GLUCOSE: GeneraLized and COntextualized Story Explanations

Nasrin Mostafazadeh[*] **Aditya Kalyanpur** **Lori Moon** **David Buchanan**[†]
Lauren Berkowitz **Or Biran** **Jennifer Chu-Carroll**
Elemental Cognition
New York, NY, USA
nasrin@verneek.com
{adityak, lorim, orb, jenniferc}@elementalcognition.com
david.buchanan@quillbot.com

Abstract

When humans read or listen, they make implicit commonsense inferences that frame their understanding of what happened and why. As a step toward AI systems that can build similar mental models, we introduce GLUCOSE, a large-scale dataset of implicit commonsense causal knowledge, encoded as causal mini-theories about the world, each grounded in a narrative context. To construct GLUCOSE, we drew on cognitive psychology to identify ten dimensions of causal explanation, focusing on events, states, motivations, and emotions. Each GLUCOSE entry includes a story-specific causal statement paired with an inference rule generalized from the statement. This paper details two concrete contributions. First, we present our platform for effectively crowdsourcing GLUCOSE data at scale, which uses semi-structured templates to elicit causal explanations. Using this platform, we collected a total of ˜670K specific statements and general rules that capture implicit commonsense knowledge about everyday situations. Second, we show that existing knowledge resources and pretrained language models do not include or readily predict GLUCOSE's rich inferential content. However, when state-of-the-art neural models are trained on this knowledge, they can start to make commonsense inferences on unseen stories that match humans' mental models.

1 Introduction

Humans make countless implicit commonsense inferences about everyday situations. For example, consider the following short story from the ROCStories corpus (Mostafazadeh et al., 2016): *Gage was riding his bike. A car turned in front of him. Gage turned his bike sharply. He fell off of his bike. Gage skinned his knee.* When even young children read this story, they construct a coherent representation of what happened and why, combining information from the text with relevant background knowledge (Kintsch and Van Dijk, 1978). For example, they can construct the causal chain that explains how the car's unexpected turn ultimately led to Gage falling, describe how Gage's emotion and location changed throughout the story, and even hypothesize that he likely shouted for help after falling.

Though humans build such mental models with ease (Zwaan et al., 1995), AI systems for tasks such as reading comprehension and dialogue remain far from exhibiting similar commonsense reasoning capabilities. Two major bottlenecks have been acquiring commonsense knowledge and successfully incorporating it into state-of-the-art AI systems. To address the first bottleneck, we have built an effective platform to acquire causal commonsense knowledge at scale. To address the second, we show that pre-trained neural models can start to make similar inferences when trained on such rich curated data.

We introduce the GLUCOSE[1] (GeneraLized and COntextualized Story Explanations) dataset. Given a short story and a sentence X in the story, GLUCOSE captures ten dimensions of causal explanation related to X. These dimensions, inspired by human cognitive psychology, cover often-implicit causes and effects of X, including events, location, possession, and other attributes, the vast majority of which are not captured by existing resources

[*] Current affiliation Verneek, Inc.

[†] Current affiliation QuillBot Inc.

[1] Human brain functions such as thinking, memory, and learning are closely linked to the glucose levels and how efficiently the brain uses this fuel source (Mergenthaler et al., 2013). If there is not enough glucose in the brain, neurotransmitters are not produced and communication between neurons breaks down. We are calling this resource GLUCOSE, since we believe AI brains need this source of fuel to enable their basic thinking and fill in their reasoning gaps!

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4569–4586,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Dimension	**Semi-structured Specific Statement and Inference Rule**: antecedent *connective* consequent
1: Event that directly causes or enables X	$\underbrace{\text{A car}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{in front of}}_{\text{preposition}}\ \underbrace{\text{him}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$ $\underbrace{\text{Something}_A}_{\text{subject}}\ \underbrace{\text{turns}}_{\text{verb}}\ \underbrace{\text{in front of}}_{\text{preposition}}\ \underbrace{\text{Something}_B\text{ (that is Someone}_A\text{'s vehicle)}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{turns}}_{\text{verb}}\ \underbrace{\text{Something}_B}_{\text{object1}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object2}}$
2: Emotion or basic human drive that motivates X	$\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{wants}}_{\text{verb}}\ \underbrace{\text{safety}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$ $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{wants}}_{\text{verb}}\ \underbrace{\text{safety}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{moves}}_{\text{verb}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{Something}_A\text{ (that is dangerous)}}_{\text{object}}$
3: Location state that enables X	$\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{was}}_{\text{verb}}\ \underbrace{\text{close to}}_{\text{preposition}}\ \underbrace{\text{a car}}_{\text{object}}$ *Enables* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object1}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{the car}}_{\text{object2}}$ $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{is}}_{\text{verb}}\ \underbrace{\text{close to}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object}}$ *Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{moves}}_{\text{verb}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object}}$
4: Possession state that enables X	$\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{possesses}}_{\text{verb}}\ \underbrace{\text{a bike}}_{\text{object}}$ *Enables* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$ $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{possesses}}_{\text{verb}}\ \underbrace{\text{Something}_A}_{\text{object}}$ *Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{moves}}_{\text{verb}}\ \underbrace{\text{Something}_A}_{\text{object}}$
5: Other attributes enabling X: N/A (the dimension is not applicable for this example)	
6: Event that X directly causes or enables	$\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{He}}_{\text{subject}}\ \underbrace{\text{fell off}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object}}$ $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{turns}}_{\text{verb}}\ \underbrace{\text{Something}_B\text{ (that is Someone}_A\text{'s vehicle)}}_{\text{object}}$ *Causes/Enables* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{falls off}}_{\text{verb}}\ \underbrace{\text{Something}_B}_{\text{object}}$
7: An emotion that is caused by X: N/A	
8: A change in location that X results in	$\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{turned}}_{\text{verb}}\ \underbrace{\text{his bike}}_{\text{object1}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{the car}}_{\text{object2}}$ *Results in* $\underbrace{\text{Gage}}_{\text{subject}}\ \underbrace{\text{was}}_{\text{verb}}\ \underbrace{\text{further}}_{\text{object1}}\ \underbrace{\text{from}}_{\text{preposition}}\ \underbrace{\text{the car}}_{\text{object2}}$ $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{moves}}_{\text{verb}}\ \underbrace{\text{away from}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object}}$ *Results in* $\underbrace{\text{Someone}_A}_{\text{subject}}\ \underbrace{\text{is}}_{\text{verb}}\ \underbrace{\text{further from}}_{\text{preposition}}\ \underbrace{\text{Something}_A}_{\text{object}}$
9: A change of possession that X results in: N/A	
10: Other changes in property that X results in: N/A	

Table 1: Entries in the GLUCOSE dataset that explain the Gage story around the sentence X= *Gage turned his bike sharply*. White and gray rows show specific statements and general rules, respectively. The syntactic slots used for constructing each semi-structured entry are shown underneath it.

and models. Importantly, GLUCOSE encodes commonsense knowledge in the form of semi-structured inference rules[2] (mini-theories about the world), each grounded in a specific story. As the examples in Table 1 demonstrate, the specific statements exemplify how the general rules can be grounded in a particular context.

To facilitate acquisition at scale, we designed an effective multi-stage crowdsourcing platform and used it to acquire more than 670K GLUCOSE annotations in the context of children's stories. Our analysis shows that these explanations extend substantially beyond the scope of the existing knowledge resources.

Given the breadth of commonsense knowledge needed for real-world inference tasks, no static knowledge source is expected to provide sufficient coverage. GLUCOSE's key contribution is enabling models to dynamically produce general inference rules to explain novel scenarios. To systematically evaluate such models, we present an evaluation task where given a story S, a sentence X, and dimension d, a model predicts relevant specific and general rules as captured in GLUCOSE. We evaluate on the task using a curated test set, based on novel stories not used for any training purposes. We show a strong correlation between human and automatic evaluation metrics, which makes systematic and reliable evaluation of models feasible. We show that pre-trained neural models perform poorly on the task; however, when fine-tuned on GLUCOSE data, they are able to generate commonsense explanations that rival humans'.

[2] We will use "inference rule" and "explanation" interchangeably: the "explanations" we are interested in are inference rules that explain a given sentence's causes and effects.

This finding supports our hypothesis that a promising recipe for giving machines commonsense is to use quality-monitored crowdsourced commonsense knowledge for training neural models that have preexisting lexical and conceptual knowledge.

2 Related Work

Recently, there has been a renewed interest in commonsense reasoning (Talmor et al., 2019; Tandon et al., 2019; Rashkin et al., 2018a; Zellers et al., 2018), further fostered by the increasing need for explainable AI systems (Yang et al., 2018).

One well-known type of commonsense knowledge is script knowledge, defined by Schank and Abelson (1977) as structured knowledge about stereotypical event sequences and their participants. However, manual encoding of such knowledge is notoriously unscalable and brittle. A more recent line of work is unsupervised learning of "narrative schemas" (Chambers and Jurafsky, 2008, 2009; Balasubramanian et al., 2013; Sha et al., 2016), where common event sequences are automatically induced from large corpora. While promising, this approach has not produced high-quality knowledge usable for downstream tasks at scale (Mostafazadeh et al., 2016). Furthermore, since commonsense knowledge is often implicit, such corpus-based methods are unlikely to induce implicit commonsense inferences (Gordon and Van Durme, 2013). In contrast, our data collection framework has enabled us to acquire high-quality and robust commonsense knowledge, including often unstated rules such as "Someone$_A$ gives Someone$_B$ Something$_A$ *Results in* Someone$_B$ possesses Something$_A$" or "Someone$_A$ is at Somewhere$_A$ *Enables* Someone$_A$ puts Something$_A$ at Somewhere$_A$".

The most fruitful efforts to date for acquiring commonsense knowledge have been crowdsourced knowledge resources. ConceptNet (Speer et al., 2017), a partially-crowdsourced resource, is a relational knowledge graph that connects short natural-language phrases via semantic edges. Most ConceptNet knowledge is taxonomic, consisting of factoids like "apple *is a* fruit", however, it also includes some causal relations, e.g., "kill *is motivated by* revenge." Despite its broad coverage, ConceptNet has been found to be noisy (Zhou et al., 2019). Its knowledge also lacks context, hampering accurate application at inference time, e.g., "kill *requires* eat breakfast" is hard to make sense of without more context.

A more directly relevant resource is ATOMIC (Sap et al., 2019), which consists of 877K textual descriptions of if-then knowledge. Each entry describes a likely cause/effect of one of 24K+ events. ATOMIC entries are organized into nine categories such as xIntent (PersonX's intention) and xEffect (effect on PersonX). For instance, "PersonX makes PersonY's coffee xEffect PersonX gets thanked". ATOMIC is a great step forward in acquiring high-quality inferential knowledge. However, it has two main shortcomings. First, ATOMIC is non-contextual and conflates knowledge about an event that may have occurred under different scenarios, which hinders interpreting and applying the knowledge in context. For example, the event "PersonX arrives the next day" has xIntents "to go on vacation" and "to attend a reunion," and xEffects "get time to relax" and "meet some friends." Although each xIntent should be associated with only one of the xEffects, such dependencies are not encoded in ATOMIC. As a result, ATOMIC cannot be used to determine which xEffect is more likely given an xIntent. GLUCOSE addresses this by grounding each piece of inferential knowledge to a particular story context consistent across dimensions.

Second, events and relations in ATOMIC are person centric; agentless events are not covered, and each relation is either about PersonX or PersonY. As a result, ATOMIC cannot describe events involving common entity types such as places, things, or groups of people, nor can it encode causes and effects other than to PersonX and their peers. In GLUCOSE, sentence X can describe any event/state, and GLUCOSE general rules can refer to indexed variables such as "Someone$_A$" or "Somewhere$_C$." Beyond these major shortcomings, ATOMIC also does not cover many commonsense knowledge types in GLUCOSE, including change of attributes such as location, which will be further discussed in Section 4.3.

3 The Knowledge Model of GLUCOSE

GLUCOSE has a unique take on explaining story events. As illustrated in Table 1, each story is explained through ten causal dimensions. The semi-structured explanation for each dimension includes both a specific statement and a general rule.

3.1 Causal Dimensions of Explanation

One of our main contributions is the identification of ten causal dimensions of explanation in the context of narratives, for which we can reliably collect high quality data from lay crowd workers. Cognitive psychology research on human comprehension of narratives (Kintsch and Van Dijk, 1978; Zwaan and Radvansky, 1998; Grazzani et al., 2018) suggests that humans primarily focus on events, their timeline, locations of entities throughout the story, causes and motivations of events, and emotional trajectory of characters.

Based on this research, GLUCOSE dimensions are designed to focus on causal reasoning around events and states, eliciting event causal chains, character motivations, emotions, naive psychology, and change of attributes such as location and possessions to core story entities. For an event or state X stated in a sentence, we categorize the dimensions of causality into events and states happening *before* X and those occurring *after* X. Each category includes five dimensions, as shown in Table 1. The precise definition and scope of these ten dimensions are the result of multiple pilot studies with crowd workers to identify intuitive and distinguishable causal dimensions, so that the overlap among dimensions is minimized and the agreement among workers is maximized.

3.2 Semi-structured Inference Rules

To uncover what constitutes a good explanation, we ran several pilot studies exploring how people define, generate, and present explanations about short stories. We concluded that in order to achieve some consensus among explanations and to facilitate further processing and evaluation, the explanations should not be entirely free-form. Instead, we represent them as semi-structured inference rules whose expressivity lies between free text and logical forms. Each rule takes the form "antecedent *connective* consequent," where the antecedent and consequent are composed by filling in syntactic slots for subject, verb, object(s), and preposition(s). For some dimensions, slot-filling involves choosing from a predefined list, e.g., dimension 2, which states a motivating emotion or basic human drive, limits its verb choices to *feel, want*, and *like*. Details regarding the slots can be found in Appendix A.

To eliminate the need for pronoun resolution when applying our general rules, variables are indexed, such as "Someone$_A$" and "Something$_A$ and Something$_B$", to refer to the same entities on both sides of the rule. Each variable can be further elaborated using an *attribute phrase* in the form of a relative clause, e.g., "Somewhere$_C$ (that is Someone$_A$'s location)." Our studies indicate that this format gives the explainers sufficient expressivity to convey their reasoning, yet constrains the resulting explanations enough to identify commonalities between them. Note that the semi-structured rules are deterministically converted to natural language form by simply concatenating all the filled slots. Table 1 shows examples of semi-structured GLUCOSE explanations.

3.3 Generalized and Contextualized

Each GLUCOSE explanation is stated both as a specific statement (grounded in a given context) and a corresponding general rule (applicable to other contexts). Research in cognitive psychology suggests that humans typically choose which of an event's many causes to cite based on its relevance to the context (Miller, 2019). Hence, grounding explanations in context is crucial for acquiring accurate explanations. Furthermore, it has been shown that human explanations take situation-specific information and link it to pre-existing knowledge about the world; people explain by appealing to broader theories that enable generalization (Lombrozo, 2006). Also, there is evidence that explanations and generalizations help scaffold cognitive development in humans (Busch et al., 2018), which can potentially play a role in the learning capabilities of AI systems as well. By explicitly stating general rules as mini-theories of how the world works, GLUCOSE seeks to enable better generalization and causal reasoning in future AI systems.

4 The GLUCOSE Dataset

4.1 Data Acquisition Platform

To enable developing models that can build mental models of narratives, we aimed to crowdsource a large, quality-monitored dataset. Beyond the scalability benefits, using crowd workers (as opposed to a small set of expert annotators) ensures diversity of thought, thus broadening coverage of a commonsense knowledge resource.

The annotation task is complex: it requires annotators to understand different causal dimensions in a variety of contexts and to come up with generalized theories beyond the story context. For strict quality control, we designed a three-stage knowledge acquisition pipeline for crowdsourcing the GLUCOSE dataset on the Amazon Mechanical Turk (Mturk) Platform. The workers first go through a qualification test[3] where they must score at least 90% on 10 multiple-choice questions on select GLUCOSE dimensions. Next, qualified workers can work on the main GLUCOSE data collection task: given a story S and a story sentence X, they are asked to fill in (allowing for non-applicable) all ten GLUCOSE dimensions, getting step-by-step guidance from the GLUCOSE data acquisition UI.[4] To ensure data consistency, the same workers answer all dimensions for an S, X pair. Finally, the submissions are reviewed by an expert who rates each worker on a scale from 0 to 3, and provides feedback on how to improve. Our final UIs are the result of more than six rounds of pilot studies, iteratively improving the interaction elements, functionality, dimension definitions, instructions, and examples.[5] See Appendix B for more details on our crowdsourcing pipeline.[6]

4.2 Dataset Composition and Statistics

Our source of stories for the GLUCOSE dataset is ROCStories (Mostafazadeh et al., 2016). ROCStories consists of crowdsourced five-sentence everyday stories rich in causal and temporal relations, making them ideal for acquiring commonsense knowledge. We focus on children's stories due to their simpler language and concepts. We computed an estimated target age[7] for each story and sampled from the 5–8 age group. To ensure diverse viewpoints and hypotheses, each S, X pair was assigned to three workers. Data collection statistics are shown in Table 2 and Figure 1.

As Figure 1 shows, the causal dimensions (1 and 6) have the most representation (18.1% and 16.4%, respectively). As our examples in Table 1 show, specific statements for these dimensions sometimes define a causal connection over paraphrases of story sentences[8], rather than introduce novel non-story content in either the antecedent or the consequent. To estimate how prevalent this phenomenon is, we manually evaluated 100 random samples of specific rules for each of dimensions 1 and 6. We found that for 66% and 63% of the samples, for dimensions 1 and 6 respectively, at least one of the annotators contributed statements that contained inferences with non-story content. The new content includes events that are likely to follow from the story as well as world knowledge about story entities.

# total annotations	~670K
# total pair of rules	~335K
# total unique stories S	4,881
# workers participated	371
Avg # of submissions by a worker	130.7
Max # of submissions by a worker	3,757
Avg minutes of work time / submission	8.78
Avg payment / submission	\$1.60
Avg # of dimensions filled in / submission	4.5

Table 2: Statistics about the GLUCOSE dataset.

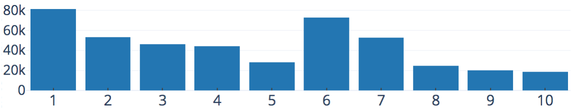


Figure 1: Number of rules collected for each dimension. Dimensions 1 and 6 have the most representation, while dimensions 9 and 10 are most often marked as not applicable.

4.3 Comparison to Other Resources

To assess the novelty of GLUCOSE knowledge, we compared its coverage against that of the two most relevant commonsense resources: ConceptNet and ATOMIC.[9] We performed a best-effort mapping from GLUCOSE dimensions to relations in ConceptNet and ATOMIC. For example, GLUCOSE dimensions 1 and 6 are mapped to ConceptNet's *Causes*, *HasSubevent*, *HasPrerequisite*, and to ATOMIC's *xEffect* and *oEffect*. For all mappings see Appendix A.

Since all three resources contain mostly natural-language entries, it is not possible to automatically quantify their precise overlap, so we adopted a

[3] GLUCOSE qualification UI: https://bit.ly/34Pej0N

[4] GLUCOSE main knowledge acquisition UI: https://bit.ly/2R8XcTt

[5] Our pilot studies helped narrow our dimensions from 18 down to 10 which workers could reliably distinguish. Notably, we collapsed Enable and Cause on which workers had significant disagreement.

[6] Additional information about the pipeline and data quality management can be found at https://tinyurl.com/y2pn5cgl

[7] Target age of individual stories was judged by age-of-acquisition and readability tests: Flesch-Kincaid Grade Level, the Coleman-Liau Index, and the Dale-Chall formula (Kuperman et al., 2012). It is important to note that this method depends on vocabulary and does not ensure that all content is appropriate for children in this age group.

[8] It is important to note that, even if the antecedent and consequent are both in the story, making the causal link between them explicit is considered to have fulfilled the purpose of providing common sense knowledge.

[9] Note that (Rashkin et al., 2018a) and (Rashkin et al., 2018b) are in essence a subset of ATOMIC, and hence, have even lower coverage compared with GLUCOSE.

Dimension	1	2	5	6	7	10
ConceptNet	1.2%	0.3%	0%	1.9%	0%	0%
ATOMIC	7.8%	1.2%	2.9%	5.3%	1.8%	4.9%

Table 3: Ceiling overlap between GLUCOSE and other resources. Omitted dimensions had no overlap.

lenient evaluation scheme. For each GLUCOSE general rule[10] A *relation* B, we queried each target resource for tuples $R'(A', B')$, where R' is the resource's mapped equivalent of *relation*, and A' and B' consist of just the main verbs in A and B. Using fuzzy matching on A' and B', we retrieved a large number of hits for the query, then filtered to those with $>$50% lexical overlap with the GLUCOSE rule.

The results, shown in Table 3, represent a ceiling in overlap with other resources. The results indicate that GLUCOSE captures extensive commonsense knowledge unavailable in existing resources. Note that GLUCOSE's knowledge model is a superset of ATOMIC's. GLUCOSE is designed to encompass all nine categories of inferential commonsense knowledge that ATOMIC covers, which are captured across different GLUCOSE dimensions. Note that there are definitely some individual pieces of knowledge that have been acquired in ATOMIC which do not exist in GLUCOSE, since some ATOMIC events may not have appeared in the GLUCOSE stories.

5 Empirical Evaluation Task

We set up a standalone evaluation task for evaluating models that predict GLUCOSE explanations: given a story S, a story sentence X, and a dimension d, provide an explanation in both specific and general forms.

Test Set Curation For a test set on commonsense reasoning to offer accurate and reliable evaluation, it should contain unambiguous examples with clear gold answers. This led to a curation process that identifies examples on which humans have high agreement, as follows: we sampled S, X pairs annotated by any three workers with the highest quality rating. A dimension d for S, X was allowed into the test set if 1) d was annotated by all three workers, and 2) the three specific statements had a round-robin average sentence-level BLEU (Lin and Och, 2004) score[11] above 0.75. Finally, two in-house annotators manually removed cases with typographical or core content errors, resulting in a test set of 500 story/sentence pairs, each with 1-5 dimensions answered.

Human and Automatic Evaluation Human evaluation is crucial for any language generation task. We crowdsourced our human evaluation on MTurk, using a dedicated UI,[12] asking three of our top-rated crowd workers from the main GLUCOSE crowdsourcing job to rate the predictions. We set up the following evaluation process to ensure calibrated judgments: the judge first reads a story with a highlighted sentence X, then reads a question about X corresponding to a GLUCOSE dimension. Next, they are shown a shuffled list of candidate answers, each produced by a different system. Finally, the judge rates each candidate answer on a four-point Likert scale: "completely incorrect," "almost incorrect," "almost correct," and "completely correct." To compare system performance, the ratings are mapped to numerical scores of 0–3, which are then averaged.

Automatic evaluation for tasks involving language generation has been a major bottleneck for research (Liu et al., 2016; Hashimoto et al., 2019). BLEU's ease of replicability has made it a popular automated metric, but its correlation with human judgement has proven weak on various tasks (Novikova et al., 2017; Gatt and Krahmer, 2018). For automatic evaluation, we use SacreBLEU (Post, 2018) with equal weights up to 4-grams at corpus-level on the three-reference test set. Using pairwise correlation analysis, we found strong correlation between human and BLEU scores on our test set, with correlation coefficients Spearman = 0.891, Pearson = 0.855, and Kendall's τ = 0.705, all with p-value < 0.001. The high correlation is due to various design choices, including 1) semi-structured inference rules in GLUCOSE are designed to be evaluable, where the structure constrains the variability of the rules, and 2) we minimized the noise in our human evaluation by designing a UI that could collect calibrated ratings from human judges educated about the task. The strong correlation suggests that BLEU is a viable metric for reporting future results on the GLUCOSE test set.

[10] We evaluated GLUCOSE's specific statements against ConceptNet, with nearly identical results to those in Table 3.

[11] We averaged the BLEU scores obtained, in round-robin fashion, by taking one rule as candidate and the other two as references. We used BLEU with equal weights up to 4-grams.

[12] GLUCOSE evaluation UI: https://bit.ly/2rJWFwy

6 Models

We developed several models for tackling the prediction task described in Section 5. The train and development sets for each model consisted of the initial 440K total annotations[13] (in the context of 3,360 stories) in the GLUCOSE dataset, minus the entries that share the context story with the test instances.

Due to their superior performance in sequence prediction, all our neural models use transformer blocks (Vaswani et al., 2017), which use multi-headed attention and fully connected layers to encode sequences. For decoding, all models use top-k random sampling (Fan et al., 2018). Details on all the models we experimented with can be found in Appendix C.

6.1 Pretrained Language Model (PT-LM)

PT-LM tests what GLUCOSE-like knowledge is captured by the pretrained 774M-parameter GPT-2 (Radford et al., 2019) language model. We elicit commonsense explanations from GPT-2 by prompting it with the story followed by sentence X and a dimension-specific trigger word like "because", and allowing the model to complete the sentence. For best results, we implemented "constrained decoding" by conditioning the GPT-2 model on the input S, X as context, then generating the next token for a dimension d as follows: if dimension d's template specifies a set of allowable words at the current position—e.g., locative prepositions for dimensions 3 and 8—sample from the options based on their likelihood as conditioned on the preceding tokens. Otherwise, allow sampling freely from the entire vocabulary. See Appendix C for a list of all templates used.

6.2 Models Trained on GLUCOSE

6.2.1 Language Models

We finetuned separate language models for specific and general rules. Each model monolithically covers all ten GLUCOSE dimensions: it generates rules given a dimension indicator as input.[14] Rules are sampled from the learned distribution $p(s) = \prod_{i=1}^{n} p(s_i \mid s_1, \ldots, s_{i-1})$, where s is the concatenation of input and output sequences. For all models in this section, we finetuned the PT-LM model described above.

One-sided Generation (1S-LM) One side of a GLUCOSE rule—the antecedent or the consequent, depending on the dimension—is always a paraphrase and/or a generalization of sentence X. In the one-sided model, we use X as is for this side of the specific statement; the model generates only the *target* side. Each training example is a text sequence *S#X#d#answer#EOS*, where d is the dimension number and *answer* is the target side. At test time, the model generates answer characters until it produces an EOS token.

Full Rule Generation (Full-LM) Full-LM learns to produce the complete rule, including the connective and the paraphrase of X. Instead of just the target side of the rule, the training examples have the full rule as the *answer* portion of the sequence. This allows the model to produce more human-like rules, including paraphrasing and/or generalizing X appropriately.

6.2.2 Encoder-Decoder Model (Enc-Dec)

Our most complex model is an encoder-decoder transformer model that jointly predicts the specific and general rules. It maximizes $p(y \mid x) = \prod_{i=1}^{n} p(y_i \mid x; y_1, \ldots, y_{i-1})$, where x is the input and y is the answer. We obtained the best results by formulating the input as *#d:* $S^*[X]$, where d is the dimension and $S^*[X]$ is the story S with sentence X surrounded by asterisks. We chose to finetune the state-of-the-art T5 model (with 770M-parameters, to be comparable to the size of the LM model), using the same hyperparameters as in (Raffel et al., 2020).

7 Results and Discussion

Table 4 shows the results from the models described in Section 6, evaluated as per Section 5. It shows that Enc-Dec uniformly outperforms all other models, confirming that full visibility into context[15] helps an architecture better learn the intricacies of GLUCOSE rules.

In fact, Enc-Dec performs competitively with humans in many dimensions. The strength of this model's performance in predicting both specific

[13] Table 2 shows the statistics of the final dataset, whereas all training for the models in the paper were conducted before the crowdsourcing of the dataset was finished.

[14] We experimented with training separate models for each dimension, which yielded much worse results.

[15] A clear drawback of language models is that the model's representation of the ith item depends only on items preceding i, and not the full input context. We show that better predictions can be made given full visibility into the entire input sequence.

	Human evaluation scores for dimension...										BLEU scores for dimension...									
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PT-LM	0.7	1.0	1.2	1.0	0.6	0.6	0.6	0.9	0.7	1.1	40.7	36.5	31.3	31.4	30.2	32.1	23.1	37.0	40.9	53.1
1S-LM	2.1	2.3	2.2	2.5	2.1	2.1	2.4	2.5	2.1	1.8	55.1	59.6	50.7	65.2	53.1	57.4	55.4	71.7	56.8	67.2
Full-LM	1.8	2.0	2.0	2.2	1.7	2.0	2.1	2.2	1.6	2.1	54.7	55.3	51.0	64.4	50.5	58.8	66.2	73.4	32.7	67.0
	1.6	1.6	1.8	2.1	1.8	1.9	1.9	2.1	1.1	1.5	56.4	55.8	57.5	62.7	59.6	59.0	65.8	67.7	53.7	56.2
Enc-Dec	**2.7**	**2.7**	**2.6**	**2.7**	**2.5***	**2.6**	**2.7**	**2.8**	**2.2**	**2.5***	**72.5**	**73.9**	**73.8**	**79.3**	**70.5**	**80.2**	**81.1**	**86.6**	**71.7**	**66.9**
	2.3	**2.3**	**2.4**	**2.5**	**2.3**	**2.4**	**2.5**	**2.7**	**1.9**	**1.7***	**66.4**	**67.6**	**68.5**	**73.0**	**69.8**	**77.6**	**76.8**	**86.8**	**68.6**	**57.5**
Human	2.8	2.7*	2.8	2.9	2.5*	2.8	2.8	2.8	2.9*	3.0	N/A									
	2.5	2.6	2.4	2.6	2.4	2.6	2.6	2.6	2.6*	2.7	N/A									

Table 4: Evaluation results for GLUCOSE models. Human evaluation scores are out of 3; BLEU scores are out of 100. Gray and regular rows show results on general and specific rules, respectively. Human model's performance was computed by showing judges a randomly selected answer from the three gold references. We performed paired sample t-tests on the human evaluation scores for each dimension for Full-LM against Enc-Dec, and then again for Enc-Dec against Human. The vast majority of differences are statistically significant at $p < 0.05$, with the exceptions noted in asterisk. Note that the dimensions where performance differences are not statistically significant strongly correlate with those with the least amount of data, as shown in Figure 1.

Model	**Dim 3: A location state that *Enables* X**	**Dim 6: An event that X *Causes/Enables***
Full-LM	Karen is at home *Enables* Karen made a pan of lasagna and brought it to the party	Karen made lasagna *Causes/Enables* Karen ate lasagna
	Someone$_A$ is in Somewhere$_A$ *Enables* Someone$_A$ makes Something$_A$ (that is edible)	Someone$_A$ cooks Something$_A$ (that is food) *Causes/Enables* Some People$_A$ to be turned away because of Something$_A$ (that is food)
Enc-Dec	Karen is in the kitchen *Enables* Karen makes a pan of lasagna	Karen makes a pan of lasagna *Causes/Enables* Karen eats it for a week
	Someone$_A$ is in a kitchen *Enables* Someone$_A$ cooks Something$_A$	Someone$_A$ makes Something$_A$ (that is food) *Causes/Enables* Someone$_A$ eats Something$_A$
Human	Karen is in the kitchen *Enables* Karen made a pan of lasagna	Karen made a pan of lasagna *Causes/Enables* She brought it to a party
	Someone$_A$ is in a kitchen *Enables* Someone$_A$ prepares Something$_A$ (that is a dish)	Someone$_A$ prepares Something$_A$ (that is a dish) *Causes/Enables* Someone$_A$ takes Something$_A$ to Something$_B$ (that is an event)

Table 5: Example model generations for the input story: *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.* (Sentence X is underlined.) Note that all test stories are unseen in the train or validation set.

and general rules is a testament to the high quality of the GLUCOSE training data. Its worst performance is on general rules for dimensions 5 and 10, which have the lowest number of training points and are the most diverse in content.

Other models perform as expected. PT-LM's poor performance shows that finetuning on our dataset significantly improves the commonsense inference capabilities of LMs. 1S-LM, which only predicts half of an inference rule, outperforms Full-LM in predicting specific statements, but lacks the ability to generalize them. We also tested various other baselines, including an ATOMIC-trained transformer model (Bosselut et al., 2019), retrieval of K-nearest-neighbors, and non-contextual variants of the presented models, all of which significantly underperformed the results in Table 4, and are presented in Appendix C.

Our results also show that our best models perform noticeably better on specific statements than on general rules. This is because generating a specific statement involves paraphrasing a story sentence and predicting an antecedent/consequent, while a general rule requires further generalizing the paraphrase and the antecedent/consequent appropriately such that the rule remains a generally valid statement about the world.

Although rule generalization can sometimes be as simple as replacing a named entity (e.g., *Gage*) with a typed variable (*Someone*$_A$), more often more complex transformations are needed, such as generalizing the action and producing type constraints on variables in the form of attribute phrases. For example, take into account the Enc-Dec results in Table

5. For dimension 3, the generalization of the story sentence, *Karen makes a pan of lasagna*, included generalizing *Karen* to *Someone$_A$* and *makes a pan of lasagna* to *cooks Something$_A$*. Note that sentence generalizations are dimension-specific: For dimension 6, the generalization of same sentence retains the verb *make* but adds a type constraint to the object, *Something$_A$ (that is a food)*, which is required for making the rule generally valid. Table 1 shows another complex transformation example where *turning his bike* is generalized into *moves away from Something (that is dangerous)*, that takes into account story context.

Overall, our evaluation results show that the state-of-the-art pre-trained models finetuned on the GLUCOSE dataset are well capable of dynamically producing GLUCOSE-like inference rules on the fly, which is the ultimate usecase of the GLUCOSE dataset. It is important to note that there is still a consistent performance gap between the best-performing model and human's on generating specific statements and general rules, which indicates that there is still a large headroom for improvement on designing better models for generalizable commonsense reasoning.

Note that in our current evaluation setup, we have made the simplifying assumption of evaluating each dimension for each sentence individually, without consideration for consistency across dimensions or across sentences. Joint prediction of all the dimensions and sentences across the story is a considerably more challenging task that can potentially yield more accurate predictions for a downstream task. We encourage the future work to focus on building models that perform joint predictions, which can be readily evaluated using our test-set. It is important to note that static test sets are inherently narrow and prone to hidden curation biases (Sharma et al., 2018; Belinkov et al., 2019). We believe that the ultimate evaluation for models that show GLUCOSE-like commonsense reasoning capabilities should be on naturally-occurring arbitrary stories and through our presented human evaluation process. As future work, we are planning to show the value of incorporating GLUCOSE-trained models in other downstream NLP tasks such as reading comprehension and dialog.

8 Conclusions

We introduced GLUCOSE, a large-scale dataset of implicit commonsense knowledge, encoded as explanatory mini-theories grounded in a narrative context. The theories are categorized into ten causal dimensions, inspired by cognitive psychology.

We presented our multi-stage pipeline for acquiring semi-structured causal explanations at scale from lay workers, resulting in ˜670K annotations in the context of everyday children's stories. We demonstrated the utility of GLUCOSE data in two ways. First, our analysis showed that GLUCOSE rules capture knowledge not available in existing resources or pre-trained models. Second, in order to evaluate how well AI models can predict GLUCOSE knowledge on novel inputs, the ultimate value of such a dataset, we defined a standalone evaluation task for predicting specific and general inference rules given a story/sentence pair and a dimension. We curated a doubly-vetted test set, developed a platform to facilitate human judgment of system outputs, and validated BLEU as a strong automated evaluation metric. We show that training on GLUCOSE data improves model performances significantly on unseen stories.

Our results validate our hypothesis that a promising approach for imbuing machines with commonsense is to use carefully-crafted data, as in GLUCOSE, to train neural architectures that have a wide range of lexical and conceptual knowledge encoded, as in models pretrained on large corpora. Together with this paper, we release our dataset[16] and models[17], which we hope will enable the AI research community to explore effective approaches to incorporate commonsense reasoning capabilities into various downstream tasks.

Acknowledgments

We would like to thank the hundreds of amazing crowdworkers whose dedication made this work possible. We thank David Ferrucci for his valuable insights and support throughout the GLUCOSE project. We thank Andy Beck for the discussions around the GLUCOSE knowledge model and Jesse Dunietz for his discussions on the paper. We are grateful for the invaluable comments of the anonymous reviewers, Niranjan Balasubramanian, and Owen Rambow on this paper.

[16] The GLUCOSE dataset is available for download at https://tinyurl.com/yyeo92pt.

[17] The trained models and the details on the GLUCOSE data files can be found through https://github.com/ElementalCognition/glucose/.

References

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, Seattle, Washington, USA. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Justin TA Busch, Aiyana K Willard, and Cristine H Legare. 2018. Explanation scaffolds causal learning and problem solving in childhood. In *Active Learning from Infancy to Childhood*, pages 113–127. Springer.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Ilaria Grazzani, Veronica Ornaghi, Elisabetta Conte, Alessandro Pepe, and Claudia Caprin. 2018. The relation between emotion understanding and theory of mind in children aged 3 to 8: The key role of language. *Frontiers in Psychology*, 9:724.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

P. Mergenthaler, U. Lindauer, G. A. Dienel, and A. Meisel. 2013. Sugar for the brain: the role of glucose in physiological and pathological brain function. *Trends in neurosciences*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

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 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

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.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Lei Sha, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. Joint learning templates and slots for event schema induction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 428–434, San Diego, California. Association for Computational Linguistics.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Shaohua Yang, Qiaozi Gao, Sari Sadiya, and Joyce Chai. 2018. Commonsense justification for action explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2627–2637, Brussels, Belgium. 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*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Yilun Zhou, Steven Schockaert, and Julie Shah. 2019. Predicting conceptnet path quality using crowdsourced assessments of naturalness. In *The World Wide Web Conference*, WWW '19, pages 2460–2471, New York, NY, USA. ACM.

Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5):292–297.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.

Appendix A: The Knowledge Model for Collecting GLUCOSE data

Semi-structured Inference Rules

The knowledge represented in GLUCOSE is captured in the form of semi-structured inference rules that are accompanied by a specific statement that grounds the rule in the context of a specific story. Each specific statement and its corresponding general rule use the common template of *antecedent connective consequent*. The antecedent and consequent are each composed by filling in a few syntactic slots, namely, subject, verb, object(s), and preposition(s). In order to further shape the semantics of the acquired knowledge, some of these slots have a pre-defined list of options to choose from.

Table 6 lists the pre-defined options for filling in the syntactic slots per GLUCOSE dimension[18]. Some of the slots allow adding a custom entry to the list of options, hence soft constraints, and some do not, hence hard constraints. Note that beyond the options listed in this table, the general rule slots across all the dimensions have pre-defined options for subject and object slots such as $Someone_A$ or $Some\ People_C$.

Comparison to Other Resources

To assess the value of the GLUCOSE dataset, we compared its coverage against the two most relevant commonsense knowledge resources: ConceptNet and ATOMIC. Table 7 shows our best-effort mapping among knowledge dimensions of GLUCOSE and relations in ConceptNet and ATOMIC.

Appendix B: Data Collection Pipeline

To ensure obtaining our desired quality, we designed a three-stage knowledge acquisition pipeline for crowdsourcing the GLUCOSE dataset on the Amazon Mechanical Turk (Mturk): The qualification test, the main task, and the expert review. In this Section we provide more detail about each stage and its designated UI design.

Qualification Test The qualification test contained questions testing workers' understanding in three areas: Identifying correct use of the UI slots for composing their answers (Figure 2), recognizing the right level of generalization (Figure 3), and identifying causes and effects with proper temporal understanding of the stories (Figure 4). Understanding generalization is the most difficult, and the most important, aspect of our task. Assessing the prospective workers' understanding of generalization was done through curating questions demonstrating under-generalization or over-generalization. The full Qualification UI, along with all the detailed instructions that were visible to the workers, is accessible here https://bit.ly/34Pej0N.

Main Task The qualified workers were able to access large batches of data with no limit. The main task starts with a page like the one shown in the Figure 5. The user loops through each of the 10 dimensions of GLUCOSE data collection, in order, presented as questions. Note that the user could answer the question by simply marking the dimension as not applicable and skipping it. If they choose to answer, as shown in Figure 6, they will be presented with the structured rule slots to input their answers. The full Main GLUCOSE UI, along with all the detailed instructions that were visible to the workers, is accessible here https://bit.ly/2R8XcTt.

Expert Review For work contributed through the main UI, data quality was controlled through daily monitoring of a percentage of incoming submissions and statistics on average dimensions filled out. For managing this process, we built a specialized UI for reviewing the incoming structured data. The percentage of answers reviewed by an in-house expert were used to update worker ratings. Workers enter the task with a score of "1", then advance to "2" as they become more proficient, getting a bonus increase. The top rating is "3". Select workers with a "3" rating were also moved into "top rated" batches that paid more per HIT and included higher bonuses and incentives. If work quality dropped, workers' ratings were adjusted accordingly. If their work was at a risk of degrading the quality of the dataset, they were disqualified from the task.[19]

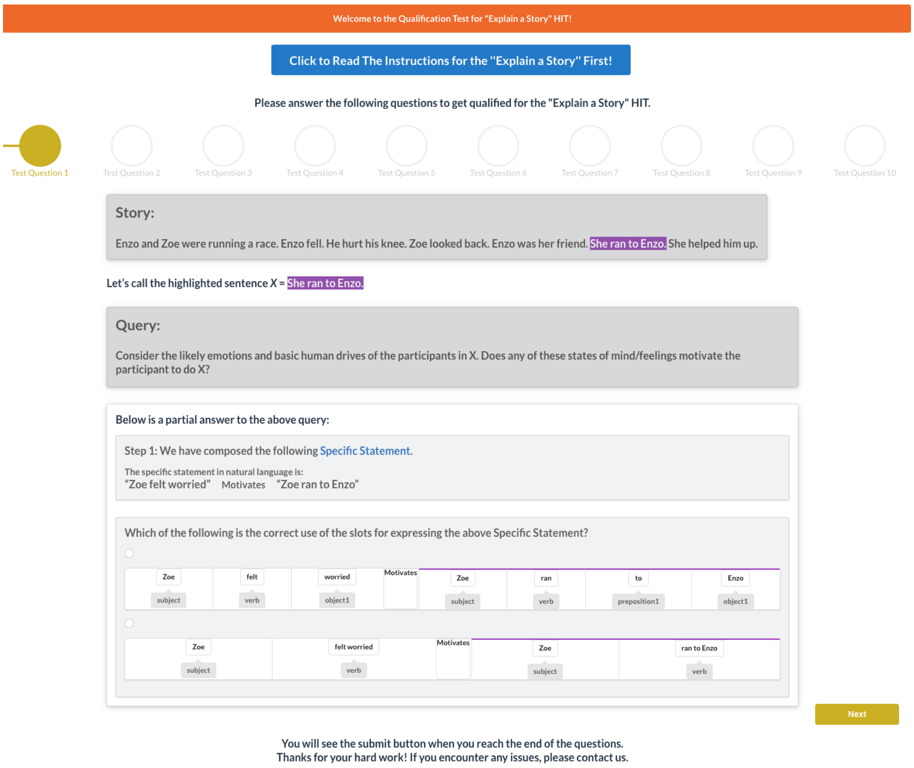
Appendix C: Details on the Models

ATOMIC-trained Model

This model is a transformer language model, specifically GPT-1 architecture, fine-tuned on ATOMIC resource. The language model is fine-tuned to generate triplet sequences such as 'PersonX goes to

[18] A sample of the semi-structured rules in GLUCOSE can be found through https://bit.ly/2LFuwOt.

[19] Additional information on the data and data quality management can be found at https://tinyurl.com/y2pn5cgl.



Welcome to the Qualification Test for "Explain a Story" HIT!

Click to Read The Instructions for the "Explain a Story" First!

Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Enzo and Zoe were running a race. Enzo fell. He hurt his knee. Zoe looked back. Enzo was her friend. She ran to Enzo. She helped him up.

Let's call the highlighted sentence *X* = She ran to Enzo.

Query:

Consider the likely emotions and basic human drives of the participants in X. Does any of these states of mind/feelings motivate the participant to do X?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

The specific statement in natural language is:
"Zoe felt worried" Motivates "Zoe ran to Enzo"

Which of the following is the correct use of the slots for expressing the above Specific Statement?

○

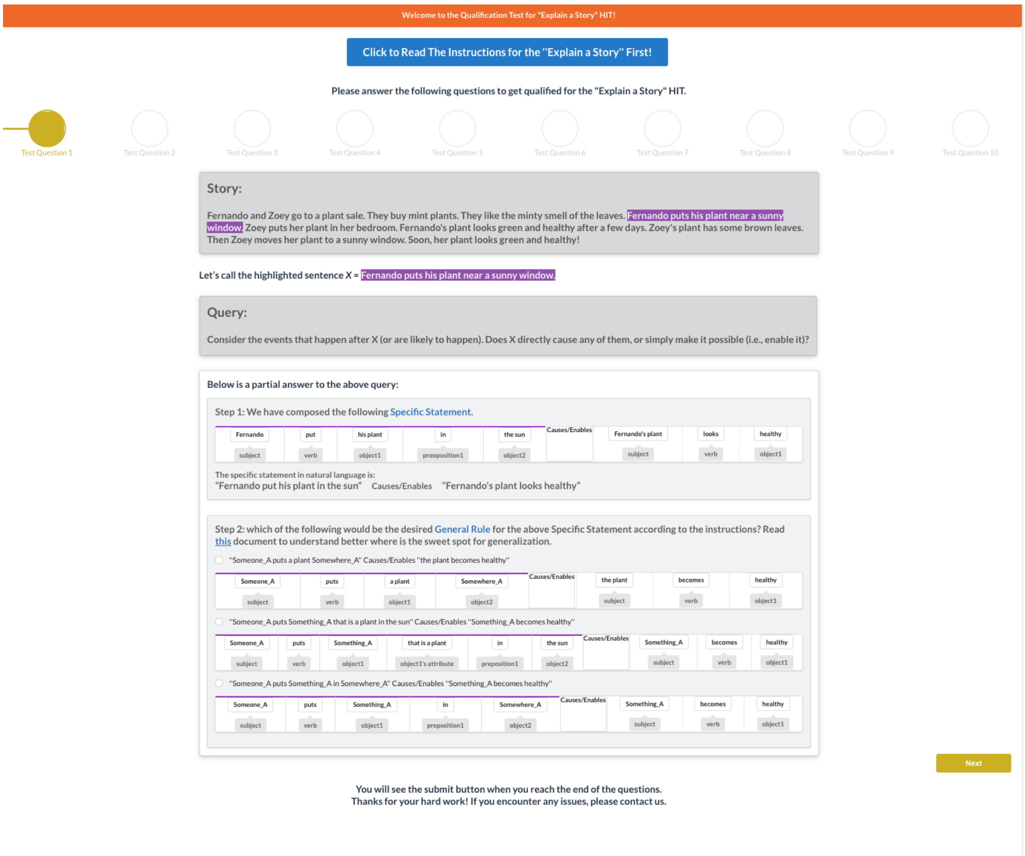
Zoe	felt	worried	Motivates	Zoe	ran	to	Enzo
subject	verb	object1		subject	verb	preposition1	object1

○

Zoe	felt worried	Motivates	Zoe	ran to Enzo
subject	verb		subject	verb

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 2: Example qualification question about the correct use of the slots.



Welcome to the Qualification Test for "Explain a Story" HIT!

Click to Read The Instructions for the "Explain a Story" First!

Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Fernando and Zoey go to a plant sale. They buy mint plants. They like the minty smell of the leaves. Fernando puts his plant near a sunny window. Zoey puts her plant in her bedroom. Fernando's plant looks green and healthy after a few days. Zoey's plant has some brown leaves. Then Zoey moves her plant to a sunny window. Soon, her plant looks green and healthy!

Let's call the highlighted sentence *X* = Fernando puts his plant near a sunny window.

Query:

Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

Fernando	put	his plant	in	the sun	Causes/Enables	Fernando's plant	looks	healthy
subject	verb	object1	preoposition1	object2		subject	verb	object1

The specific statement in natural language is:
"Fernando put his plant in the sun" Causes/Enables "Fernando's plant looks healthy"

Step 2: which of the following would be the desired General Rule for the above Specific Statement according to the instructions? Read this document to understand better where is the sweet spot for generalization.

○ "Someone_A puts a plant Somewhere_A" Causes/Enables "the plant becomes healthy"

Someone_A	puts	a plant	Somewhere_A	Causes/Enables	the plant	becomes	healthy
subject	verb	object1	object2		subject	verb	object1

○ "Someone_A puts Something_A that is a plant in the sun" Causes/Enables "Something_A becomes healthy"

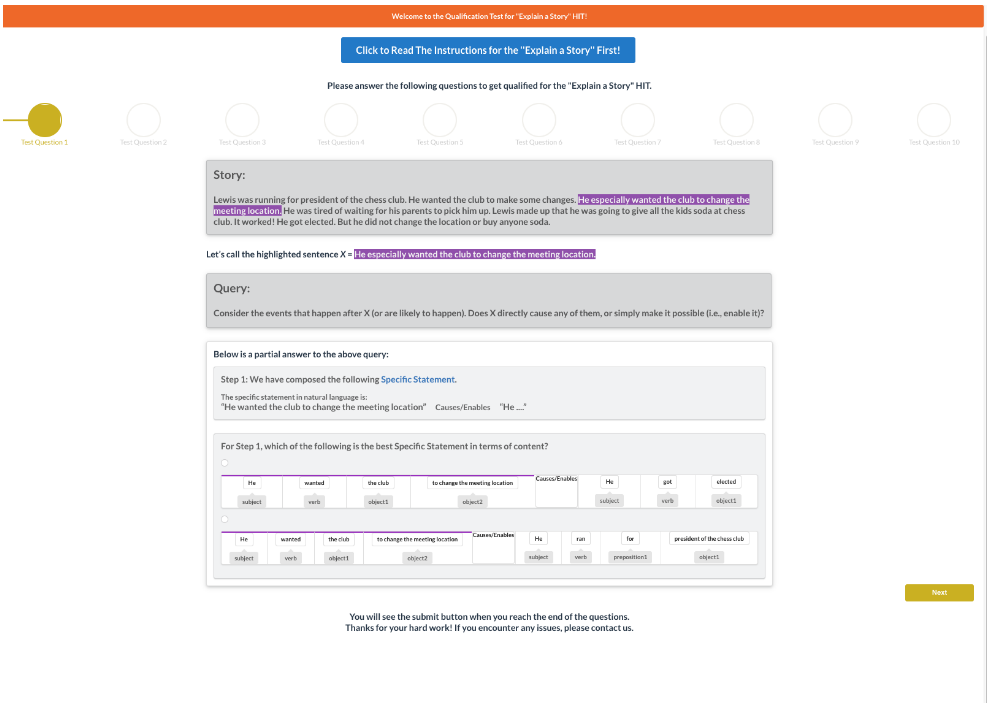
Someone_A	puts	Something_A	that is a plant	in	the sun	Causes/Enables	Something_A	becomes	healthy
subject	verb	object1	object1's attribute	preposition1	object2		subject	verb	object1

○ "Someone_A puts Something_A in Somewhere_A" Causes/Enables "Something_A becomes healthy"

Someone_A	puts	Something_A	in	Somewhere_A	Causes/Enables	Something_A	becomes	healthy
subject	verb	object1	preposition1	object2		subject	verb	object1

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 3: Example qualification question about the correct level of generalization.



Welcome to the Qualification Test for "Explain a Story" HIT!

Click to Read The Instructions for the "Explain a Story" First!

Please answer the following questions to get qualified for the "Explain a Story" HIT.

Story:

Lewis was running for president of the chess club. He wanted the club to make some changes. He especially wanted the club to change the meeting location. He was tired of waiting for his parents to pick him up. Lewis made up that he was going to give all the kids soda at chess club. It worked! He got elected. But he did not change the location or buy anyone soda.

Let's call the highlighted sentence *X* = He especially wanted the club to change the meeting location.

Query:

Consider the events that happen after X (or are likely to happen). Does X directly cause any of them, or simply make it possible (i.e., enable it)?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

The specific statement in natural language is:
"He wanted the club to change the meeting location"   Causes/Enables   "He ...."

For Step 1, which of the following is the best Specific Statement in terms of content?

○

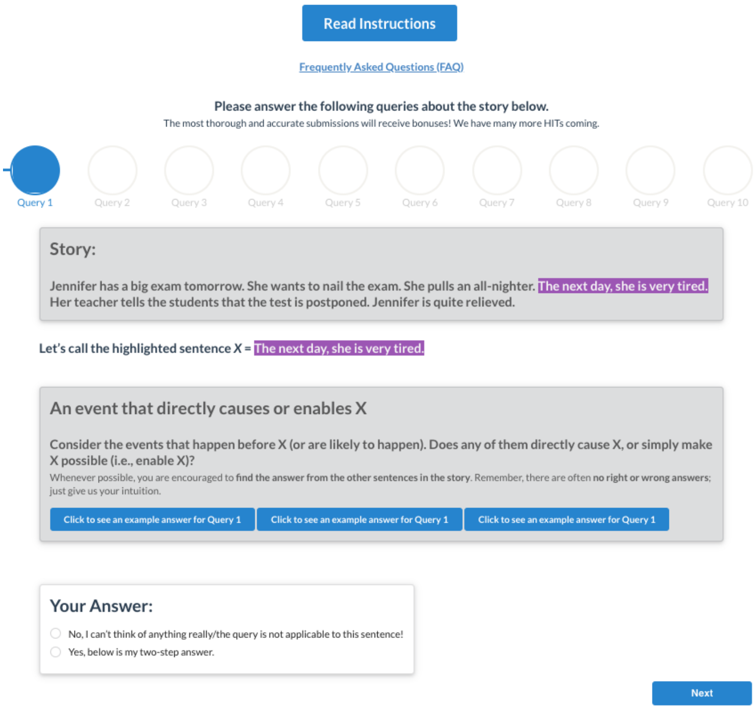
He	wanted	the club	to change the meeting location	Causes/Enables	He	got	elected
subject	verb	object1	object2		subject	verb	object1

○

He	wanted	the club	to change the meeting location	Causes/Enables	He	ran	for	president of the chess club
subject	verb	object1	object2		subject	verb	preposition1	object1

You will see the submit button when you reach the end of the questions.
Thanks for your hard work! If you encounter any issues, please contact us.

Figure 4: Example qualification question about understanding causal relations between events.



Please answer the following queries about the story below.
The most thorough and accurate submissions will receive bonuses! We have many more HITs coming.

Story:

Jennifer has a big exam tomorrow. She wants to nail the exam. She pulls an all-nighter. The next day, she is very tired. Her teacher tells the students that the test is postponed. Jennifer is quite relieved.

Let's call the highlighted sentence X = The next day, she is very tired.

An event that directly causes or enables X

Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

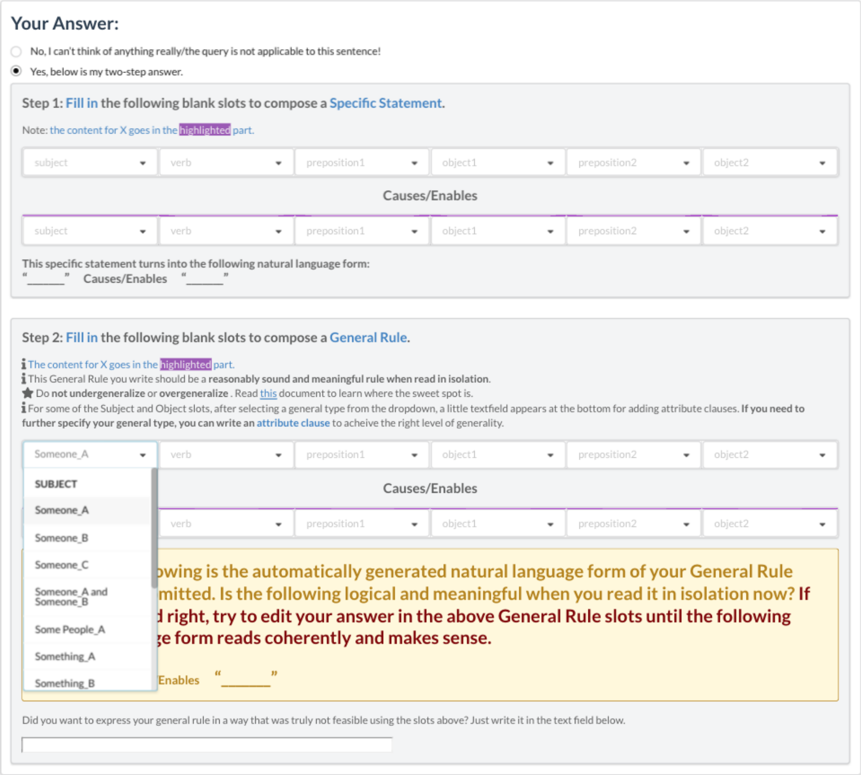
Whenever possible, you are encouraged to **find the answer from the other sentences in the story**. Remember, there are often **no right or wrong answers**; just give us your intuition.

Click to see an example answer for Query 1
Click to see an example answer for Query 1
Click to see an example answer for Query 1

Your Answer:

○ No, I can't think of anything really/the query is not applicable to this sentence!
○ Yes, below is my two-step answer.

Figure 5: The preview page of the Main UI for GLUCOSE data collection, which can be accessed via https://bit.ly/2R8XcTt.



Your Answer:

○ No, I can't think of anything really/the query is not applicable to this sentence!

● Yes, below is my two-step answer.

Step 1: Fill in the following blank slots to compose a Specific Statement.

Note: the content for X goes in the highlighted part.

Causes/Enables

This specific statement turns into the following natural language form:
"______" Causes/Enables "______"

Step 2: Fill in the following blank slots to compose a General Rule.

ℹ The content for X goes in the highlighted part.
ℹ This General Rule you write should be a **reasonably sound and meaningful rule when read in isolation**.
★ Do **not undergeneralize** or **overgeneralize** . Read this document to learn where the sweet spot is.
ℹ For some of the Subject and Object slots, after selecting a general type from the dropdown, a little textfield appears at the bottom for adding attribute clauses. **If you need to further specify your general type, you can write an** attribute clause to acheive the right level of generality.

Someone_A

Causes/Enables

owing is the automatically generated natural language form of your General Rule mitted. Is the following logical and meaningful when you read it in isolation now? If d right, try to edit your answer in the above General Rule slots until the following ge form reads coherently and makes sense.

Enables "______"

Did you want to express your general rule in a way that was truly not feasible using the slots above? Just write it in the text field below.

Figure 6: The answer-entry part of the main UI. When "Yes" is selected for "Your Answer" on the main UI for GLUCOSE data collection, the workers can input answers to the dimension in question.

the mall <xIntent >to buy clothes'. We use the same exact model trained for (Bosselut et al., 2019). This model is only applicable to General Rule prediction. The results from this model were significantly worse than the PT-LM model, which is the worst-performing model presented in the main paper. This was expected, given the little overlap that exists between the ATOMIC dataset and the GLUCOSE knowledge, as presented in the main paper under "Comparison to Other Resources" Section.

K-Nearest Neighbor (KNN)

For a given test pair S, X, the KNN baseline retrieves the K most similar training instances and returns one as the prediction. It uses BERT (Devlin et al., 2019) sentence embeddings to compute cosine similarity between a candidate and each retrieved training instance. We tuned three parameters on the development set: K, min_sim, and max_sim. If a candidate has a similarity score above max_sim, it is emitted as the prediction. Otherwise, candidates scoring below min_sim are dropped, and the centroid among the remaining pool is emitted. We evaluate KNN only for general rules, since it is not meaningful to retrieve specific statements from the training set. The results from this model were significantly worse than the PT-LM model, which is the worst-performing model presented in the main paper. The performance of the KNN model highlights the importance of generalizing beyond the training data.

Pretrained Language Model (PT-LM)

We experimented with prompting the pretrained language models, specifically GPT-2, as is, for predicting GLUCOSE dimensions. Table 8 shows the list of particular templates used for decoding. We used 774M-parameter GPT-2 model, with top-K random sampling for decoding, with K = 15. The decoding for this model was done on CPU.

1S-LM and Full-LM

This model uses the exact model as with PT-LM. These models were finetuned on 8 NVIDIA Tesla V100 GPUs for 10K steps.

Enc-Dec Model

We finetuned the 770M-parameter pre-trained T5 model using the exact same hyperparameters as in (Raffel et al., 2020). We have used top-K random

Dimension	Connective	Slot Constraints
Dim 1: An event that directly causes or enables X	Causes/Enables	None
Dim 2: An emotion or basic human drive that motivates X	Motivates	*Verb slot hard constraints*: feels, wants, likes; *Object slot soft constraints*: curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic.
Dim 3: A location state that enables X	Enables	*Verb slot hard constraints*: am, is, are; *Preposition slot hard constraints*: above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of.
Dim 4: A possession state that enables X	Enables	*Verb slot hard constraints*: possess(es).
Dim 5: Other attribute that enables X	Enables	*Verb slot hard constraints*: am, is, are, has, have, want, wants, need, needs.
Dim 6: An event that is directly caused or enabled by X	Causes/Enables	None
Dim 7: An emotion that is caused by X	Causes	*Verb slot hard constraints*: feels, wants, likes; *Object slot soft constraints*: curiosity, independence, competition, honor, approval, power, status, romance, success, friendship, belonging, health, safety, livelihood, happy, stressed, angered, disgusted, sad, surprised, fearful, trusting, love, obedient, amazed, disappointment, regret, worthless, aggression, optimistic.
Dim 8: A change of location that X results in	Results in	*Verb slot hard constraints*: am, is, are; *Preposition slot hard constraints*: above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of.
Dim 9: A change of possession that X results in	Results in	*Verb slot hard constraints*: possess(es)
Dim 10: Other change in attribute that X results in	Results in	*Verb slot hard constraints*: am, is, are, has, have, want, wants, need, needs.

Table 6: The list of pre-defined options for filling in the syntactic slots per GLUCOSE dimension.

Glucose	ConceptNet Rel	ATOMIC Rel
Dims 1 & 6	HasSubevent HasFirstSubevent HasLastSubevent HasPrerequisite	xEffect/oEffect
Dim 2	Desires CausesDesire MotivatedByGoal	xAttr ("feels") xIntent (otherwise)
Dim 7	Same as dim2	xReact/oReact ("feels")
Dims 5 & 10	Desires CausesDesire	xAttr/xWant oWant

Table 7: Mappings between GLUCOSE dimensions and ConceptNet/ATOMIC relations. ConceptNet "Causes" applies to all GLUCOSE dimensions. Omitted GLUCOSE dimensions have no mapping in ATOMIC.

sampling for decoding, with K = 15. We did the training and decoding for this model on Google TPU v3-8. We trained this model for 500k steps after pre-training, which took about 72 hours.

We also experimented with non-contextual version of all the models presented in the main paper. For non-contextual models, the story S is simply removed from the input. The non-contextual models all underperformed their contextual counterparts. This further validates the importance of using context in making commonsense inferences.

Dimension	**Connective**	**Natural Language Template**
Dim 1 An event that directly causes or enables X	Causes/Enables	[because, since]
Dim 2 An emotion or basic human drive that motivates X	Motivates	[because, since]+ [he, she, they, I, you, we]+ [feels, wants, likes]
Dim 3 A location state that enables X	Enables	[because, since]+ [he, she, they, I, you, we]+ [is, was, were]+ [above, across from, between, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
Dim 4 A possession state that enables X	Enables	[because, since]+[he, she, they, it, I, you, we]+ [has, have]
Dim 5 Other attribute that enables X	Enables	[because, since]+[he, she, they, it, I, you, we]+ [am, is, are, has, have, want, wants, need, needs]
Dim 6 An event that is directly caused or enabled by X	Causes/Enables	[causes, caused, results in , . This causes, . As a result]
Dim 7 An emotion that is caused by X	Causes	[. As a result]+ [he, she, they,I, you, we]+[feels]
Dim 8 A change of location that X results in	Results in	[. As a result]+ [he, she, they, it, I, you, we]+ between, [is, was, were]+ [above, across from, at, below, far from, in, in front of, inside of,near, next to, on top of, outside of]
Dim 9 A change of possession that X results in	Results in	[. As a result] + [he, she, they, it, I, you, we]+ [has, have]
Dim 10 Other change in attribute that X results in	Results in	[. As a result] + [he, she, they, it, I, you, we]+ [am, is, are, has, have, want, wants, need, needs]

Table 8: Templates used for turning the ten dimensions for GLUCOSE data into natural language statements for decoding proper sequences from the pre-trained language models.