Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding

Shane Storks* **Qiaozi Gao†** **Yichi Zhang*** **Joyce Chai***
*Computer Science and Engineering Division, University of Michigan
†Department of Computer Science and Engineering, Michigan State University
{sstorks, zhangyic, chaijy}@umich.edu
gaoqiaoz@msu.edu

Abstract

Large-scale, pre-trained language models (LMs) have achieved human-level performance on a breadth of language understanding tasks. However, evaluations only based on end task performance shed little light on machines' true ability in language understanding and reasoning. In this paper, we highlight the importance of evaluating the underlying reasoning process in addition to end performance. Toward this goal, we introduce Tiered Reasoning for Intuitive Physics (TRIP), a novel commonsense reasoning dataset with dense annotations that enable multi-tiered evaluation of machines' reasoning process. Our empirical results show that while large LMs can achieve high end performance, they struggle to support their predictions with valid supporting evidence. The TRIP dataset and our baseline results will motivate verifiable evaluation of commonsense reasoning and facilitate future research toward developing better language understanding and reasoning models.

1 Introduction

Recent years have seen a surge of research activities toward commonsense reasoning in natural language understanding. Dozens of relevant, large-scale benchmark datasets have been developed, and online leaderboards encourage broad participation in solving them. In the last few years, extraordinary performance gains on these benchmarks have come from large-scale language models (LMs) pre-trained on massive amounts of online text (Peters et al., 2018; Radford et al., 2018a,b; Raffel et al., 2020; Brown et al., 2020). Today's best models can achieve impressive performance and have surpassed human performance in challenging language understanding tasks, including benchmarks for commonsense inference (Bowman et al., 2015; Zellers et al., 2018; Bhagavatula et al., 2020). This rapid period of growth and progress has been an undoubtedly exciting time for NLP.

Despite these exciting results, it is a subject of scrutiny whether these models have a deep understanding of the tasks they are applied to (Bender and Koller, 2020; Linzen, 2020). A key concern is widespread bias in language benchmarks leading to superficial correlations between context and class labels (Schwartz et al., 2017; Gururangan et al., 2018; Poliak et al., 2018), allowing systems to bypass reasoning and achieve artificially high performance (Niven and Kao, 2019; McCoy et al., 2019). Consequently, it remains unclear whether the problems are truly solved, and whether machines can perform verifiable reasoning as humans do.

In this work, we first introduce Tiered Reasoning for Intuitive Physics (TRIP), a benchmark targeting physical commonsense reasoning. TRIP poses a high-level end task for story plausibility classification, a common proxy task for commonsense reasoning problems (Roemmele et al., 2011; Mostafazadeh et al., 2016; Sap et al., 2019b; Bisk et al., 2020b). Notably, however, it includes dense annotations for each story capturing multiple tiers of reasoning beyond the end task. From these annotations, we propose a tiered evaluation, where given a pair of highly similar stories (differing only by one sentence which makes one of the stories implausible), systems must jointly identify (1) the plausible story, (2) a pair of conflicting sentences in the implausible story, and (3) the underlying physical states in those sentences causing the conflict. The goal of TRIP is to enable a systematic evaluation of machine coherence toward the end task prediction of plausibility. In particular, we evaluate whether a high-level plausibility prediction can be *verified* based on lower-level understanding, for example, physical state changes that would support the prediction.

We further present several baseline systems powered by large LMs. Our empirical results show that while large LMs can achieve high end task performance (up to 78% accuracy), they struggle to

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4902–4918
November 7–11, 2021.

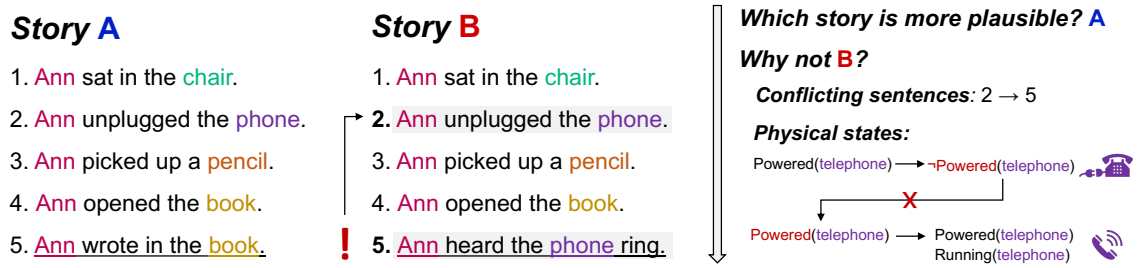


Figure 1: Story pair from TRIP, along with the tiers of annotation available to represent the reasoning process.

jointly support their predictions with the proper evidence (only up to 11% of examples supported with correct physical states and conflicting sentences). Consequently, the predictions from these powerful systems are overwhelmingly not accountable to their understanding of how the world works.

The contributions of this work are the first-of-its-kind dataset TRIP and new metrics that facilitate quantitative evaluation of coherent reasoning in commonsense language understanding. Our detailed analysis by applying large LMs on this dataset demonstrates key disconnections between low-level and high-level predictions in the reasoning process. This dataset and our baseline results motivate future work to develop systems that are capable of *verifiable* language understanding and reasoning.

2 Tiered Reasoning for Intuitive Physics

Physical commonsense reasoning, also referred to as naïve physics (Davis and Marcus, 2015) or intuitive physics (Lake et al., 2017), has recently gained attention in the NLP community (Gao et al., 2016; Forbes and Choi, 2017; Mishra et al., 2018; Bosselut et al., 2018; Forbes et al., 2019; Bisk et al., 2020b). From a young age, humans possess commonsense knowledge and reasoning skills about a wide variety of physical phenomena, such as movement, rigidity, and balance (Bliss, 2008). This problem is consequently thought to be especially challenging for machines because physical commonsense is considered obvious to most humans, and suffers from reporting bias (Forbes and Choi, 2017). As NLP systems are typically trained only on written communications, it remains unclear whether they can learn this (Bisk et al., 2020a). We have developed a dataset in English to target this domain and shed more light on this question.

2.1 TRIP Dataset

The Tiered Reasoning for Intuitive Physics (TRIP) is a benchmark for physical commonsense reasoning that provides traces of reasoning for an end task of plausibility prediction. The dataset consists of human-authored stories, such as those in Figure 1, describing sequences of concrete physical actions. Given two stories composed of individually plausible sentences and only differing by one sentence (i.e., Sentence 5), the proposed task is to determine which story is more plausible. To understand stories like these and make such a prediction, one must have knowledge of verb causality[1] and precondition[2], and rules of intuitive physics.[3]

Plausible stories were crowd-sourced from Amazon Mechanical Turk.[4] To convert each story into several implausible stories, we hired separate workers to each write a new sentence to replace a sentence in the original story, such that the new story after replacement is no longer realistic in the physical world. To ensure quality, these workers flagged stories which were incoherent or did not describe realistic actions. We eliminated those stories and performed a manual round of validation to remove any remaining bad stories and correct typos.

2.2 Controlled Data Curation

TRIP was carefully curated and restricted to support probing of reasoning abilities possessed by text classifiers. Compared to current benchmark trends, this dataset has the following unique properties.

Objectivity in physical commonsense. As commonsense knowledge differs between humans based on region, culture, and other factors (Davis, 2017), plausible reasoning tasks can become ambiguous and subjective, for example, in open-domain commonsense reasoning problems (Zhang et al., 2017; Bhagavatula et al., 2020). To address

[1]For example, *cutting* an object causes it to be in pieces, and *melting* an object causes it to be in liquid form.

[2]For example, to *cut* an object, it must be in solid form, but to *stir* an object, it must be in liquid form.

[3]For example, the constraint that an object inside of a container moves when its container moves.

[4]https://www.mturk.com/

this issue, we directed story authors to write sentences involving concrete actions, which can be unambiguously visualized in the physical world, while avoiding mental actions such as to *think* or *like*. We limit stories to typical household happenings by directing annotators to write stories in one of six possible "rooms" seen in everyday life.

To further reduce subjectivity and block other confounding factors that may result from complex use of language, we encourage crowd workers to write sentences in a simple declarative form, typically starting with the agent of the story, followed by a verb, a direct object, and an optional indirect object. The simplicity of language use would additionally allow us to focus less on linguistic processing and semantic phenomena, and more on investigating machines' reasoning ability.

Plausibility in longer context. Many benchmarks for plausible reasoning only (or most frequently) provide one sentence of context, with similarly short choices to complete the context (Roemmele et al., 2011; Zellers et al., 2018; Bisk et al., 2020a). In TRIP, we imposed several restrictions to require reasoning over multiple sentences with associated physical state changes. First, we required annotators to write stories at least five sentences long. Further, when collecting new sentences to convert plausible stories into implausible stories, we required that the new sentence should be plausible in isolation, and only become implausible when considering the world state implied by other sentences in the story. This constraint encourages stories to be rich in interesting action dynamics rather than nonsense sentences such as "Mary fried eggs on the printer" or "Tom ate the spoon," which may be easier to recognize through distributional biases. As this new sentence can conflict with any other sentence(s) in the story, solving the task requires reasoning over the entire context.

Multi-tier annotation. To enable a systematic investigation of a system's reasoning process, we manually provided three levels of annotation. As shown in Figure 1, the first level is the *end task label* to indicate which of the two story choices are more plausible. By design, most implausible story choices have exactly one pair of *conflicting sentences*, e.g., Sentences 2 and 5 in the example. The second level of annotation identifies these sentences in each story. On a random set of 100 implausible stories from the training data, a second annotator labeled these pairs of sentences, reaching a near-perfect Cohen's κ (Cohen, 1960) of 0.929, supporting the objectivity of these labels. The third level justifies the implausibility with labels for the underlying *physical states*, giving a detailed account of the physical changes associated with each sentence. In our example, unplugging *the phone* in Sentence 2 causes it to lose power, while Sentence 5 requires that the phone is powered in order to *ring*.

Measure	Train	Val.	Test	All
# plausible stories	370	152	153	675
# implausible stories	799	322	351	1472
avg. # sentences	5.1	5.0	5.1	5.1
avg. sentence length	8.3	8.0	8.5	8.3
# story authors	97	57	62	134
avg. # stories/author	3.8	2.7	2.5	5.0
avg. # conflicting sentence pairs	1.2	1.2	1.2	1.2
# physical state labels	18.8k	8.74k	9.09k	36.6k

Table 1: Statistics of the TRIP dataset. Implausible stories in each partition are generated from and paired with the plausible stories in the same partition.

In order to generate these rich annotations, we defined a space of 20 physical attributes (5 for humans, 15 for objects) which capture most conflicts found in the stories. This was collected in part from related attribute spaces proposed by Gao et al. (2016) and Bosselut et al. (2018), and chosen based on a random set of implausible training stories, specifically the nature of their conflicts and physical changes objects underwent during the stories. For each entity in each sentence in the dataset, we annotate the implied values of these attributes before (precondition) and after (effect) the events of the sentence take place. This step of the annotation was a substantial effort. Note that while relevant entities in each sentence are provided in the data for convenient evaluation, these can be fairly reliably extracted using the noun chunk parser from spaCy.[5] To verify the quality of annotations, we measured inter-annotator agreement on a representative subset of 157 sentences from 31 stories in the training data, finding a substantial Cohen's κ of 0.7917. A detailed description of this annotation process can be found in Appendix A.

Table 1 lists the overall statistics of the resulting dataset. While this dataset is small by today's standards, our goal is depth, not breadth. Rather than training models on a surplus of data to simply achieve high accuracy on the end task, we aim

[5] https://spacy.io/

to use our deep, multi-tiered annotations to probe the capability of NLP models to perform coherent reasoning toward the end task.

2.3 Proposed Tasks

From the TRIP dataset, we propose several tiered tasks as shown in Figure 1. Together, these tasks form a human-interpretable reasoning process supported by a chain of evidence.

Physical state classification. From our physical state annotations, we propose two tasks for each sentence-entity pair in each story choice: precondition and effect state classification. For example, consider the entity *potato* in the sentence "John cut the cooked potato in half." First, we should predict that the potato was solid in order to be *cut*, i.e., the precondition label for the `solidity` attribute is *true*. Second, we should predict that the potato was in pieces as a result of being *cut*, i.e., the effect label for the `in pieces` attribute is *true*.

Conflict detection. Next, we define the task of conflict detection as identifying a pair of sentences in the form $S_i \rightarrow S_j$. S_j is a *breakpoint*, i.e., the point where the story first becomes implausible given the context so far, while S_i serves as *evidence* that explains the breakpoint, usually causing a conflicting world state. For example, in Figure 1, Sentence 5 is a breakpoint, while Sentence 2 is the evidence that explains why the story becomes implausible after Sentence 5. Note that it is possible that a story may have multiple pairs of conflicting sentences beyond the breakpoint and evidence pair. However, across the dataset, the average number of conflicting sentence pairs is only 1.2, so one conflicting sentence pair is a sufficient and simpler explanation for the conflict (albeit not exhaustive).

Story classification. Lastly, the end task is to determine which of two stories is the plausible one. This should be determined based on any conflicts detected within the two stories.

2.4 Benchmark Goals

It is important to note that while one can treat these tasks separately, the goal of this benchmark is to solve them jointly to form a coherent reasoning chain: physical state classification explains conflict detection, which further explains story classification. Unlike most existing benchmarks in this area, which assess language understanding ability through some high-level end tasks, the goal of our benchmark is to enable development of systems for interpretable and consistent reasoning toward language understanding. Our baseline models (Section 3) and evaluation metrics (Section 4.1) are developed to serve this purpose.

It is also worth noting that although data bias is an issue for high-level benchmark tasks where systems are not required to justify their predictions, we are not directly targeting this issue. Recent work has attempted to remove biases from benchmark data and thus prevent exploitation of them in performing high-level tasks (Zellers et al., 2018; Nie et al., 2020). In contrast, our framing of language understanding as being built from the ground up (i.e., from low-level to high-level tasks) provides systems with the proper supporting evidence toward high-level tasks, and thus can potentially mitigate some of the problems around data bias.

3 A Tiered Baseline for TRIP

Figure 2 displays a high-level view of our proposed baseline system to solve TRIP. It individually embeds each sentence-entity pair in each story, classifies physical precondition and effect states, then identifies conflicting sentences from these. Given a pair of stories, it aggregates conflict predictions for each story to decide which is more plausible.

3.1 Module Implementations

Each module is implemented as some kind of neural network architecture. Here, we describe some details of the implementations.

Contextual Embedding. The Contextual Embedding module is implemented as a pre-trained, transformer-based language model. Generally, this module takes as input a sentence and the name of an entity from a story, following an entity-first input formulation (Gupta and Durrett, 2019), and outputs a dense, contextualized numerical representation.

Precondition and Effect Classifiers. The Precondition and Effect Classifiers are implemented as typical feedforward classification heads for contextual embeddings, with one precondition classifier and one effect classifier for each of the 20 physical attribute tracked in the dataset. Softmax is applied to the output for classification. Altogether, the predictions from these classifiers label physical states of each entity in each sentence of the story.

Conflict Detector. For each entity and its predicted physical states over all sentences in a story,

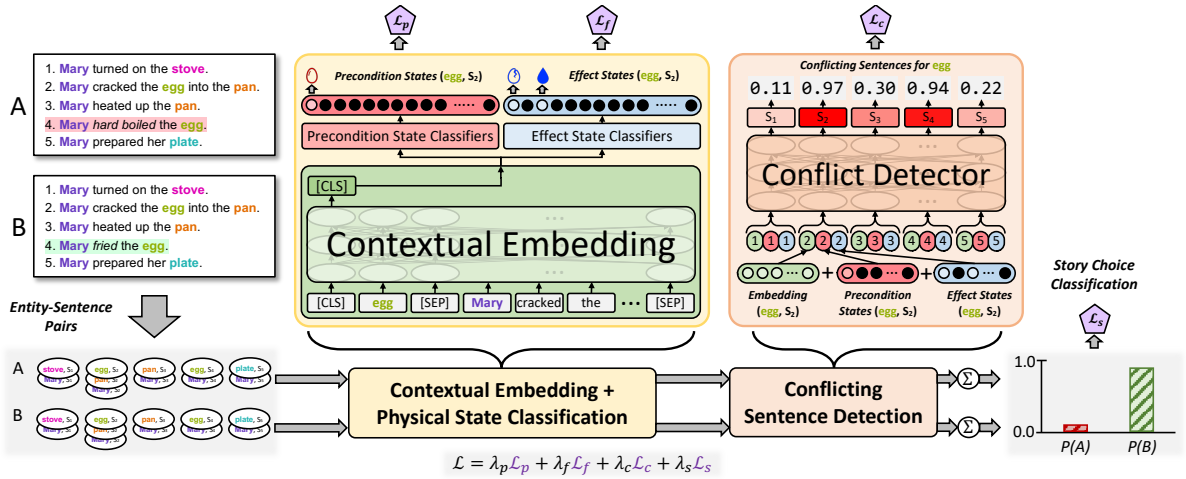


Figure 2: Proposed tiered reasoning system with loss functions $\mathcal{L}_p$ for precondition state classification, $\mathcal{L}_f$ for effect state classification, $\mathcal{L}_c$ for conflicting sentence detection, and $\mathcal{L}_s$ for story choice classification. The model is trained end-to-end by optimizing the joint loss $\mathcal{L}$, a weighted sum of these loss functions.

the Conflict Detector predicts whether there is some conflict in the entity's physical states, specifically flagging a pair of conflicting sentences through multi-label classification. We use another transformer for this module, but model the high-level sequence of sentences in a story rather than the low-level sequence of tokens in a sentence. For each sentence-entity pair, we input the contextual embedding, as well as the classification logits behind all physical state predictions. We apply an additional feedforward classification layer and sigmoid function to the generated hidden states in order to model the belief probability of each sentence conflicting with another sentence in the story.

Story choice prediction. Given any detected conflicts, we lastly select which of the two given stories is plausible. As each Conflict Detector output represents a belief that the physical states of an entity in a particular sentence conflict with that of another sentence, we can simply sum the negative outputs for each story and apply softmax to determine which story is least likely to have a conflict.

3.2 Model Training

We train the architecture's parameters through gradient descent on the overall loss $\mathcal{L}$:

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

$\mathcal{L}$ sums individual cross-entropy loss functions $\mathcal{L}_p$ for precondition classification, $\mathcal{L}_f$ for effect classification, $\mathcal{L}_c$ for conflict detection, and $\mathcal{L}_s$ for story choice classification, each balanced by respective weights λ_p, λ_f, λ_c, λ_s summing to 1.

4 Experiments

Using TRIP, we evaluate several variations of the proposed reasoning system powered by selected pre-trained language models: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021).[6] These models offer a range in design choices such as model complexity and size of pre-training data. We begin with an evaluation from the perspective of the end task, then take a detailed look at the lower-level tasks.

4.1 Evaluation Metrics

To enable a better understanding of machines' ability in coherent reasoning toward end task performance, we apply the following evaluation metrics.

Accuracy. The traditional metric of end task accuracy, i.e., the proportion of testing examples where plausible stories are correctly identified.

Consistency. The proportion of testing examples where not only the plausible story is correctly identified, but also the conflicting sentence pair for the implausible story is correctly identified. This is to demonstrate the consistency with identified conflicts when reasoning about plausibility.

Verifiability. The proportion of testing examples where not only the plausible story and the conflicting sentence pair for the implausible story are correctly identified, but also underlying physical states (i.e., preconditions and effects) that contribute to the conflict are correctly identified.[7] This is to

[6] We use the "large" configurations of BERT (355M parameters) and RoBERTa (355M parameters), and the "base" configuration of DeBERTa (140M parameters).

[7] At least one nontrivial, i.e., non-default, positive-class physical state label must be predicted in the preconditions of

demonstrate that the detected conflict can be verified by a correct understanding of the underlying implausible change of physical states.

It is worth noting that this notion of verifiability, although different, is motivated by the notion of *verification* in software engineering (Pierce, 1996). This term refers to determining whether a given software solution satisfies its architectural and design requirements, and is built from the correct sub-components. Along this line, our notion of verifiability can be seen as a method to evaluate whether a language understanding system's reasoning process is built up from the correct components.

Each successive metric dives deeper into the coherence of reasoning that supports the end task prediction. Consequently, if accuracy is a, consistency is b, and verifiability is c, then $a \geq b \geq c$. A system that reliably produces a coherent chain of reasoning is demonstrated by $a \approx b \approx c$.

4.2 Results

Recall that we consider four loss functions for training the tiered system: $\mathcal{L}_p$ for precondition classification, $\mathcal{L}_f$ for effect classification, $\mathcal{L}_c$ for conflicting sentence detection, and $\mathcal{L}_s$ for story choice classification. To investigate how each loss affects model performance, we train instances using several combinations of them. The results of this study on the validation set are listed in Table 2.

The role of end task supervision. In the first section of Table 2, we train the system jointly on all four loss functions. Here, we see low verifiability and consistency for all three LMs, while the end task accuracy is relatively high, reaching 78.3% when using BERT. When we omit the story classification loss in the second section, however, we see sharp gains in verifiability and consistency for all models, with RoBERTa jumping from 0.9% verifiability and 6.8% consistency to 10.6% and 22.4%, respectively. This comes at a slight cost of end task accuracy for BERT and RoBERTa.

This suggests that while fine-tuning systems based on a high-level classification loss targeting the end task can improve the end task accuracy, this drastically reduces the interpretability of the underlying reasoning process. One potential explanation for this is that this loss drives the system to exploit spurious statistical cues in order to further increase the end task accuracy. This gives us motivation to move away from using over-simplified end tasks to train and evaluate language understanding. In fact, if we fine-tune RoBERTa's contextual embedding directly on the end task of TRIP without intermediate classification layers, we can achieve up to 97% accuracy, but have no insight toward verifiability or consistency of the system. This raises questions about the validity of such a result.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	47.8	11.3	0.0
All Losses			
BERT	**78.3**	2.8	0.0
RoBERTa	75.2	6.8	0.9
DeBERTa	74.8	2.2	0.0
Omit Story Choice Loss $\mathcal{L}_s$			
BERT	73.9	**28.0**	9.0
RoBERTa	73.6	22.4	**10.6**
DeBERTa	75.8	24.8	7.5
Omit Conflict Detection Loss $\mathcal{L}_c$			
BERT	50.9	0.0	0.0
RoBERTa	49.7	0.0	0.0
DeBERTa	52.2	0.0	0.0
Omit State Classification Losses $\mathcal{L}_p$ and $\mathcal{L}_f$			
BERT	75.2	17.4	0.0
RoBERTa	71.4	2.5	0.0
DeBERTa	72.4	9.6	0.0

Table 2: End and tiered task metrics for tiered classifiers on the validation set of TRIP trained on varied combinations of loss functions. Random baseline (averaged over 10 runs) makes tiered predictions at random.

Natural emergence of intermediate predictions. In the third and fourth sections of Table 2, we respectively omit conflict detection loss and state classification losses to explore whether conflicting sentences or physical states would emerge naturally in the reasoning process. When omitting conflict detection loss, all metrics degrade to near or below random performance. Clearly, conflict detection is not implicitly learned from the downstream story classification loss, and since the story choice classification directly depends on the conflict detection output, the end task accuracy drops as well.

Meanwhile, when omitting physical state classification loss, verifiability unsurprisingly drops to zero, but high accuracy on the end task can still be achieved by all models (up to 75.2%). Notably, this suggests that reasonable supporting evidence is not required in order to achieve high accuracy on the end task. This casts further doubt that existing state-of-the-art results on other commonsense lan-

the breakpoint sentence and effects of the evidence sentence, and all such predictions must be correct.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	49.5	10.7	0.0
BERT	70.9	21.9	**8.3**
RoBERTa	**75.2**	18.8	5.7
DeBERTa	72.9	**22.2**	6.6

Table 3: Metrics for the best tiered systems on the test set of TRIP. Compared to random baseline.

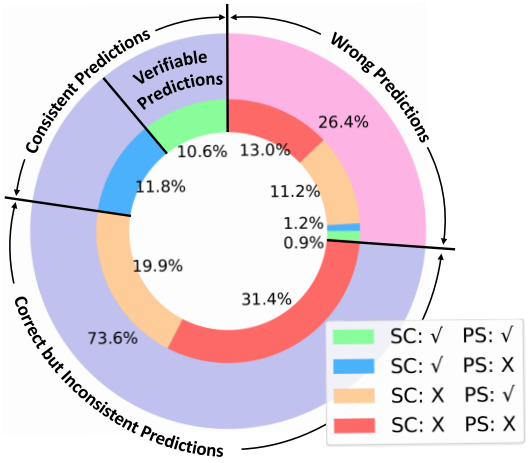


Figure 3: Distribution of RoBERTa successes and failures on TRIP. SC (sentence conflict) and PS (physical state) denote whether the predicted conflicting sentences or physical states are correct (✓) or not (×).

guage understanding benchmarks possess any kind of coherent reasoning beyond end classification tasks which over-simplify the problem.

In Table 3, we present the testing results for the best loss function configuration of the system, i.e., omitting story choice classification loss. Compared to the validation set results in Table 2, we see slight drops in consistency and verifiability, further demonstrating the difficulty of this problem.

4.3 Analysis

Given the poor performance along our proposed metrics, we next consider the connections between the tiered tasks, and what goes wrong in unverifiable end task instances. We focus our analysis on the systems achieving the highest verifiability on the validation set in Section 4.2.

Failure mode distribution. Figure 3 provides a detailed breakdown of the combinations of failure modes on the validation set. Of the 73.6% of validation instances that are classified correctly on the end task, almost half of these (31.4% overall) are entirely unverified, with incorrect physical states and conflicts predicted by the system. Similarly,

Model	Prec. F1 (%)	Eff. F1 (%)	Confl. F1 (%)
BERT	**54.9**	57.2	66.3
RoBERTa	51.2	51.2	**69.6**
DeBERTa	52.8	**57.3**	63.6

Table 4: Macro-F1 scores of best tiered systems on aggregate precondition, effect, and conflicting sentence classification. Scores averaged over all attributes for physical state classification.

of the 26.4% of instances with *incorrect* end task predictions, about half (13% overall) have incorrect physical state and conflict predictions. Meanwhile, a combined 31.1% of instances correctly predict physical states in the conflicting sentences of the implausible story, but fail to detect a conflict in those sentences (19.9% are correct at the end task, while 11.2% are not). These instances, represented by orange wedges in the graph, are a significant disconnect in the reasoning process.

Low-level task performance. To further address this disconnect, we examined system performance from the perspective of physical state classification and conflict detection. First, Table 4 lists the validation metrics for our best baselines on the tasks of precondition and effect classification (by sentence-entity pair), as well as conflicting sentence detection (by end task instance). Across the board, we find reasonable performance on all tasks.

The best performing baseline from Table 2 is trained using loss functions for both physical state classification and conflict detection. Given this configuration, we further examined how each task is learned. Figure 4 shows training curves for the loss functions of physical state classification (averaged for precondition and effect), conflicting sentence detection, and story choice classification. Notably, though story choice classification is not used as a training objective, this end task is learned fairly well (albeit overfitting), with training and validation losses generally decreasing through training. This shows that learning to reason from the lower-level tasks is successful to some degree. However, the lower-level tasks appear challenging to learn. For physical state classification, losses decrease steadily, but slowly. For conflict detection, the losses also decrease slowly, and the model begins overfitting the training data, perhaps indicating a need for more training data at this challenging step. Future work may consider automatic data augmentation techniques to resolve this.

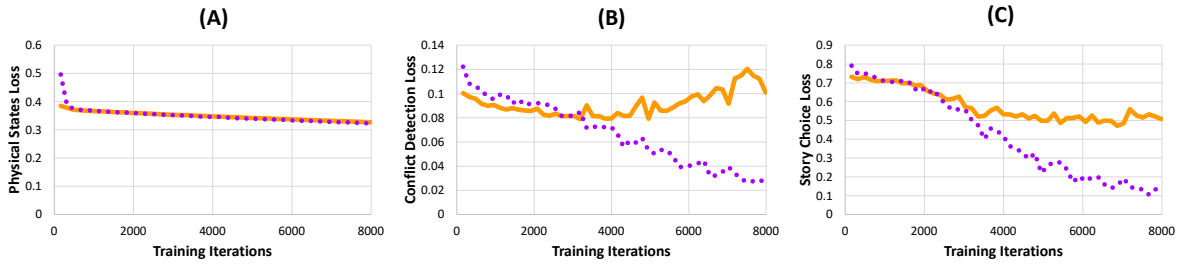


Figure 4: Training (purple, dotted) and validation (orange, solid) losses for best tiered RoBERTa system trained on TRIP for 10 epochs. Uses the best configuration of the loss functions (as found in Section 4.2) for (A) physical state classification, (B) conflict detection, and (C) story choice classification. Validation loss recorded 4 times per epoch, with training loss averaged over the trained batches since the previous recording.

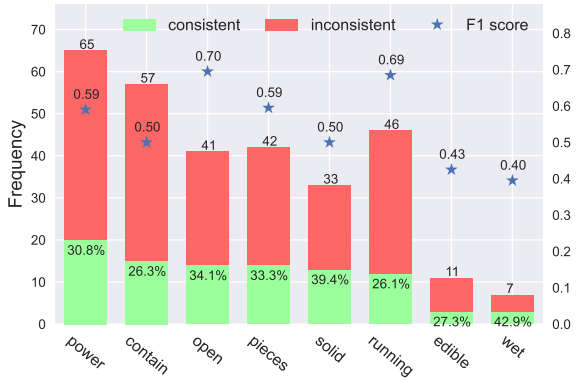
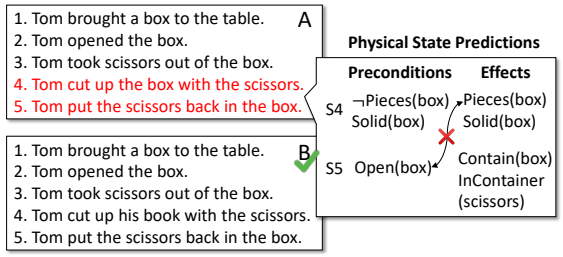


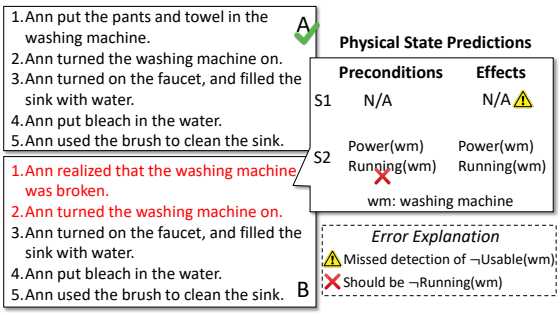
Figure 5: Contribution of correct RoBERTa-predicted physical states to consistency evaluation for selected attributes. The macro-F1 score of precondition and effect predictions is shown by blue stars. Among all correctly predicted states (for both effects and preconditions), the bar regions indicate whether these states appear in successfully detected conflicting sentences.

Connecting states to conflicts. To dig deeper into the connection between physical states and plausibility conflicts, we next examined correct physical state predictions by attribute in Figure 5. In the graph, we indicate the percentage of predictions supporting a successfully detected conflict, which may be interpreted as a *utility* measure of each attribute toward conflict detection. We find that some attributes, like whether an electrical object is `running`, rarely contribute to successful conflict detections (only 26.1%) despite having reasonably high F1 score (0.69). Other attributes, like `wet`, are more likely to appear in successful conflict detections when predicted correctly, even though their overall classification performance is lower. This provides strong insights for targeted improvement, for example, to better take advantage of lower-level predictions toward high-level tasks.

Sample system outputs. Figure 6 presents sample outputs from the tiered RoBERTa system. In Example (a), the prediction is entirely verifiable.



(a) A verifiable prediction.



(b) A consistent but not verifiable prediction.

Figure 6: Sample outputs from the baseline system. The detected conflicting sentences are in red, and physical state predictions are shown on the right.

The system correctly chooses the plausible story, identifies Sentences 4 and 5 as the conflicting sentences in the implausible story, and even predicts that the *box* is `in pieces` after Sentence 4, and thus cannot become `open` in Sentence 5. In Example (b), the prediction is consistent but unverifiable, as the system identifies a conflict between Sentences 1 and 2, but cannot support the conflict with correct underlying physical states in either sentence. Although some relevant attributes are identified for the breakpoint sentence, e.g., `power` and `running`, they are not quite right. Meanwhile, no states are predicted for the evidence sentence.

5 Related Work

Physical commonsense. There exist a few NLP datasets around physical commonsense reason-

ing which offer various classification tasks. ProPara (Mishra et al., 2018) tracks existence and location of entities in each sentence, similar to TRIP's physical state classification, but in a more restricted state space. Physical Interaction Question Answering (PIQA) from Bisk et al. (2020b) provides a similar high-level end task of multiple-choice text plausibility classification targeting physical commonsense. Other benchmarks focus on specific domains of physical reasoning, such as temporal reasoning (Zhou et al., 2019) and spatial reasoning (Mirzaee et al., 2021). Visual (Johnson et al., 2017; Bakhtin et al., 2019) and multimodal (Hudson and Manning, 2019; Das et al., 2018; Anderson et al., 2018; Shridhar et al., 2020) benchmarks also investigate systems' commonsense understanding of the physical world through perception and interaction. Different from these existing benchmarks, TRIP is the first dataset of its kind with dense annotation to support evaluation of verifiable reasoning toward the end task prediction.

Robust language inference. In the face of statistical bias enabling artificially high performance in NLP models, several works have explored ways to evaluate and enable robust language inference. Several probing studies have examined how well surface-level syntactic and semantic phenomena are captured in contextual language embeddings (Adi et al., 2017; Ettinger et al., 2018; Tenney et al., 2018; Hewitt and Manning, 2019; Jawahar et al., 2019; Tenney et al., 2019). For stronger evaluation of potentially biased systems, others have explored specialized natural language inference tasks (Welleck et al., 2019; Uppal et al., 2020) and logic rules (Li et al., 2019; Asai and Hajishirzi, 2020) to support and evaluate consistency of models across instances of the end task. Some approaches have been proposed to instead remove biases from language by filtering out data too easily discriminated by state-of-the-art text classifiers (Zellers et al., 2018; Nie et al., 2020), and to improve robustness of systems against exploiting various types of biases (Belinkov et al., 2019; Clark et al., 2019; Min et al., 2020). Recent work has attempted to compile large amounts of semi-structured commonsense knowledge (Sap et al., 2019a; Mostafazadeh et al., 2020) and inject this knowledge into pre-trained language models (Bosselut et al., 2019; Zhang et al., 2019) in order to enable knowledge-supported language understanding and on-the-fly explanation. Different from these efforts, this paper enables direct training and evaluation of consistent and verifiable language inference by providing a dataset that makes explicit the underlying evidence chains behind a high-level text classification task.

6 Conclusion and Discussion

In this work, we proposed TRIP, a tiered benchmark dataset for physical commonsense reasoning posing a new challenge of jointly solving low-level to high-level tasks to form a coherent reasoning process. We experimented with several variations of tiered systems to solve the tasks. Our results show that in many cases, *supervising large LMs based on high-level classification tasks in order to learn commonsense language understanding leads to inconsistent and unverifiable reasoning*, and inability to capture intermediate evidence toward the end task. Instead, we should train systems to jointly incorporate multiple types of lower-level evidence to solve reasoning tasks coherently.

Our detailed analysis of results offers strong intuition for future progress toward this goal. As such, TRIP and our baselines provide an important first step toward verifiable, human-aligned commonsense language understanding, and a direction for development of AI systems in this area.[8]

Broader impact. We use physical commonsense reasoning as an example in this work, but expect that a similar approach can apply to many aspects of language understanding. Our results have shown that a new challenge for the future will be to build machines that can reason logically and coherently, similar to what we expect from human reasoning. As these machines ultimately will work with humans, such alignment in reasoning is critical, as it will improve accountability and transparency in human-machine enterprise.

Acknowledgements

This work was supported in part by the National Science Foundation (IIS-1617682 and IIS-1949634) and the DARPA XAI program through UCLA (N66001-17-2-4029). We thank Bri Epstein and Haoyi Qiu for their assistance in annotation and hyperparameter tuning during this work, and the anonymous reviewers for their helpful comments and suggestions.

[8] Our source code and data are publicly available at https://github.com/sled-group/Verifiable-Coherent-NLU.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *arXiv:1608.04207 [cs]*.

Aida Amini, Antoine Bosselut, Bhavana Dalvi, Yejin Choi, and Hannaneh Hajishirzi. 2020. Procedural reading comprehension with attribute-aware context flow. In *Conference Automated Knowledge Base Construction (AKBC 2020)*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, Salt Lake City, UT, USA. IEEE.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020a. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020b. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Inteligence (AAAI-20)*, New York, NY, USA. AAAI Press.

Joan Bliss. 2008. Commonsense reasoning about the physical world. *Studies in Science Education*, 44(2):123–155.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating Action Dynamics with Neural Process Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

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 (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Samuel R. 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 (EMNLP 2015)*, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv: 2005.14165*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. 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 2019)*, Hong Kong, China. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA. IEEE.

Ernest Davis. 2017. Logical Formalizations of Commonsense Reasoning: A Survey. *Journal of Artificial Intelligence Research*, 59:651–723.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, Minneapolis, MN, USA.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing Composition in Sentence Vector Representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, NM, USA. Association for Computational Linguistics.

Maxwell Forbes and Yejin Choi. 2017. Verb Physics: Relative Physical Knowledge of Actions and Objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, BC, Canada. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do Neural Language Representations Learn Physical Commonsense? In *Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci 2019)*, Montreal, QC, Canada.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical Causality of Action Verbs in Grounded Language Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany. Association for Computational Linguistics.

Aditya Gupta and Greg Durrett. 2019. Effective Use of Transformer Networks for Entity Tracking. 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 2019)*, Hong Kong, China. Association for Computational Linguistics.

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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654*.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, Minneapolis, MN, USA. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA. IEEE.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. IEEE.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural 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)*, Hong Kong, China. Association for Computational Linguistics.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 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: 1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. SpartQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2021)*, Online. Association for Computational Linguistics.

Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking State Changes in Procedural Text: A Challenge Dataset and Models for Process Paragraph Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, New Orleans, LA, USA. 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 Framework for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016)*, San Diego, CA, USA. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized Story Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online. Association for Computational Linguistics.

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 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018)*, New Orleans, LA, USA. Association for Computational Linguistics.

P. Pierce. 1996. Software verification and validation. In *IEEE Technical Applications Conference. Northcon/96. Conference Record*, pages 265–268.

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*, New Orleans, LA, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018b. Improving Language Understanding with Unsupervised Learning. *OpenAI*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford, CA, USA.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*, Vancouver, BC, Canada.

Maarten Sap, Ronan LeBras, 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 Thirty-Third AAAI Conference on Artificial Inteligence (AAAI-19), Honolulu, HI, USA. AAAI Press.

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

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language (CoNLL 2017)*, Vancouver, BC, Canada. Association for Computational Linguistics.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, Online.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada.

Shagun Uppal, Vivek Gupta, Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. Two-step classification using recasted data for low resource settings. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

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

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, Brussels, Belgium. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-Sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy. Association for Computational Linguistics.

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 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, Hong Kong, China. Association for Computational Linguistics.

A Physical State Annotations

To collect our physical state annotations, we defined a space of 20 physical attributes (5 for humans, 15 for objects) which capture most conflicts found in the stories, collected in part from related attribute spaces proposed by Gao et al. (2016) and Bosselut et al. (2018). For humans, we track *location*, *hygiene*, and whether a human is *conscious*, *dressed*, or *wet*. For objects, we consider *location* and whether or not an object *exists*, is *clean*, connected to *power*, *functional*, *in pieces*, *wet*, *open*, *hot*, *solid*, *occupied* (i.e., containing another object), *running* (i.e., turned on), *movable*, *mixed*, or *edible*.

The values of these attributes each represent directions of physical state change (e.g., attribute became true or attribute became false), as listed in Appendix A.1. In the training data, we manually labeled each entity in the sentence with these attributes and values. For the other partitions, we used a semi-automatic approach described in Appendix A.2.

A.1 Physical Annotation Label Space

When labeling entities for directions of physical state changes in sentences, we adopted the label space in Table 5. For predicting precondition and effect in non-location attributes as done in this work, it is straightforward to collapse this space into *true*, *false*, or *unknown* for each. For human location labels, we use the full label space for predicting both precondition and effects for simplicity. Meanwhile, for object location labels, we simplify the problem by mapping them to smaller precondition and effect label spaces. While this does not significantly affect verifiability, this should be expanded in a full solution for better interpretability. For more detailed explanations, future work may consider tracking spans of text describing entity locations along the lines of Amini et al. (2020).

A.2 Completing Physical State Annotations

To expand our manual physical state annotations to the validation and testing data, we used the existing annotations to train classifiers to predict values for each attribute given a sentence-noun pair. First, each story was broken down into all possible sentence-noun pairs, using spaCy[9] to identify noun phrases. These sentence-noun pairs were passed into the physical state classifier,[10] implemented as 20 parallel branches of RoBERTa, one for each physical attribute, as shown in Figure 7. For efficiency, we use the pretrained DistilRoBERTa$_{\text{BASE}}$ parameters (82M), distilled from RoBERTa$_{\text{BASE}}$ by Liu et al. (2019) with a small performance reduction (Sanh et al., 2019). Using this module, we generated candidate physical state annotations for the remaining data, then manually revised them. As a different annotator completed this work from the annotator who completed the training data, we measured inter-annotator agreement on a representative subset of 157 sentences from 31 stories in the training data, finding a substantial Cohen's κ (Cohen, 1960) of 0.7917.

Label	Human Location	Object Location	Other Attributes
0	irrelevant	irrelevant	irrelevant
1	disappeared	disappeared	*false* → *false*
2	moved	picked up	*true* → *true*
3	–	put down	*true* → *false*
4	–	put on	*false* → *true*
5	–	removed	___ → *no*
6	–	put in container	___ → *true*
7	–	taken out of container	*false* → ___
8	–	moved	*true* → ___

Table 5: Label space and meanings for human location, object location, and other attributes. Each label represents a specific physical change (or lack of change).

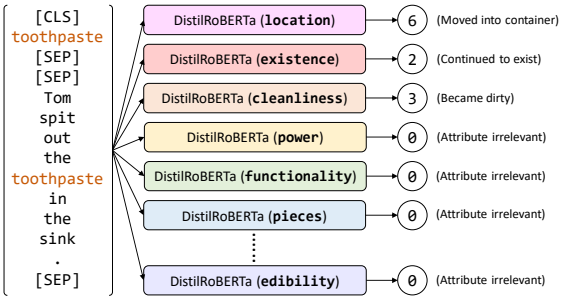


Figure 7: Proposed structure of the physical state classifier, consisting of 20 parallel instances of DistilRoBERTa. Each instance outputs an integer representing a particular kind of change (or lack of change) in the corresponding attribute.

[9] https://spacy.io/

[10] Followed Gupta and Durrett (2019) for formatting the input in order to generate entity-centric embeddings.

B Model Implementation Details

Each module in our tiered systems is implemented as some kind of neural network architecture. Here, we describe low-level details of the implementations.

Contextual Embedding. The Contextual Embedding module is implemented as a pre-trained transformer language model. Generally, this module takes as input a sentence and the name of an entity from a story, and outputs a dense numerical representation. We follow Gupta and Durrett (2019) in using an entity-first input to the language model to generate entity-centric embeddings. While there are some model-specific variations in special tokens, given an entity e and a sentence $t_1, t_2, \cdots, t_n$, we structure the input sequence as " [CLS] e [SEP] t_1 t_2 $\cdots$ t_n [SEP]," where [CLS] is a special token meant for input to classification layers, and [SEP] is a special separator token for multi-text inputs.

Precondition and Effect Classifiers. The Precondition and Effect Classifiers are implemented like typical classification heads for contextual embeddings, with one precondition classifier and one effect classifier for each of the 20 physical attribute tracked in the dataset. Specifically, each classifier is made up of two feedforward layers, each preceded by a dropout layer (using model specific defaults for dropout probability), with $\tanh$ activation in between them. The first layer performs a linear transformation on an input contextual embedding, while the second layer projects the hidden state to the size of the label space for the corresponding attribute. Argmax is applied to the output for classification. Altogether, the predictions from these classifiers label physical states of each entity in each sentence of the story.

Conflict Detector. For each entity and its predicted physical states over all sentences in a story, the Conflict Detector predicts whether there is some conflict in the entity's physical states, specifically flagging a pair of conflicting sentences through multi-label classification. Again, we use a transformer (6 additional layers with 8 attention heads) for this module, but model the high-level sequence of sentences in a story rather than the low-level sequence of tokens in a sentence. For each sentence-entity pair, we consider the contextual embedding generated earlier, as well as the logits for all predicted precondition and effect states. We project both representations through linear layers to the same size, then concatenate them to form an entity dynamics representation. This representation for each sentence is input to the transformer, and the resulting hidden states are concatenated. Lastly, we use a feedforward layer followed by sigmoid activation to transform the hidden state to a belief probability of each sentence conflicting with another sentence in the story.

Story choice prediction. Given the output from the Conflict Detector, we lastly need to select which of the two given stories is plausible. As each Conflict Detector output represents the belief that a particular sentence conflicts with another sentence, we can simply sum the negative outputs for each story and apply softmax to determine which story is least likely to have a conflict.

Loss function details. To jointly train these various modules, we must balance several loss functions. The loss functions are weighted by corresponding scalar weights λ_p, λ_f, λ_c, and λ_s. In preliminary experiments, we found the best balance between state classification and the other tasks with the following assignment of weights: $\lambda_p = \lambda_f = \frac{0.4}{|A|}$, $\lambda_c = \lambda_s = 0.1$, where $|A|$ is the number of attributes tracked, i.e., 20. When omitting different loss functions, we rebalance the weights by ensuring $\lambda_c + \lambda_s = 0.2$, or $\lambda_c = \lambda_s$ where state classification losses are omitted.

C Model Training Details

The RoBERTa, BERT, and DeBERTa models are built from HuggingFace's `Transformers` library (Wolf et al., 2020), particularly their implementation for multiple-choice classification, and the pre-trained $\text{BERT}_{\text{LARGE}}$ parameters (336M), $\text{RoBERTa}_{\text{LARGE}}$ parameters (355M), and $\text{DeBERTa}_{\text{BASE}}$ parameters (140M) respectively. For all models, we use the AdamW optimizer (Loshchilov and Hutter, 2018). Batch size is fixed at 1 story pair for all models, the maximum allowed by our GPU memory. To select the optimizer learning rate and number of training epochs, all models are trained by grid search over these two, maximizing the validation set verifiability as defined in Section 4.1. Learning rate is selected from the set $\{1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$, while the maximum number of epochs is fixed at 10. Ties are broken first by validation accuracy on the end plausibility classification

Model	Learning Rate	Epochs
Table 2, All Losses		
BERT	5e-6	5
RoBERTa	1e-5	8
DeBERTa	5e-6	6
Table 2, Omit Story Choice Loss		
BERT	5e-5	9
RoBERTa	1e-5	6
DeBERTa	5e-5	8
Table 2, Omit Conflict Detection Loss		
BERT	1e-6	2
RoBERTa	5e-6	9
DeBERTa	1e-6	4
Table 2, Omit State Classification Loss		
BERT	1e-5	4
RoBERTa	1e-6	8
DeBERTa	5e-6	10

Table 6: Selected learning rate (LR), number of training epochs, and validation verifiability and accuracy for all results presented in the paper.

task, then by selecting the model instance trained for fewer epochs (to avoid overfitting). The selected learning rate and number of epochs for each model presented in the main paper are listed in Table 6.

D Supplementary Results

Lastly, we provide additional results that were omitted from the main paper.[11]

D.1 Conflict Detector Ablations

The Conflict Detector module takes in two types of inputs: 1) contextual embeddings of sentence-entity pairs, and 2) physical state logits from the Precondition and Effect Classifiers. To determine the impact of each, we present ablations omitting them for the best-performing instances from the previous section, i.e., those not considering story choice classification loss. Table 7 presents these results for the validation set, while Table 8 presents these results for the test set.

Without including the physical state inputs, we see a slight drop in consistency and verifiability of some models. For example, RoBERTa drops from 9.7% verifiability and 23.4% consistency to 4.6% and 17.7%, respectively. Meanwhile, DeBERTa increases from 8.0% verifiabiliy and 20.2% consistency to 11.4% and 24.5%. While RoBERTa seems to depend slightly on the predicted physical states in performing conflict detection, DeBERTa favors the contextual embedding.

Without including the contextual embeddings, we see a drastic drop across the board to below-random performance, with RoBERTa dropping to 0% verifiability and consistency, and DeBERTa to 2.3% and 6.6% respectively. This suggests that while forcing the model to track physical states enables greater explanation, they are not sufficient for models to learn conflict detection, or they are not incorporated successfully into the higher-level predictions. The contextual embedding, which is fine-tuned on physical state classification and conflict detection jointly, seems to be most powerful for solving the end task. Future work should further explore how to harness the rich information provided by the physical states to improve system performance and interpretability.

[11] Note that the results in this appendix use a slightly simpler label space for `location` state classification, and thus are not directly comparable to the results presented in the main paper.

Model	Verif. (%)	Acc. (%)	Prec. F1 (%)	Eff. F1 (%)	Confl. F1 (%)
Contextual Embeddings + Physical States					
BERT	9.6	70.2	74.4	66.7	65.1
RoBERTa	**12.1**	**77.0**	72.3	62.7	70.9
DeBERTa	11.2	72.7	77.0	71.1	68.2
Contextual Embeddings Only					
BERT	**10.9**	72.7	75.9	69.3	66.7
RoBERTa	9.6	**76.1**	72.5	61.6	70.3
DeBERTa	9.9	**76.1**	77.3	71.3	68.6
Physical States Only					
BERT	0.6	54.7	60.5	59.9	51.1
RoBERTa	0.0	43.2	38.4	37.8	49.5
DeBERTa	2.2	58.1	81.0	79.0	53.0

Table 7: Validation set performance of best models in Table 2 when ablating inputs to the Conflict Detector.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Contextual Embeddings + Physical States			
BERT	63.2	15.7	7.4
RoBERTa	**76.6**	23.4	9.7
DeBERTa	72.9	20.2	8.0
Contextual Embeddings Only			
BERT	70.7	16.8	6.8
RoBERTa	**76.6**	17.7	4.6
DeBERTa	74.1	**24.5**	**11.4**
Physical States Only			
BERT	56.1	3.4	0.3
RoBERTa	42.2	0.0	0.0
DeBERTa	59.3	6.6	2.3

Table 8: Validation set performance of best models in Table 2 when ablating inputs to the Conflict Detector.

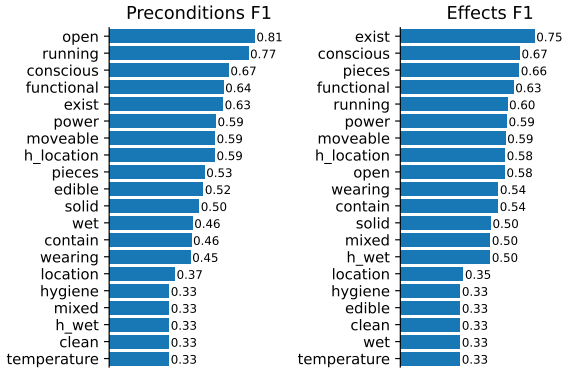


Figure 8: Precision and recall of predictions for each attribute from our best RoBERTa model on the validation set.

D.2 State Classification Performance by Attribute

Figure 8 breaks down the F1 score for predicting precondition and effect states by attribute across the TRIP dataset. We find that for preconditions, openness and whether objects are running, i.e., activated, are best captured, and for effects, existence and consciousness are. Meanwhile, wetness and temperature are challenging for predicting both preconditions and effects.