RICA: Evaluating Robust Inference Capabilities
Based on Commonsense Axioms

Pei Zhou   Rahul Khanna   Seyeon Lee   Bill Yuchen Lin   Daniel Ho
Jay Pujara   Xiang Ren
Department of Computer Science and Information Sciences Institute
University of Southern California
{peiz,rahulkha,seyeonle,yuchen.lin,hsiaotuh,jpujara,xiangren}@usc.edu

Abstract

Pre-trained language models (PTLMs) have achieved impressive performance on commonsense inference benchmarks, but their ability to employ commonsense to make robust inferences, which is crucial for effective communications with humans, is debated. In the pursuit of advancing fluid human-AI communication, we propose a new challenge, RICA: **R**obust **I**nference using **C**ommonsense **A**xioms, that evaluates robust commonsense inference despite textual perturbations. To generate data for this challenge, we develop a systematic and scalable procedure using commonsense knowledge bases and probe PTLMs across two different evaluation settings. Extensive experiments on our generated probe sets with more than 10k statements show that PTLMs perform no better than random guessing on the zero-shot setting, are heavily impacted by statistical biases, and are not robust to perturbation attacks. We also find that fine-tuning on similar statements offer limited gains, as PTLMs still fail to generalize to unseen inferences. Our new large-scale benchmark exposes a significant gap between PTLMs and human-level language understanding and offers a new challenge for PTLMs to demonstrate commonsense.[1]

1 Introduction

Smooth and effective communication requires the ability to make various forms of commonsense inferences (Clark and Brennan, 1991). When a friend texts, "*I'm going to perform in front of thousands tomorrow,*" you may reply reassuringly, "*Deep breaths, you'll do great!*" Implicit to this communication is a commonsense logical inference that a person performing in front of a crowd may feel anxious, and that a reassuring remark helps ease anxiety (Figure 1). A growing body of literature (Bosselut et al., 2019; Petroni et al., 2019)
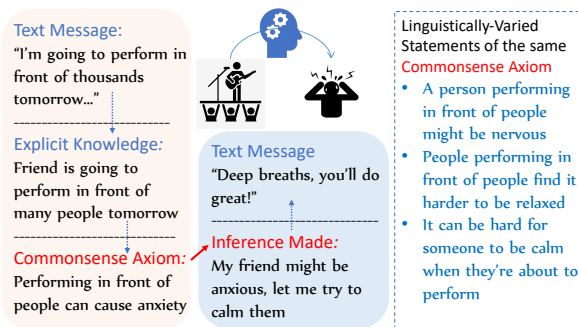

Figure 1: Human communication requires commonsense inferences. RICA evaluates such inferences via commonsense axioms with many linguistic variations.

shows pre-trained language models (PTLMs) are able to catalog the types of commonsense relationships necessary for fluid communication. However, as we show in this paper, PTLMs have a shocking inability to leverage such commonsense knowledge to make robust inferences.

Here we focus on two specific characteristics crucial to human-AI communications: (1) combining commonsense knowledge with information expressed in natural language to make inferences and (2) producing consistent inferences amidst logically-equivalent yet linguistically-varied paraphrases. We focus on *commonsense axioms*, such as "*Performing in front of people can cause anxiety*", and exploit the flexibility of language to express the same axiom in many forms — *e.g.*, "*Performing in front of people makes it hard to stay calm.*" We test these characteristics by generating self-contained commonsense statements involving novel entities ("*Prindag is going to perform in front of a crowd, so prindag is more likely to feel nervous.*") and adapt them to two evaluation settings.

Unfortunately, these two capabilities have largely been overlooked by existing natural language inference (NLI) benchmarks (Williams et al., 2018) and knowledge probing studies for transformer-based PTLMs (Vaswani et al., 2017;

---

[1]Our code and data are public at https://sites.google.com/usc.edu/rica.

Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020; Petroni et al., 2019). Most existing commonsense reasoning-focused datasets (Zhang et al., 2017; Williams et al., 2018; Ostermann et al., 2019; Zhou et al., 2021a; Talmor et al., 2019) do not systematically evaluate robustness against linguistic variations, meaning we cannot preclude the possibility that models are learning spurious patterns to solve the needed task.

To fill this gap, we introduce RICA, a challenge to evaluate a model's **R**obust **I**nference using **C**ommonsense **A**xioms in English. RICA draws on linguistic and cognitive science research (Schank and Abelson, 1977; Alshawi and van Eijck, 1989) suggesting humans translate language to logical representations and reason using these abstract representations. RICA consists of a set of natural language statements in the "premise-conclusion" format that require reasoning using latent (implicit) commonsense relationships. We formulate these abstract commonsense relations between entities in first-order logic and refer to them as *commonsense axioms* (see Fig. 1). To insulate from PTLM biases and test human-like acquisition ability on new words (Carey and Bartlett, 1978), RICA uses *novel entities*, which are unseen strings used to ground axioms into natural language. Finally, we introduce a set of *linguistic perturbations* that paraphrase a commonsense axiom into natural language in various forms.

Each component of RICA is generalizable, providing a systematic procedure to generate myriad commonsense statements. In this paper, we generate 257k commonsense statements capturing 43k axioms comprising different types of commonsense, such as physical, material, and social properties. To demonstrate the quality of RICA, we create a manually-curated set of 1.6k probes based on commonsense axioms, and also undertake a large-scale, crowdsourced verification of 10k generated statements with multiple human annotators. RICA is built by leveraging existing commonsense knowledge bases such as ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019a) to support easy expansion. Furthermore, RICA's statements can be posed as popular PTLM tasks such as masked word prediction or sentence probability, making our benchmark widely applicable. RICA provides an extensible platform for evaluating commonsense reasoning in a variety of PTLMs.
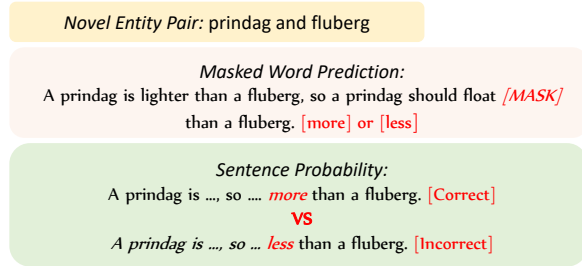

Figure 2: Illustration of two evaluation settings with a pair of novel entities used by RICA probes.

When evaluating state-of-the-art transformer-based PTLMs on the RICA probes following a *zero-shot* setting (*e.g.*, predicting "*more*" vs. "*less*" in the first example in Fig. 2), we consistently discover their performance is on par with *random guessing*. Even after fine-tuning with large amounts of labeled examples, PTLMs exhibit a significant gap relative to human performance. We drill down into this finding through (1) zero-shot, (2) low-resource, (3) high-resource, and (4) noisy training settings and find that even with appreciable performance gains on automatically generated probes in high resource settings, PTLMs still remain on par with random guessing on difficult, human-curated RICA probes. To better understand these results, we identify a pervasive intrinsic bias in PTLMs that demonstrates positivity bias in human languages (Dodds et al., 2015).

2 The RICA Challenge

The RICA challenge is posed as a set of textual statements (sentences), each expressing a latent commonsense relationship in the "premise-conclusion" format (see Stage 5 in Fig. 3 for examples). These statements use generated novel entities such as "*prindag*" and "*fluberg*" instead of real-world entities such as "*thimble*" and "*elephant*" to separate factual recalling from reasoning. Each statement can be viewed as an instantiation of a commonsense principle, such as "*smaller objects cannot contain larger objects.*"

We express these commonsense principles in first-order logic, further generalizing statements through the use of general predicates for object properties (*e.g.*, size) and object-object relations (*e.g.*, containment). We turn these logical formulae into the associated textual statements using a set of perturbation operators and a conversion module, which together produce a logically-equivalent set of commonsense statements. In the rest of this sec-
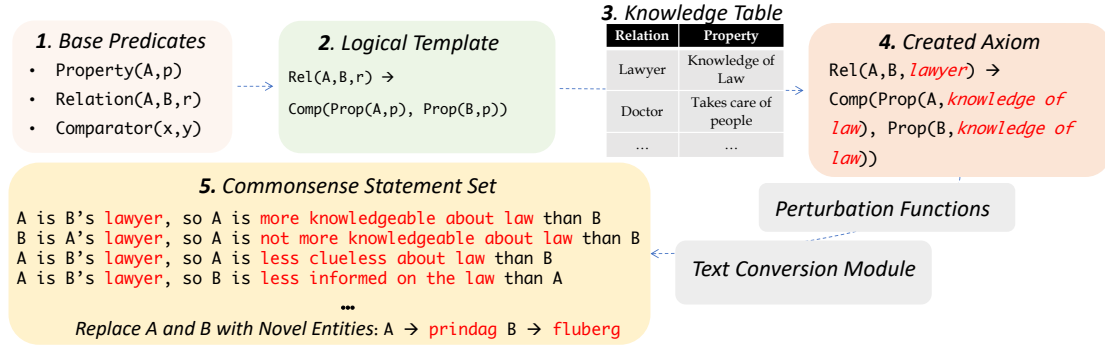
Figure 3: **Overview of the workflow of our statement construction process.** The output is a set linguistically-diverse of masked sentences that follow the same reasoning template.

TERMINOLOGY	Description
Logical Template (LT)	General FOL formula constructed from predicates and logical connectives
Arguments	Specific entities and relations to fill predicates in LTs
Axiom	Commonsense relationship expressed in FOL by filling a LT with arguments
Commonsense Statement	Natural language sentence after converting an axiom using a TT
Statement Set	Statements that inform the same axiom after applying perturbations
Evaluation Instances/Probe	A set of statements after adapting to an evaluation task

Table 1: Description of terminology used in RICA.

tion, we first provide a formal definition of RICA challenge, then provide a detailed description of the statement construction process.

2.1 Challenge Formulation

Formally, we define a commonsense axiom a_i, expressed via a first-order-logic (FOL) formula, as a relationship between entities that can be inferred using commonsense knowledge (see Stage 4 in Fig. 3). To test whether PTLMs understand an axiom a_i, as well as examine their robustness to linguistic variations, we instantiate the axiom a_i by a set of m syntactically-different commonsense statements $\{s_1^i, s_2^i, ..., s_m^i\}$, each expressing the *same* logic as the axiom. Each statement takes the form of an inferential implication with a premise and conclusion. Finally, depending on the PLTM, we select an appropriate task (Section 3), transform each statement in the set into its task-specific *probe*, and evaluate how well the PTLM can leverage the logic of a_i to solve each of a_i's corresponding probes. We deem a model "successful" on the challenge (or, understands the axioms) only if it can perform like humans on all probes of the axioms.

2.2 Statement Set Construction Process

This subsection introduces our proposed procedure for the construction of commonsense inference statement sets for the challenge. A list of terminologies and descriptions can be found in Table 1 and an overview of our workflow is shown in Figure 3.

Stage 1. Define Predicates. In FOL, predicates are used to denote a property of objects or a relation between objects and every predicate symbol comes with an arity larger or equal to 1. We define three general high-level predicates that serve as the backbone for the logical formulations of our axioms: *Property*, *Comparator* and *Relation*. (1) $\textbf{PROP}(A, p)$ represents that entity A has a certain property p. "$\text{PROP}(A, glass)$" indicates that A is made of glass. (2) $\textbf{REL}(A, B, r)$ represents that A and B have a certain relation r. "$\text{REL}(A, B, lawyer)$" indicates that A is B's lawyer. (3) $\textbf{COMP}(x, y)$ represents a comparative relationship between values x and y, where "COMP" will be replaced with comparison words like "better," "more," or "easier." We will later define multiple sub-types of these predicates to crawl from Knowledge Bases (KBs) to ensure a wide coverage of common knowledge.

Stage 2. Compose Logical Templates. We manually create first-order logical formulae, referred to as *logical templates* (LT), using the predicates defined in Stage 1. Each formula takes the form of an implication, expressing an inference based on commonsense knowledge. For example, $\text{REL}(A, B, r) \rightarrow \text{COMP}(\text{PROP}(A, p), \text{PROP}(B, p))$ expresses a logical inference comparing a property of two entities, A and B, based on a relationship between them. An instantiated version of this template can be $\text{REL}(A, B, lawyer) \rightarrow \text{MORE}(\text{PROP}(A, know\_law), \text{PROP}(B, know\_law))$.

Stage 3. Populating Knowledge Tables. Materializing the abstract relationships in a logical template requires connecting abstract logic to commonsense knowledge. We define a structure called *knowledge table* (KT) that contains valid arguments to populate a specific LT and form a FOL representation of the axiom. KTs are generated by crawlining commonsense KBs such as ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019a). The first step of the crawling process is to narrow down the predicates to specific types. For example, PROP is general enough to capture an entity's capabilities (e.g., *knowledge of law*) or its intrinsic properties (e.g., *hardness*). We pre-define several type constraints for both properties (PROP) and relations (REL). For PROP, we consider *Capability, Attribute, and Condition*. For REL, we consider *Role and Action*. Note that these categories can be extended for wider coverage of knowledge and allow our LTs to be adapted to a broader range of KB schemas. After specifying type constraints, we specify steps for crawling the arguments either from commonsense KBs such as Concept and ATOMIC or general web KB such as Wikipedia. In our example in Fig. 3, we can crawl occupations from Wikipedia, and then query ConceptNet for triples with the occupation as the subject and *CapableOf* as the relationship to create a KT with professions and capabilities. We show all crawling steps for KTs in Appendix A.

Stage 4. Creating Axioms. Combining knowledge tables and logical templates allows us to generate commonsense axioms at scale, which are partially-filled LT formulae. For example in Fig. 3 Stage 3, the arguments of predicates REL, PROP, and COMP are set in order to reflect the commonsense relationship between *lawyer* and *knowledge of law*, while leaving the entities A and B ungrounded. Once the predicates are instantiated, we call this partially-filled LT a *commonsense axiom*.

Stage 5. Generate Statement Sets. After filling the logical templates, each partially-filled LT represents one commonsense axiom. To comprehensively challenge models' understanding of an abstract axiom, we construct a *statement set* expressing the *same axiom* with *different phrasings*, i.e., logically-equivalent yet linguistically-varied. We define several *perturbations* to apply on the *arguments* from knowledge tables.

(1) *Linguistic Operators.* We define seven types of linguistic operators to facilitate and formalize

LINGUISTIC OPERATOR	EXAMPLE
NEGATION	NEG(fit into) = not fit into
ANTONYM	ANT(fit into) = contain
PARAPHRASE	PARA(fit into) = put into
PARAPHRASE INVERSION	PARA(ANT(fit into)) = Para(contain) = hold inside
NEGATION ANTONYM	NEG(ANT(fit into)) = NEG(contain) = not contain
NEGATION PARAPHRASE	NEG(PARA(fit into)) = NEG(put into) = not put into
NEGATION PARA_INV	NEG(PARA(ANT(fit into))) = NEG(PARA(contain))= NEG(hold inside) =not hold inside

Table 2: Linguistic operators, logic, and examples.

perturbations, shown in Table 2. We construct the last four operators by combining some of the single operators listed in the first three rows. Note that for NEGATION, ANTONYM, PARAPHRASE INVERSION, and NEGATION PARAPHRASE types, the logic of the original phrase is changed, so words in the statements have to be changed accordingly. For example, if we apply ANTONYM to "*fit into*" in the probe "*A is smaller than B, so A is more likely to fit into B*," we will get "*A is smaller than B, so A is less likely to contain B*." (2) *Asymmetry Operator.* Most of our logical templates use several strongly-ordered comparisons and relationships allowing us to introduce asymmetries that preserve meaning. For example, $\text{MORE}(A, B) \rightarrow \neg\text{MORE}(B, A)$ and $\text{REL}(A, B, parent) \rightarrow \neg\text{REL}(B, A, parent)$. Using this invariant, we can swap the positions of two entities for these predicates and the logic will also be negated, so we denote this perturbation as $\text{ASYM}(\text{P}(A, B)) \rightarrow \text{P}(B, A) = \neg\text{P}(A, B)$.

We apply the defined operators to the arguments in the predicates to first form a set of partially-filled LTs (axioms) and use for a conversion module to convert axioms to statements with diverse perturbations. In practice, this module can be a sequence-to-sequence (seq2seq) model (that takes in FOL and outputs natural language text), or human-written templates. Finally, commonsense axioms are general logical relationships that hold for all entities. To formulate specific commonsense statements, we generate specific *novel entities*. These entities are randomly generated character strings from length 3 to 12 that are not seen in the training data of the PTLMs. Using novel entities enables us to avoid conflating fact-based recall with commonsense reasoning when evaluating PTLMs.

7563

LT1: `Prop(A,p)∧Prop(B,q)→Comp(Prop(A,m),Prop(B,m))`
E.g.: A is made of glass, B is made of stone, so A is less opaque than B
LT2: `Rel(A,B,r)→ Comp(Prop(A,m),Prop(B,m))`
E.g.: A is B's priest, so A spends more time praying than B
LT3: `Prop(A,p)∧¬Prop(B,p)→Comp(Prop(A,m),Prop(B,m))`
E.g.: A makes the varsity team but not B, so A is more skilled than B
LT4: `Comp(Prop(A,m),Prop(B,m))→Comp(Prop(A,n),Prop(B,n))`
E.g.: A is able to concentrate more than B , so A is more effective than B
LT5: `Prop(A,p)→ Comp(Prop(A,m),Prop(B,m))`
E.g.: A turned on the heater, so A was cold before turning on the heater

Table 3: Example first-order logical templates we construct for our probes and an example for each template.

3 Experiment Setup

3.1 Probing Tasks

To examine transformer-based PTLMs' performance on RICA challenge, we draw conclusions from evaluation results on two distinct probing tasks shown in Figure 2, described as follows.

Masked Word Prediction (MWP) Inspired by the masked word prediction objective in BERT (Devlin et al., 2019), we examine if the models can recover masked-out keywords in the statement given the remaining context. Since RICA's statements take the form of implications, we mask words in the consequent to evaluate the inference performance, given the premise. Specifically, we choose to mask the comparative words (from COMP) such as '*more/less*" and "*better/worse*" as shown in Figure 2, since they not only capture the commonsense relationship, but also focus on masking positions where only a few options are appropriate logically and syntactically.

Sentence Probability (SP) evaluates if PTLMs assign higher probability for statements that express commonsense axioms versus contradictory statements. RICA statements are input to PTLMs, computing SP by taking the product of each word's probability conditioned on previous words, *i.e.*, the left-to-right language modeling loss. For each RICA statement, we pair it with an incorrect (non-commonsense) statement by swapping the comparative word (*i.e.*, the masked word in MWP) with its opposite word, as shown in Figure 2.

3.2 Probing Data Details

Raw Set Following the process in Section 2, we use the three high-level predicates to generate five LTs as shown in Table 3. Then we construct knowledge tables to fill in each template by crawling from two commonsense KBs: ConceptNet (Liu

and Singh, 2004) and ATOMIC (Sap et al., 2019a). Specifically, for each LT, we design 1 to 4 crawling strategies based on the type constraints we impose on the predicates so that it covers multiple aspects of commonsense knowledge (for all strategies please see Table 4 in Appendix A). For example, the example shown for LT1 in Table 3 is about inference of physical properties based on the material of two objects as we constrain PROP in the premise to be materials. However, we can also constrain PROP in the premise to be animals so that we can use the same template to examine inference of properties based on the animal types of A and B, e.g., "A is a fish, B is a horse, so A is more likely to be in the bottom of the sea than B."

We have 11 type-constrained LTs and we populate the KTs using 11 human-designed crawling strategies shown in Appendix A, resulting in around 43k axioms. Then we apply the perturbation operators as described before to form a set of 257k perturbed axioms. For this large set, we apply negation and asymmetry operators automatically by adding negation and switching the order of entities. To convert FOL axioms to text, we train a seq2seq model based on BART (Lewis et al., 2020a) on 200 manually converted axiom-text pairs covering each type-constrained LT and each perturbation type. Finally, we replace entities to unseen entities to form a set of 257k commonsense statements.

Quality Check. To check for language quality of the generated probes from BART, we randomly sample 5% of the 10k set and ask a native English speaker to check for the naturalness. We found that only *4 out of 500* (0.8%) probes contain grammar or fluency issues. Since all probes follow a premise-conclusion format, we find that using 200 pairs of first-order logic (FOL) and aligned text for fine-tuning BART is sufficient to convert FOL into text, both from our manual inspection and the crowdsourcing verification of the generated probes. We tried increasing the training set size and didn't observe a clear difference in quality.

Human-Verified Set To ensure the quality of crawled data, we conduct human evaluation using Amazon Mechanical Turk (AMT) on 10k of our collected 257k statements covering 1.7k different commonsense axioms. We present a pair of statements by flipping the comparative term in the original statement to its opposite, and ask two annotators to choose the one that follows commonsense. If they disagree, we subject the pairs to
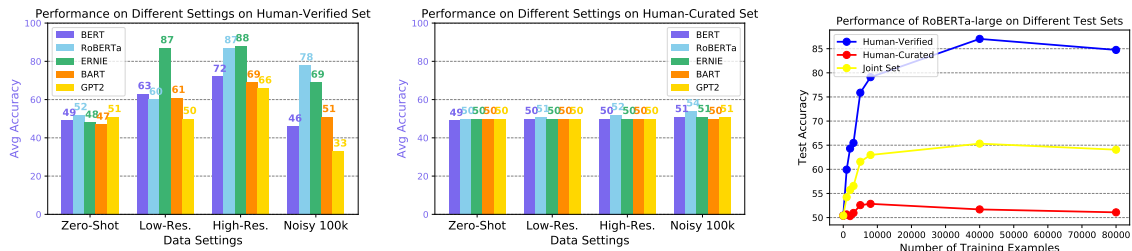
(a) Performance on Human-Verified Set (b) Performance on Human-Curated Set (c) Fine-tuning curve for RoBERTa-large

Figure 4: Performance of different transformer-based models on different settings of our data. BERT, RoBERTa, ERNIE, and BART are evaluated using masked word prediction and GPT2 is evaluated using sentence probability. Zero-shot performance is no better than random guessing. More data helps greatly for human-verified test set (10k) although noisy training hinders the improvement. Increasing data does not help at all for our human-curated set.

a second round of turking with three annotators, and use majority voting to validate the statement. When annotators prefer a flipped statement over the original, we replace the statement accordingly.

Quality Check. The Fleiss-kappa agreement (Fleiss, 1971) on the two rounds of turking is 0.72 and 0.52, indicating that some statements are difficult for humans to verify. Of 10k statements in the verified set, we sample 10% (1k with 170 axioms) that 2 annotators agree on in the first or second round to form our *Human-Verified Test Set*.

Human-Curated Set To further challenge models on more flexible forms of text, we ask humans to write perturbed axioms. Specifically, given an axiom in FOL, a human annotator is asked to provide input that perturbs the conclusions following *all* 7 types of linguistic perturbations as shown in Table 2, including compositional combinations that are hard to generate using automated approaches, Then we apply the asymmetry operators either on the premises or conclusions. Thus we have in total of 24 types of perturbations, including the unperturbed one. We focus on 80 axioms covering physical, social, and temporal commonsense knowledge and create 1.6k statements. We show examples of all perturbations for one probe in Appendix Table 7 and sampled 60 probes in Appendix Table 8.

Joint Test Set Combines the Human-Curated and Human-Verified sets, for a total of 2.6k statements.

3.3 Evaluation Settings

Using the collected probe data introduced above, we consider four evaluation settings to examine models' capabilities to perform robust inference on our dataset.

1. Zero-Shot: In the zero-shot setting, we test models without any exposure to training data.

2. Low-Resource: For the low-resource setting, we fine-tune the models on 1k (10%) of the verified 10k set to determine how a small amount of in-domain text influences PTLM performance.

3. High-Resource: We use 90% of the verified training set (8k for training, 1k for validation). We further increase the number of training instances by introducing 5 different novel entities for each statement, yielding 40k training instances that include 5 repetitions of each probe with different novel entities, providing models more opportunities to learn patterns in the training set.

4. Raw Large-Scale Training: Finally, to analyze the effects of training on an even larger but noisier set with the similar format. Starting from the raw set of 257k crawled statements, we sample 100k statements from 17k axioms ensuring no overlap with the test set.

3.4 Baseline Methods

We evaluate multiple state-of-the-art transformer-based PTLMs covering both masked and generative language models. For the masked word prediction task, we consider BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ERNIE, a knowledge enhanced LM (Zhang et al., 2019), and BART (Lewis et al., 2020b). For sentence probability, we consider GPT-2 (Radford et al.), a unidirectional language model for left-to-right language generation.

4 Results and Analysis

We examine the performance of multiple language models on each evaluation setting on our probe data, including zero-shot and fine-tuning on various splits, and present ablation studies to analyze

(a) Results on Positivity Bias   (b) Ablation on Novel Entities   (c) Results per Linguistic Perturbation
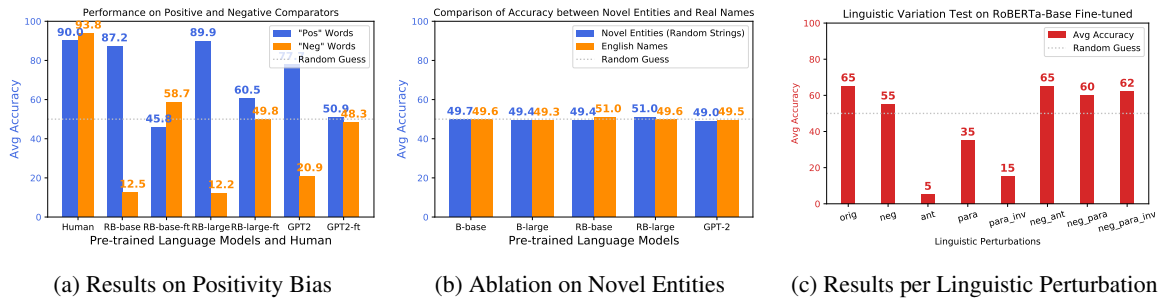
Figure 5: Results of fine-tuning and the ablation study on novel entities. Shows that (a) models are biased to positive words, requiring fine-tuning to correct (b) poor performance persists after replacing novel entities with real names—indicating the use of random strings is not hindering PTLMs' abilities, (c) fine-tuning mitigates the bias towards positive words, but the inconsistency issue for linguistic variation become obvious.

4.1 Zero-Shot Performance

As shown in the first group of bars in Figures 4a and 4b, the average binary accuracies of all five models (we show the large version) on both MWP and SP tasks are around 0.5, regardless of the test data. A random baseline that chooses between the two comparative words would have an accuracy of 0.5. This shows that the tested models barely beat a random guessing baseline without training.

Is Knowledge-Augmented Model Better? To see if adding knowledge during training helps, we also test a knowledge-enhanced LM, ERNIE (Zhang et al., 2019). However, as we can see in Figures 4a and 4b, ERNIE also performs on par with random guessing, demonstrating that simply adding more knowledge does not help with the robust inference capability.

Human Performance To benchmark human performance, we sampled 5% of our joint test set consisting of both human-verified and human-curated data and gathered answers from 20 subjects (annotators) with diverse backgrounds who were not involved in the probe construction process. We consider this as zero-shot testing for humans as they have not seen the training set before. Humans obtained 91.7% accuracy, taking a majority vote for each probe, with a substantial inter-annotator agreement of 0.768 Kappa (Cohen, 1960)).

4.2 Fine-tuning Performance

To study if poor performance in §4.1 is from a lack of exposure to RICA's probe sets, we conduct experiments to fine-tune baseline language models. As in §3.3, we consider training on low-resource

data by sampling a subset of the verified set, on high-resource by filling multiple novel entities in the verified set, and the noisy 100k data. We fine-tune BERT, RoBERTa, ERNIE, and BART using the same masking approach as MWP evaluation, and fine-tune GPT-2 on the causal language modeling task. Details for training are in the appendix.

More Data Helps on Human-Verified Set Figure 4a shows fine-tuning on our probe set helps the model on the human-verified set, especially for RoBERTa and ERNIE, where the high-resource setting almost reaches 90% accuracy. This demonstrates with enough data, PTLMs are able to reach near-human performance on generated axioms. The low-resource (except for ERNIE) and noisy training settings, however, pose an enduring challenge for most models.

Diversity of Curated Set Stumps All. Evaluating models fine-tuned on human-verified data on the human-curated set, where human editors provide greater diversity in probes, tells a different story. The model accuracy (Figure 4b) remains near 50%, on par with random guessing, for all models in all settings. This indicates that exposing these models to numerous linguistically similar sentences does not improve robust inference ability. Furthermore, we evaluate training data sensitivity for both the human-verified and human-curated set (Figure 4c). We vary training set size from 0 to 80k for RoBERTa-large. Our results show that performance on the human-verified set saturates around 80% accuracy after 10k instances, but human-curated accuracy remains close to 50% throughout. This casts doubt on the model's generalizability and whether the improved performance may be due to pattern-matching seq2seq generation, not commonsense acquisition. An inability to im-

prove on reasoning tasks after fine-tuning supports the challenging nature of RICA, which cannot be trivially solved by fine-tuning.

4.3 Performance Analysis

Positivity Bias in PTLMs. We find a pattern that when PTLMs are asked to infer a comparative relationship between the property of two entities, the model is heavily biased towards predicting words that evoke positive emotions (positive valence) regardless of what commonsense axiom is embedded in the statement. Figure 5a shows that the accuracy for "*positive valence*" words such as "*more*" and "*easier*" is much higher than "*negative valence*" words such as "*less*" and "*harder*". Fine-tuning on our probes, which have a balanced number of sentences containing positive and negative comparatives, helps mitigate this bias for RoBERTa (base and large) and GPT-2. We conjecture that this may be due to the frequency difference between positive valence words and negative valence words related to reporting bias in language (Gordon and Van Durme, 2013). Dodds et al. (2015) shows a universal positivety bias in human languages and to check if our comparators also possess it, we use Google Ngram Viewer [2] to find frequencies for the masked words, and confirm that the positive valence words are about 5 times more frequent than their negative counterparts. This correlation supports the claim that PTLMs do not reason as humans do, but are guided by statistical patterns. Our challenge clearly reveals this bias in PTLMs and suggests a potential mitigation using RICA.

Ablation of Novel Entities In order to ensure novel entities used in RICA did not impact PTLM performance, we conducted an ablation study on 4,800 of our human-curated set (each statement is repeated for 3 times). These probes involved social commonsense, where novel entities took the place of names. We conduct an ablation by choosing common names instead of novel entities, producing probes containing only previously-seen words. As Figure 5b shows, the performance of all models in three settings did not change significantly, strongly suggesting that novel entities are not critical to PTLM performance. We conclude novel entities do not introduce helpful or distracting sub-words.

Impact of Linguistic Perturbations Before fine-tuning, a heavy bias for positive valence words

interfered with the perturbations analysis, since each perturbation has a balanced number of positive and negative valence words. After fine-tuning, however, the bias is mitigated and we find significant variations in performance for different perturbation types (Figure 5c). This shows that language variation greatly affects a model's capability to make inference on our commonsense probes, while suggesting models do not comprehend the axioms. Interestingly, the composite perturbation types such as NEGATION ANTONYM are not necessarily harder for PTLMs, even though performance on ANTONYM is the lowest. We speculate that the model is exploiting some pattern in NEGATION ANTONYM that is not present for just ANTONYM.

5 Related Work

Commonsense Reasoning has a long history in AI, with classical work primarily focusing on executing symbolic rules as hand-crafted programs for machines to learn (Mccarthy, 1960). The majority of recent commonsense reasoning benchmarks (Zellers et al., 2018; Talmor et al., 2019; Bisk et al., 2020; Sap et al., 2019b; Lin et al., 2021c,a,b; Sakaguchi et al., 2020) test a model's ability to choose the correct option given a context and a question; PTLMs have reached high performance on these benchmarks after fine-tuning. We differ from these benchmarks by focusing on robustness to linguistic variation via our linguistically-varied commonsense statements. RICA also challenges PTLMs on two evaluation tasks to better probe the PTLMs' representations.

Reasoning-focused Inference There have been many benchmarks that focus on reasoning abilities in multiple tasks such as reading comprehension (Huang et al., 2019; Yu et al., 2020), dialogue systems (Cui et al., 2020), and NLI (Williams et al., 2018), that involve inferences on language. Recent work also aims to probe models in these tasks to see if reasoning is actually achieved (Richardson and Sabharwal, 2020; Richardson et al., 2020).

Robustness to Linguistic Variations Previous work has also examined model robustness against paraphrases by producing linguistically-varied sentences for different tasks such as NLI (Liu et al., 2020; Li et al., 2020), question answering (Weller et al., 2020), and sentiment analysis (Ribeiro et al., 2018), just to name a few. Our work distinguishes from them as we look into robustness in regard to commonsense reasoning and develops a systematic

---

[2] https://books.google.com/ngrams

procedure to generate paraphrases using first-order logic.

Probing PTLMs Prior works in analyzing the (commonsense) reasoning ability of PTLMs have primarily focused on creating probing tasks by generating ad-hoc masked sentences either from knowledge bases (Petroni et al., 2019; Davison et al., 2019; Lin et al., 2020) or existing datasets (Zhou et al., 2020; Talmor et al., 2020; Kwon et al., 2019; Zhou et al., 2021b). This first line of works aim to test if PTLMs can work as knowledge bases, i.e. can they retrieve factual knowledge; our work focuses on implicit commonsense relations, not facts. We differ from the second line of work by proposing a systematic procedure to generate probes and evaluate for robustness. Clark et al. (2020) shows that PTLMs can emulate deductive reasoning given explicit rules, but we focus on unstated commonsense relations.

6 Conclusion

We design RICA as an AI challenge to test robust inference capabilities on linguistically-varied probes covering different commonsense axioms. RICA is built on a systematic process to construct probes using FOL formulae, perturbation operators, and novel entities. Following this approach, we generate and verify more than 10k statements from 1.7k axioms and test multiple PTLMs in various settings. We find that PTLMs perform on par with random guessing on zero-shot setting, have strong positivity bias, and are not robust under linguistic perturbations.

Acknowledgments

Ethical Considerations

Our work aims to pose a new challenge to improve effective human-AI communications by collecting new data in English, which benefits English speakers more. We have conducted human evaluation using Amazon Mechanical Turks. We pay turkers around $11 per hour, above the national minimum wage and engage in constructive discussions if they have concerns about the process. We also give each annotation instance enough time so that we do not pressure annotators.

Our data construction process makes uses of available public resources: Wikipedia, Concept-Net (Liu and Singh, 2004), and ATOMIC (Sap et al., 2019a), which could contain societal biases as shown by Mehrabi et al. (2021). Although our probes do not involve specific demographics, we admit the possibility that biases in knowledge resources are included in our data. We have provided detailed descriptions about our data construction process to minimize potential confusions.

References

Hiyan Alshawi and Jan van Eijck. 1989. Logical forms in the core language engine. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890. ijcai.org.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 1406–1416, Online. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pre-trained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

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

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Sunjae Kwon, Cheongwoong Kang, Jiyeon Han, and Jaesik Choi. 2019. Why do masked neural language models still need common sense knowledge? *ArXiv preprint*, abs/1911.03024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (lit): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. 2021b. Differentiable open-ended commonsense reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4611–4625, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021c. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515, Online. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608.

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

John W. Mccarthy. 1960. Programs with common sense.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *ArXiv preprint*, abs/2103.11320.

Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark. 2019. Commonsense inference in natural language processing (COIN) - shared task report. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 66–74, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog 1.8 (2019): 9.*

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *AAAI*, pages 8713–8721.

Kyle Richardson and Ashish Sabharwal. 2020. What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

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

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Roger C Schank and Robert P Abelson. 1977. Scripts, plans, goals and understanding: An inquiry into human knowledge structures.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *ICML Workshop Workshop on Visualization for Deep Learning.*

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 1361–1375.

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.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021a. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

Pei Zhou, Pegah Jandaghi, Bill Yuchen Lin, Justin Cho, Jay Pujara, and Xiang Ren. 2021b. Probing causal common sense in dialogue response generation. *In Findings of EMNLP*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. *AAAI*.

A Probing Data Details

A.1 Raw Set Collection

We define 1-4 combinations of type constraints on the predicates for each LT and designe crawling strategies accordingly using resources: Concept-Net, ATOMIC, and Wikipedia. Descriptions for each of the 11 strategies are included in Table 4. All data and code for crawling strategies is included in the supplementary materials.

A.2 Turking Details for Human-Verified Set

We present a pair of statements by flipping the comparative term in the original statement to its opposite, and ask two annotators to choose the one that follows commonsense. The AMT page for turkers to annotate is shown in Figure 8. If they disagree, we then take the pairs and do a second round of turking by asking three annotators and use majority voting to decide what is the right sentence in the pair. We replace the original statement with the opposite one if there are more annotators think that the other one in the pair follows more commonsense. In total, around 2500 pairs are sent to the second round and 300 pairs are flipped to the opposite according to annotators. The estimated time for completing each instance is around 20 seconds and we pay each instance $0.06, which translates to around $11 per hour.

A.3 Human-Curated Set Details

We show all perturbations for one probe in Table 7 and 60 of our human-curated set's unperturbed statement in Table 8 (for temporal refer to supplementary material). Full data is included in the supplementary material.

B Experimental Details

Model Detail We test our probes on in total 10 models, with the number of parameters and other details in Table 6. For RoBERTa-base, RoBERTa-large, RoBERTa-large-MNLI, and BART-large-MNLI, we use the fairseq implementation [3]. For BERT-base-uncased, BERT-large-uncased, AL-BERT, and GPT-2, we use the huggingface transformers library [4]. For COMET trained on Concept-Net and ATOMIC, we follow their github repo [5]. We use ERNIE from their original github [6].

Fine-tuning Details We fine-tune BERT-base-uncased, BERT-large-uncased, RoBERTa-base, and RoBERTa-large based on HappyTransformers [7] framework, using a consistent learning rate of 1e-5. We fine-tune GPT-2 based on huggingface transformers library's example code [8], using their default parameters. We train them on one NVIDIA Quadro RTX 6000 GPU for 10 epochs and after each epoch we test the fine-tuned model on our validation set, and save the model with the highest validation set performance. Fine-tuning RoBERTa-base and GPT-2 takes around 30 minutes for each epoch and RoBERTa-large takes around 1 hour. The best validation performance for RoBERTa-base is the fourth epoch, with perplexity 1.3378140926361084 and evaluation loss': 0.2910370217429267. For RoBERTa-large, the best is epoch 5, with perplexity 1.3949965238571167 and evaluation loss 0.3328918993473053. For GPT-2, the best is epoch 3, with perplexity 1.2786548795017285.

Interpretation Details We use the AllenInterpret demo [9]. To identify important context words, we run the algorithm over the same probe for 5 times, each with different entity names, and select the words that are ranked in the top 5 most important words at least 3 times. We find that the interpretations are not very consistent as the most important words change when we input the same sentence for multiple times and will also change when different names are used, so we conduct 5 trials with different names for each probe and pick the words that appear in the majority of the trials.

C Additional Studies

Does explicitly providing commonsense knowledge help? Shocked by the severe bias observed in PTLMs, we construct an easier set of probes, where we explicitly state all knowledge needed to make the correct logical inference. We have two

---

Logical Template	Type Constraint	Crawling Strategy	Example Axiom (Adjusted for readability)
1	Attribute-Material (126)	Get a list of materials, and find properties in ConceptNet using HasProperty; then find a second material using NotMadeOf from the previous property.	Material(A, glass) and Material(B, wood), so More(clear(A), clear(B))
	Attribute-Grade (132)	Input an ordered list of numbers and form pairs and comparative relations following the orders	Grade(A, first) and Grade(B, third), so More(young(A), young(B))
	Condition-Location (1k)	Get a list of places with descending latitude from Wikipedia and form pairs by the relation (higher latitude is colder than lower latitude), considering both hemispheres.	Location(A,equator) and Location(B, north pole), so More(living in hot condition(A), living in hot condition(B))
	Attribute-Animal (10k)	Get a list of animals from Wikipedia and find properties in ConceptNet using CapableOf and LocateAt	Animal(A, fish) and Animal(B, horse), so More(locate at the bottom of the sea(A), locate at the bottom of the sea(B))
2	Role (1.2k)	Input a list of occupations from Wikipedia and find properties in ConceptNet using CapableOf	Priest(A,B), so More(Pray(A), Pray(B))
	Action (10k)	For every event in ATOMIC that involves two people, we find properties by following the Attribute edge in ATOMIC	Forces upon(A, B), so More(pushy(A), pushy(B))
3	Action (10k)	For each event in ATOMIC that involves people, we find properties by following the Attribute edge in ATOMIC, note that we replace PersonX with "themself" and PersonY with "another person" to sound natural	Assesses patient(A) and not Assesses patient(B), so More(analytical(A), analytical(B))
	Capability-Physical (100)	Input a list of adjectives describing objects, we find properties by following UsedFor edge in ConceptNet	Tie knot(A) and not Tie knot(B), so More(elastic(A), elastic(B))
4	Action (10k)	Similarly to LT3-Event	More(Concentrate(A), Concentrate(B)), so More(Effective(A), Effective(B))
	Capability-Physical (100)	Simiarly to LT3-Physical	More(square(A), square(B)), so Better(divide two space(A), divide two space(B))
5	Attribute-Temporal (100)	Manually come up with temporal ordered-events, included in Human-Curated Set	entered the building(A),so before(outside(A))

Table 4: Crawling strategies for 11 type-constrained KT crawling for our Raw Set.

CATEGORY	EXAMPLE
Physical (30%)	A is smaller than B, so A is easier to put into a box than B.
Material (30%)	A is made out of glass and B is made out of stone, so A is more transparent than B.
Social (30%)	A makes the varsity team while B does not, so A is more skilled than B.
Temporal (10%)	A was eating dinner, so A was hungry before eating dinner.

Table 5: Different types of commonsense axioms included in our human-curated probe set

Model	Details
BERT-base-uncased	12-layer, 768-hidden, 12-heads, 125M parameters
BERT-large-uncased	24-layer, 1024-hidden, 16-heads, 355M parameters
RoBERTa-base	12-layer, 768-hidden, 12-heads, 125M parameters
RoBERTa-large	24-layer, 1024-hidden, 16-heads, 355M parameters
ALBERT	12 repeating layer, 128 embedding, 4096-hidden, 64-heads, 223M parameters
GPT-2	12-layer, 768-hidden, 12-heads, 117M parameters.
COMET-Concept	GPT-2 config + Traning on ConceptNet
COMET-ATOMIC	GPT-2 config + Traning on ATOMIC
RoBERTa-L-MNLI	24-layer, 1024-hidden, 16-heads, 355M parameters
BART-L-MNLI	24-layer, 1024-hidden, 16-heads, 406M parameters + a classification head

Table 6: Models tested and details.

settings for this test, one where parroting the now-provided commonsense fact is all that is needed to correctly answer the probe, and the other where a simple negation switch of the commonsense fact is needed to solve the probe:

- A is made of glass, B is made of stone, *and glass is more transparent than stone*, so A is [MASK] transparent than stone. (parrot)

- A is made of glass, B is made of stone, *and glass is more transparent than stone*, so A is **not** [MASK] transparent than stone. (negation switch)

We do this so to investigate whether RoBERTa

is actually able to use the provided commonsense fact, or is it possibly just pattern matching.

We add this piece of background knowledge to the 60 original (unperturbed) statements along with their corresponding negated statements to form an "easier" setting of our task. As shown in Figure 6, we find two patterns PTLMs exhibit. For RoBERTa, ALBERT, and GPT-2, there is a stark difference in performance between the two settings. When they are being asked to parrot the commonsense fact, the performances jump up to near perfect scores, however when all they have to do is the equivalent of applying a negation operator on the fact, they fail even worse than when they are not provided the fact. These results suggest that in the parrot easier setting, it is likely RoBERTa, ALBERT, and GPT-2 are just parroting the commonsense fact they see in the sentence and not utilizing some sort of reasoning ability, as when asked to perform the simplest of logical operations they fail. The other pattern we notice is that providing background knowledge does not help or hurt the performances for COMET and models tested on the textual entailment task. For COMET models, this may be due to the fact that COMET is trained on triplets from knowledge bases: given a head entity and a relation, predict the tail entity, so it is not used to taking auxiliary knowledge into its input. As for models fine-tuned on MNLI, the performance stays unchanged because they still think most of the sentence pairs of our probes are neutral, failing to grasp the embedded logical inference step.

Case Study on Contextual Clues To gain a better understanding on model behaviors, we conduct analysis to identify context words that the model relies on when solving our probes. We use
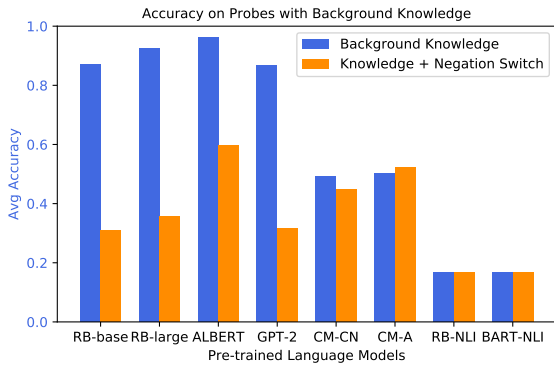
Figure 6: Results of average performance of PTLMs when we provide background knowledge in our probes. For RoBERTa, ALBERT, and GPT-2, we notice a huge increase in accuracy when provided knowledge. However, we find that they are merely parroting what appears in the context since when we apply a negation in the probe, which should change the prediction, they are simply predicting the same as the context shows, resulting in performance drop. For COMET moddels and models tested on the NLI setting, we do not observe the same pattern and it seems that adding knowledge does not help or hurt.

the SmoothGrad (Smilkov et al., 2017) algorithm from AllenNLP Interpret (Wallace et al., 2019) for masked word prediction on our probes with real people's names (the same set as our ablation study) using BERT. Aggregated across all probe sets, we find that the three words BERT finds most important are: "*than*", "*not*", and "*so*", which make sense as they are indicators for comparison, negation, and causality, respectively.

"*Not*" and "*so*" are also the textual forms of the logical connectives ¬ and →, which we use to construct LTs.

Furthermore, we find that BERT also regards *argument* words (inputs into LTs' predicates via a knowledge table, such as "*lawyer*" or "*knowledge of law*") important. The model finds on average 3.4 words as contextual clues and 1.5 out of them are knowledge-specific argument words. This finding shows that a PTLM is able to recognize words specific to the commonsense axiom tested. However, noticing all these clues does not necessarily aid in a PTLM's ability to understand their logical implications, as evidenced by their performances. In other words, a PTLM, in this case BERT, knows that these words are important when making a decision, but it does not know how to properly answer RICA's questions based on these lexical signals.
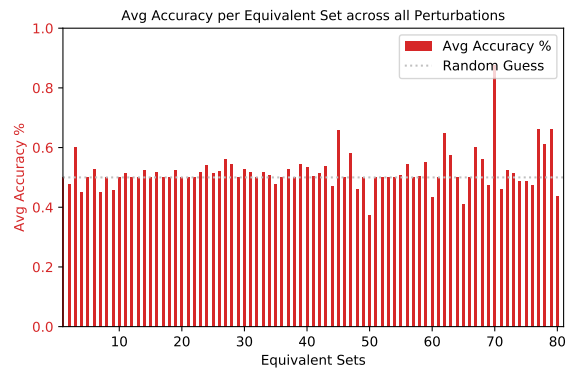

Figure 7: Results of average accuracy of RoBERTa-large on MWP. We can see that the PTLM makes random-guessing like predictions across all sets.

linguistic perturbation	asymmetric perturbation	probe
original	original	A is wider than B, so A finds it harder to slip through cracks than B
original	asymmetric_premise	B is wider than A, so A finds it easier to slip through cracks than B
original	asymmetric_conclusion	A is wider than B, so B finds it easier to slip through cracks than A
negation	original	A is wider than B, so A does not find it easier to slip through cracks than B
negation	asymmetric_premise	B is wider than A, so A does not find it harder to slip through cracks than B
negation	asymmetric_conclusion	A is wider than B, so B does not find it harder to slip through cracks than A
antonym	original	A is wider than B, so A finds it easier to be blocked by cracks than B
antonym	asymmetric_premise	B is wider than A, so A finds it harder to be blocked by cracks than B
antonym	asymmetric_conclusion	A is wider than B, so B finds it harder to be blocked by cracks than A
paraphrase	original	A is wider than B, so A is worse at fitting into openings than B
paraphrase	asymmetric_premise	B is wider than A, so A is better at fitting into openings than B
paraphrase	asymmetric_conclusion	A is wider than B, so B is better at fitting into openings than A
paraphrase_inversion	original	A is wider than B, so A is more impeded by small openings than B
paraphrase_inversion	asymmetric_premise	B is wider than A, so A is less impeded by small openings than B
paraphrase_inversion	asymmetric_conclusion	A is wider than B, so B is less impeded by small openings than A
negation_antonym	original	A is wider than B, so A does not find it harder to be blocked by cracks than B
negation_antonym	asymmetric_premise	B is wider than A, so A does not find it easier to be blocked by cracks than B
negation_antonym	asymmetric_conclusion	A is wider than B, so B does not find it easier to be blocked by cracks than A
negation_paraphrase	original	A is wider than B, so A is not better at fitting into openings than B
negation_paraphrase	asymmetric_premise	B is wider than A, so A is not worse at fitting into openings than B
negation_paraphrase	asymmetric_conclusion	A is wider than B, so B is not worse at fitting into openings than A
negation_paraphrase_inversion	original	A is wider than B, so A is not less impeded by small openings than B
negation_paraphrase_inversion	asymmetric_premise	B is wider than A, so A is not more impeded by small openings than B
negation_paraphrase_inversion	asymmetric_conclusion	A is wider than B, so B is not more impeded by small openings than A

Table 7: An example probe set—24 logically equivalent, but semantically different statements.
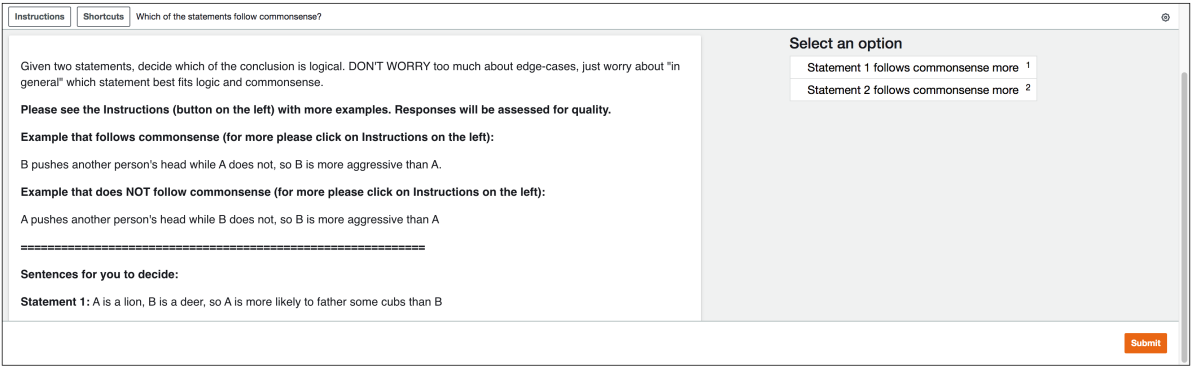
Figure 8: AMT annotation user interface for human verification on collected set.

template	probe
1	A is made out of glass and B is made out of stone, so A is more transparent than B
1	A is made out of cotton and B is made out of glass, so A is less sharp than B
1	A is made out of concrete and B is made out of paper, so A should be more heavy than B
1	A is made out of metal and B is made out of rubber, so A should float worse than B
1	A is made out of glass and B is made out of copper, so A is more fragile than B
1	A is made out of steel and B is made out of wool, so A is less soft than B
1	A is made out of wood and B is made out of glass, so A is more combustible than B
1	A is made out of sponge and B is made out of nylon, so A is worse for water resistance than B
1	A is made out of copper and B is made out of concrete, so A is more ductile than B
1	A is made out of metal and B is made out of cloth, so A is less foldable than B
1	A is made out of chocolate and B is made out of metal, so A is harder to keep frozen than B
1	A is made out of metal and B is made out of dirt, so A is a better electrical conductor than B
1	A is made out of stone and B is made out of helium, so A has a harder time flying than B
1	A is made out of honey and B is made out of water, so A is more viscous than B
1	A is made out of titanium and B is made out of rubber, so A is less elastic than B
1	A is made out of water and B is made out of methane, so A is more safe to store than B
1	A is made out of mercury and B is made out of oxygen, so A is worse for your health to consume than B
1	A is made out of wood and B is made out of fur, so A will more easily expand when heated than B
1	A is made out of concrete and B is made out of wood, so A is less penetrable than B
1	A is made out of glass and B is made out of tar, so A will reflect light better than B
3	A makes the varsity team while B does not, so A is more skilled than B
3	A is going to perform for people while B is not, so A finds it harder to be relaxed than B
3	A won the competition while B did not, so A finds it easier to be happy than B
4	A is able to concentrate more than B, so A finds it easier to be productive than B
3	A bullies people while B does not, so A is less kind than B
2	A is B's boss, so A commands more respect than B
4	A has more work than B, so A finds it harder to be at ease than B
2	A has a crush on B, so A finds it harder to be relaxed around B
4	A has more dedication than B, so A will have a harder time failing than B
2	A is B's parent, so A initially takes more care of B
2	A is B's doctor, so A takes more care of B
2	A hurt B's feelings, so A must be more insensitive than B
2	A is B's priest, so A spends less time sinning than B
2	A is B's lawyer, so A is less ignorant of the law than B
4	A has a lot less money than B, so A is less financially secure than B
4	A watches more tv shows than B, so A is more capable of understanding pop-culture references than B
2	A always loses to B in tennis, so A is a less proficient tennis player than B
2	A makes B late, so A has less reason to be annoyed at B
4	A is a better friend than B, so A is more thoughtful than B
2	A is B's teacher, so A should be more informed than B
4	A is smaller than B, so A is easier to put into a box than B
4	A is heavier than B, so A is better at sinking than B
4	A is denser than B, so A should withstand piercing more easily than B
4	A is wider than B, so A finds it harder to slip through cracks than B
4	A is hotter than B, so A should be easier to melt than B
4	A is more elastic than B, so A should bounce better than B
4	A is tougher than B, so A is harder to rip apart than B
4	A is harder than B, so A is less comfortable than B
4	A is taller than B, so A will cast a more lengthy shadow than B
4	A is lighter than B, so A finds it harder to support weight than B
4	A has less momentum than B, so A has a worse ability to damage on impact than B
4	A is more luminous than B, so A is more dangerous to look at than B
4	A is more soluble than B, so A is harder to discern in water than B
4	A is more pungent than B, so A is easier to detect than B
4	A is smaller than B, so A finds it harder to displace liquid in a tub than B
4	A is shorter than B, so A is worse for keeping things out of reach than B
4	A is larger than B, so A is more difficult to carry than B
4	A is more taut than B, so A is worse at withstanding additional force than B
4	A is much hotter than B, so A will be more painful to hold onto than B
4	A is more magnetic than B, so A is harder to separate from another magnet than B

Table 8: Sixty probes and their corresponding logical templates

			BERT-base			BERT-large	
		Easy Set	Hard Set	Joint Set	Easy Set	Hard Set	Joint Set
Zero-shot		49.32	49.7	49.56	49.15	49.35	49.27
Low resource	10%	56.38	49.85	52.37	63.08	50.18	55.16
	20%	59.3	50.5	53.89	65.64	50.37	56.26
	30%	55.09	50.22	52.1	65.3	50.27	56.07
	50%	60.62	50.33	54.3	63.48	50.25	55.35
Full resource	with 1 novel entity	54.26	50.5	51.95	61.02	49.46	53.97
	with 5 novel entities	64.48	50.58	55.94	69.92	49.72	57.51
	with 10 novel entities	82.58	51.18	63.3	85.93	50.64	64.74
100k		45.4	50.09	46.03	46.14	50.93	47.02

			RoBERTa-base			RoBERTa-large	
		Easy Set	Hard Set	Joint Set	Easy Set	Hard Set	Joint Set
Zero-shot		53.6	49.39	55.32	51.81	49.69	58.79
Low resource	10%	59.08	49.36	53.1	59.93	50.65	54.23
	20%	60.53	49.41	53.7	64.31	50.31	55.71
	30%	60.79	49.93	54.21	65.47	50.95	56.55
	50%	64.11	49.4	55.07	75.88	52.58	61.57
Full resource	with 1 novel entity	64.24	49.89	55.43	79.1	52.84	62.97
	with 5 novel entities	82.32	49.93	62.43	87.03	51.69	65.32
	with 10 novel entities	85.63	50.64	64.14	84.74	51.08	64.07
100k		72.35	50.02	70.06	78.13	53.92	73.71

			GPT2	
		Easy Set	Hard Set	Joint Set
Zero-shot		51.27	49.6	50.1
Low resource	10%	50.57	49.91	50.29
	20%	48.22	49.33	49.01
	30%	48.18	49.2	48.97
	50%	50.44	49.87	49.96
Full resource	with 1 novel entity	55.95	49.34	52.16
	with 5 novel entities	66.3	49.53	55.98
	with 10 novel entities	71.6	49.91	58.25
100k		32.94	49.16	35.21

		BART	
		EASY	HARD
Zeroshot		47.46	50.37
Low resource	10%	60.55	49.72
	20%	61.58	49.91
	30%	63.54	49.96
	50%	61.65	49.68
Full resource	with 1 novel entity	63.94	49.31
	with 5 novel entities	69.05	49.73
	with 10 novel entities	79.73	50.52
100k		50.87	49.95

		ERNIE	
		EASY	HARD
Zeroshot		47.97	50.31
Low resource	1%	63.11	50.22
	3%	72.91	48.89
	5%	87.42	49.70
Full resource	with 1 novel entity	87.66	50.06
	with 5 novel entities	87.65	50.03
	with 10 novel entities	87.36	49.77
100k		69.25	50.95