mental model. E.g, "egg white both surrounds and is surrounded by the egg shell." The second component "constraint reasoning" then applies a constraint satisfaction layer on top of these raw predictions to choose a subset of these relation tuples that are maximally probable and minimally conflicting with each other. Note that ParRoT-Con is a zero-shot approach, where both probing LMs and constraint reasoning steps do not require any task-specific fine-tuning or re-training.

5.1 Probing a Pre-trained Language Model

We use the following pre-trained language models for our study: GPT-3 (Brown et al., 2020) and Macaw[9] (Tafjord and Clark, 2021). We probe them using True/False questions of type: "Judge whether this statement is true or false: In an <everyday thing>, <part1 relation part2>." For each query q, we record an answer $a \in \{True, False\}$, and the model's beliefs about the likelihood of the relation being "True" as

$$\frac{p(True|q)}{p(True|q) + p(False|q)}.$$

5.2 Constraint Reasoning

We observed a significant amount of inconsistency in raw predictions from these LMs by considering the following constraints:

- **Symmetric relations:** This constraint ensures symmetric relations like "directly connected to" and "next to" hold both ways.
  i.e. x rln $y \leftrightarrow y$ rln x

[7]First obtain the output of "stochastic *parrots*," (Bender et al., 2021) then apply constraints to reason on top of the output.

[8]See Appendix D Figure 8 for an illustration.

[9]A SOTA T5-11B based question-answering system that outperforms GPT-3 on some QA tasks.

- **Asymmetric relations:** For asymmetric relations like part of, has part, inside, contains, in front of, behind, above, below, surrounds, surrounded by, requires, required by, this constraint makes sure that both "x rln y" and "y rln x" cannot be true at the same time.
  i.e. $\neg(x \text{ rln } y) \vee \neg(y \text{ rln } x)$

- **Inverse relations:** For a set of inverse relations e.g. above vs below, this constraint makes sure that (x above y) and (y below x) have the same truth value.
  i.e. x rln $y \leftrightarrow y$ inverse(rln) x

- **Transitive relations:** For relations like inside, contains, in front of, behind, above, below, surrounds, surrounded by, this constraint will impose transitivity.
  i.e. x rln $y \wedge y$ rln $z \rightarrow x$ rln z

In this step, we try to resolve inconsistencies in LMs' raw predictions by solving a MaxSAT constraint satisfaction problem where each (x, relation, y) tuple is represented as a variable with confidence value from the LM used as its weight (soft clause). We introduce 4 types of hard constraints (listed above) between these variables as hard clauses and any constraint violation results in an extremely high penalty. Given a WCNF formula with these, a weighted MaxSAT solver tries to find an optimal assignment of truth values to relation tuples that maximizes the sum of weights of satisfied soft clauses and satisfies all the formula's hard clauses. We use the RC2 MaxSAT solver (Ignatiev et al., 2018b) in PySAT (Ignatiev et al., 2018a).

6 Results and Analysis

6.1 Evaluation Metrics

We evaluate the parts mental models produced by the two LMs in terms of accuracy and consistency:

Accuracy: We compute the True/False accuracy of parts mental models based on the 11.7K gold relation tuples present in ParRoT.

Consistency: Following Kassner et al. (2021); Mitchell et al. (2022), we adapt the Conditional Violation (τ) (Li et al., 2019a) metric to measure inconsistency across the 4 types of constraints defined in Section 5.2. For constraints $L(x) \rightarrow R(x)$ imposed on samples $x \in D$, where D is the dataset, we calculate conditional violation as:

$$\tau = \frac{\sum_{x \in D} \left[\bigvee_{(L,R)} \neg(L(x) \rightarrow R(x)) \right]}{\sum_{x \in D} \left[\bigvee_{(L,R)} L(x) \right]}.$$

6.2 Results

Q1: How consistent are LMs when they answer questions about everyday things?
We measure the consistency of parts mental models constructed by LMs based on 4 types of constraints described in Section 5.2. This measurement is purely based on LMs' predictions and is independent of relations in the gold mental models acquired for the everyday things. Table 3 shows that LMs contradict themselves (19-43% conditional violation) when we ask them multiple questions about parts of the same everyday thing to probe for their parts mental model. E.g., in Appendix D, the LM believes that in an egg, "yolk surrounds the shell" and "shell surrounds the yolk" are both True. Table 3 also breaks down the LMs' inconsistency across 4 types of constraints. We observe that GPT-3 struggles with maintaining consistency for symmetric and inverse relations, whereas Macaw-11B finds it most challenging to satisfy constraints for asymmetric relations.

Q2: Do language models have accurate mental models of everyday things?
Next, we investigate how accurate are these parts mental models when compared to gold mental models in our ParRoT dataset. Table 4 shows that such queries pertaining to parts of everyday things are challenging for even SOTA models, with an average accuracy of 54-59%. This is barely better than the majority class baseline at 59% and random chance at 50%.

The LMs' low performance shows that ParRoT is a challenging dataset, which is expected given the fact that this dataset queries for commonsense knowledge about everyday things (e.g. spatial relationship between parts of a device) that are often omitted in text, and hence less likely seen during pre-training. Further, by construction, our queries minimally differ e.g. for relations between parts of a tree, the edit distance between a statement with true relation "the leaves are above the roots" and false relation "the leaves are below the roots" is just 1 word. This makes our task even more challenging

	%True tuples	%Conditional Violation (lower is better) Symmetric relations	Asymmetric relations	Inverse relations	Transitive relations	Avg. (macro)	Avg. (micro)
GPT-3 (text-davinci -003)	12.64	66.37 (1,987/2,994)	23.01 (4,699/20,422)	71.14 (13,869/19,495)	32.18 (6,550/20,354)	48.17	42.84 (27,105/63,265)
Macaw-11B	57.77	29.98 (3,089/10,305)	64.97 (42,170/64,910)	33.63 (21,642/64,361)	10.08 (44,121/437,746)	34.66	19.23 (111,022/577,322)

Table 3: Parts mental models constructed by LMs are significantly inconsistent with respect to their own predictions, violating basic commonsense constraints. In brackets, we indicate (# violations) / (# constraints fired).

	# params	Base LM (%)	ParRoT-Con (%)	Improve (%)
GPT-3 (text-davinci-003)	175B	53.83	70.26	16.42
Macaw-11B	11B	59.45	79.28	19.84

Table 4: Comparing the accuracy of parts mental models before and after constraint reasoning on ParRoT dataset.

as the models need to understand the semantics of relational phrases to give the correct answer.

Q3: Does ParRoT-Con, our proposed constraint reasoning approach, help create more accurate mental models?

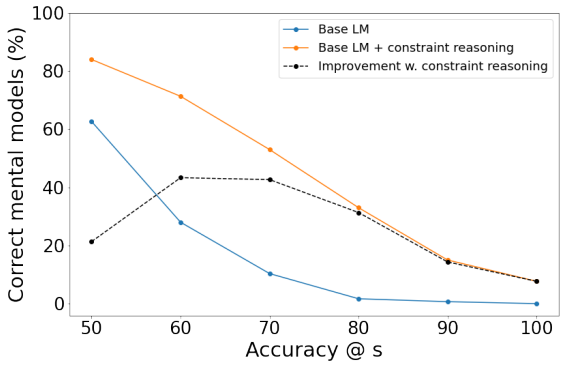
Our proposed approach, ParRoT-Con, utilizes the inherent inconsistency in LMs' raw predictions to self-correct their own parts mental models. It finds an optimal assignment of truth values to relation tuples that accounts for both the model's original beliefs (about the likelihood of each relation statement being True or False), and the 4 types of commonsense constraints imposed. By imposing the commonsense constraints as hard constraints, our proposed method produces perfectly consistent mental models for all LMs with respect to the imposed constraints i.e. % conditional violation becomes 0 for all columns in Table 3. Using these basic commonsense constraints, ParRoT-Con improves parts mental model accuracy significantly by 16-20% on ParRoT (Table 4).

6.3 Further analysis

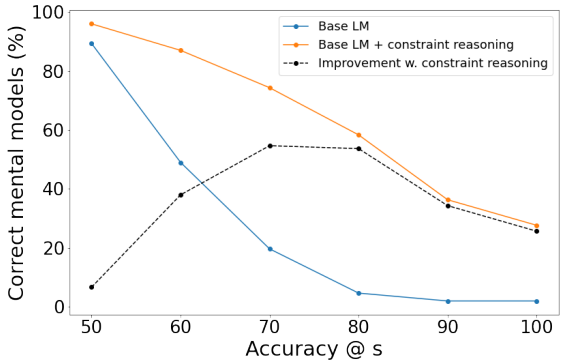
Most effective range We analyze what is the quality range of mental models that ParRoT-Con is most effective on. We quantify the quality of parts mental models by defining accuracy@s, a metric that says a mental model is correct if the proportion of correct relations is at least s%. We then plot the percentage of mental models (out of 300) that are correct vs accuracy@s for different values of s, where $s \in \{50, 60, 70, 80, 90, 100\}$. Figure 3 shows that ParRoT-Con not only effectively increases the percentage of mental models that are approximately correct (s = 50, 60) but also the percentage of mental models that are (almost) totally correct (s = 90, 100). The improvements with constraint reasoning are even more prominent when it comes to increasing the percentage of mental models that are at least 60-80% accurate. This is likely attributed to the improvement in mental models that have enough signals from LMs' raw predictions and also enough margin to improve.

Accuracy of parts mental models per relation Figure 4 shows that the base LMs are more accurate in predictions for queries containing relationships like 'part of' which is more likely to be stated in text than spatial relations like 'above', 'below', and 'behind' which are lower-level physical details often not mentioned in text. Different models also differ in which relationships they perform better on: e.g. GPT-3 performs poorly on bi-directional relations like 'connects' and 'next to', with accuracy way below chance level, while Macaw-11B achieves around 70% accuracy for queries involving these relations.

Success and failure across models per everyday thing LMs show both similarities and differences in what everyday things they have better mental models of. For each model, Figure 5 shows the top 20 everyday things that the models performed *best* on in terms of base LM accuracy. Both GPT-3 and Macaw-11B perform well on the following everyday things: sandwich, kayak, dog, kite, bird, rat, cat, pencil sharpener, tree, cable car, and butterfly. It is interesting to see that both models perform well on several natural living things like animals (e.g. dog, bird, rat, cat), insect (e.g. butterfly), and plant (e.g. tree). Figure 6 shows the top 20 everyday things that the models performed *worst* on in terms of base LM accuracy. We observe that

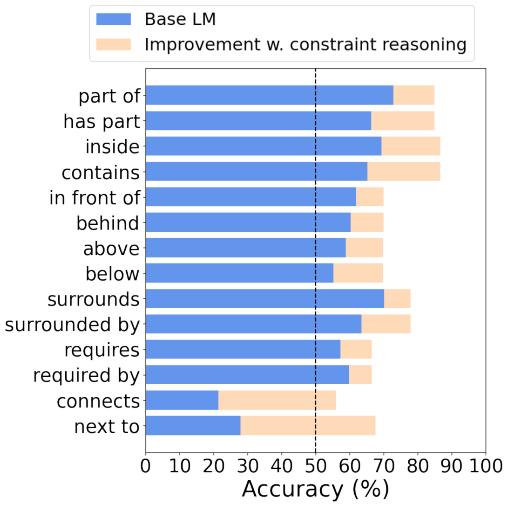


(a) GPT-3

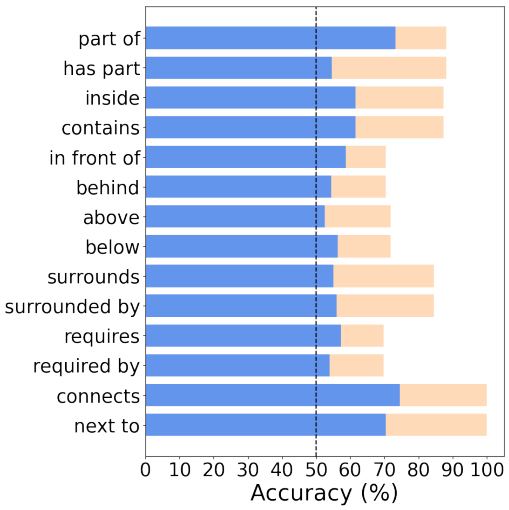


(b) Macaw-11B

Figure 3: Percentage of correct mental models vs accuracy@s shows that for both GPT-3 and Macaw-11B, there is a higher percentage of correct mental models after constraint reasoning (orange) as compared to raw LM predictions (blue), no matter the threshold for considering a mental model to be correct is lower or higher. For improvements from constraint reasoning (black), we observe the highest increase in percentage of mental models that are at least 60-80% accurate.



(a) GPT-3



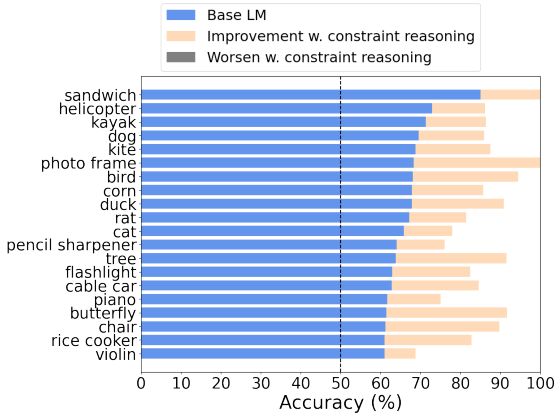
(b) Macaw-11B

Figure 4: Accuracy of base LM and improvement achieved through constraint reasoning on different relations in ParRoT dataset.

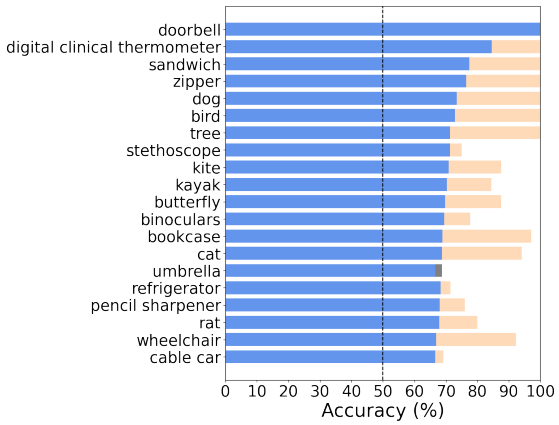
entities like typewriter, bed, air conditional, and computer are challenging for both models to form accurate mental models of. Although the models share some similarities in what everyday things they have better/worse mental models of, they also show differences, especially for man-made devices: e.g. GPT-3 does well but Macaw-11B performs poorly on forming an accurate parts mental model of piano; Macaw-11B does well, but GPT-3 performs poorly on devices like doorbell, digital clinical thermometer, and binoculars.

7 Conclusion

Do language models have coherent mental models of everyday things? To systematically study this question, we present a benchmark dataset, ParRoT, consisting of 300 human-constructed mental models for 100 everyday objects, including over 2K parts and 11.7K relationships between these parts. Our experiments reveal that even SOTA LMs generally have poor mental models (inaccurate and violating basic commonsense constraints) of everyday things, thus providing insight into their apparent knowledge and behavior not previously explored. We apply constraint reasoning on top of base LM predictions to construct more coherent mental models. Our method, ParRoT-Con, improves both accuracy (up to 20% improvement) and consistency (up to 43% improvement) of such parts mental models. This suggests a broader cognitive architecture (LM + reasoner) for future systems, to construct more coherent mental models than using the LM alone.

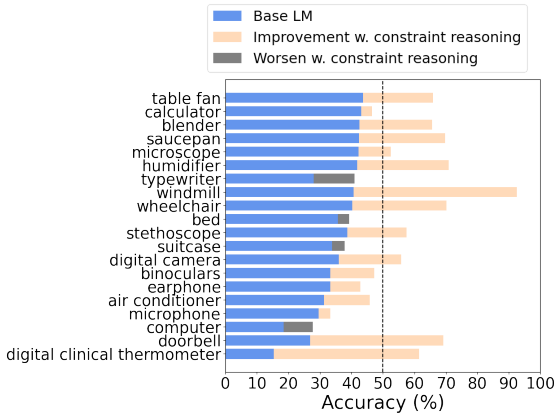


(a) GPT-3

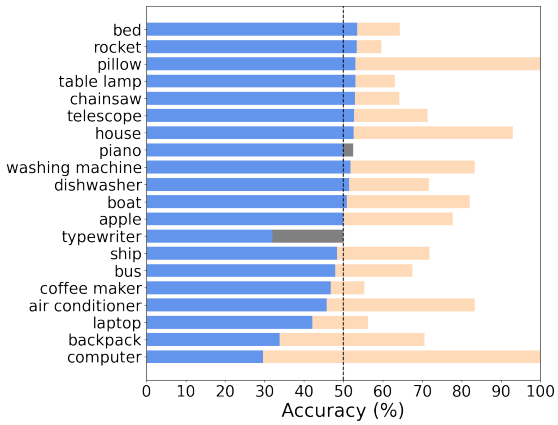


(b) Macaw-11B

Figure 5: 20 everyday things that each model achieved **best performance** on, based on models' raw predictions (i.e. Base LM). In almost all cases, constraint reasoning boosts the accuracy of the parts mental models produced by the base LM, pushing it even closer to 100%.



(a) GPT-3



(b) Macaw-11B

Figure 6: 20 everyday things that each model achieved **worst performance** on, based on models' raw predictions (i.e. Base LM). In many of these cases, the accuracy of the parts mental models produced by the base LM is at around or below chance level and constraint reasoning boosts accuracy to beyond 50%.

Limitations

Common everyday things change over the years. While we try to choose ones that are in children's vocabulary, over decades, devices evolve and humans change in which things they interact with more frequently, affecting which relationships would be more prominent in an average person's mental model. So the parts mental models in such a dataset may not stay constant over time (e.g. some entities may be less familiar and certain relations may be less salient to annotators of the future). It would be interesting to use our ParRoT dataset as a point of comparison when studying mental models of everyday things in the future to reveal interesting insights on how humans' mental models of everyday things evolve over time.

Other important future directions include to explore how more coherent mental models can help in complex reasoning tasks about everyday things, combine these parts mental models with mental models along other dimensions e.g. Gu et al. (2022a,b), as well as using our dataset of commonsense queries about everyday things as a source of follow-up questions for existing QA tasks e.g., PIQA (Bisk et al., 2020) and CSQA (Talmor et al., 2019).

This paper only focuses on relationships (spatial orientation, connectivity, and functional dependency) between parts of everyday things. However, our approach ParRoT-Con is easily extensible to other applications such as:

- spatial relations in other domains e.g. for geographical distances, we can similarly impose constraints on inverse relations like *closer* and *further*
- temporal relations e.g. on a timeline, if event A occurred *before* event B, then event B cannot have occurred *before* event A (*before* is asymmetric)

We leave the demonstration of the generalizability of our approach to future works.

Ethics Statement

All annotators that participated in the data collection process have been anonymized. The only personal information we collect is the worker IDs from Amazon Mechanical Turk, which we will not release. No personally identifiable information is contained in our dataset or otherwise released. We took great care to pay fair wages, and were responsive to feedback and questions throughout the data collection process. This study involves the use of large-scale language models. We only use them to generate True/False answers to questions about parts of everyday things, therefore we do not foresee any substantial ethical issues with their use for research presented in this submission.

Acknowledgements

We thank the anonymous ACL reviewers, as well as Ernest Davis, Chris Callison-Burch and members of the Aristo team at AI2 for their valuable feedback on an earlier draft.

References

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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.

Wonder House Books. 2018a. *My First 100 Things that move*. Wonder House Books.

Wonder House Books. 2018b. *My First Library : Boxset of 10 Board Books for Kids*. Wonder House Books.

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.

Kenneth James Williams Craik. 1943. *The nature of explanation*, volume 445. Cambridge University Press.

Soham Dan, Hangfeng He, and Dan Roth. 2020. Understanding spatial relations through multiple modalities. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2368–2372, Marseille, France. European Language Resources Association.

Valorie Fisher. 2019. *Now You Know How It Works*. Scholastic.

Steve Graham, Karen R. Harris, and Connie Loynachan. The Basic Spelling Vocabulary List. https://www.readingrockets.org/article/basic-spelling-vocabulary-list. Accessed: 2022-09-23.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022a. DREAM: Improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022b. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Gunning, Vinay K Chaudhri, Peter E Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David D McDonald, Sunil Mishra, et al. 2010. Project halo update—progress toward digital aristotle. *AI Magazine*, 31(3):33–58.

David R Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint*, abs/1803.10122.

Graeme S. Halford. 1993. *Children's Understanding: The Development of Mental Models*. Lawrence Erlbaum Associates, Inc.

S. J. Hespos and E. S Spelke. 2004. Conceptual precursors to language. In *Nature*. Nature.

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. 2021. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. In *Advances in Neural Information Processing Systems*, volume 34, pages 17427–17440. Curran Associates, Inc.

Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. 2018a. PySAT: A Python toolkit for prototyping with SAT oracles. In *SAT*, pages 428–437.

Alexey Ignatiev, Antonio Morgado, and Joao Marques-Silva. 2018b. Rc2: a python-based maxsat solver. *MaxSAT Evaluation*, 2018:22.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

P. Johnson-Laird. 1983. *Mental Models : Towards a Cognitive Science of Language, Inference and Consciousness*. Harvard University Press.

P. Johnson-Laird. 2006. *How we reason*. Oxford University Press.

David H. Jonassen and Philip Henning. 1996. Mental models: Knowledge in the head and knowledge in the world. *Educational Technology archive*, 39:37–42.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019a. A logic-driven framework for consistency of neural models. *arXiv preprint arXiv:1909.00126*.

Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. 2019b. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Eric Mitchell, Joseph J. Noh, Siyan Li, William S. Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D. Manning. 2022. Enhancing self-consistency and performance of pretrained language models with nli. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Donald A. Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. In *Advances in Neural Information Processing Systems*, volume 34, pages 25192–25204. Curran Associates, Inc.

Jochen Renz, editor. 2002. *The Region Connection Calculus*, pages 41–50. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pritish Sahu, Michael Cogswell, Yunye Gong, and Ajay Divakaran. 2022. Unpacking large language models with conceptual consistency. *arXiv preprint arXiv:2209.15093*.

Murray Shanahan. 2022. Talking about large language models. *arXiv preprint*, abs/2212.03551.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with Macaw. *arXiv preprint arXiv:2109.02593*.

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.

A Source of everyday things

We compiled a list of 100 everyday things from:

1. Children's books
   (a) My First Library series (Books, 2018b)
   (b) Now you know how it works (Fisher, 2019)
   (c) My first 100 things that move (Books, 2018a)

2. Vocabulary lists
   (a) Grade 1-5 vocabulary list (Graham et al.)
   (b) Select from all the nouns from an 8th-grade vocabulary list that were also under either "artifact" or "device" in WordNet (Miller, 1994)

3. Online web search

B Details on mental model annotation task

Mechanical Turk task instructions:

Instructions (click here to collapse/expand instructions)

NOTE: To complete this HIT, you need a Google account (to upload your work, step 3). If you don't have one, you can easily create a temporary one by clicking here.

We are wanting to understand the parts and relationships that come to mind, when people think of an everyday object, e.g., a book. This will help us understand how people picture everyday objects, and allow us to compare that with a computer's picture of those objects.

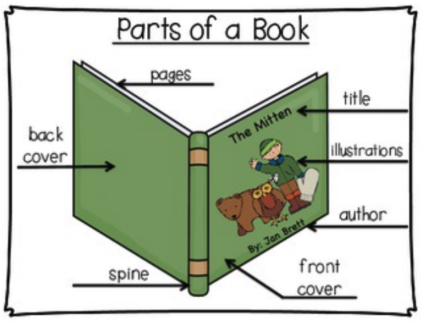
The HIT is a little unusual: you simply draw a graph, then email a photo/PDF of it to us. We will approve all reasonable graphs (but not spam) within 30 hours of submission. Please carefully read through the do's and don'ts below and make sure your graph follows these instructions.

Here's how it works: First we will give you the name of an everyday object, e.g., "book", and a list of some of its parts (e.g., "spine" "cover" "pages"). Your job is then to sketch a graph showing how those parts are connected, then submit a photo/PDF of that graph to us! You can either draw the graph physically (and legibly) with a pen and paper, or sketch it on the computer, as you like.

Example 1:

Consider the below everyday thing:

- Everyday thing: **book**
- Parts list (as a guide): **title, author, front cover, pages, back cover, spine, illustrations**
- Diagram (as a guide):



Now:

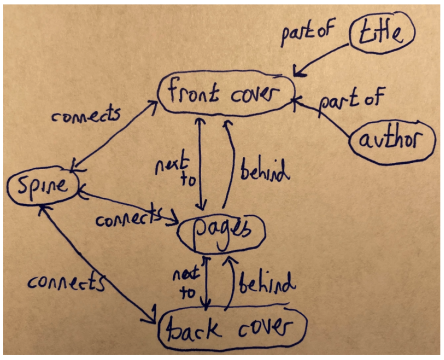
1. (Thinking) First think about this object placed in a setting that is most common/natural to you.
2. (Sketching) Now, get a pencil and paper (or a sketching tool) and **sketch a graph** where:
   1. generally, each node is one of the parts above.
   2. each edge shows a relationship that holds between two parts.

Comments:

- You don't have to use all the parts in the "Parts list" or diagram, and you can also add in extra parts that seem important. The goal is to generally use the "Parts list", but to produce a graph which seems **natural** to you.
- Try to sketch **all** the relations that hold between two parts, not just one relation. (e.g. "next to" and "behind" can be true at the same time in this example)

Here are the eight relationships to use. Six of them below point in a particular direction:
-part-of→
-inside→
-behind→
-above→
-surrounds→
-requires→ ("X -requires→ Y" if X cannot perform its main function without B. e.g., "flashlight bulb" -requires→ "battery")
and two others are bi-directional (work in both directions):
←connects→ (Use "connects" as a shorthand for "directly connected to")
←next-to→

Here is an example of a sketch you might produce. (This one was sketched by hand on paper). Note that the nodes (in circles) are mainly the list of **parts** shown earlier above.



Comment: Note that the nodes in the graph are mainly the parts in the **"Parts list"** earlier, except the worker here decided to ignore the part "illustrations", even though it's in the "Parts list", because it did not seem a salient part. In general you can ignore parts in the "Parts list" if they seem unimportant, or add extra parts that seem important.

Also note that the ←connects→ and ←next-to→ arrows are bidirectional, i.e., have two arrows, so holds in both directions.

3. (Send us the graph) **When you've created your graph** (either on paper or electronically), **upload** a photo or PDF of your graph to **this Google Form** for us to check and process. We will approve all reasonable graphs we receive (but not spam graphs) within 30 hours.

Our participants were recruited on the Amazon Mechanical Turk platform. The workers met minimum qualification in AMT: 95% approval rate. They were from US locations and rated at Amazon's Masters Level. Workers were paid at a rate of ≈$15/hr.

C Unanimity and diversity in parts mental models

People vary greatly in how they construct mental models, but the underlying reasoning is often structurally similar i.e. in accordance with commonsense constraints (Halford, 1993; Jonassen and Henning, 1996). In our ParRoT dataset, similarly, contradictions amongst crowdworkers (e.g., for guitar, one worker annotated that the neck is part of the fingerboard, while another annotated that the fingerboard is part of the neck) are extremely rare. There are only 80 instances out of 11720 in total in our entire dataset (0.68%) – less than 1%.

We also looked at relations overlapped across workers in our dataset to analyze if workers pay attention to similar or different aspects of everyday things. To do so, we gathered a set of (p1, rln, p2) relations that are common across all 3 annotators for each everyday thing. These relationships are ones that achieved full agreement across all the 3 assigned annotators for that everyday thing in terms of the spatial/connectivity/functional relationship annotated and the parts involved. Together, we refer to this set as the ParRoT++ dataset. Table 5 summarizes the number of such high-agreement relationships for each everyday thing. Everyday things with few or no high-agreement relationships (refer Figure 7 for an example) imply higher diversity among annotators in terms of which spatial/connectivity/functional relationship and what parts they decided to include in their annotations. There are a total of 508 overlapped relations in ParRoT++, out of the 11720 in ParRoT, suggesting that attention is often paid to different aspects of everyday things.

In Table 6, we present accuracy on ParRoT++, revealing similar results for relationships that achieved full agreement across all assigned annotators. Using basic commonsense constraints, ParRoT-Con improves parts mental model accuracy significantly by 16-22% on ParRoT++. These trends are similar to that obtained for ParRoT, illustrating that the results hold across all gold-standard parts relations, regardless of whether they are more unanimous or diverse across annotators.

# full-agreem. relations	Everyday thing(s)
36	coffee maker, fish
28	rabbit
18	deer
16	egg, electric stove, tree
14	ink pen
12	laptop, sandwich, rice cooker, airplane, table
10	fire extinguisher, bird
8	elevator, flashlight, stroller, dishwasher, kayak, ship, teapot, telescope, corn, hot air balloon, microwave
6	wheelchair, barbeque grill, kite, microphone, computer, duck, helicopter
4	pillow, truck, washing machine, door, hair dryer, rocket, screw, toaster, butterfly, chair, knife, photo frame, shoe, baby bottle, bed, bird cage, car, chainsaw, electric tea kettle, humidifier, piano
2	binoculars, digital camera, zipper, apple, digital clinical thermometer, earphone, flower, windmill, backpack, dog, doorbell, lightbulb, bat, cat, umbrella, stethoscope, tent
0	air conditioner, bicycle, blender, boat, glider, guitar, house, pencil sharpener, table fan, dryer, pencil, suitcase, telephone, microscope, refrigerator, space heater, typewriter, violin, wall clock, window, bookcase, bus, cable car, calculator, saucepan, train, cow, rat, table lamp

Table 5: Number of relationships that achieved full agreement across all the 3 assigned annotators for each everyday thing. Higher number of such relations indicates more unanimous parts mental model annotations, whereas lower number reflects more diversity.

	# params	Base LM (%)	ParRoT-Con (%)	Improve (%)
GPT-3 (text-davinci-003)	175B	55.51	71.13	15.62
Macaw-11B	11B	60.04	82.41	22.38

Table 6: Comparing the accuracy of parts mental models before and after constraint reasoning on ParRoT++ dataset.

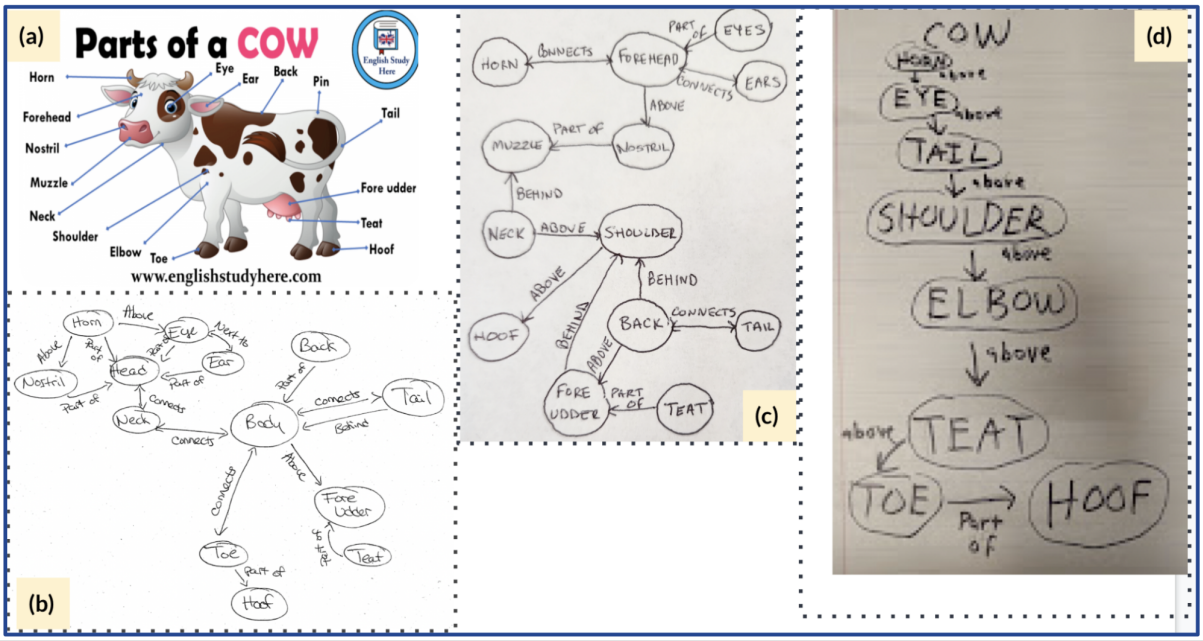


Figure 7: Example parts mental model annotations from ParRoT: (a) we provide the crowdworkers a diagram of cow retrieved from the Web. (b), (c), (d) are parts mental model sketches by 3 different crowdworkers. Note that all 3 models are accurate but there is some divergence in terms of (1) part names: e.g., 'head' vs 'forehead' and (2) which relation tuples they consider salient. Similar forms of diversity have been reported in Ji et al. (2022), for instance, as part naming divergence and segmentation divergence.

D Pictorial illustration of ParRoT-Con

Our proposed approach, ParRoT-Con, is illustrated in Figure 8 with an example everyday entity "egg".

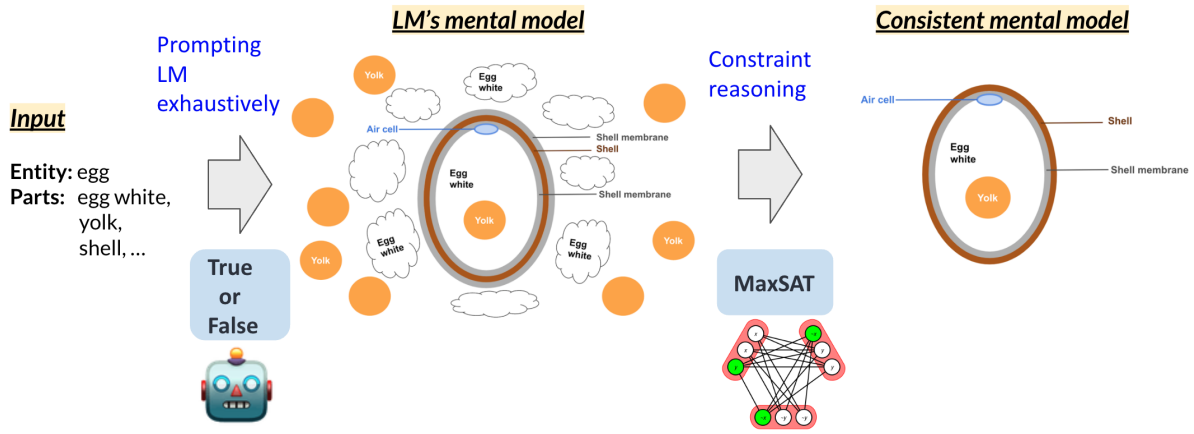


Figure 8: When asked about relationships between parts of an everyday thing, LMs can produce inconsistent relations. E.g., GPT-3 believes that in an egg, "yolk surrounds the shell" and "shell surrounds the yolk" are both True. Our proposed neuro-symbolic method, ParRoT-Con, applies constraint reasoning over raw LM predictions to produce more accurate and consistent mental models of everyday things.

E Accuracy on different everyday things

Table 7 gives example prompts and GPT-3's responses (includes both correct and incorrect) for entity "tree". Top 20 and bottom 20 everyday things that each model achieved best and worst performance on are shown in Figures 5 and 6 respectively. Further, Figure 11 demonstrates everyday things with 21st to 80th ranking in terms of the base LM accuracy.

Model	Prompt	Model's Answer
GPT-3	Judge whether this statement is true or false: In a tree, twig is directly connected to the branches.	True (correct)
GPT-3	Judge whether this statement is true or false: In a tree, trunk is above the roots.	False (incorrect)
GPT-3	Judge whether this statement is true or false: In a tree, roots are surrounded by the trunk.	True (incorrect)
GPT-3	Judge whether this statement is true or false: In a tree, trunk is below the roots.	False (correct)

Table 7: Example prompts and GPT-3's responses for an everyday entity "tree".

F Use of models for inference

For all experiments in this paper we used existing models/toolkits without any re-training or fine-tuning. We used GPT-3 text-davinci-003 and Macaw (T5-11B based) as representative LMs for our experiments. To probe GPT-3 text-davinci-003, we used their web API which took around 30 to 60 msec per relation tuple (one T/F question). To probe Macaw, we used two 48GB GPUs and it takes around 10.4 msec per relation tuple. We also run a MaxSAT solver for each everyday entity's parts mental model. To solve a constraint satisfaction problem per parts mental model takes a few msec up to around 3 minutes depending on the WCNF formula involved.

G On the use of our dataset and code

We have made all data and code used in this paper publicly available. Our dataset and code are released for research purposes only.

H FAQs

Q: Does ChatGPT do better?

From informal tests, we find that ChatGPT is not devoid of mistakes either. We provide some examples to illustrate how the lack of coherent mental models of everyday things may also appear for other models of the GPT-3.5 family, like ChatGPT in Figure 9. Others have also found ChatGPT responses that convey ridiculous interactions with everyday things e.g. it generates that "When you fry an egg, the white and the yolk are both held together by the eggshell." (See Figure 10)

Q: GPT-3 and ChatGPT models are often updated, when were the models accessed for your experiments?

In our experiments with GPT-3, we used the text-davinci-003 model and queried the API on December 16, 2022 (during the period of time between 12 PM to 3.30 PM PST). ChatGPT as in Figure 9 was accessed on December 17, 2022 (at around 9.30 PM PST). It would be interesting for researchers to investigate if future versions of the systems can construct better parts mental models of everyday things.

Q: How do you ensure high-quality mental models are acquired via crowdsourcing?

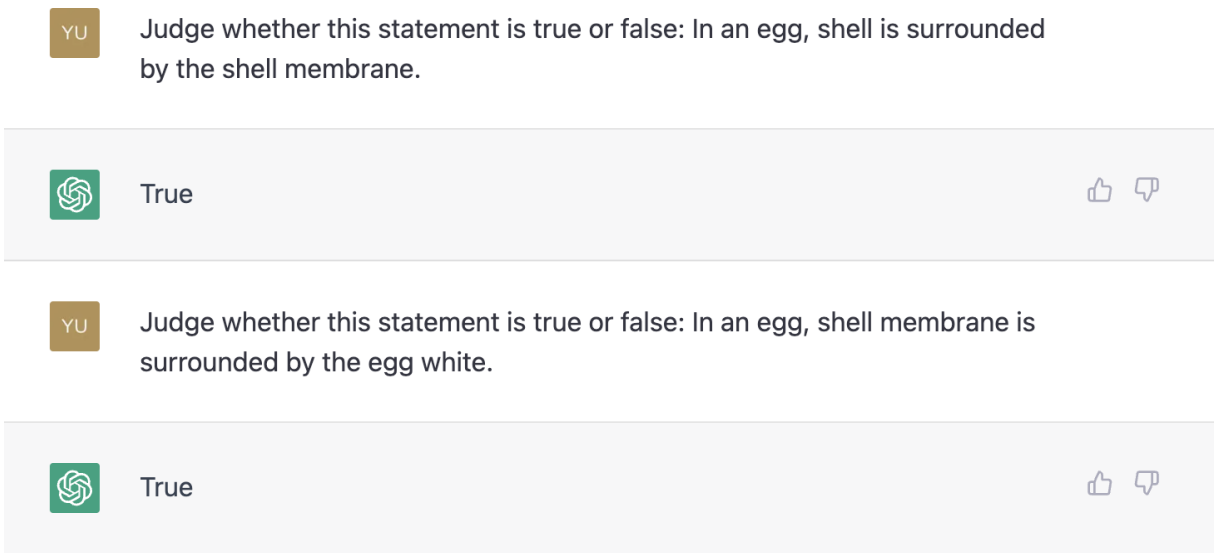
We enforced a set of manual and automated checks during data acquisition which includes collecting mental model sketches and transcribing them into relation tuples.
Manual checks: We randomly sampled 15 mental model sketches and made sure that the transcription of relation tuples was accurate i.e. all the relations tuples in mental model sketches drawn by crowdworkers were precisely added to our dataset. We also checked the quality and format of sketches ('.png' files) which will be released with our dataset.
Automated checks: After enriching with implied relations, we also programatically checked that all individual mental models (total of 11.7K relations) in ParRoT are fully consistent (based on the 4 commonsense constraints described in Section 5.2).

Q: Do similar trends apply to smaller models?

Experiments on Macaw-3B, Macaw-large, UnifiedQA-large pointed towards the same trends. We also make our code and data fully accessible at `https://github.com/allenai/everyday-things` for interested researchers to experiment with other models of interest to them.

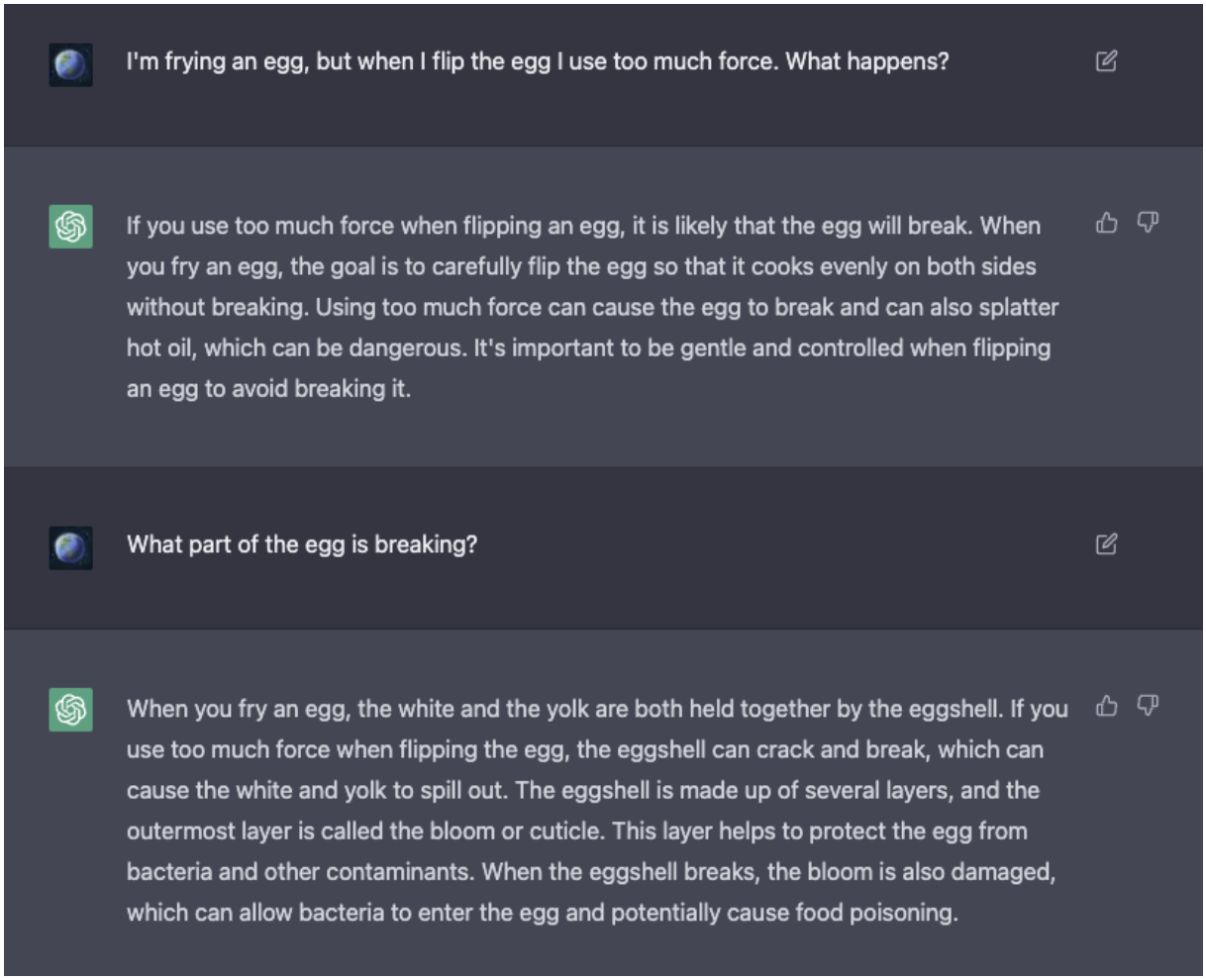
Q: Can ParRoT-Con be applied to other languages?

While our dataset is in English, relationships between parts of everyday things could indeed be authored for/ translated into other languages. We made our code and data publicly available, so others could use the infrastructure to apply the technique to other languages.



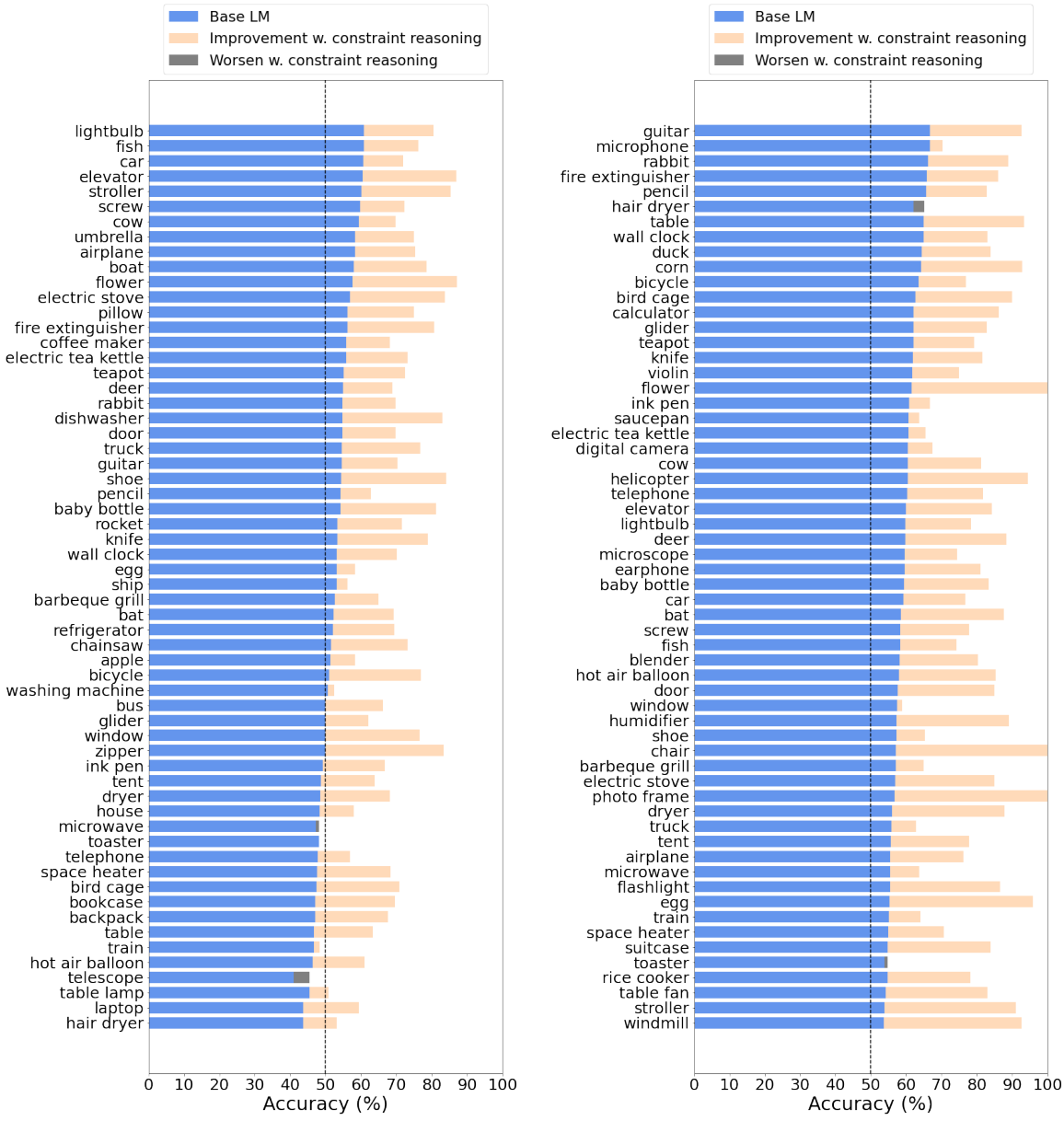
ChatGPT

Figure 9: Like GPT-3 (text-davinci-003), ChatGPT also seems to have incoherent mental pictures of everyday things.



ChatGPT

Figure 10: ChatGPT provides ridiculous responses regarding daily life activities such as frying an egg, illustrating poor mental models of everyday things and interactions with them. (Example by @bio_bootloader, posted on Twitter https://twitter.com/bio_bootloader/status/1599131249553330176/photo/1 at 11:59 AM Dec 3, 2022.)



(a) GPT-3

(b) Macaw-11B

Figure 11: Performance on other everyday things. Accuracy of base LM and improvement achieved through constraint reasoning on different everyday things in our dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, we discussed the limitations of our work in the "Limitations" section.

☑ A2. Did you discuss any potential risks of your work?
Yes, we discussed the potential risks of our work in the "Ethics Statement" section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, the abstract at the start and section 1 introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
No.

B ☑ Did you use or create scientific artifacts?

Section 4 provides details on the dataset we created. Section 5 discusses how we use existing language models.

☑ B1. Did you cite the creators of artifacts you used?
Yes, we cited the models used in Section 5.1. We explained who helped with the creation of the dataset (Section 4 and Appendix B on crowdworkers and instructions given to them).

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

☑ 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 G. "Our dataset and code are released for research purposes only. "

☑ 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?
"Ethics Statement" section discusses that we removed any personally identifiable information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We provide details on domain of our data (Section 4), crowdworker demographics (Appendix B on crowdworkers)

☑ 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 4.3

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

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Results table in Section 6. Appendix F.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 discusses the experimental setup in detail. But no hyperparameter search is needed for our purposes.

☑ 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?
Section 4 dataset statistics and Section 6 results.

☐ 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.)?
Not applicable. Left blank.

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

Section 4 and Appendix B on crowdworkers and instructions given to them.

☑ 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 B on crowdworkers and instructions given to them.

☑ 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 B on crowdworkers.

☑ 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 B. We explained why are we collecting this data and how the data would be used.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

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