

# RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms

Pei Zhou   Rahul Khanna   Bill Yuchen Lin   Daniel Ho  
Jay Pujara   Xiang Ren

Department of Computer Science and Information Sciences Institute  
University of Southern California

{peiz, rahulkha, yuchen.lin, hsiaotuh, jpujara, xiangren}@usc.edu

## Abstract

Pre-trained language models (PTLM) have impressive performance on commonsense inference benchmarks, but their ability to practically employ commonsense to communicate with humans is fiercely debated. Prior evaluations of PTLMs have focused on factual world knowledge or the ability to reason when the necessary knowledge is provided explicitly. However, effective communication with humans requires inferences based on implicit commonsense relationships, and robustness despite paraphrasing. In the pursuit of advancing fluid human-AI communication, we propose a new challenge, RICA, that evaluates the capabilities of making commonsense inferences and the robustness of these inferences to language variations. In our work, we develop a systematic procedure to probe PTLMs across three different evaluation settings. Extensive experiments on our generated probe sets show that PTLMs perform no better than random guessing (*even with fine-tuning*), are heavily impacted by statistical biases, and are not robust to perturbation attacks. Our framework and probe sets can help future work improve PTLMs’ inference abilities and robustness to linguistic variations bringing us closer to more fluid communication.<sup>1</sup>

## 1 Introduction

Smooth and effective communication requires the ability to make various forms of commonsense inferences. 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 as shown in Figure 1. A

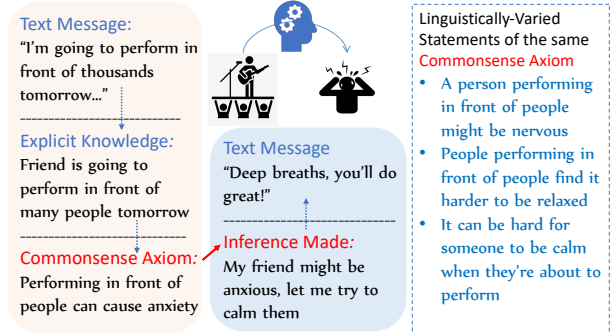


Figure 1: Humans often make commonsense inferences during communication. This work evaluates such inferences through commonsense axioms expressed with many linguistic variations.

growing body of literature (Bosselut et al., 2019; Petroni et al., 2019) shows pre-trained language models (PTLMs) are able to catalog these 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 inferences.

More specifically, our study focuses 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 “Larger objects can contain smaller objects.” Furthermore, we exploit the flexibility of language to express the same axiom in many forms, e.g., “Smaller objects fit inside larger objects.” We test these characteristics by generating self-contained commonsense statements involving novel entities (“a prindag is smaller than a fluberg, so a fluberg is more likely to contain a prindag.”) and adapt them to three evaluation settings.

Unfortunately, these two capabilities have largely been overlooked by existing natural lan-

<sup>1</sup>Our data and code are submitted as supplementary material and will be public upon publication.

guage understanding (NLU) benchmarks (Williams et al., 2018; Levesque et al., 2012) and probing studies for transformer-based pre-trained language models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020; Petroni et al., 2019). Most existing natural language inference (NLI) datasets (Williams et al., 2018) do not focus on commonsense inferences, while commonsense reasoning benchmarks (Ostermann et al., 2019; Talmor et al., 2019b) do not systematically evaluate robustness against linguistic variations, meaning we cannot preclude the possibility that models are just pattern matching to solve the needed task.

To fill this gap, we introduce RICA, a challenge to evaluate a model’s **Robust Inference** capability based on **Commonsense Axioms**. RICA draws on linguistic and cognitive science research (Schank and Abelson, 1977; Johnson-Laird, 1983, 2013) that suggests humans translate language to logical representations and reason using these abstract representations. RICA consists of a set of statements that require reasoning using a latent commonsense relationship. We formulate these abstract relations between entities in first-order logic and refer to them as *commonsense axioms*. To insulate from PTLM biases and test coverage, RICA uses *novel entities*, 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 present results on 16,000 commonsense statements, capturing 80 different axioms from four types of commonsense knowledge: physical, material, social, and temporal. RICA is designed to leverage existing commonsense knowledge bases such as ConceptNet (Liu and Singh, 2004) to support easy expansion. Furthermore, RICA’s statements can be formulated as evaluation instances, which we refer to as probes, for three common benchmark tasks for PTLMs: masked word prediction, textual entailment, and sentence probability. RICA provides an extensible platform for evaluating commonsense reasoning in a broad variety of PTLMs.

Shockingly, when evaluating state-of-the-art (SOTA) PTLMs on the RICA challenge probes, we consistently discovered performance that is on par with *random guessing*. Even when fine-tuned, the average performance hovers around 50%, with high

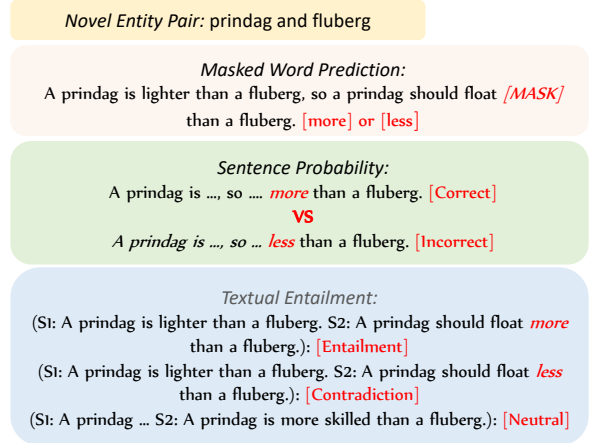


Figure 2: Illustration of three evaluation settings with a pair of novel entities used by RICA probes.

variance under perturbation attacks. We identify a pervasive intrinsic bias in PTLMs that correlates with word frequencies. Finally we use a gradient-based interpretation model to find contextual clues that PTLMs use to inference on RICA.

Our contributions are summarized as follows: **(1)** We propose a new textual inference challenge, RICA. Our challenge tests PTLMs’ ability to use commonsense axioms in many different linguistic forms, and can be framed in any of three popular tasks for PTLMs. **(2)** We propose a system that allows for the expansion of our challenge and showcase its usefulness by generating probes for RICA. **(3)** Our findings reveal that current PTLMs do not beat out a random baseline on our probes, are heavily impacted by statistical biases, and are not robust to linguistic perturbations. We will release the code and the probe dataset for future research.

## 2 The RICA Challenge

The RICA challenge is posed as a set of commonsense statements—*i.e.*, sentences expressing a latent commonsense relationship, e.g., “*a prindag is smaller than a fluberg, so a prindag is less likely to contain a fluberg.*” These sentences use generated novel entities such as “*prindag*” and “*fluberg*” instead of real-world concepts such as “*thimble*” and “*elephant*” to separate factual recall from reasoning. Each statement can be viewed as a generalization 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 relationships (e.g., containment). We link these logical formulae to the associated statements using

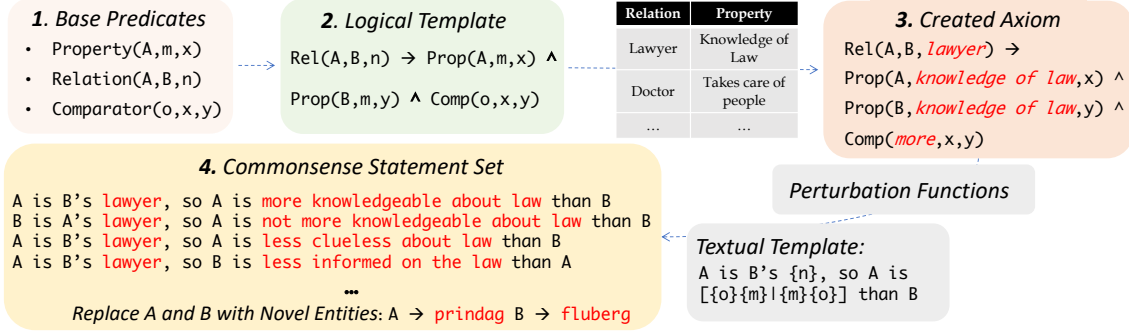


Figure 3: **Overview of the workflow of our statement construction process.** The output is a set linguistically-diverse of masked sentences that follow the same reasoning template.

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

Table 1: Description of terminology used in RICA.

a set of textual templates and perturbation operators, which together produce a logically-equivalent set of commonsense statements. In this section, we first provide a formal definition of our challenge, then provide a detailed description of this process.

## 2.1 Challenge Formulation

We define a commonsense axiom,  $a_i$ , as a relationship between entities that can be inferred using commonsense knowledge expressed via a first-order-logic (FOL) formula. 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  $\{s_1^i, s_2^i, \dots, s_m^i\}$ , each expressing the logic of the axiom. Each statement takes the form of an implication with a premise and conclusion, to easily test inference. 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 call that a model is successful on our challenge (e.g., understand an axiom) only if it can perform well on all probes of the axiom.

## 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.** We propose three logical predicates that serve as the backbone for the logical formulations of our axioms: *Property*, *Comparator* and *Relation*, as follows: (1) **PROP**( $A, m, x$ ) indicates that the property  $m$  (e.g., “knowledge of law”) of entity  $A$  takes the value of  $x$ . A value can be numerical, categorical, or boolean. (2) **REL**( $A, B, n$ ) indicates that  $A$  and  $B$  have a relation  $n$  (e.g., “lawyer”). (3) **COMP**( $o, x, y$ ) denotes that the value  $x$  is  $o$  than  $z$ , where  $o$  is a comparison word like “better”, “more” or “easier”.

While using three predicates allow us to compose a wide variety of FOL formulae, our efforts are just a first step in developing a comprehensive representation of commonsense axioms in FOL.

**Stage 2. Compose Logical Templates.** We create first-order logical formulae, referred to as *logical templates* (LT), using the predicates defined above. Each formula takes the form of an implication, expressing an inference based on commonsense knowledge. For example,  $\text{REL}(A, B, n) \rightarrow \text{PROP}(A, m, x) \wedge \text{PROP}(B, m, y) \wedge \text{COMP}(o, x, y)$  expresses the logical inference that can be made based on relation  $n$  about two entities  $A$  and  $B$  and the comparison of their common property  $m$ . We also manually construct a *textual template* (TT) for each LT, which is a natural language expression of the logical relationship, used later in our process.

**Stage 3. Create Axioms.** We use LTs to generate commonsense axioms, which are partially-grounded LT formula. For example in Figure 3,  $n$ ,  $m$ , and  $o$  are set in order to reflect the commonsense relationship between a lawyer and knowledge

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.

of law, while leaving the entities and exact values ungrounded. Once the needed arguments are set, we call this partially-filled LT a *commonsense axiom*. We use *knowledge tables* to fill the arguments of the predicates in the LT to form a FOL representation of the axiom. Figure 3 shows candidates of concepts used to generate specific instances. The generality of the predicates in LTs allows RICA to use commonsense KBs such as ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019a) to automatically populate knowledge tables. For example, one can input occupations for the LT relation, query ConceptNet using edge types CAPABLEOF and HASPROPERTY to get properties.

**Stage 4. Generate Statement Sets.** After filling the logical templates, we get one commonsense statement for each axiom. However, one statement is not enough to comprehensively challenge models’ understanding of an abstract axiom, so we construct a *statement set* embedding the same axiom with different phrasings, i.e., logically-equivalent yet linguistically-varied. We define several *perturbations* to apply on the *arguments* we fill in the predicates using knowledge tables.

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

<b>LT 1:</b> $\text{PROP}(A, m, x) \wedge \text{PROP}(B, m, y) \rightarrow \text{PROP}(A, n, x) \wedge \text{PROP}(B, n, y) \wedge \text{COMP}(o, x, y)$ <b>Example:</b> A is made out of glass, B is made out of stone, so A is more transparent than B
<b>LT 2:</b> $\text{REL}(A, B, n) \rightarrow \text{PROP}(A, m, x) \wedge \text{PROP}(B, m, y) \wedge \text{COMP}(o, x, y)$ <b>Example:</b> A is B’s priest, so A spends more time praying than B
<b>LT 3:</b> $\text{PROP}(A, m, x) \wedge \neg \text{PROP}(B, m, x) \rightarrow \text{PROP}(A, n, x) \wedge \text{PROP}(B, n, y) \wedge \text{COMP}(o, x, y)$ <b>Example:</b> A makes the varsity team while B does not, so A is more skilled than B
<b>LT 4:</b> $\text{PROP}(A, m, x) \wedge \text{PROP}(B, m, y) \wedge \text{COMP}(o, x, y) \rightarrow \text{PROP}(A, n, x) \wedge \text{PROP}(B, n, y) \wedge \text{COMP}(o, x, y)$ <b>Example:</b> A is able to concentrate more than B, so A is more effective than B
<b>LT 5:</b> $\text{PROP}(A, m, x) \rightarrow \text{PROP}(A, n, y) \wedge \text{COMP}(o, x, y)$ <b>Example:</b> 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.

parisons and relationships allowing us to introduce asymmetries that preserve meaning. For example,  $\text{MORE}(A, B) = \neg \text{MORE}(B, A)$  and  $\text{REL}(A, B, \text{parent}) = \neg \text{REL}(B, A, \text{parent})$ . 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  $\text{ASYM}(\text{P}(A, B)) = \text{P}(B, A) = \neg \text{P}(A, B)$ . We apply the defined operators to the arguments in the predicates and adopt the same TT for converting axioms to statements with diverse perturbations. We have eight forms of linguistic perturbations including the original one, and we can apply the asymmetry operators (if there are two entities involved) either on the premise or conclusion or keep the original ordering. Thus we have in total of 24 types of perturbations.

(3) *Novel Entities.* 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.



### 3 RICA Evaluation Setup

#### 3.1 Probing Tasks

To comprehensively examine transformer-based PTLMs’ inference abilities, we propose three tasks. We argue that each task evaluates some capability of the inference ability based on commonsense axioms, but does not provide a comprehensive picture. We draw conclusions from experimental results on three distinct evaluation tasks, aiming to provide convincing and comprehensive probing insights. A general illustration of three tasks is shown in Figure 2 and described in the following paragraphs.

**Masked Word Prediction (MWP)** We adopt the masked LM pre-training objective from BERT (Devlin et al., 2019) to examine if the models can recover key words in the statement given the context. Since RICA’s commonsense statements take the form of implications, we mask words in the consequent to focus on the inference performance given the premise. Additionally, we choose to mask the words that fill the “o” argument in the predicate COMP(o,x,y), such as “*more/less*” and “*better/worse*” since they not only capture the key commonsense inference, but also restrict masking to words where only a few options are appropriate logically and syntactically. For example, in the statement “*A is B’s parent, so A is more likely to care for B*”, we mask “*more*”.

**Sentence Probability (SP)** evaluates if PTLMs output higher probability for statements that express commonsense axioms versus contradictory statements. We input complete commonsense statements into the model, assigning probabilities by multiplying each word’s probability conditioned on the previous words, i.e., the causal language modeling loss. For each statement, we pair it with an incorrect statement that does not follow commonsense axioms by swapping the comparison term (the masked word in the above evaluation setting) with its opposite. In the example above, we create that probe’s pair as: “*A is B’s parent, so A is less likely to care for B*”.

**Textual Entailment/NLI** evaluates a multi-class classification of the relationship between two sentences, in RICA’s case an implication’s premise and conclusion. We separate a statement’s premise and conclusion and label the pair as entailment. For neutral pairs, each premise of the original statement is paired with a randomly sampled conclusion from other statement sets. For contradiction pairs, we

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 4: Different types of commonsense axioms included in our probe sets.

separate premises and conclusions of the incorrect statements from the SP setting.

#### 3.2 Probing Data Details

Following the process in Section 2, we generate five logical templates as shown in Table 3. Then we use knowledge tables to fill in each template and finally apply the perturbation operators as described before to form a final set of 1,600 linguistically-varied statements. For each masked statement, we randomly generate 10 novel entities to fill in positions of *A* and *B* (if any), yielding 16,000 probes in total for our probing experiments.

#### 3.3 Evaluation Metrics

**Accuracy** For MWP and SP, we evaluate by simply comparing the binary rankings of [sentences containing] the original comparative word from the commonsense statement and its antonym. For NLI, we calculate the accuracy of models’ predictions for 3 classes: entail, neutral, and contradict.

**Confidence Ratio** To provide more accurate measurement on the model’s confidence gap between true and false options, we further propose a metric called confidence ratio. This only applies to MWP and SP because in these two tasks we have true and false options for each instance. We calculate our ratio score using:  $\frac{(score_{right} - score_{wrong})}{(score_{right} + score_{wrong})}$ , where  $score_{right}$  indicates the model’s output confidence/probability score for the correct masked word (MWP) or sentence (SP) and similarly for  $score_{wrong}$ . The more positive the final score is, the better the performance according to this metric, and vice versa. For NLI, the model directly predicts a label for a pair of sentences, so we are not given a confidence score to calculate the ratio.

#### 3.4 Baseline Methods

We evaluate a wide range of state-of-the-art transformer-based PTLMs covering both masked

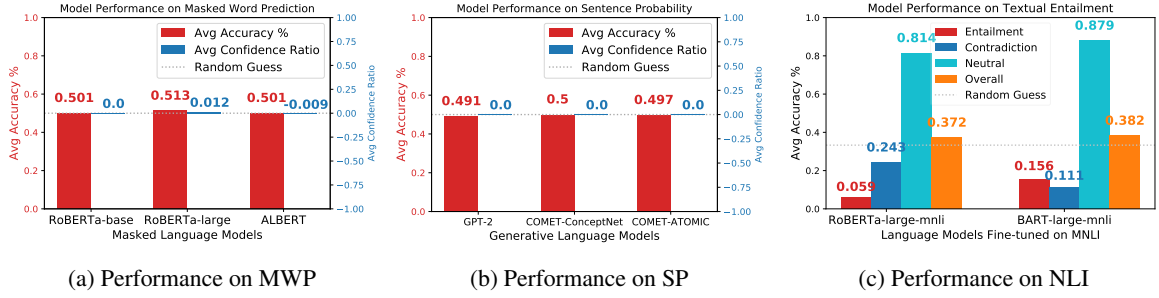


Figure 4: Performance of different transformer-based models on the three task settings. For MWP and SP, the models’ performance is no better than random guessing. For NLI, the overall is slightly higher than random guessing with the accuracy for neutral a lot higher than the two other labels.

and unidirectional language models. For the masked word prediction task, we consider RoBERTa base and large (Liu et al., 2019) and ALBERT (Lan et al., 2019), two recent BERT-based (Devlin et al., 2019) masked language models that show strong results on many benchmarks. For sentence probability, we consider GPT-2 (Radford et al.), a large language model for left-to-right language generation. And COMET (Bosselut et al., 2019), a generative model for knowledge graph completion whose backbone is GPT-2, but is further trained on large knowledge bases such as ConceptNet (Liu and Singh, 2004) or ATOMIC (Sap et al., 2019a) (we test both) we consider (and anecdotally observe) COMET to possess knowledge of our commonsense axioms. For the task of textual entailment, we use one masked language model RoBERTa and one sequence-to-sequence model BART (Lewis et al., 2019), both fine-tuned on MultiNLI (Williams et al., 2018).

## 4 Results and Analysis

### 4.1 Performance without Fine Tuning

We first examine the general performance of multiple language models on each task, show fine-tuning results on our probes, and present ablation studies to analyze performance more thoroughly.

**Results on MWP and SP.** As shown in Figures 4a and 4b, the average binary accuracies of all three masked language models on MWP and all three generative models on the SP task are around 0.5. Similarly, the average confidence value of the predicted answer is close to 0, as a result of the model having similar confidence in correct and incorrect predictions. A random baseline that chooses between the two comparative words would have an accuracy of 0.5 and a confidence ratio of 0. This shows that the tested models barely beat a random guessing baseline. Noticeably, COMET performs on par with vanilla GPT-2, demonstrating the dis-

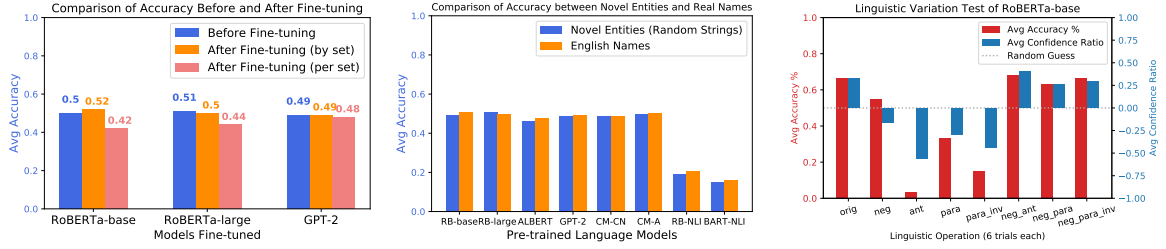
tinction between *storing* commonsense axioms and *reasoning* with axioms.

**Human Performance on MWP.** To benchmark human performance on the MWP task (we assume humans will perform similarly on other tasks), we sampled 5% of our MWP probes and gathered answers from 20 subjects with diverse backgrounds who were not involved in probe construction. Humans obtained 91.7% accuracy, taking a majority vote for each probe, with substantial inter-annotator agreement (0.768 Kappa (Cohen, 1960)).

**Results of Textual Entailment.** Figure 4c presents the average accuracies (for all labels and separated by class) of two models fine-tuned on MNLI dataset and tested on the textual entailment task. We find that RoBERTa and BART, after fine-tuning on MNLI, have a significant variance in the performance on different labels: accuracy for neutral labels is much higher than that for entailment and contradiction. The overall average is slightly above random guessing (33%). Examining their predictions, we find both are mostly predicting the neutral label, failing to discriminate between axioms and contradictions.

### 4.2 Performance of Fine-tuned Models

To study if poor performance in §4.1 is due to a lack of exposure to RICA’s probe sets, we conduct an experiment to fine-tune baseline language models. We split the constructed statements into training/development/test sets with an 80/10/10 partition, with two types of sampling strategies: (1) *sample by set*, where we partition the probe sets so that the axioms in test set remain unseen in training set; (2) *sample per set*, where we randomly sample some from each set and axioms in test overlap with those in training data. We fine-tune RoBERTa-base and RoBERTa-large using the same masking approach as the MWP evaluation, and fine-tune GPT-2 on the causal language modeling task. De-



(a) Fine-Tuning Results

(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) even after fine-tuning, the models still perform like random guessing, (b) consistent poor performance 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.

tails for the training are in the appendix.

The model accuracy after fine-tuning, as presented in Figure 5a, remains near 0.5, on par with random guessing. Surprisingly, fine-tuning by holding out axioms (sampling by set) performs better than exposing the models to all axioms during training. Even exposing RoBERTa to linguistically similar sentences does not improve inference ability, and improved performance on unseen axioms may suggest improvements are due to pattern-matching, not commonsense acquisition. An inability to improve on reasoning tasks after fine-tuning supports the challenging nature of RICA, which cannot be trivially solved by fine-tuning.

### 4.3 Performance Analysis

**Study of Model Prediction Biases.** 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 property we are comparing i.e. regardless of the logic in the statement. Table 5 shows that the accuracy for “positive valence” words such as “more”, “better”, and “easier” is much higher than “negative valence” words such as “less”, “worse”, 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 GPT-2. We conjecture that this may be due to the frequency difference between positive valence words and negative valence words. To check this hypothesis, we use Google Ngram Viewer<sup>2</sup> 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 cor-

Model	Acc on “Pos”	Acc on “Neg”
RoBERTa-base	0.872	0.125
RoBERTa-large	0.899	0.122
ALBERT	0.734	0.246
<b>RoBERTa-base-ft</b>	<b>0.458</b>	<b>0.587</b>
RoBERTa-large-ft	0.775	0.158
GPT-2	0.777	0.209
<b>GPT-2-ft</b>	<b>0.509</b>	<b>0.483</b>
COMET-ConceptNet	0.698	0.294
COMET-ATOMIC	0.664	0.332
<b>Human</b>	<b>0.900</b>	<b>0.938</b>

Table 5: We see a clear gap between the accuracy for probes containing positive versus negative valence words. Fine-tuning (by set) mitigates the bias in some models (in bold). Humans do not exhibit such bias.

relation supports the claim that PTLMs do not reason as humans do, but are guided by statistical patterns.

**Effect 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 probes, around one third of our full dataset. These probes involved social commonsense, where novel entities took the place of names. We conducted 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 per-

<sup>2</sup><https://books.google.com/ngrams>

turbation 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.

#### 4.4 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 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  $\neg$  and  $\rightarrow$ , 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.

### 5 Related Work

**Machine Commonsense** 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., 2019b; Bisk et al., 2020; Sap et al., 2019b) 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 not only examining reasoning abilities, but also looking into robustness to linguistic variation via our linguistically-varied commonsense statements. RICA also challenges PTLMs on three evaluation tasks (not one) to better probe the PTLMs’ representations.

**Natural Language Inference** NLI has been well studied and several benchmarks have been proposed (Bowman et al., 2015; Williams et al., 2018), which have a larger scope than just commonsense inference. This suggests a strong performance on an NLI dataset does not imply strong commonsense reasoning abilities. Studies have also found that neural models that perform well on NLI datasets tend to adopt heuristics and are prone to adversarial attacks (McCoy et al., 2019; Nie et al., 2020).

**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; Feldman et al., 2019) or existing datasets (Zhou et al., 2020; Talmor et al., 2019a; Kwon et al., 2019). 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 probe 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 16,000 statements from 80 sets of probes and test 8 transformer-based LMs on three different evaluation tasks. We find that PTLMs perform on par with random guessing (even with fine-tuning), have a strong statistical bias towards positive words, and are not robust under linguistic perturbations. We conclude that there is much work to be done to enable fluid AI-human conversation, but we hope RICA aids in this pursuit.



## References

- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *AAAI*.
- 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.
- Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *IJCAI 2020*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.
- Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pre-trained models. In *EMNLP/IJCNLP*.
- Philip N Johnson-Laird. 2013. *Human and machine thinking*. Psychology Press.
- Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.
- Sunjae Kwon, Cheongwoong Kang, Jiyeon Han, and Jaesik Choi. 2019. Why do masked neural language models still need common sense knowledge? *arXiv preprint arXiv:1911.03024*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.
- Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Hugo Liu and Push Singh. 2004. Conceptnet: a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- John W. McCarthy. 1960. Programs with common sense.
- R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *ACL*.
- 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.
- 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.
- 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.
- 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 4453–4463.
- 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. 2019a. olympics—on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019b. 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allenlp 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.
- 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. Association for Computational Linguistics.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. *AAAI*.

## A Probing Data Examples

We show all perturbations for one probe in Table 5 and 60 of our probe set’s unperturbed statement in Table 6 (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 8 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<sup>3</sup>. For ALBERT and GPT-2, we use the huggingface transformers library<sup>4</sup>. For COMET trained on ConceptNet and ATOMIC, we follow their github repo<sup>5</sup>.

**Fine-tuning Details** We fine-tune RoBERTa-base and RoBERTa-large based on HappyTransformers<sup>6</sup> framework, using a consistent learning rate of 1e-5. We fine-tune GPT-2 based on huggingface transformers library’s example code<sup>7</sup>, 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<sup>8</sup>. To identify important context words, we

<sup>3</sup><https://github.com/pytorch/fairseq/tree/master/examples/roberta>, <https://github.com/pytorch/fairseq/tree/master/examples/bart>

<sup>4</sup>[https://huggingface.co/transformers/model\\_doc/albert.html](https://huggingface.co/transformers/model_doc/albert.html), [https://huggingface.co/transformers/model\\_doc/gpt2.html](https://huggingface.co/transformers/model_doc/gpt2.html)

<sup>5</sup><https://github.com/atcbosselut/comet-commonsense>

<sup>6</sup><https://github.com/EricFillion/happy-transformer>

<sup>7</sup><https://github.com/huggingface/transformers/tree/master/examples/language-modeling>

<sup>8</sup><https://demo.allennlp.org/masked-lm>

Model	Details
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 + Training on ConceptNet
COMET-ATOMIC	GPT-2 config + Training 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.

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 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 7, 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,

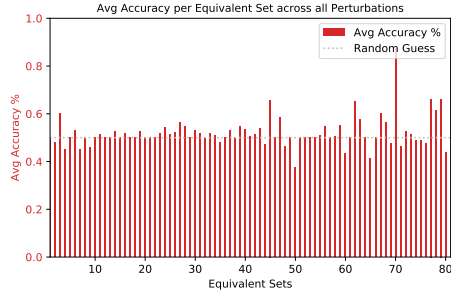


Figure 6: Results of average accuracy of RoBERTa-large on MWP. We can see that the PTLM makes random-guessing like predictions across all sets.

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.

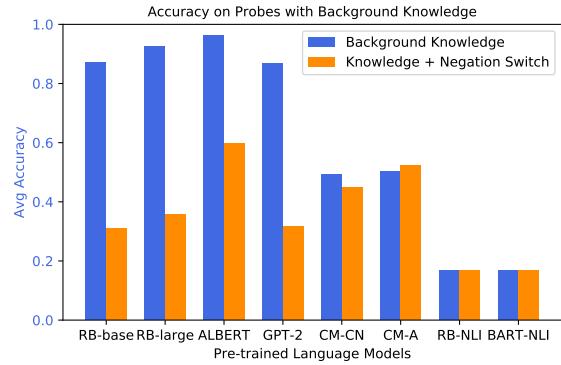


Figure 7: Results of average performance of PTLMs when we provide background knowledge in our probes. For RoBERTa, ALBERT, and GPT-2, we notice a huge increase in accuracy when provided knowledge. However, we find that they are merely parroting what appears in the context since when we apply a negation in the probe, which should change the prediction, they are simply predicting the same as the context shows, resulting in performance drop. For COMET models 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.



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 set24 logically equivalent, but semantically different statements.

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