

# CURIE: An Iterative Querying Approach for Reasoning About Situations

Dheeraj Rajagopal<sup>\*</sup>, Aman Madaan<sup>\*</sup>, Niket Tandon<sup>†</sup>, Yiming Yang,  
Shrimai Prabhume, Abhilasha Ravichander, Peter Clark<sup>†</sup>, Eduard Hovy

Language Technologies Institute, Carnegie Mellon University

<sup>†</sup> Allen Institute for Artificial Intelligence

{dheeraj, amadaan, yiming, sprabhun, aravicha, hovy}@cs.cmu.edu

{nikett, peterc}@allenai.org

## Abstract

Predicting the effects of unexpected situations is an important reasoning task, e.g., would cloudy skies help or hinder plant growth? Given a context, the goal of such situational reasoning is to elicit the consequences of a new situation (*st*) that arises in that context. We propose CURIE, a method to iteratively build a graph of relevant consequences explicitly in a structured situational graph (*st* graph) using natural language queries over a fine-tuned language model. Across multiple domains, CURIE generates *st* graphs that humans find relevant and meaningful in eliciting the consequences of a new situation (75% of the graphs were judged correct by humans). We present a case study of a situation reasoning end task (WIQA-QA), where simply augmenting their input with *st* graphs improves accuracy by 3 points. We show that these improvements mainly come from a hard subset of the data, that requires background knowledge and multi-hop reasoning.

## 1 Introduction

A long-standing challenge in reasoning is to model the consequences of an unseen situation in a context. In the real world unexpected situations are common. Machines capable of situational reasoning are crucial because they are expected to gracefully handle such unexpected situations. For example, when eating leftover food, would it be more safer from virus if we microwave the food? - answering this requires understanding the complex events *virus contamination* and *effect of heat on virus*. Much of this information remains implicit (by Grice’s maxim of quantity (Grice, 1975)), thus requiring inference.

Recently, NLP literature has shown renewed interest in situational reasoning with applications in qualitative reasoning (Tandon et al., 2019;

<sup>\*</sup> authors contributed equally to this work. Ordering determined by dice rolling.

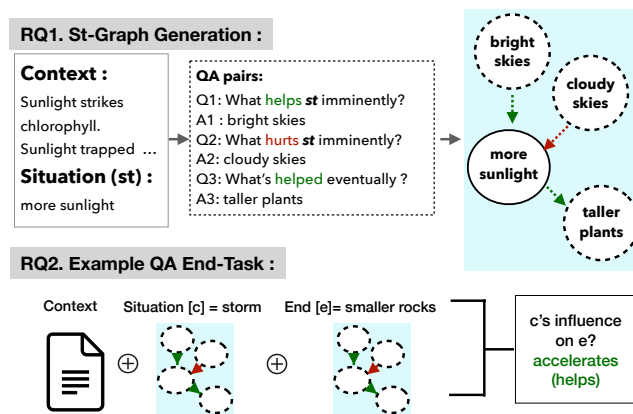


Figure 1: **RQ1:** CURIE generates situational graphs by iteratively querying a model, making explicit the model’s knowledge of effects of influences (+ve / -ve). **RQ2:** Situational graphs improve situational reasoning QA when appended to the question context.

Tafjord et al., 2019), physical commonsense reasoning (Sap et al., 2019; Bisk et al., 2020), and defeasible inference (Rudinger et al., 2020). These tasks take as input a context providing background information, a situation (*st*), and an ending, and predict the reachability from *st* to that ending. However, these systems have three limitations: (i) systems trained on these tasks are often domain specific, (ii) these tasks do not require a supporting structure that elicits the dynamics of the reasoning process, and (iii) these tasks are addressed as a classification problem restricting to a closed vocabulary setting.

To address these limitations, we propose CURIE—a system to iteratively query pretrained language models to *generate* an explicit structured graph of consequences, that we call a *situational reasoning graph* (*st*-graph). The task is illustrated in Figure 1: given some context and situation *st* (short phrase), our system generates a *st*-graph based on the contextual knowledge. CURIE supports the following kinds of reasoning:

- If a situation *st* occurs, which event is

more/less likely to happen imminently/ eventually?

- Which event will support/ prevent situation  $st$  from happening imminently/ eventually?

As shown in Figure 1, our approach to this task is to iteratively compile the answers to questions 1 and 2 to construct the  $st$ -graph where imminent/eventual capture multihop reasoning questions. Compared to a free-form text output obtained from an out-of-the-box sequence-to-sequence model, our approach gives more control and flexibility over the graph generation process, including arbitrarily reasoning for any particular node in the graph. The generated  $st$ -graphs are of high quality as judged by humans for correctness. In addition to human evaluation, we also show that a downstream task that requires reasoning about situations can compose natural language queries to construct a  $st$ -reasoning graph via CURIE. The resulting  $st$ -graph can be simply augmented to their input to achieve performance gains, specifically on the subset of hard questions that require background knowledge and multihop reasoning. In summary, this paper addresses the following research questions:

- RQ1:** Given a context and a situation, how can we generate a situational reasoning ( $st$ ) graph? To answer RQ1, we present CURIE, the first domain-agnostic situational reasoning system that takes as input a context and a situation  $st$  and iteratively generates a situational reasoning graph (§2). Our system is effective at situational reasoning across three datasets as validated by human evaluation and automated metrics.
- RQ2:** Can the  $st$ -graphs generated by CURIE improve performance of a downstream task? To answer RQ2, we show that  $st$  graphs generated by CURIE improve a  $st$ -reasoning task (WIQA-QA) by 3 points on accuracy by simply augmenting their input with our generated situational graphs, especially for a hard subset that requires background knowledge and multi-hop reasoning (§4).

## 2 CURIE for Situational Reasoning

CURIE provides both a general framework for situational reasoning and a method for constructing  $st$ -reasoning graphs from pretrained language models.

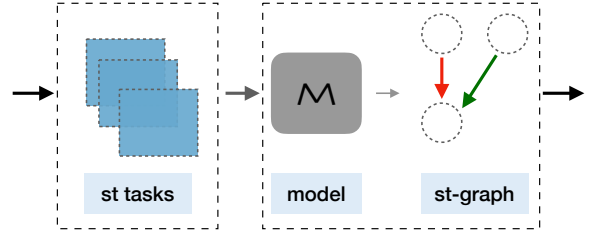


Figure 2: CURIE framework consists of two components: (i) a formulation that adapts datasets that allow  $st$ -reasoning for pretraining (ii) a method to iteratively build structured  $st$ -graphs using natural language queries over a fine-tuned language model ( $\mathcal{M}$ ).

Figure 2 shows the overall architecture of CURIE. CURIE framework consists of two components: (i) *st-reasoning task formulation* : a formulation that adapts datasets that allow situational reasoning (ii) *st-graph construction* : a method to fine-tune language model  $\mathcal{M}$  to generate the consequences of a situation and iteratively construct structured situational graphs (shown in Figure 1). In this section, we present (i) our task formulation (§2.1), (ii) adapting existing datasets for CURIE task formulation (§2.2), (iii) the learning procedure (§2.3), and (iv) the  $st$ -graph generation process (§2.4).

### 2.1 Task Formulation

We describe the general task formulation for adapting pretraining language models to the  $st$ -reasoning task. Given a context  $T = \{s_1, s_2, \dots, s_N\}$  comprising of  $N$  sentences, and a situation  $st$ , our goal is to generate an  $st$ -graph  $G$  that captures the effects of situation  $st$ .

An  $st$ -graph  $G(V, E)$  is an unweighted directed acyclic graph. A vertex  $v \in V$  is an event or a state that describes a change to the original conditions in  $T$ . Each edge  $e_{ij} \in E$  is labeled with a relationship  $r_{ij}$ , that indicates whether  $v_i$  *positively* or *negatively* influences  $v_j$ . Positive influences are represented via **green** edges comprising one of  $\{entails, strengthens, helps\}$  and negative influences represented via **red** edges that depict one of  $\{contradicts, weakens, hurts\}$ . Our relation set is general and can accommodate various  $st$ -reasoning tasks. Given two nodes  $v_i, v_k \in V$ , if a path from  $v_i$  to  $v_k$  has more than one edge, we describe the effect  $c$  as *eventual* and a direct effect as *imminent*.

We derive the training data by transforming a repository of (context  $T$ ,  $st$ -graph  $G$ ) tuples into a set of question-answer pairs. Each pair of vertices  $v_s, v_t \in G$  that are connected by a path contribute

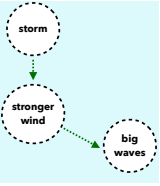
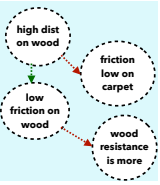
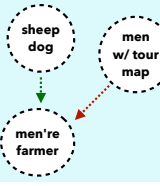
Dataset	Original formulation	Original <i>st</i> graph	Iterative formulation ( <i>st</i> )
WIQA	<p><i>context</i>: Wind creates waves.. Waves wash on beaches...</p> <p><i>ques</i>: If there is storm, how will it affect bigger waves?</p> <p><i>explanation</i>: storm → stronger wind → bigger waves</p> <p><i>answer</i>: helps bigger waves</p>		<p>Given <i>context</i> and <i>st</i>: there is a storm</p> <p>Q1: What does <i>st</i> <b>help</b> <i>imminently</i> ?</p> <p>A1: stronger wind</p> <p>Q2: What does <i>st</i> <b>help</b> <i>eventually</i> ?</p> <p>A2: bigger waves</p>
QUAREL	<p><i>context</i>: Car rolls further on wood than on thick carpet</p> <p><i>ques</i>: what has more resistance?</p> <p>(a) wood (b) the carpet</p> <p><i>simplified logical form of context, ques</i>: distance is higher on wood → (a) friction is higher in carpet (or) (b) friction is higher in wood</p> <p><i>answer</i>: (b) the carpet</p>		<p>Given <i>context</i> and <i>st</i>: distance is higher on wood</p> <p>Q1: What does <i>st</i> <b>entail</b> <i>imminently</i> ?</p> <p>A1: friction is lower in wood</p> <p>Q2: What does <i>st</i> <b>contradict</b> <i>imminently</i> ?</p> <p>A2: friction is lower in carpet</p> <p>Q3: What does <i>st</i> <b>entail</b> <i>eventually</i> ?</p> <p>A3: wood has more resistance</p>
DEFEAS	<p><i>context</i>: Two men and a dog are standing among the green hills.</p> <p><i>hypothesis</i>: The men are farmers.</p> <p><i>update1</i>: The dog is a sheep dog strengthens hypothesis</p> <p><i>update2</i>: Men with tour map weakens hypothesis</p>		<p>Given <i>context</i> and <i>st</i>: dog is a sheep dog</p> <p>Q1: What does <i>st</i> <b>strengthen</b> <i>imminently</i> ?</p> <p>A1: The men are farmers</p> <p><i>st</i>: men are studying tour maps</p> <p>Q2: What does <i>st</i> <b>weaken</b> <i>imminently</i>?</p> <p>A2: The men are farmers</p>

Table 1: The datasets used by CURIE and how we re-purpose them for *st* reasoning graph generation task. As explained in §2.1, the **green** edges set depicts relation (*r*) (entail, strengthen, helps) and **red** edges depict one of (contradict, weaken, hurts). The { *imminent*, *eventual* } effects (*c*) are used to support multihop reasoning. DEFEAS = DEFEASIBLE, *chain* refers to reasoning chain. Some examples are cut to fit. The key insight is that an *st*-graph can be decomposed into a series of QA pairs, enabling us to leverage seq-to-seq approaches for *st*-reasoning.

one question-answer pair to the training data for CURIE, such that every question comprises of: i) context  $T$ , ii) a *st*-vertex  $v_s$ , iii) a relation  $r$ , and iv) the nature of the effect  $c$  and the answer is the target node  $v_t$ . An example is shown in Figure 1. Compared to an end-to-end approach to graph generation, our approach gives more flexibility over the generation process, enabling reasoning for any chosen node in the graph. Thus the training data consists of tuples  $(\mathbf{x}_i, \mathbf{y}_i)$ , with  $\mathbf{x}_i = (T, v_s, r, c)_i$  and  $\mathbf{y}_i$  is the target situation  $v_t$ .

## 2.2 Generalizing Existing Datasets

Despite theoretical advances, lack of a large-scale general situational reasoning dataset presents a challenge to train seq-to-seq language models. We describe how we generalize existing diverse datasets towards *st*-reasoning towards finetuning a language model  $\mathcal{M}$ . If a reasoning dataset contains

a context, a *st*-situation and can describe the influence of *st* in terms of **green** and/or **red** edges, it can be seamlessly adapted to CURIE framework. Due to the lack of existing datasets that directly support our task formulation, we adapt the following three diverse datasets - WIQA, QUAREL and DEFEASIBLE for CURIE (dataset statistics in Table 3).

**WIQA**: WIQA task studies the effect of a perturbation in a procedural text (Tandon et al., 2019). The context  $T$  is a procedural text describing a physical process, and *st* is a perturbation i.e., an external situation deviating from  $T$ , and the effect of *st* is either **helps** or **hurts**. See Table 1 for examples.

**QUAREL**: QUAREL dataset (Tafjord et al., 2019) contains qualitative story questions where  $T$  is a narrative, and *st* is a qualitative statement.  $T$  and *st* are also expressed in a simpler, logical form, which we use as it highlights the reasoning challenge. The effect of *st* is **entails** or **contradicts** (see Table 1).

Research question	Training dataset	Test dataset	Task	Metrics
Can we generate good <i>st</i> graphs? (§3)	WIQA- <i>st</i>	WIQA- <i>st</i>	generation	ROUGE, BLEU
	QUAREL- <i>st</i>	QUAREL- <i>st</i>	generation	ROUGE, BLEU
	DEFEASIBLE- <i>st</i>	DEFEASIBLE- <i>st</i>	generation	ROUGE, BLEU
Can we improve downstream tasks? (§4.1)	WIQA- <i>st</i> , WIQA-QA	WIQA-QA	finetuned QA	accuracy

Table 2: Overview of experiments

Dataset	train	dev	test
WIQA	119.2k	34.8k	34.8k
QUAREL	4.6k	1.3k	652
DEFEASIBLE	200k	14.9k	15.4k

Table 3: Dataset wise statistics, we maintain the splits

**DEFEASIBLE:** The DEFEASIBLE reasoning task (Rudinger et al., 2020) studies inference in the presence of a counterfactual. The context  $T$  is a premise describing an everyday context, and the situation  $st$  is an observed evidence which either **strengthens** or **weakens** the hypothesis. We adapt the original abductive setup as shown in Table 1. In addition to commonsense situations, DEFEASIBLE-*st* also comprises of social situations, thereby contributing to the diversity of our datasets.

### 2.3 Learning to Generate *st*-graphs

To reiterate our task formulation (§2.1), for a given context and  $st$ , we first specify a set of questions and the resulting outputs for the questions is then compiled to form a *st*-graph.

The training data consists of tuples  $(\mathbf{x}_i, \mathbf{y}_i)$ , with  $\mathbf{x}_i = (T, st, r, c)_i$  where  $T$  denotes the context,  $st$  the situation,  $r$  is the edge (**green** or **red**),  $c$  indicates the nature of the effect (imminent or eventual), and  $\mathbf{y}_i$  is the output (a short sentence or a phrase depicting the effect). The output of  $N_Q$  such questions is compiled into a graph  $G = \{\mathbf{y}_i\}_{1:N_Q}$  (Fig. 1).

We use a pretrained language model  $\mathcal{M}$  to estimate the probability of generating an answer  $\mathbf{y}_i$  for an input  $\mathbf{x}_i$ . We first transform the tuple  $\mathbf{x}_i = \langle x_i^1, x_i^2, \dots, x_i^N \rangle$  into a single query sequence of tokens by concatenating its components i.e.  $\mathbf{x}_i = \text{concat}(T, st, r, c)$ , where `concat` is string concatenation. Let the sequence of tokens representing the target event be  $\mathbf{y}_i = \langle y_i^1, y_i^2, \dots, y_i^M \rangle$ , where  $N$  and  $M$  are the lengths of the query and the target event sequences. We model the conditional

**Algorithm 1:** ITERATIVEGRAPHGEN (IGEN): generating *st* graphs with CURIE

**Given:** CURIE language model  $\mathcal{M}$ .

**Given:** Context  $T$ , situation  $st$ , a set  $R = \{(r_i, c_i)\}_{i=1}^{N_Q}$  of  $N_Q$   $(r, c)$  tuples.

**Result:** *st* graph  $G$ :  $i^{\text{th}}$  node is generated with relation  $r_i$ , effect type  $c_i$ .

**Init:**  $G \leftarrow \emptyset$

**for**  $i \leftarrow 1, 2, \dots, N_Q$  **do**

    /\* Create a query \*/

$\mathbf{x}_i = \text{concat}(T, st, r_i, c_i)$ ;

    /\* Sample a node from  $\mathcal{M}$  \*/

$\mathbf{y}_i \sim \mathcal{M}(\mathbf{x}_i)$ ;

    /\* Add sampled node, edge \*/

$G = G \cup (r_i, c_i, \mathbf{y}_i)$ ;

**end**

**return**  $G$

probability  $p_\theta(\mathbf{y}_i | \mathbf{x}_i)$  as a series of conditional next token distributions parameterized by  $\theta$ : as  $p_\theta(\mathbf{y}_i | \mathbf{x}_i) = \prod_{k=1}^M p_\theta(y_i^k | \mathbf{x}_i, y_i^1, \dots, y_i^{k-1})$ .

### 2.4 Inference to Decode *st*-graphs

The auto-regressive factorization of the language model  $p_\theta$  allows us to efficiently generate target event influences for a given test input  $\mathbf{x}_j$ . The process of decoding begins by sampling the first token  $y_j^1 \sim p_\theta(y | \mathbf{x}_j)$ . The next token is then drawn by sampling  $y_j^2 \sim p_\theta(y | \mathbf{x}_j, y_j^1)$ . The process is repeated until a specified *end-symbol* token is drawn at the  $K^{\text{th}}$  step. We use nucleus sampling (Holtzman et al., 2019) in practice. The tokens  $\langle y_j^1, y_j^2, \dots, y_j^{K-1} \rangle$  are then returned as the generated answer. To generate the final *st*-reasoning graph  $G$ , we combine all the generated answers  $\{\mathbf{y}_i\}_{1:N_Q}$  that had the same context and  $st$  pair  $(T, st)$  over all  $(r, c)$  combinations. We can then use generated answer  $st' \in \{\mathbf{y}_i\}_{1:N_Q}$ , as a new input to  $\mathcal{M}$  as  $(T, st')$  to recursively expand the *st*-graph to arbitrary depth and structures (Al-

gorithm 1). One such instance of using CURIE *st* graphs for a downstream QA task is shown in §4.

### 3 RQ1: Establishing Baselines for *st*-graph Generation

This section reports on the quality of the generated *st* reasoning graphs and establishes strong baseline scores for *st*-graph generation. We use the datasets described in section §2.2 for our experiments.

Model ( $\mathcal{M}$ )	BLEU	ROUGE
WIQA- <i>st</i>		
LSTM Seq-to-Seq	7.51	18.71
GPT $\sim$ (w/o $T$ )	7.82	19.30
GPT-2 $\sim$ (w/o $T$ )	10.01	20.93
GPT	9.95	19.64
GPT-2	<b>16.23</b>	<b>29.65</b>
QUAREL- <i>st</i>		
LSTM Seq-to-Seq	13.05	24.76
GPT $\sim$ (w/o $T$ )	20.20	36.64
GPT-2 $\sim$ (w/o $T$ )	26.98	41.14
GPT	25.48	42.87
GPT-2	<b>35.20</b>	<b>50.57</b>
DEFEASIBLE- <i>st</i>		
LSTM Seq-to-Seq	7.84	17.50
GPT $\sim$ (w/o $T$ )	9.91	20.63
GPT-2 $\sim$ (w/o $T$ )	9.17	9.43
GPT	10.49	<b>21.79</b>
GPT-2	<b>10.52</b>	21.19

Table 4: Generation results for CURIE with baselines for language model  $\mathcal{M}$ . We find that context is essential for performance (w/o  $T$ ). We provide these baseline scores as a reference for future research.

#### 3.1 Baseline Language Models

To reiterate, CURIE is composed of (i) task formulation component and (ii) graph construction component, that uses a language model  $\mathcal{M}$  to construct the *st*-graph. We want to emphasize that any language model architecture can be a candidate for  $\mathcal{M}$ . Since our *st*-task formulation is novel, we establish strong baselines over the three datasets. Our experiments include large-scale language models (LSTM and pretrained transformer) with varying parameter sizes and pre-training, and the corresponding ablation studies. Our choices for  $\mathcal{M}$  are:

**LSTM Seq-to-Seq:** We train an LSTM (Hochreiter and Schmidhuber, 1997) based sequence to sequence model (Bahdanau et al., 2015) which uses global attention described in (Luong et al., 2015).

We initialize the embedding layer with pre-trained 300 dimensional Glove (Pennington et al., 2014)<sup>1</sup>. We use 2 layers of LSTM encoder and decoder with a hidden size of 500. The encoder is bidirectional.

**GPT:** We use the original design of GPT (Radford et al., 2018) with 12 layers, 768-dimensional hidden states, and 12 attention heads.

**GPT-2:** We use the medium (355M) variant of GPT-2 (Radford et al., 2019) with 24 layers, 1024 hidden size, 16 attention heads. For both GPT and GPT-2, we initialize the model with the pre-trained weights and use the implementation provided by Wolf et al. (2019).

We use Adam (Kingma and Ba, 2014) for optimization with a learning rate of  $5e - 05$ . All the dropouts (Srivastava et al., 2014) were set to 0.1. We found the best hyperparameter settings by searching the space using the following hyperparameters.

1. embedding dropout = {0.1, 0.2, 0.3}
2. learning rate = {1e-05, 2e-05, 5e-05, 1e-06}

We compare the *st*-graphs generated by various language models with the gold-standard reference graphs. To compare the two graphs, we first flatten both the reference graph and the *st*-graph as text sequences and then compute the overlap between them. Due to a lack of strong automated metrics, we use the commonly used evaluation metrics for generation BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004)<sup>2</sup>. Our results shown in Table 4 indicate that the task of *st* generation is challenging, and suggests that incorporating *st*-reasoning specific inductive biases might be beneficial. At the same time, Table 4 shows that even strong models like GPT-2 achieve low BLEU and ROUGE scores (specifically on WIQA and DEFEASIBLE), leaving a lot of room for model improvements in the future.

We also show ablation results for the model with respect to the context  $T$  (§2.1), by fine-tuning without the context. We find that context is essential for performance for both GPT and GPT-2 (indicated with w/o  $T$  in Table 4). Further, we note that the gains achieved by adding context are higher for GPT-2, hinting that larger models can more effectively utilize the context<sup>3</sup>.

<sup>1</sup><https://github.com/OpenNMT/OpenNMT-py>

<sup>2</sup><https://github.com/Maluuba/nlg-eval>

<sup>3</sup>More qualitative examples shown in appendix B

Error category	%	Example question	Reference	Predicted
Polarity	7%	What does ‘oil fields over-used’ help eventually ?	there is not oil refined	more oil is refined
Linguistic Variability	27%	What does ‘rabbits will not become pregnant’ hurt imminently ?	more rabbits	more babies
Related Event	23%	What does ‘inhaling more air from the outside’ hurt imminently ?	there will be less oxygen in your blood	you develop more blood clots in your veins
Wrong	40%	What does ‘nutrients are unavailable for plants’ hurt eventually ?	more plants	more wine being produced
Erroneous Reference	3%	What does ‘rabbit are not mating’ hurt imminently?	less rabbits	more babies

Table 5: Canonical examples per error category. Error analysis is only shown for the incorrect outputs. For polarity errors, we use guidelines shown in appendix A.1

### 3.2 Human Evaluation

N-gram metrics such as BLEU and ROUGE are known to be limited, specifically for reasoning tasks. Further, we observe from Table 4 that context is crucial for generation quality. To better understand this effect, we perform human evaluation on a random sample from the dev set to compare GPT-2- w/o  $T$  and GPT-2 models. Our goal is to assess quality of generations, and the importance of grounding generations in context. Four human judges annotated 100 unique samples for *correctness*, *relevance* and *reference*, described next.

**Correctness:** We conducted a human evaluation to evaluate the correctness of the generated graphs where we aggregated nodes for a given  $st$ . The user interface for the annotation (shown in Figure 3) displayed the context  $T$  and the corresponding graph  $G$  generated by GPT-2 using Algorithm 1. The human judges were asked to annotate the nodes, edges, and the overall graph for correctness. A graph was labeled as correct if either a) all the nodes and edges were correct, or b) the graph had a minor issue that the judges deem not detrimental to the overall correctness. The inter-annotator agreement on graph correctness was substantial with a Fleiss’ Kappa score (Fleiss and Cohen, 1973) of 0.69. Table 6 shows that human judges rated  $>75\%$  of the graphs to be correct given the context, showing that CURIE generates high-quality graphs for a diverse set of contexts.

**Relevance:** The annotators are provided with the context  $T$ , the situation  $st$ , and the relational ques-

Attribute	Node	Edge	Graph
% Correct	79.71	77.78	75.36

Table 6: Human Analysis of Graph Correctness. About 75% of the graphs were deemed as *correct*.

tions. The annotators were asked, “Which system (A or B) is more accurate relative to the background information given in the context?” They could also pick option C (no preference). The order of the references was randomized. Table 7 (row 1) shows that GPT-2 outperforms GPT-2 (w/o  $T$ ), confirming our hypothesis that context is important as GPT-2 generates target events that are grounded in the passage and source events.

Task	GPT-2 (w/o $T$ )	GPT-2	No Preference
Relevance	23.05	46.11	30.83
Reference	11.67	31.94	56.39

Table 7: Results of human evaluation. The numbers show the percentage(%) of times a particular option was selected for each metric.

**Reference:** We measure how accurately each system-generated event reflects the reference (true) event. Here, the annotators saw only the reference sentence and the outputs of two systems (A and B) in a randomized order. We asked the annotators, “Which system’s output is closest in meaning to the reference?” The annotators could pick the options A, B, or C (no preference). Table 7 (row 2) illus-

C left	<input type="text" value="-- select a quality --"/>	<input type="text" value="['plants dont die and organic material isnt created']"/>	<input type="text" value="['more plants die and become organic material']"/>	C right	<input type="text" value="-- select a quality --"/>
edge CS l	<input type="text" value="-- select a quality --"/>	↓	✓	edge CS r	<input type="text" value="-- select a quality --"/>
S left		<input type="text" value="['if the organic material is increased']"/>	<input type="text" value="['']"/>	S right	
edge SM l		↓	↘	edge SM r	
M left	<input type="text" value="-- select a quality --"/>	<input type="text" value="['tress wont be able to grow']"/>	<input type="text" value="['more tress will grow']"/>	M right	<input type="text" value="-- select a quality --"/>
edge MH l		↓	✗	edge MH r	
H left	<input type="text" value="-- select a quality --"/>	<input type="text" value="['less forest formation']"/>	<input type="text" value="['more forest formation']"/>	H right	<input type="text" value="-- select a quality --"/>
Overall graph quality	<input type="text" value="-- select a quality --"/>				

Figure 3: User interface for graph correctness evaluation. The human judges were asked to rate the if the generated nodes, edges, and the overall graph are correct for the given context. The paragraph for this example was: *Grass and small plants grow in an area. These plants die. The soil gains organic material. The soil becomes more fertile. Larger plants are able to be supported. Trees eventually grow.*

trates that the output generated by GPT-2 is closer in meaning to the reference compared to GPT-2 (w/o  $T$ ) reinforcing the importance of context.

Both the models (with and without context) produced similarly grammatically fluent outputs.

### 3.3 Error Analysis

The reference and relevance task scores together show that GPT-2 does not generate target events that are exactly similar to the reference target events, but are correct in the context of the passage and source event. To investigate this, we analyze a random sample of 100 points from the dev set. Out of the erroneous samples, we observe the following error categories (shown in Table 5):

- **Polarity (7%)**: Predicted polarity was wrong but the event was correct.
- **Linguistic Variability (27%)**: Output was a linguistic variant of the reference.
- **Related event (23%)**: Output was related but different reference expected.
- **Wrong (40%)**: Output was fully unrelated.
- **Erroneous reference (3%)**: Gold annotations themselves were erroneous.

### 3.4 Consistency Analysis

Finally, we measure if the generated  $st$ -graphs are consistent. Consider a path of length two in the generated  $st$ -graph (say,  $A \rightarrow B \rightarrow C$ ). A consistent graph would have identical answers to *what does A help eventually* i.e., “C”, and *what does B help imminently* i.e., “C”. To analyze consistency, we

manually evaluated 50 random generated length-two paths, selected from WIQA- $st$  dev set. We observe that 58% samples had consistent output w.r.t the generated output. We also measure consistency w.r.t. the gold standard (the true outputs in the dev set), and observe that the system output is  $\approx 48\%$  consistent. Despite being trained on independent samples,  $st$ -graphs show reasonable consistency and improving consistency further is an interesting future research direction.

### 3.5 Discussion

In summary, CURIE allows adapting pretrained language models to generate  $st$ -graphs that humans meaningful and relevant with a high degree of correctness. We also perform an in-depth analysis of the errors of CURIE. We establish multiple baselines with diverse language models to guide future research. We show that context is more important than model size for  $st$ -reasoning tasks.

## 4 RQ2: CURIE for Downstream Tasks

In this section, we describe the approach for augmenting  $st$  graphs for downstream reasoning tasks. We first identify the choice of tasks ( $st$ -tasks) for domain adaptive pretraining (Gururangan et al., 2020) and obtain CURIE language model  $\mathcal{M}$  (based on GPT-2). The downstream task then provides input context,  $st$  and (relation, type) tuples of interest, and obtains the  $st$ -graphs (see Algorithm 1) from CURIE. We describe one such instantiation in §4.1.

### 4.1 CURIE augmented WIQA-QA

We examine the utility of CURIE-generated graphs in the WIQA-QA (Tandon et al., 2019) downstream question answering benchmark. Input to this task

is a context supplied in form of a passage  $T$ , a starting event  $c$ , an ending event  $e$ , and the output is a label  $\{\textit{helps}, \textit{hurts}, \textit{or no\_effect}\}$  depicting how the ending  $e$  is influenced by the event  $c$ .

We hypothesize that CURIE can augment  $c$  and  $e$  with their influences, giving a more comprehensive scenario than the context alone. We use CURIE trained on WIQA-*st* to augment the event influences in each sample in the QA task as additional context. We obtain the influence graphs for  $c$  and  $e$  by defining  $R_{fwd} = \{(\textit{helps}, \textit{imminent}), (\textit{hurts}, \textit{imminent})\}$  and  $R_{rev} = \{(\textit{helped by}, \textit{imminent}), (\textit{hurt by}, \textit{imminent})\}$ , and using algorithm 1 as follows:

$$G(c) = \text{IGEN}(T, c, R_{fwd})$$

$$G(e) = \text{IGEN}(T, e, R_{rev})$$

We hypothesize that WIQA-*st* graphs are able to generate reasoning chains that connect  $c$  to  $e$ , even if  $e$  is not an immediate consequence of  $c$ . Following Tandon et al. (2019), we encode the input sequence  $\text{concat}(T, c, e)$  using the BERT encoder  $E$  (Devlin et al., 2019), and use the [CLS] token representation ( $\hat{\mathbf{h}}_i$ ) as our sequence representation.

We then use the same encoder  $E$  to encode the generated effects  $\text{concat}(G(c), G(e))$ , and use the [CLS] token to get a representation for augmented  $c$  and  $e$  ( $\hat{\mathbf{h}}_a$ ). Following the encoded inputs, we compute the final loss as:  $\mathbf{l}_i = \text{MLP}_1(\hat{\mathbf{h}}_i)$ , and  $\mathbf{l}_a = \text{MLP}_1(\hat{\mathbf{h}}_a)$  and  $\mathcal{L} = \alpha \times \mathcal{L}_i + \beta \times \mathcal{L}_a$ , where  $\mathbf{l}_i, \mathbf{l}_a$  represent the logits from  $\hat{\mathbf{h}}_i$  and  $\hat{\mathbf{h}}_a$  respectively, and  $\mathcal{L}_i$  and  $\mathcal{L}_a$  are their corresponding cross-entropy losses.  $\alpha$  and  $\beta$  are hyperparameters that decide the contribution of the generated influence graphs and the procedural text to the loss. We set  $\alpha = 1$  and  $\beta = 0.9$  across experiments.

**QA Evaluation Results** Table 8 shows the accuracy of our method vs. the vanilla WIQA-BERT model by question type and number of hops between  $c$  and  $e$ . We also observe from Table 8 that augmenting the context with generated influences from CURIE leads to considerable gains over WIQA-BERT based model, with the largest improvement seen in 3-hop questions (questions where the  $e$  and  $c$  are at a distance of three reasoning hops in the influence graphs). The strong performance on the 3-hop question supports our hypothesis that generated influences might be able to connect two event influences that are farther apart in the reasoning chain. We also show in Table 8 that augmenting with CURIE improves performance on the difficult

Query Type	WIQA-BERT + CURIE	WIQA-BERT
1-hop	<b>78.78</b>	71.60
2-hop	<b>63.49</b>	62.50
3-hop	<b>68.28</b>	59.50
Out-of-para	<b>64.04</b>	56.13
In-para	73.58	<b>79.68</b>
No effect	<b>90.84</b>	89.38
Overall	<b>76.92</b>	73.80

Table 8: QA accuracy by number of hops, and question type. WIQA-BERT refers to the original WIQA-BERT results reported in Tandon et al. (2019), and WIQA-BERT + CURIE are the results obtained by augmenting the QA dataset with the influences generated by CURIE.

Out-of-para category of questions, which requires background knowledge.

#### Source of improved performance: *st* graphs?

Since CURIE uses GPT-2 model to generate the graphs, we perform an additional experiment to verify whether simply using GPT-2 classifier for WIQA would achieve the same performance gains. To establish this, we train a GPT-2 classifier, and augment it with CURIE graphs to compare their relative performances on WIQA. Table 9 shows that augmenting CURIE graphs to both WIQA-BERT and GPT-2 classifiers provides consistent gains, suggesting the effectiveness of CURIE graphs.

Model	Accuracy
WIQA-BERT	73.80
WIQA-BERT + CURIE	<b>76.92*</b>
GPT-2	72.70
GPT-2 + CURIE	<b>74.33*</b>

Table 9: WIQA-QA results for both WIQA-BERT and GPT-2 augmented with CURIE graphs. Across both classifiers, augmenting CURIE graphs shows performance gains. \*-indicates statistical significance

WIQA-BERT scores are slightly lower than the GPT-2 scores for WIQA classification despite having similar parameter size. We hypothesize that this is due to the pretrained classification token ([CLS]) in WIQA-BERT, while GPT-2 uses the pooling operation over the sequence for classification. In summary, the evaluation highlights the value of CURIE as a framework for improving performance on downstream tasks that require coun-



terfactual reasoning and serves as an evaluation of the ability of CURIE to reason about *st*-scenarios.

## 4.2 Discussion

In summary, we show substantial gains when a generated *st*-graph is fed as an additional input to the QA model. Our approach forces the model to reason about influences within a context, and then answer the question, which proves to be better than answering the questions directly.

## 5 Related Work

**Language Models for Knowledge Generation:** Using large scale neural networks to generate knowledge has been studied under various task settings (Sap et al., 2019; Bosselut et al., 2019; Shwartz et al., 2020; Bosselut et al., 2021; Malaviya et al., 2019). Another line of querying language models (LMs) aims to understand the type of knowledge LMs contain. Davison et al. (2019) explore whether BERT prefers true or fictitious statements over ConceptNet (Speer et al., 2017). Logan et al. (2019) observe that the LM over-generalize to produce wrong facts, while Kassner and Schütze (2019) show that negated facts are also considered valid in an LM.

Our work closely aligns with Tandon et al. (2019), Bosselut et al. (2019), and Bosselut et al. (2021). Compared to Bosselut et al. (2019), CURIE gives a method that can naturally incorporate context and reason about situation via hops and nature of the influence. Additionally, any node can be arbitrarily expanded via the iterative procedure, producing complete graphs for situations. We reformulate the task of studying event influence from a QA task (Tandon et al., 2019) to a generation task. Our framework is similar in spirit to Bosselut et al. (2019), but extend it for situational reasoning with LMs. Bosselut et al. (2021) aim to generate events that can aid commonsense tasks. In contrast, our focus is context-grounded *st* graph generation. To this end, our formulation includes multiple forward/backward reactions, imminent and eventual edges, and an algorithm to compile the individual nodes to a complete graph (Algorithm 1).

**Situational reasoning :** There has been immense interest in extracting event chains (as causal graphs) in stories and news corpora in both unsupervised (Chambers and Jurafsky, 2008) and supervised (Rudinger et al., 2015; Liu et al., 2018; Asghar, 2016; Dunietz et al., 2017; Nordon et al.,

2019; Zhao et al., 2017) settings. Such approaches often depend on events that are explicitly mentioned in the input text, thereby unable to generate events beyond the input text.

Recently, there has been interest in *st* reasoning from a retrieval setting (Lin et al., 2019) and also generation setting, attributed partially to the rise of neural generation models (Yangfeng Ji and Celikyilmaz, 2020) as knowledge bases (Petroni et al., 2019; Roberts et al., 2020; Talmor et al., 2020; Shwartz et al., 2020; Sap et al., 2019). Qin et al. (2019) present generation models to generate the path from a counterfactual to an ending in a story. Current systems make some simplifying assumptions, e.g. that the ending is known. Multiple *st* (e.g., more sunlight, more pollution) can happen at the same time, and these systems can only handle one situation at a time. All of these systems assume that *st* happens once in a context. Our framework strengthens this line of work by not assuming that the ending is given during deductive *st* reasoning.

## 6 Conclusion

We present CURIE, a situational reasoning that: (i) is effective at generating *st*-reasoning graphs, validated by automated metrics and human evaluations, (ii) improves performance on two downstream tasks by simply augmenting their input with the generated *st* graphs. Further, our framework supports recursively querying for any node in the *st*-graph. Our future work is to design models that seek consistency, and study recursive *st*-reasoning as a bridge between dialog and reasoning.

## Acknowledgments

We would like to thank Peter Clark for the thoughtful discussions and useful feedback on the draft. We also want to thank the anonymous reviewers for valuable feedback. This material is partly based on research sponsored in part by the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

## References

- Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *ArXiv*, abs/1605.07895.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439.
- Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.
- 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.
- Joe Davison, Joshua Feldman, and Alexander Rush. 2019. [Commonsense knowledge mining from pre-trained 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)*, pages 1173–1178, Hong Kong, China. 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.
- Jesse Dunietz, Lori S. Levin, and J. Carbonell. 2017. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133.
- Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.
- H. Grice. 1975. Logic and conversation syntax and semantics. In *Logic and conversation Syntax and Semantics*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Nora Kassner and Hinrich Schütze. 2019. Negated lama: Birds cannot fly. *arXiv preprint arXiv:1911.03343*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *MRQA@EMNLP*.
- Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018. [Narrative modeling with memory chains and semantic supervision](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–284, Melbourne, Australia. Association for Computational Linguistics.
- Robert Logan, Nelson F Liu, Matthew E Peters, Matt Gardner, and Sameer Singh. 2019. Barack’s wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.
- Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2019. Exploiting structural and semantic context for commonsense knowledge base completion. *arXiv preprint arXiv:1910.02915*.
- Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, and Kira Radinsky. 2019. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1102–1109.

- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *EMNLP*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. [URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf).
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.
- Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1681–1686.
- Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. [Thinking like a skeptic: Defeasible inference in natural language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.
- Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olympics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.
- 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 6078–6087.
- Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.
- Thomas Wolf Yangfeng Ji, Antoine Bosselut and Asli Celikyilmaz. 2020. [The amazing world of generation](#). *EMNLP tutorials*.
- Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, T. Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*.

## A Appendix

### A.1 Polarity Words

To compute polarity for error analysis, we use the following words as guidelines.

**Increasing words** *helps, more, higher, increase, increases, stronger, faster, greater, longer, larger, helping*

**Decreasing words** *hurts, less, lower, decrease, decreases, weaker, slower, smaller, hurting, softer, fewer*

### B Examples from CURIE

In table 12, we show some qualitative QA examples from CURIE. Here, *Topic Matches* signifies whether the generated answers is relevant to the context.

<b>Path Length</b>	<b>Question</b>	<b>Answer</b>	<b>Predicted Answer</b>	<b>Topic Matches</b>
2	hurts more magma is released	LESS igneous rock forming	LESS or SMALLER eruptions.	Yes
2	hurts more water in solid or liquid phase	MORE evaporation?	LESS vapor being made.	Yes
2	hurts less flowers for bees to visit	MORE flowers being pollinated?	a LARGER cake?.	No
1	helps changes in the atmosphere	if the moon orbits slower	MORE/LARGER solar power being generated?.	No
2	helps less time goes by	there are fewer fissures	LESS or SMALLER eruptions.	Yes
3	hurts there is less water in the lakes	a MORE INTENSE water cycle?	MORE salt being removed from the water.	No
2	helps tectonic plates rub less	SMALLER mountains	LESS or SMALLER eruptions.	Yes
1	hurts Better crop watering	LESS crops	LESS vegetables.	Yes
3	hurts weather is accommodating for frogs	LESS frogs	LESS or SMALLER Frogs.	Yes
1	helps there will be more eggs	MORE fish?	MORE birds.	No
3	helps Urine levels increase	LESS waste being removed from the body	MORE waste being removed from the body.	Yes
3	helps Having normal body functions	HURTING the DNA to replicate properly	MORE cells being damaged.	Yes
2	helps not having a water source nearby	vines will grow faster	LESS or SMALLER grapes.	Yes
1	hurts bees may become extinct	more bees are collecting pollen	LESS butterflies.	Yes
3	hurts more bees fly around	LESS flowers being pollinated	a SMALLER flowers.	Yes
2	hurts the earths core gets warmer	less volcanic mountains are formed	LESS or SMALLER valleys being formed.	Yes
3	hurts more land available	LESS igneous rock forming	less magma cools.	Yes
3	hurts there is less water on the ground	a MORE INTENSE water cycle	MORE salt being removed from the water.	No

Table 10: Sample Generations. Topic matches captures whether the topic of the generated event matches with the context. (section 3). Path length = 1 refers to the immediate effects, and Path length > 1 refers to eventual effects.

<b>Path Length</b>	<b>Question</b>	<b>Answer</b>	<b>Predicted Answer</b>	<b>Topic Matches</b>
3	hurts more conservation achieved	LESS flowers being pollinated	less nectar available.	Yes
2	hurts the eggs become food for other fish	MORE fish	Less larvae eat and grow.	Yes
2	helps more magma inside volcano	more magma changes in pressure	MORE/GREATER eruptions?.	Yes
2	helps less commercial fishing	more fry emerge	LESS damage by acid rain.	
2	hurts more stormy weather occurs	less plant growth occurs	MORE vegetables.	Yes
2	helps more pumpkin seeds planted	MORE or LARGER pumpkins	more water used for more flowers.	No
2	hurts more Global warming causes extreme temperatures	Rains are plentiful and more regular	MORE vegetables?.	Yes
2	helps warmer weather evaporates more water	a MORE INTENSE water cycle	MORE/STRONGER storms?.	Yes
2	helps dry hot environment evaporates water	LESS frogs	MORE or LARGER frogs.	Yes
3	helps stronger heat source	MORE evaporation	more heat causes the molecules to increase in energy.	Yes
2	helps living in a rain forest	more water collects in the bodies of water	MORE salt being removed from the water.	No
2	hurts there is no tadpole from the egg	MORE frogs	MORE ELABORATE swimming.	No
1	helps more pulling and stretching of tectonic plates	more cracks in earths crust	MORE or STRONGER earthquakes.	Yes
2	hurts less animals that hunt frogs	less tadpoles loses their tails	more fish grow bigger.	No
2	hurts both kidneys are present and functioning	less waste is removed from the blood	less waste is removed in the blood.	Yes

Table 11: Sample Generations. Topic matches captures whether the topic of the generated event matches with the context. (section 3). Path length = 1 refers to the immediate effects, and Path length > 1 refers to eventual effects. (section 3).

<b>Path Length</b>	<b>Question</b>	<b>Answer</b>	<b>Predicted Answer</b>	<b>Topic Matches</b>
2	helps the bees have a very hairy leg gene	the bees would carry more pollen away from the flower	a LARGER nectar star.	Yes
2	hurts If more eggs are layed	MORE frogs	the mouth will grow smaller.	No
1	hurts bees are imported	fewer bees land on flowers	a SMALLER hive.	No
1	hurts more adolescent fish grow to adulthood	fewer fish can lay more eggs	LESS damage by acid rain.	No
2	helps the heat rises	greater precipitations will happen	MORE/STRONGER .	Yes
2	helps All the eggs were eaten	There were few eggs laid	less eggs are laid..	Yes
1	hurts plates move away from each other	edges of plates crumple more	MORE or GREATER eruptions.	Yes
1	hurts more proteins available	less help occurs	less endowment of nucleotides.	Yes

Table 12: Sample Generations. Topic matches captures whether the topic of the generated event matches with the context. Path length = 1 refers to the immediate effects, and Path length > 1 refers to eventual effects. (section 3).