COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences

Shikhar Singh[*1], **Nuan Wen**[*1], **Yu Hou**[1], **Pegah Alipoormolabashi**[2], **Te-Lin Wu**[3], **Xuezhe Ma**[1], **Nanyun Peng**[3]
[1]University of Southern California, [2] Sharif University of Technology,
[3]University of California, Los Angeles
{ssingh43,nuanwen,houyu,xuezhe.ma}@usc.edu palipoor976@gmail.com
{telinwu,violetpeng}@cs.ucla.edu

Abstract

Commonsense reasoning is intuitive for humans but has been a long-term challenge for artificial intelligence (AI). Recent advancements in pretrained language models have shown promising results on several commonsense benchmark datasets. However, the reliability and comprehensiveness of these benchmarks towards assessing model's commonsense reasoning ability remains unclear. To this end, we introduce a new commonsense reasoning benchmark dataset comprising natural language true/false statements, with each sample paired with its complementary counterpart, resulting in 4k sentence pairs. We propose a pairwise accuracy metric to reliably measure an agent's ability to perform commonsense reasoning over a given situation. The dataset is crowdsourced and enhanced with an adversarial model-in-the-loop setup to incentivize challenging samples. To facilitate a systematic analysis of commonsense capabilities, we design our dataset along the dimensions of knowledge domains, reasoning scenarios and numeracy. Experimental results demonstrate that our strongest baseline (UnifiedQA-3B), after fine-tuning, achieves ~71% standard accuracy and ~51% pairwise accuracy, well below human performance (~95% for both metrics). The dataset is available at https://github.com/PlusLabNLP/Com2Sense.

1 Introduction

The capability of acquiring and reasoning over commonsense knowledge plays a crucial role for artificial intelligence (AI) systems that interact with humans and accomplish tasks in the real world. For example, given a situation where *someone is asleep*, an agent should choose to *broom* instead of *vacuum* to clean the room, as the latter would be noisy. Likewise, a personal assistant should be able to infer that one is probably unavailable if they are *at work*. This ability to contextualize and draw upon implicit knowledge, and generalize to novel situations, requires commonsense reasoning.

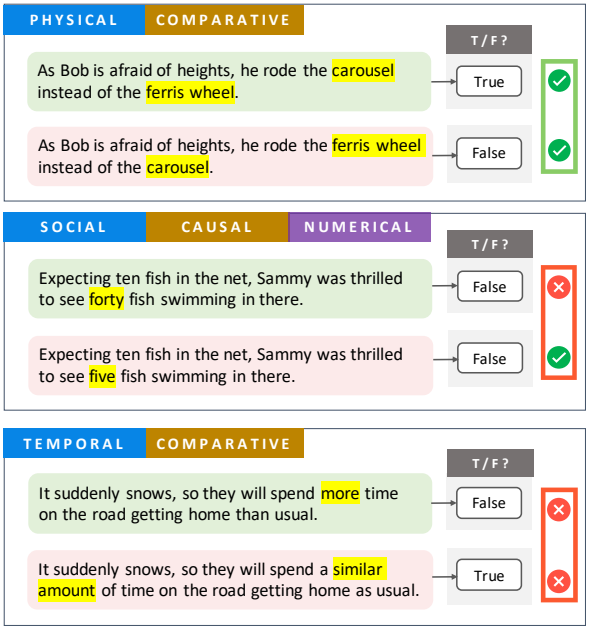


Figure 1: Complementary sentence pair samples from COM2SENSE defined along knowledge domains (*e.g. physical*), reasoning scenarios (*e.g. comparative*) and numeracy attributes. Each sentence within a pair is either true (green boxes) or false (red boxes), followed by model predictions and annotations of whether the predictions are correct. A *standard* accuracy is computed by the percentage of correctly judged sentences (50% for these three pairs), while the *pairwise* accuracy requires both *individual* judgements to be correct in each pair (33% for these three pairs).

While humans are able to intuitively acquire commonsense knowledge from everyday experience and make sound inferences, whether current AI systems also possess such capabilities remains an open question. Recent advancements in natural language processing (NLP) has led to a surge in

[*] indicates equal contributions

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 883–898
August 1–6, 2021.

new benchmark datasets towards evaluating commonsense reasoning. Specifically, existing benchmarks are formulated as natural language inference (NLI) (Bhagavatula et al., 2020), *multiple choice* (MC) question answering (Talmor et al., 2019a; Zellers et al., 2019; Bisk et al., 2020), and machine reading comprehension (Huang et al., 2019) tasks.

While recent state-of-the-art models (Liu et al., 2019; Raffel et al., 2020; Khashabi et al., 2020) have quantitatively demonstrated near human-level performance on these benchmarks, the exploitation of certain spurious patterns (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019) in the datasets can be partly attributed to such achievements. Consider the examples in Figure 1, where each sentence is true/false, and is paired with a similar (with a few modifications) complementary counterpart such that the answer is flipped. Humans can infer each statement independently with confidence, but models on the other hand struggle to give consistent judgements for the complementary pairs. This indicates that models are able to *guess* the correct answer without a thorough understanding of the given input. If we formulate this as a multiple choice task, where *only the true* sentence needs to be *singled out* given the pairs, the models have higher chances to get it correct, as they are only required to select the *relatively better* option.

Furthermore, most existing commonsense benchmarks focus on the *factual* aspects of commonsense (Talmor et al., 2019a; Bisk et al., 2020), and generally do not explicitly concern with *reasoning* (Singer et al., 1992), *i.e*. the mental manipulation of factual knowledge, which we hypothesize is crucial for generalizing to novel situations. While some prior works investigate commonsense reasoning in the context of social intelligence and coreference resolutions (Sap et al., 2019; Sakaguchi et al., 2020), the reasoning components are implicit. Existing benchmarks fail to provide a systematic and comprehensive means of analyzing different aspects of commonsense knowledge and reasoning.

To address these challenges, we introduce the **Com**plementary **Com**mon**sense** (COM2SENSE) benchmark dataset which contains 4k complementary true/false sentence pairs. Each pair is constructed with minor perturbations to a sentence to derive its complement such that the corresponding label is inverted (see Figure 1). This *pairwise* formulation provides a more reliable evaluation metric, where a model is considered correct *only if it succeeds on both statements*. We employ an adversarial crowdsourcing framework to collect human created samples via a *gamified* machine-in-the-loop process: A strong pretrained model is setup to provide instant feedbacks, thereby incentivizing challenging samples that can *fool* the model.

Broadly inspired by the *Theory of Core Knowledge*, *i.e*. the ability to reason about objects, places, numbers and the social world (Spelke and Kinzler, 2007), we design our dataset along the following dimensions: **knowledge domains** (*physical*, *social*, *temporal*), and **reasoning scenarios** (*causal*, *comparative*). Additionally, concurrent to a recent work (Lin et al., 2020a) on studying numerical commonsense, we include a third dimension of **numeracy**, which extends the factual focus of Lin et al. (2020a) (*e.g*. *"Ants have **six** legs."*) to *numerical reasoning* (*e.g*. the *ten fish* versus *forty fish* in Figure 1). To the best of our knowledge, we are the first to explicitly introduce these dimensions in a commonsense benchmark dataset, thereby facilitating a more detailed and systematic probing of models' commonsense understanding.

Our experiments demonstrate that the best performing pretrained language models achieve ~71% standard and ~51% pairwise accuracy, well below human performance. Additionally, we provide ablation studies on effect of training size on model performance, and the transferrability across the reasoning scenarios. We summarize our contributions as follows: 1) We introduce a commonsense reasoning dataset which we position as a challenging *evaluation benchmark* (instead of a training resource) for NLP models. 2) We propose a pairwise evaluation metric featured by our complementary pair formulation for a more reliable assessment of commonsense reasoning abilities. 3) We benchmark state-of-the-art models that highlight significant gaps (>45%) between model and human performances.

2 Dataset

We introduce COM2SENSE, a dataset for benchmarking commonsense reasoning ability of NLP models. We use crowdsourcing to collect the dataset and supplemented with an adversarial *model-in-the-loop* approach. The key features of our development process are: 1) qualification quiz to filter and familiarize workers, 2) *gamified* creation tasks, and 3) quality check by experts. The details of dataset formulation and collection pro-

Domain	Scenario	Numeracy	Example	Complement
Physical	Comparative	No	*If we dropped **milk** on the floor, it is better to clean with a mop rather than a broom.*	*cereal*
Physical	Causal	No	*To **read books** at night, one should turn on the lights.*	*see stars*
Social	Comparative	No	*Sam robbed a store, while Tim jumped the lights. People will likely be more **forgiving** towards Tim.*	*chastising*
Social	Causal	Yes	*Given his $1500 monthly income and no savings, he can afford an apartment rent of **$500**.*	*$3000*
Temporal	Comparative	Yes	*Tim needs to return home in 2 hours, so he would prefer to **hit the gym** rather than **go hiking**.*	`swap`
Temporal	Causal	Yes	*If Leo earns $100 per day, then by working from Monday to **Friday** his weekly income will be $500.*	*Wednesday*

Table 1: Data samples from different categories in COM2SENSE. Each example is labelled as true, while its complement (false) is generated by substituting or swapping the words in **bold** (in green or red font).

cedure, along with statistics are provided in the following sections.

2.1 Formulation

COM2SENSE seeks to measure a comprehensive commonsense understanding of everyday events and entities. The task requires one to judge whether a given sentence is true or false. For each sentence in the dataset, we also compose its complementary counterpart by modifying a few words, such that the answer is inverted. The key advantages of using complementary pairs are two-folds: 1) it provides a more robust way of evaluating models' commonsense reasoning ability by requiring both sentences to be correctly judged, and 2) the complements naturally highlight the salient words which may be useful in probing model behaviors.

Furthermore, to facilitate a systematic study of commonsense, we design our dataset across the following three dimensions:

1. **Knowledge Domain:** We categorize commonsense knowledge into *physical*, *social* and *temporal* domains. The physical domain emphasizes on an intuitive understanding of physical properties (*e.g.* weight, shape, motion, space) and object affordances. The social domain encapsulates interactions (*e.g.* intent, emotion, reaction), activities, and societal norms. The temporal domain captures the notion of time, particularly attributes such as duration, frequency and order of events. While domains may not always be strictly exclusive (*e.g.* choice of transport and duration), our complementary pair setup naturally places emphasis on the intended domain.

2. **Reasoning Scenario:** We define two types of inferential reasoning scenarios: 1) The *causal* scenario requires the ability to infer whether a cause explanation or a subsequent event (cause-effect) is correct. 2) The *comparative* scenario requires the ability of determining the most plausible hypothesis between two or more competing ones.

3. **Numeracy:** Refers to the basic understanding of numbers, arithmetic, ratios, statistics, etc. With the objective of linking numeracy to commonsense, we particularly focus on "number sense" – an intuitive understanding of numbers, their magnitude and relationships, rather than computational and numerical precision.

Therefore, each sample in our dataset should fall into a category defined by a combination of the above dimensions, as exemplified in Table 1.

2.2 Dataset Creation

COM2SENSE is developed through crowdsourcing on Amazon Mechanical Turk (MTurk) with the goal of collecting complementary sentence pairs. The creation tasks are constructed for each *category* defined by the combination of domain, scenario and numeracy attributes. An overview of the data collection workflow is illustrated in Figure 2. In order to participate, the workers are required to pass a **qualification quiz** designed to familiarize them with the key aspects of our dataset.

Creation: During the creation phase, to orient and aid workers' creativity, they are provided with five examples of complementary pairs that belong to a particular category as reference. We also share a list of verbs and topics pertinent to the current

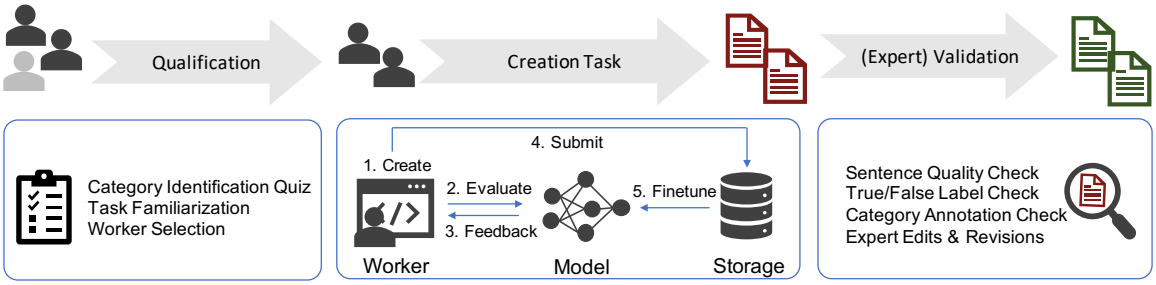


Figure 2: Data collection workflow: 1) qualification quiz to instruct the key aspects of our creation task and filter unqualified workers, 2) interactive *model-in-the-loop* creation process to incentivize challenging samples via model feedback, and 3) data validation according to our guidelines and category descriptions.

domain, as an optional resource. While our examples serve as a reference towards creating complementary pairs, the workers have the freedom to construct their sentences as they deem appropriate.

We employ an adversarial *model-in-the-loop* approach to provide workers with immediate feedback (*i.e.* model predictions) on each created sentence. After entering the inputs and labels, they may choose to evaluate and revise their inputs. If the sentence successfully fools our model to answer incorrectly, workers are awarded with an additional amount for each input[1].

To further incentivize worker creativity, we offer bonuses if the inputs are qualitatively regarded as creative during the validation stage. Such *gamified* process may continue for a few rounds until the workers are satisfied with their monetary rewards.

Model: We deploy a RoBERTa-large based model for binary sentence classification, finetuned on SemEval-2020 Task 4 (Wang et al., 2020) given its true/false format and broad coverage of commonsense knowledge. After the first phase of collection (2k pairs), the model weights are updated by finetuning on our dataset with 60% train, 20% dev and 20% test splits. This will naturally help diversify our dataset samples, as the model is unlikely to be fooled with repetitive knowledge and sentence structures.

Validation: To ensure high quality, the samples are validated by internal members to look for inconsistencies with regard to the category-type *i.e.* follows the domain, scenario and numeracy requirements, and inferential ambiguities that may arise due to insufficient context, specialized concepts, grammatical errors, etc. Furthermore, annotators may choose to revise the samples to fix any of the aforementioned issues. Each sample is validated by three annotators and the final outcome is decided through a majority vote. The inter-annotator agreement score is 0.989 measured using Fleiss' Kappa. Additionally, pairs in which neither input could fool the model are discarded during this stage.

The dataset is developed with the help of 173 workers. To ensure that workers are proficient in English, the demographic pool of the workers is initially limited to the United States. However, we removed this criteria to avoid cultural biases in the dataset. Additionally, to understand the utility of our adversarial model feedback setting, we analyze the data on number of revisions made by workers in order to successfully fool the model. We find that the average number of revisions is 1.36, while the median is zero. This suggests that for majority of samples, workers find our reference material sufficient and are also able to leverage model feedback to aid their creations. Additional details on dataset development are in Appendix Section A.

2.3 Dataset Statistics

Given that COM2SENSE is primarily a benchmark dataset, it is partitioned into train[2] (20%), development (10%), and test (70%) set, respectively. There are in total 4k of statement pairs in our dataset. Complementary statements from the same pair are distributed to the same partition. Table 2 gives the essential statistics of our dataset across different splits. Note that due to the complementary pair formulation, the type-token ratio is approximately reduced by a factor of two, and the dataset is naturally balanced along the true and false labels.

Table 3 gives the breakdown of percentage of samples from each category defined by a combination of the three dimensions. The distribution of most frequent nouns in the dataset is visualized

[1]Base pay = $0.05 – $0.1 and bonus pay = $0.5 – $0.9.

[2]As a resource to adapt models for our task.

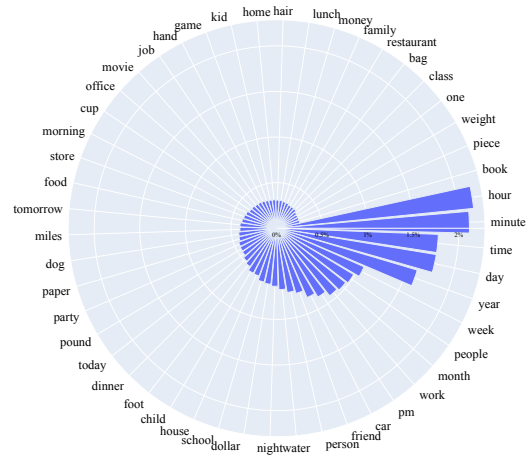


Figure 3: Top-50 frequent nouns in the dataset.

Figure 4: Top-50 frequent topics in the dataset.

in Figure 3. Likewise, the distribution of most frequent topics – lexical categories generated using the *Empath*[3]tool, is provided in Figure 4.

3 Experimental Setup

The experiments are designed to meet the following objectives: 1) benchmark state-of-the-art NLP models along the standard and pairwise formulation; 2) analyze the model performance across different categories of commonsense reasoning; 3) report the effect of training size on model performance; and 4) verify the role of reasoning types by measuring "cross-scenario" transferability.

Besides standard accuracy, we introduce a new metric called **pairwise accuracy** that evaluates as correct if both predictions within a pair are accurate.

[3]https://github.com/Ejhfast/empath-client

[4]Input lengths are computed with Spacy tokenizer

Statistic	Train	Dev	Test
# complementary pairs	804	402	2779
Avg input length	21	21	21
Max input length	68	49	67
Min input length	6	7	6
# unique tokens	2306	1541	4407
# total tokens	21116	10520	72517

Table 2: Dataset statistics across different splits[4]

	Scenario	
Domain	Causal	Comparative
Physical	17.47% (24%)	18.92% (23%)
Social	14.68% (50%)	16.51% (22%)
Temporal	16.74% (57%)	15.68% (62%)

Table 3: Category-wise breakdown (percentage) of dataset samples. The quantities in parenthesis refer to the relative proportion of samples with numeracy, under the given combination of domain and scenario.

We benchmark several state-of-the-art NLP models, specifically the ones proven preeminent in existing commonsense benchmarks, and additionally include a Bi-LSTM model as a baseline to help check for potential spurious correlations in the dataset. We consider the following baselines:

BiLSTM+GloVe A bidirectional-LSTM model (Hochreiter and Schmidhuber, 1997) taking input word embeddings from GloVe (Pennington et al., 2014).

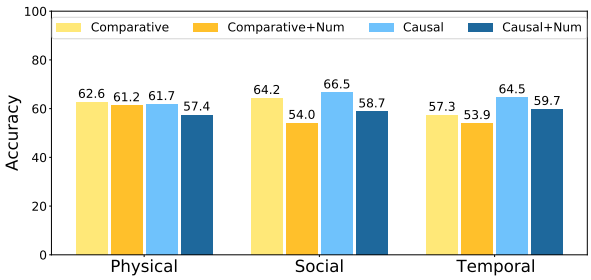
BERT The BERT-base (110M) model introduced in (Devlin et al., 2019).

RoBERTa-large A large variant (355M) of RoBERTa model (Liu et al., 2019) built upon BERT-large architecture.

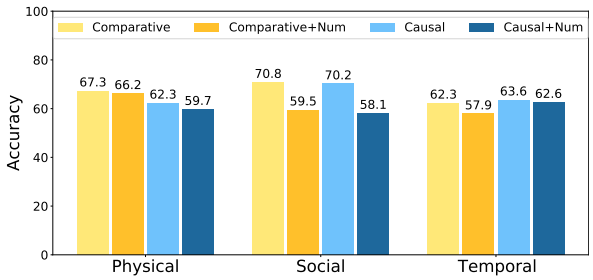
DeBERTa-large Recently He et al. (2020) proposed a novel disentangled attention mechanism that improves upon BERT and RoBERTa models. We consider the large variant (390M) as a baseline.

T5-large Similarly, the large variant (770M) of the T5 model (Raffel et al., 2020). We follow the standard prefix-based text-to-text format, and adapt it for our binary classification setup.

UnifiedQA The UnifiedQA (Khashabi et al., 2020) was originally trained on numerous datasets including several commonsense reasoning benchmarks, and performed well under zero-shot setting. We consider the variants with T5-large and T5-3B as the architecture backbone.



(a) T5-large.



(b) DeBERTa-large.

Figure 5: Model performance breakdowns across domains and scenarios, "+Num" denotes numeracy involved.

Model	Standard	Pairwise
Random	50.00	25.00
BiLSTM+GloVe	53.80	29.50
BERT-base	51.79	12.91
RoBERTa-large	59.35	33.28
T5-large	60.56	41.84
UnifiedQA-large	60.83	34.79
DeBERTa-large	63.53	45.30
UnifiedQA-3B	**71.31**	**51.26**
Human	96.50	95.00

Table 4: Test set accuracy for selected models, trained and evaluated on our dataset. Human performances are obtained with 200 randomly selected and decoupled samples from 100 pairs.

	Model			
Dataset	Random	RoBERTa	T5	Human
CQA	20.00	72.10	73.35	88.90
SWAG	25.00	89.92	88.72	88.00
SocialIQA	33.33	77.12	73.25	84.40
PIQA	50.00	77.21	79.89	94.90
WinoGrande	50.00	79.14	75.02	94.00
COM2SENSE$_{\text{standard}}$	50.00	59.35	60.56	96.50
COM2SENSE$_{\text{pairwise}}$	25.00	33.28	41.84	95.00

Table 5: Test set accuracy for selected models (RoBERTa-large and T5-large), trained and evaluated on respective datasets.

4 Results and Analysis

Human Performance: To estimate a human upper bound for our dataset, we perform a separate run with ten top performing workers that had participated in our collection phase to examine a randomly selected subset of 200 samples (*i.e.* from 100 pairs). Each worker is assigned with a set of shuffled samples with his/her own creations deliberately filtered out. The answer is determined by majority vote from *three* workers. Human performances are 96.5% with standard accuracy and 95.0% with pairwise accuracy, respectively.

4.1 Results

Benchmark results: Table 4 summarizes the baseline performances on the test set. As the Bi-LSTM model performs close to random, we claim that improvements from stronger baselines should be attributed to the models and not annotation biases that they can exploit. Among the baseline models, the UnifiedQA-3B achieves the best performance on both the standard and pairwise metric. Note that the number of learnable parameters in UnifiedQA-3B is much larger than those in the second and third best models, which are DeBERTa-large (390M) and UnifiedQA-large (770M). Our COM2SENSE benchmark remains quite challenging, as there are significant gaps between the model and the human performances.

Dataset comparisons: In order to contrast the difficulty of COM2SENSE with other related benchmarks, we report the performances of two well performing models on the following: CommonsenseQA (CQA) (Talmor et al., 2019b), SWAG (Zellers et al., 2018), SocialIQA (Sap et al., 2019), PhysicalIQA (PIQA) (Bisk et al., 2020) and WinoGrande (Sakaguchi et al., 2020).

The results in Table 5 indicate that models clearly struggle more to perform well on COM2SENSE than other datasets.

Performance across domains and scenarios: In Figure 5 we present the in-depth breakdown results for T5-large and DeBERTa-large across combinations of domain, scenario and numeracy. We observe that models consistently perform worse in categories involving numeracy, highlighting the limitations of current language models. For physical domain, both models perform worse in causal than in comparative scenario. We hypothesize that while the models may possess the required physi-

Setting	Test	Dev
Multiple-Choice	70.63	77.61
Standard (T/F)	63.53	66.29

Table 6: Performance (accuracy) of DeBERTa-large on the MC formulation of our dataset compared to the standard true/false setting. The model performs relatively worse on the latter.

	$\textbf{Model}_{\textbf{metric}}$			
Train set	$\text{DeBERTa}_{\text{std}}$	T5_{std}	$\text{DeBERTa}_{\text{pair}}$	T5_{pair}
20%	63.92	60.65	41.51	34.29
40%	67.74	62.60	48.04	38.11
60%	68.46	63.96	48.47	40.66

Table 7: Performance across different training set sizes (20% / 40% / 60% of the entire dataset) for DeBERTa-large and T5-large. `"std"` and `"pair"` stand for standard and pairwise accuracy correspondingly.

cal knowledge, they fail to generalize to a logical reasoning over known facts or grasp the implicit changes of physical properties, which is generally unseen in the pretraining corpora for NLP models. Opposite trends are observed in both social and temporal domains, where similar hypothesis can be made that causal statements are more frequent patterns in the corpora when social activities or senses of time are the subjects. Furthermore since model feedback was part of our dataset construction, we also report the category-wise difficulty in fooling the model (number of trials) during sample creation in Section A.4 for a reference.

True/False versus multiple choice setup: We further conduct an experiment with DeBERTa-large model on the same data splits with the input formulated as an **MC task** in Table 6. Under this setting, the model is provided with the sentence pair and is required to select the true sentence among the two choices, for the response to be correct. The performance is significantly higher compared to both standard (>7%) and pairwise accuracy (>25%) presented in Table 4. This result supports our intuition that it is easier for models to exploit spurious correlations in the surface patterns under the *multiple choice* question answering setup.

4.2 Analysis

The Effect of Training Data Size: To study the effect of training size on model performance, we design an experiment by varying the sample size in the training set, with fixed dev (10%) and test (30%) sets to **ensure consistency** in evaluation. We consider DeBERTa-large and T5-large models for this ablation study, and report our findings in Table 7. The results indicate a plateau in performance with increase in training samples.

Setting	Standard	Pairwise
$\text{Train-C}_{\text{exclude}}$	56.52	19.00
$\text{Train-C}_{\text{include}}$	63.54	40.49

Table 8: Test set performance of DeBERTa-large on two different setups with respect to the complementary pairs. Both setups have the same training set size, but in $\text{Train-C}_{\text{exclude}}$ only one sentence of each pair is present in the training set, while in $\text{Train-C}_{\text{include}}$ both samples in a pair are included. The results indicate the effectiveness of training the models with our formulated complementary pairs.

Role of Complementary Pairs: In previous experiments, we measure the model generalizations by distributing data samples into train and evaluation sets **by complementary pair**. This ensures the similarly constructed sentences within the same pair is not *leaked* into different data splits, and thus an "inter-pair" generalization is measured. To investigate the effectiveness of our complementary pair formulation on training models to acquire commonsense reasoning ability, we first sample a subset with identical size (20% data, 800 pairs) to that of the original train set, and then construct the following two variants (using the same subset):

- **$\text{Train-C}_{\text{exclude}}$:** One of the complementary samples (in each pair) is *excluded*, *i.e.* no two samples belong to the same pair in this train set. It comprises 800 samples from 800 pairs, with balanced true/false labels.
- **$\text{Train-C}_{\text{include}}$:** We retain half of the data samples where both sentences from a pair are *included*, which leads to 800 samples from 400 pairs in this train set.

The remaining samples from the dataset (without the excluded ones in the two settings) are then split into dev (10%) and test (70%) sets. We compare the performance of DeBERTa-large in the above two different settings in Table 8. The results show a significant decrease in performance when complementary sentences are not provided. We hypothesize that the worse performances are due to the models' tendency to pick up surface patterns and memorize the labels in the training set without really understanding the scenario. Also, model

	Evaluate			
Train	$\text{Causal}_{\text{std}}$	$\text{Comp.}_{\text{std}}$	$\text{Causal}_{\text{pair}}$	$\text{Comp.}_{\text{pair}}$
Causal	**63.64**	59.36	**35.46**	28.25
Comp.	58.47	**64.50**	26.43	**43.86**

Table 9: Performance of DeBERTa-large trained on X and evaluated on Y, where X and Y are partitions created as per a reasoning scenario *(causal, comparative)*.

generalization benefits from having complementary samples within the training set.

Cross-Scenario Generalizability: Given that knowledge domain and numeracy attributes of our dataset are intuitively distinct, we intend to quantitatively investigate if the same holds for reasoning scenarios. Our "cross-scenario" experiments with DeBERTa-large, *i.e.* trained on *causal*, evaluated on *comparative* and vice versa, indicate a poor generalization across both standard and pairwise accuracy metrics (see Table 9), underscoring the significance of having reasoning types.

5 Related Works

Commonsense Resources: As commonsense is a crucial component to the actualization of AI, there has been a surge in creating relevant benchmarks, notable ones include evaluating machines' commonsense abilities in the format of pronoun resolution (Levesque et al., 2012; Sakaguchi et al., 2020), multiple choice (Zellers et al., 2018; Talmor et al., 2019a), natural language generations (Lin et al., 2020b), story understanding (Mostafazadeh et al., 2016), and reading comprehensions (Zhang et al., 2018; Huang et al., 2019; Ning et al., 2020). Our work puts forth to create a commonsense benchmark in the format of true/false complementary pairs, where a more robust pairwise accuracy is adopted. Note that although natural language inference (NLI) can be tasked similarly to the true/false formulation, the existing commonsense NLI benchmark either is not crowdsourced with high quality (Zhang et al., 2017), or still resorts to a multiple choice setting (Bhagavatula et al., 2020). There are also benchmarks that specifically concern a type of commonsense knowledge, such as physical (Bisk et al., 2020) and social (Sap et al., 2019) intelligence, as well as temporal understanding (Zhou et al., 2019). The ability to understand and induce numerical knowledge in texts has been studied in several recent works (Dua et al., 2019; Ravichander et al., 2019), including numerical commonsense (Lin et al., 2020a). Our work differs to these works in the focus on less factual and arithmetic-precise numerical knowledge, but more on the intuitive sense of numbers, in conjunction with our defined knowledge domains and the scenarios.

It is worth noting that some prior works (Wu et al., 2017; Clark et al., 2019) also investigate the effectiveness in the binary true/false (yes/no) formulation to construct a question answering dataset, while COM2SENSE is the first to focus on commonsense reasoning.

Dataset Biases: It is a widely perceived issue that spurious statistical patterns in datasets can often be exploited by machine learning models, which can potentially lead to overoptimistic judgements on the model improvements. Particularly in NLP domain, prior works have shown that hypothesis-only baselines or syntactic heuristics perform surprisingly well in the NLI task (Gururangan et al., 2018; Glockner et al., 2018; Tsuchiya, 2018; Poliak et al., 2018; McCoy et al., 2019). Model exploiting biases or failing on simple adversarial patterns, can also be seen in sentence classification (Wieting and Kiela, 2019) and question answering (Jia and Liang, 2017; Kaushik and Lipton, 2018; Geva et al., 2019) tasks. We put forth to reduce the potential sentence-level biases by requiring the models to perform equally well on both directions in a complementary true/false pair.

Adversarial Data Collection: Removing representation biases in a dataset by adversarially filtering undesired data samples, has been frequently practiced to collect datasets more challenging to the models. Recent work *AFLite* (Sakaguchi et al., 2020; Le Bras et al., 2020), built upon the adversarial filtering (AF) method in (Zellers et al., 2018, 2019), adopted an iteratively improving model-in-the-loop approach to collect challenging commonsense benchmarks (Sakaguchi et al., 2020; Bisk et al., 2020). *Gamified* (Yang et al., 2018) or interactive (Wallace et al., 2019) approaches leverage human-in-the-loop to increase the difficulty of datasets and hence more robust model training. Counterfactual editing of data samples with human annotators (Kaushik et al., 2020; Gardner et al., 2020) is also closely related to our complementary pair construction that seeks to *invert* the model predictions for a more reliable evaluation.

Recently, several works have attempted to exploit the merits in involving both models and humans in the data creation cycle, *i.e.* human-and-

model-in-the-loop, to construct data samples that are both *new* and challenging to the models (Chen et al., 2019; Nie et al., 2020; Bartolo et al., 2020). To our best knowledge, we are the first to employ such an approach in constructing commonsense reasoning benchmark, specifically, our complementary pair formulation makes it more sophisticated as the annotators are required to not only *fool* the model but also pay attention to the salient concepts of their creations in both directions.

6 Conclusion

We present a new challenging commonsense reasoning benchmark, COM2SENSE, developed via an adversarial *gamified* model-in-the-loop approach. COM2SENSE comprises 4k *manually created* complementary true/false statement pairs, designed along three dimensions: knowledge domain, reasoning scenario, and numeracy. We propose a robust pairwise metric to evaluate models' commonsense reasoning ability based on the complementary pair formulation, and benchmark the dataset with several state-of-the-art NLP models, highlighting significant gaps well below human performances ($>$ 45% gap).

On top of providing a new commonsense reasoning benchmark, we demonstrate studies on transferrability among defined commonsense aspects, with an objective to spur future research on a more systematic probing of models' grasp of commonsense. As a potential future work drawn from these insights, we hope to inspire future model developments, specifically in two directions: 1) the ability to reason over known facts (*i.e.* reasoning scenario), and 2) acquiring the implicit knowledge that is commonsensible to humans (*i.e.* knowledge domain). Furthermore, we hope our investigation in the formulations of question answering task (*i.e.* MC setting versus our true/false complementary setting) can shed light on future explorations in identifying potential artifacts in NLP datasets.

7 Acknowledgements

We thank the anonymous reviewers for their feedback, and all the workers who have participated in the dataset creation on Amazon Mechanical Turk. We give special thanks to Peifeng Wang, Xiangci Li, and Gleb Satyukov for their great help in the early stage of annotation pipeline construction as well as providing valuable feedback in composing the guidline instructions. This work is supported by the Machine Common Sense (MCS) program under Cooperative Agreement N66001-19-2-4032 with the US Defense Advanced Research Projects Agency (DARPA). The views and the conclusions of this paper are those of the authors and do not reflect the official policy or position of DARPA.

8 Ethics and Broader Impacts

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ACM Code of Ethics* and honor the code of conduct. This work is mainly about the creation of a challenging commonsense benchmark dataset. The followings give the aspects of both our ethical considerations and our potential impacts to the community.

Dataset We collect an English dataset of commonsense complementary sentence pairs via Amazon Mechanical Turk (MTurk) and ensure that all the personal information of the workers involved (e.g., usernames, emails, urls, demographic information, etc.) is discarded in our dataset. This research has been reviewed by the **IRB board** and granted the status of an **IRB exempt**. The detailed annotation process (pay per amount of work, guidelines) is included in the appendix; and overall, we ensure our pay per task is above the the annotator's local minimum wage (~$12 USD/HR). Although commonsense can vary from different demographic areas, we primarily consider English speaking regions for the first round, and include more annotators from non English-spoken countries to diversify the dataset. Future work can include collecting a more diverse dataset across more demographics regions to incorporate more regional-dependent commonsense, while using some post editing to ensure English proficiency of the constructed data.

Techniques We benchmark the created dataset with the state-of-the-art large-scale pretrained language models, with minimum adaptation to the formulation of this dataset (*i.e.* true/false formulation). As commonsense is of our main focus, we do not anticipate production of harmful outputs, especially towards vulnerable populations, after training NLP models on our dataset.

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for

reading comprehension. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 8, pages 662–678.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 2924–2936, Minneapolis, Minnesota. 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 North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. In *Neural computation*, volume 9, pages 1735–1780. MIT Press.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics: EMNLP*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1078–1088. PMLR.

H. Levesque, E. Davis, and L. Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*.

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

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1823–1840, Online. Association for Computational Linguistics.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 839–849.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 4885–4901, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1158–1172. Association for Computational Linguistics.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*, volume 21, pages 1–67.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. In *Journal of Memory and Language*, volume 31, pages 507–524. Elsevier.

Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. In *Developmental science*, volume 10, pages 89–96. Wiley Online Library.

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

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

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *International Conference on Language Resources and Evaluation (LREC)*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. In *Transactions of the Association for Computational Linguistics (TACL)*.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations (ICLR)*.

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

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. In *Computer Vision and Image Understanding*, volume 163, pages 21–40. Elsevier.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Mastering the dungeon: Grounded language learning by mechanical turker descent. In *International Conference on Learning Representations (ICLR)*.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

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

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 5, pages 379–395.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Additional Details of COM2SENSE

A.1 Collection with MTurk

Qualification Quiz To familiarize the workers with our collection task, we design a quiz with the following types of questions: 1) examine if a given statement can be correctly judged with only commonsense or it requires specialized knowledge, 2) infer the true/false label of a given statement, and 3) select the most suitable domain and scenario where a given statement belongs to.

Human Intelligence Tasks (HITs) The general instructions of our HIT page include: **Task Overview**, **Task Payment** and **Overall Task Procedure** for each category, to engage more workers. At the end of the HIT instruction page, a link is provided to redirect the workers to the data creation page, where more detailed instructions and useful resources for the creation tasks are provided. Besides passing our qualification quiz, the workers are also required to have a *HIT Approval Rate* greater than 98% and the *Number of HITs Approved* greater than 5000. In each HIT assignment, workers are required to submit three complementary pairs. In the first phase of data collection, the base pay is \$0.6 for each assignment and workers will receive a \$0.5 bonus per sentence if it follows our instructions and fools the model; for the second phase, the base payment for each assignment is \$0.3 but we change the bonus to: \$0.5 (for either high-quality sentence or successful fooling) or \$0.9 (if both requirements are met, similar to those for the \$0.5 bonus in the first phase) to encourage workers to create higher quality data.

A.2 Details of the Creation

Tool Interface Screenshots of our creation interface are as shown in Figure 6 and Figure 7. We name our deployed model (RoBERTa-large) *Carl* to help emphasize the interactive and gamified creation set-up.

Guidelines To inspire workers and collect from more diverse topics of commonsense, we further provide: 1) some hints for having higher chances fooling the model, such as exploiting contradictory physical concepts, negations, swapping entities, etc., 2) topics pertinent to the domains, and (3) examples of low quality along with their reasons.

A.3 Details of Validation

To ensure data quality, our internal members have helped checking each pair with the validation tool we implement. For each pair received from the workers, both labels for the statements and their intended domains and scenarios are carefully verified. For statements which are ambiguous even for humans, if they can be easily fixed by adding more context or better word choices, another round of editing is conducted.

A.4 Adversarial Setting

The total number of collected complementary pairs is around 4.8k, where around .8k are discarded for not having sufficiently high quality, *e.g.* "*Frank traded a stock an hour late and lost 80 million dollars.*" and "*Frank traded a stock a second late and lost 80 million dollars.*" Among all the data we collected, the overall fooling rate is 48.55% per sentence and 78.7% per pair. For category-specific fooling rates, please refer to Figure 10.

For sentences that successfully fool the model, we report the mean time of fooling one sentence to be 3.40, the standard deviation (std) as 2.48, and the median as 2.57 (all in minutes). Please notice that the total time is directly retrieved from MTurk and is likely to be overestimated due to worker inactivity. The mean of the number of revisions per fooling sentence is 1.36 with a std as 3.95, and a median as 0 (fooling without re-attempts, requiring no revision). Noticeably, 63% of the fooling sentences are submitted with no revision.

For any potential interests, Figure 8 shows the mean and median of required time of fooling per sentence across categories, and similarly Figure 9 for the mean and median of the number of attempts which equals to the number of revisions $+1$.

Figure 11 shows the distribution of ratings during the exit survey from a total of 699 valid responses. The survey questions include: 1) how helpful is our instruction? and 2): how challenging is our task? For question 1, the mean rating is 4.66 ± 0.59 and median is 5; for question 2, the mean rating is 4.23 ± 0.92 and the median is 4, where 1-5 is from low-to-high rating.

A.5 Statistics of Workers

173 workers participated in our task, and Figure 12 shows the worker counts for the different numbers of assignments attempted by each of the workers, and Figure 13 shows the worker counts for the time

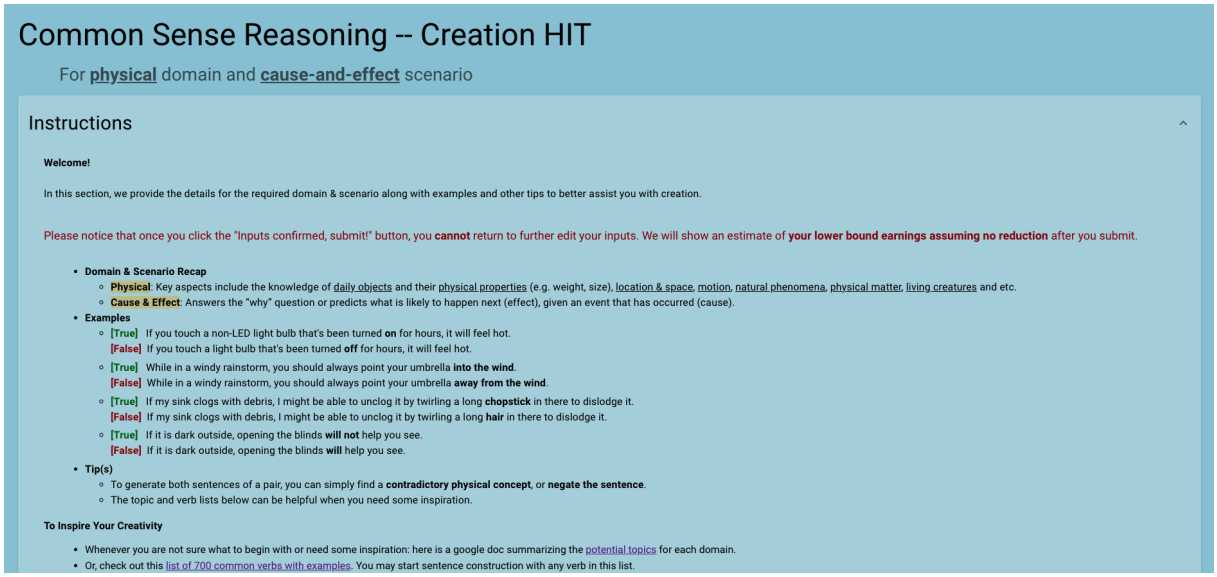


Figure 6: Screenshot of the creation interface (instruction section).

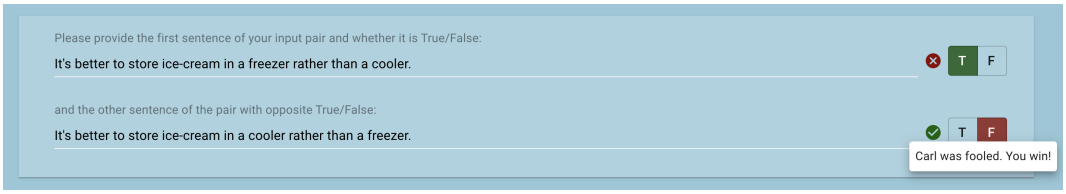
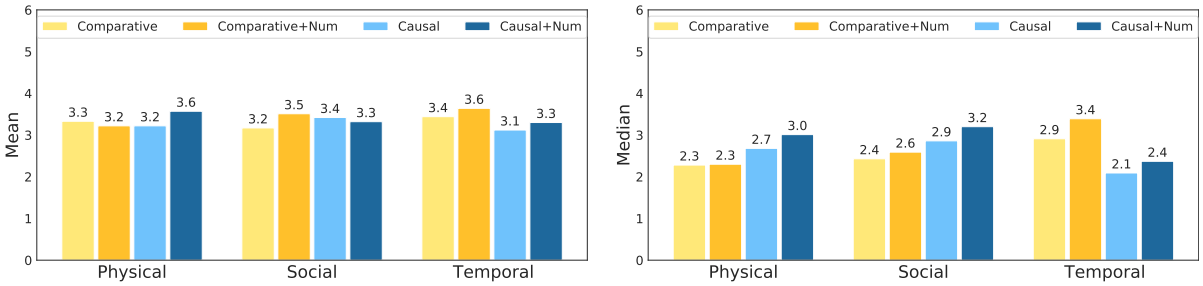
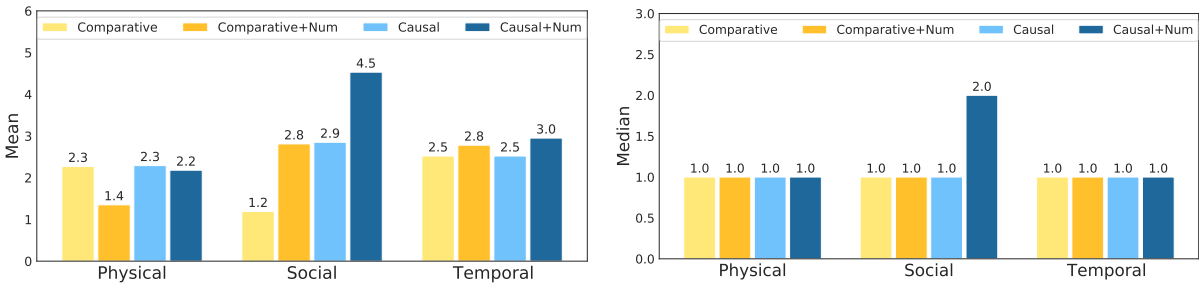


Figure 7: Screenshot of the creation interface (1/3 input section).



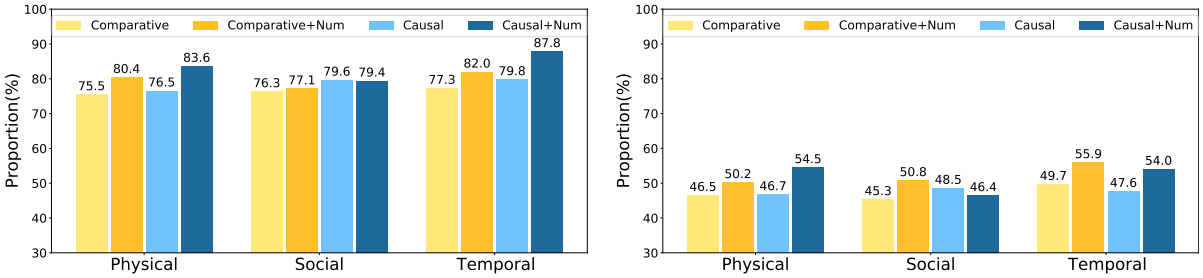
(a) Mean (b) Median

Figure 8: Mean and median of the time needed (in minutes) to fool a sentence for all categories, "+Num" denotes numeracy involved.



(a) Mean (b) Median

Figure 9: Number of attempts (*i.e.* # revisions + 1) per sentence for all categories, "+Num" denotes numeracy involved.



(a) Standard (b) Pairwise

Figure 10: Sentence fooling rates for all categories, "+Num" denotes numeracy involved.

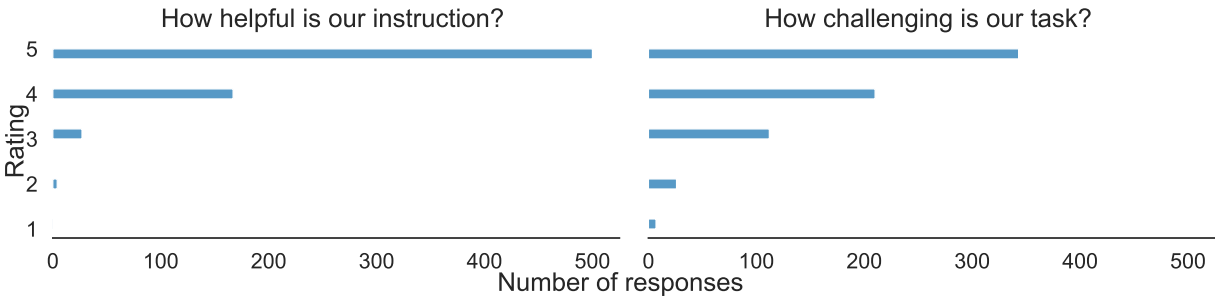


Figure 11: Rating distribution in exit survey.

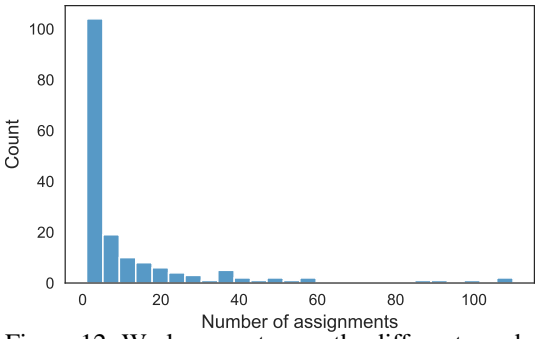


Figure 12: Worker counts over the different numbers of assignments.

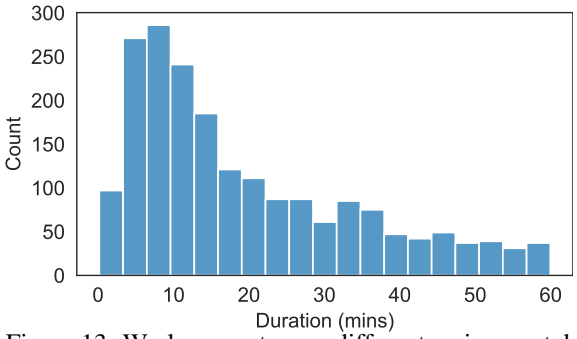


Figure 13: Worker counts over different assignment duration (in minutes).

(duration, in minutes) each worker spent on one assignment.

B Additional Details on Baseline Models

We include several essential implementation details of the benchmark models in the following:

Bi-LSTM+GloVe Our Bi-LSTM model (Hochreiter and Schmidhuber, 1997) is one-layered with a 512-dimensional hidden layer, which takes input word embeddings from 300-dimensional GloVe word embeddings (Pennington et al., 2014). We train all LSTM layers from scratch.

BERT-base Models For BERT-style architectures we employ a multi-layer-perceptron (MLP) on top of the [CLS] special token for binary prediction.

T5-large To adopt T5-large's text-to-text format to our dataset, we use the prefix *com2sense sentence:* and the labels *True* and *False* as model output.

UnifiedQA Models We use two UnifiedQA Models. One with the T5-large backbone and one with the T5-3b backbone. For these models, we use *Is the following sentence correct?* as the prefix, to create a question. Then as the answer we use *Yes* / *No*.

C More Details on the Experiments

C.1 Hyperparameters

All the essential hyperparameters used throughout this work can be referred to in Table 10. We also include the search bounds as well as the number of trials in searching for our manually-tuned hyperparameter search procedures in Table 10.

Model	# Params	Batch-Size	LR	Training Iterations	Gradient Accumulation Steps	Max. Token Length
BiLSTM+GloVe	3.5M	64	1×10^{-5}	100	4	80
BERT-base	109.5M	64	1×10^{-5}	100	4	80
RoBERTa-large	355.4M	32	1×10^{-5}	100	4	80
DeBERTa-large	405.2M	32	1×10^{-5}	100	4	80
T5-large	737.5M	8	1×10^{-5}	100	4	80
UnifiedQA-t5-large	737.5M	8	1×10^{-5}	100	4	80
UnifiedQA-t5-3b	3000M	2	1×10^{-5}	100	8	64
Bound (lower-upper)		1-64	5×10^{-5}–1×10^{-6}	10-100	1-10	
Number of trials		2-3	2-3	2-3	2-3	

Table 10: Hyperparameters used for each model during finetuning on COM2SENSE along with the search bounds for them: *LR* denotes the learning rate that does not change during the training process. All the models are trained with Adam optimizers (Kingma and Ba, 2015). We include number of parameters of each model in the first column, denoted as *# params*.

C.2 Validation Set Results

We validate all trained models on a 402-pair validation set and tune the hyperparameters accordingly. The performances on the validation set are reported in Table 11.

Model	Standard	Pairwise
Random	50.00	25.00
BiLSTM+GloVe	52.80	27.50
BERT-base	57.07	23.11
RoBERTa-large	62.81	38.30
T5-large	62.81	35.82
UnifiedQA-large	63.43	37.31
DeBERTa-large	66.29	43.03
UnifiedQA-3b	**75.12**	**56.22**

Table 11: Validation-set accuracy for selected models, trained and evaluated on respective datasets.

C.3 Performance Across Input Lengths

Although the sentence length in our dataset varies, we find no obvious relation between the length of the sentences and the difficulty for the model to comprehend, in terms of accuracy. As depicted in Figure 14, we therefore conclude that sentence length would not have significant influence on fooling models including DeBERTa-large.

C.4 Software, Hardware, & Other Details

Transformer-based models are implemented via the HuggingFace PyTorch API (Wolf et al., 2020). All the benchmarked models, except for UnifiedQA-T5-3b are trained on Nvidia GeForce 2080Ti GPUs[5] on a CentOS 7 operating system. The UnifiedQA-T5-3b is trained on NVIDIA Tesla V100 GPUs [6] on an Ubuntu 18 operating system.

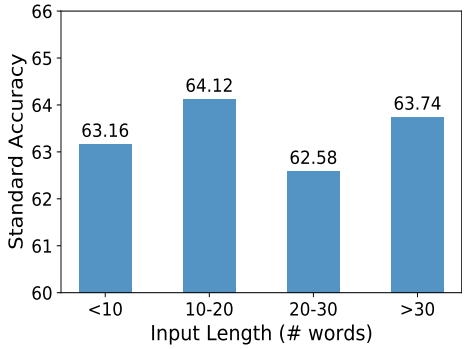


Figure 14: Standard accuracy of the DeBERTa-large model measured on subsets of data with different input lengths.

The T5-large and UnifiedQA-T5-large are trained using the model parallelism approach on two GPUs. The UnifiedQA-T5-3b is trained using model parallelism on 8 GPUS.

The maximum training time is approximately 6 hours for all the models, with the BERT-style models on the lower end of the range and the T5-style models on the higher end.

[5] https://www.nvidia.com/en-us/geforce/graphics-cards/rtx-2080-ti/

[6] https://www.nvidia.com/en-gb/data-center/tesla-v100/