A Supplementary Material

A.1 Data

Counterfactual Invariance Prediction (CIP) In this work, we define a new task that tests the capability of models to predict whether under the assumption of a counterfactual event, a (later) factual event remains invariant or not in a narrative context. The formal setup is: given the first three consecutive sentences from a narrative story s_1 (premise), s_2 (initial context), s_3 (factual event) and an additional sentence $s_2^{'}$ that is counterfactual to the initial context s_2, the task is to predict whether s_3 is invariant given s_1, $s_2^{'}$ or not. Table 11, display some examples of CIP task.

How was the data collected? (Qin et al., 2019) proposed a dataset to encourage models to learn how to rewrite stories with counterfactual reasoning. They build the dataset on top of the ROCStories corpus, which comprises of five-sentence stories S = (s_1, s_2, ..., s_5). The formal setup is: s_1 (premise), s_2 (initial context), the last three sentences $s_{3:5}$ are the original ending of story. For each story they ask crowdworkers to write a counterfactual sentence to the initial context s_2 and also re-write the ending $s'_{3:5}$ according to the counterfactual sentence. We automatically collect counterfactual invariance examples by checking if (original) s_3 == (edited) s'_3 and similarly, if (original) s_3 != (edited) s'_3 non-invariant examples from their dataset to create a balanced dataset for our proposed CIP task.

A.2 Hyperparameter

In all models the Reasoning Cell and the Knowledge Encoder are both instantiated by a Transformer with 4 attention heads and depth = 4. For each task, we select the hyperparameters that yield best performance on the dev set. Specifically, we perform a grid search over the hyperparameter settings with a learning rate in $\{1e\text{-}5, 2e\text{-}5, 5e\text{-}6\}$, a batch size in $\{4, 8\}$, and a number of epochs in $\{3, 5, 10\}$. Training is performed using cross-entropy loss. For evaluation, we measure accuracy. We report performance on the test sets by averaging results along with the variance obtained for 5 different seeds. We use Adam Optimizer, and dropout rate = 0.1. The best hyperparameter details are stated in Table 9. We experimented on GPU size of 11GB and 24GB.

Datasets	Learning Rate	Epochs	Batch
αNLI	5e-6	5	8
CIP	5e-6	5	8

Table 9: Best Hyperparameter

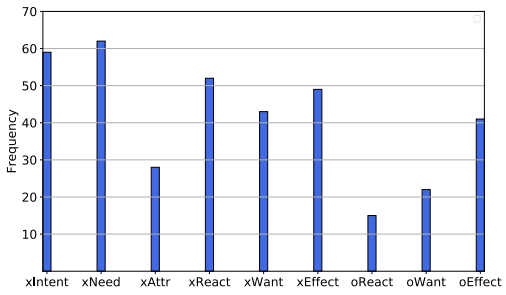

Figure 8: Distribution of relational Knowledge required per instances

A.3 Human Analysis

We conduct human evaluation to validate the effectiveness and relevance of the extracted social commonsense knowledge rules. We randomly selected 100 instances from the αNLI dev set for which the RoBERTa-Large Baseline had failed, along with gold labels. Firstly, we asked two annotators to construct the knowledge rules required to perform the αNLI task for these 100 instances, using the different dimensions of ATOMIC knowledge. We found that for 90 instances such relational knowledge is required, and on average 3 knowledge rules are important. This study suggests that models need to find a chain of rules to perform the inference task. Further, we test the robustness of the models' performance by removing random knowledge rules vs. removing knowledge rules with relations which were found most relevant by our annotators (namely, 'PersonX intent', 'PersonX's want', 'PersonX's need', 'effect on PersonX', 'effect on other', 'PersonX feels') see Figure 8.

A.4 Attention Visualization

We study the visualization the attention distributions over different (relation) dimensions of ATOMIC, produced by our MHKA (Reasoning Cell) model. It depicts the attention distribution and change in attention over multiple structured social knowledge rules (relations), and over different layers. It also allows to inspect inner working of the Reasoning cell. αNLI Examples: **Observation1:** *Dotty was being very grumpy.* **Observation2:** *She felt much better afterwards.* **Hypothesis:** *Dotty*

Relation	Question	Textual Description
xIntent	Why does X cause the event?	'because PersonX wanted'
xNeed	What does X need to do before the event?	'PersonX needed'
xAttr	How would X be described?	'PersonX is seen as'
xReact	How does X feel after the event?	'PersonX feels'
xWant	What would X likely want to do after the event?	'PersonX wants'
xEffect	What effects does the event have on X?	'effect on PersonX'
oReact	How do others' feel after the event?	others feel
oWant	What would others likely want to do after the event	others wants'
oEffect	What effects does the event have on others?	'effect on others'

Table 10: The taxonomy of if-then reasoning types from ATOMIC (Sap et al., 2019).

call some close friends to chat. The model correctly attended the relation 'xwant, xintent, xneed, effect on others'.

Context & Questions	Options
Bob had to get to work in the morning. His car battery was struggling to start the car. He called his neighbor for a jump start. **Alternatively:** *Bob had to get to work in the morning. His car won't start.*	answer [Yes]
Bill and Teddy were at the bar together. Bill noticed a pretty girl. He went up to her to flirt. **Alternatively:** *Bill and Teddy were at the bar together. Bill noticed his mom was there.*	answer [No]
I loved to eat honey with my oatmeal. One day I unexpectedly ran out of honey. I did not want to eat my oatmeal without honey. **Alternatively:** *I loved to eat honey with my oatmeal. One day I realized that maple syrup was even better with my oatmeal.*	answer [No]
I went to las vegas. I learned that i really like the slot machines. I spent a lot of time on them. **Alternatively:** I went to las vegas. I learned that slot machines are a great way to make money.	answer [Yes]

Table 11: CIP Examples

Example of ANLI:

Observation1 Jordan was playing fetch with his dog.
Observation2 His dog splashed water all over his neighbor's porch.
Hypothesis1 It was really rainy outside.
Hypothesis2 The dog jumped into the pool.

Extracted Knowledge using SRL + COMET 2.0:

{Jordan playing fetch with his dog : <**Jordan, wanted, to have fun**>, <Jordan, needed, to have a ball> , <Jordan, is seen as, playful> , <Jordan, feels, happy> , <Jordan, wants, to have fun> , <effect on, Jordan, gets exercise> , <others, feel, happy> , <others, wants, to have fun> , <effect on, others, dog runs away> }

{It was really rainy outside : <PersonX, wanted, to be dry> , <PersonX, needed, to go outside>, <PersonX. is seen as, wet>, <PersonX, feels, wet>, <PersonX, wants, to dry off>, <effect on, PersonX, gets wet>, <others, feel, wet> , <others wants to have fun> , <effect on, others, they get wet>}

{His dog splashed water all over his neighbor's porch: <dog, wanted, to have fun>, <dog, needed, to be outside>, <**dog, is seen as, careless>, <**dog, feels, guilty>, <dogs, wants, to clean up the mess>}

{The dog jumped into the pool : <**the dog, wanted, to have fun>, <**The dog, needed, to go the pool> , <The dog, is seen as, playful>, <The dog, feels, happy>, <the dog, wants, to have fun>, <effect on, the dog gets wet>, <others, feel, happy>, <others, wants, to have fun>, <**effect on, others gets, splashed with water>}**

Figure 9: Example