Abstract

Knowledge underpins reasoning. Recent research demonstrates that when relevant knowledge is provided as additional context to commonsense question answering (QA), it can substantially enhance the performance even on top of state-of-the-art. The fundamental challenge is where and how to find such knowledge that is high quality and on point with respect to the question; knowledge retrieved from knowledge bases are incomplete and knowledge generated from language models are inconsistent.

We present RAINIER¹, or Reinforced Knowledge Introspector, that learns to generate contextually relevant knowledge in response to given questions. Our approach starts by imitating knowledge generated by GPT-3, then learns to generate its own knowledge via reinforcement learning where rewards are shaped based on the increased performance on the resulting question answering. RAINIER demonstrates substantial and consistent performance gains when tested over 9 different commonsense benchmarks: including 5 datasets that are seen during model training, as well as 4 datasets that are kept unseen. Our work is the first to report that knowledge generated by models that are orders of magnitude smaller than GPT-3, even without direct supervision on the knowledge itself, can exceed the quality of commonsense knowledge elicited from GPT-3.

1 Introduction

Commonsense is a significant challenge for modern NLP models, due to the obscurity of underlying knowledge that grounds the reasoning process. While humans are generally able to introspect the underlying reasons for their conclusion (Mercier and Sperber, 2017), neural models lack the capability to verbalize the premises leading to their prediction. This hinders models’ performance and robustness on commonsense tasks, and makes it difficult to inspect their point of failure. Recent research has demonstrated that relevant knowledge can provide useful context for approaching commonsense tasks. Yet these methods either retrieve from in-domain knowledge bases (Mitra et al., 2019; Chang et al., 2020) that do not have good coverage over commonsense, or generate knowledge from neural models (Shwartz et al., 2020; Gu et al., 2022; Liu et al., 2022), which often need domain-specific engineering and very large models (e.g. GPT-3 (Brown et al., 2020)). It is still an open challenge to systematically find high-quality knowledge.

In this work, we use a novel, reinforcement-learning-based method to develop RAINIER, a generative neural model that can introspect the underlying knowledge for reasoning about commonsense questions. As illustrated in Fig 1, RAINIER is trained to generate knowledge that are both fluent natural language statements, and useful prompts that optimize the performance of a generic question answering (QA) model. Our model (1) demonstrates strong generalization to unseen benchmarks without additional engineering effort (e.g. fine-tuning), (2) produces commonsense knowledge of

¹Code, model and knowledge-extended datasets are available at http://github.com/liujch1998/rainier

Figure 1: RAINIER can introspect for commonsense knowledge that underpin the reasoning process, and is trained via reinforcement learning, where the reward is derived from the effectiveness of knowledge when prompting a frozen, generic QA model.
high quality and diversity, and (3) is substantially smaller in size compared to GPT-3, the best knowledge source reported so far.

To train RAINIER, we optimize knowledge introspection for the resulting QA, instead of direct supervision, because there are usually no gold knowledge labels on commonsense datasets. In order to ensure that our model learns to generate generically useful knowledge for a broad range of QA models, we train only RAINIER, the knowledge introspector, without finetuning the QA model. Since our desired knowledge are sequences of discrete, non-differentiable word tokens, we adapt a reinforcement learning algorithm, Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022), to optimize the knowledge introspector. Specifically, the reward is defined as the effect of RAINIER-generated knowledge on the QA model’s prediction. We train RAINIER in a multi-task setting on 8 commonsense QA datasets – encompassing general, scientific, physical, and social commonsense – to equip the model with better generalization to unseen benchmarks.

Experiments show that RAINIER substantially improves the performance of QA models on 9 commonsense benchmarks (5 datasets seen during training and 4 unseen datasets), and gives larger and more consistent gains than a few-shot GPT-3 (Liu et al., 2022) despite being 16x smaller in parameter size. It also boosts the performance on top of those QA models that it is not trained against, indicating that it generates generically useful knowledge instead of merely hacking into the reward given by a single QA model. Knowledge generated by RAINIER can even boost a QA model that is 4x larger than it, showing the promise of using model-generated knowledge as a complement to model scaling in making progress in commonsense reasoning. Our analyses show that the knowledge generated by RAINIER are of high quality, and are diverse in terms of domain (e.g. scientific, social), relation expressed (e.g. part of, member of, purpose), and syntactic property (e.g. negation, comparison). The effect of these knowledge on the QA model also aligns well with human judgments. The success of RAINIER shows that moderately-sized models can serve as source of high-quality and useful commonsense knowledge that facilitates reasoning. We publicly release the code, the trained RAINIER model, and the commonsense datasets extended with knowledge generated by RAINIER.

Algorithm 1 Training RAINIER

\begin{algorithm}
\textbf{Input}: initial policy model $\theta_0$, initial value model $\phi_0$, pretrained QA model $\psi_{\text{QA}}$
\begin{align*}
D_{\text{seen}} & \leftarrow \text{Get silver knowledge on $D_{\text{seen}}$ from GPT-3.} \\
\theta_{\text{init}} & \leftarrow \text{Optimize $\theta_0$ with Eqn 2 from $D_{\text{init}}$.} \quad \triangleright \text{Section 2.1}
\end{align*}
\begin{align*}
\theta_{\text{RAINIER}} & \leftarrow \text{REINFORCE\textsc{Learning}($D_{\text{seen}}$, $\theta_{\text{init}}$, $\phi_0$, $\psi_{\text{QA}}$)} \quad \triangleright \text{Section 2.2}
\end{align*}
\begin{algorithmic}
\Procedure{REINFORCE\textsc{Learning}}{($D_{\text{seen}}$, $\theta$, $\phi$, $\psi_{\text{QA}}$)}
\State $\theta_{\text{old}} \leftarrow \theta$, $\phi_{\text{old}} \leftarrow \phi$
\For{iterations = 1, 2, \ldots}
\State Sample a minibatch from $D_{\text{seen}}$.
\For{step = 1, 2, \ldots, $s$}
\State Compute $L_{\text{PPO}}$ on the minibatch with Eqn 3.
\State Optimize $\theta$ and $\phi$ with $L_{\text{PPO}}$ for one step.
\EndFor
\State $\theta_{\text{old}} \leftarrow \theta$, $\phi_{\text{old}} \leftarrow \phi$
\EndFor
\Return $\theta$
\EndProcedure
\end{algorithmic}
\textbf{Output}: $\theta_{\text{RAINIER}}$
\end{algorithm}

2 Method

Problem Overview. We focus on the tasks of multiple-choice commonsense QA, consisting of instances of format $x = (q, A, a^*)$, where $q$ is the question, $A$ is the set of candidate answers, and $a^* \in A$ is the correct answer. For full contextualization, we append candidate answers $A$ to the question $q$ to form the input to the QA model as follows:

$q = \langle \text{question} \rangle (A) \langle \text{choice}_A \rangle (B) \langle \text{choice}_B \rangle \ldots$

Common approaches only train supervised QA models. As a complement, we train a separate model, which we refer to as RAINIER, that can introspect question-specific knowledges that are useful to prompt a fixed QA model. RAINIER is a sequence-to-sequence language model, $p_{K}(K|q; \theta)$, and we expect it to generate knowledge statements ($k$’s) in response to the given question ($q$). However, the challenge is that we have no gold knowledge labels as supervision.

Training. Since we do not have gold knowledge to train RAINIER, we obtain this model by fine-tuning a pretrained language model in two stages: (I) imitation learning, and then (II) reinforcement learning. In Stage I (§2.1), we get silver knowledge labels on some datasets from GPT-3, and teach our model to imitate this knowledge-generating GPT-3. This equips our model with the basic functionality of knowledge generation. In Stage II (§2.2), we use reinforcement learning to continue training the model obtained in Stage I to make the generated knowledge more useful while staying fluent and
meaningful. Specially, we set the reward to be the effect of the generated knowledge on the prediction made by a fixed, generic QA model. We obtain silver knowledge and train RAINIER on the union of multiple QA datasets (which are considered seen during training), i.e., $\mathcal{D}_{\text{seen}} = \bigcup_{d=1}^{n_{\text{seen}}} \mathcal{D}_d$, where $\mathcal{D}_d = \{(q_j, A_j, a_j^\ast)\}_{j=1}^{|\mathcal{D}_d|}$. The generic QA model we use may or may not have been trained on these seen datasets. The complete training process is outlined in Algorithm 1.

Inference. The effectiveness of RAINIER is evaluated against a set of unseen QA datasets, $\mathcal{D}_{\text{unseen}}$, in addition to the seen datasets. Note that RAINIER is not trained on any unseen datasets, which means we neither get silver knowledge, nor do imitation learning or reinforcement learning on them. The generic QA model we use was not trained on any unseen datasets as well. We discuss details of inference in §2.3.

### 2.1 Training Stage I: Imitation Learning

In Stage I, we train RAINIER so that it generates fluent and meaningful natural language statements that resemble knowledge. There is no large-scale commonsense QA dataset labeled with high-quality knowledge, but GPT-3 has been shown as a good generator for relevant knowledge (Liu et al., 2022). Therefore, we get silver knowledge from GPT-3 on our seen datasets. Following Liu et al. (2022), we elicit question-related knowledge by prompting GPT-3 with a task-specific set of few-shot demonstrations (See §C for details on the prompts), and decoding $M$ knowledge for each question:

$$K(q) = \{k_m : k_m \sim p_{G}(k \mid \text{prompt}(q), q)\},$$

where $p_{G}(\cdot \mid \cdot)$ denotes GPT-3 with nucleus sampling where $p = 0.5$ (Holtzman et al., 2020). This yields a silver dataset of question-knowledge pairs:

$$\mathcal{D}_{\text{imit}} = \{(q, k) : (q, A, a^\ast) \in \mathcal{D}_{\text{seen}}, k \in K(q)\}.$$  

(1)

We then train RAINIER, starting from a pre-trained sequence-to-sequence language model, on this silver dataset with standard supervised loss:

$$L^{\text{train}}(\theta) \propto \sum_{(q,k) \in \mathcal{D}_{\text{imit}}} - \log p_{K}(k \mid q; \theta).$$  

(2)

The parameterization of the resulting model is denoted as $\theta_{\text{imit}}$.

### 2.2 Training Stage II: Reinforcement Learning

As we will see in the empirical results, the imitation model obtained in Stage I does not provide the most beneficial knowledge. Therefore, in Stage II, we continue optimizing RAINIER to generate knowledge that best prompts the QA model, by directly maximizing the reward given by this QA model.

Knowledge generation as reinforcement learning. Since knowledge statements ($k$’s) are discrete and thus non-differentiable, we adopt a reinforcement learning approach, and consider knowledge generation as a sequential decision making process over the natural language vocabulary space. We consider the generation of knowledge statement $k$ with $T$ tokens as an episode of length $T$. At step $t \in [1, T]$, the state $s_t = (q, k_{<t})$ is the combination of the question and the knowledge decoded up to the $(t-1)$-th token; the action $a_t = k_t$ would be the $t$-th token to decode. The RAINIER model, $p_{K}(k_t \mid q, k_{<t}; \theta)$, is the policy model that we optimize. We define a reward function $r(x, k)$ that characterizes the effect of the knowledge on the QA model’s prediction, and discuss the definition of this reward function in §2.2.1.

To ensure that the generated knowledge stay fluent and meaningful, we would like the learned policy model not to move too far from the initial imitation model. Therefore, we add to the reward an (approximate) KL penalty between the learned policy and the initial policy (Ouyang et al., 2022),

$$R(x, k) = r(x, k) - \beta \log \frac{p_{K}(k \mid q; \theta)}{p_{K}(k \mid q; \theta_{\text{imit}})}.$$  

Since this reward is computed based on the full knowledge statement, we assign it to the last step of the episode. Non-terminal steps are assigned zero rewards. Formally,

$$r_T = R(x, k) \quad (\text{where } T = |k| \text{ and } k_T = [\text{EOS}]);$$

$$r_t = 0 \quad (\text{where } 1 \leq t < T).$$

We employ Proximal Policy Optimization\(^2\) (PPO) (Schulman et al., 2017) as our reinforcement learning algorithm, and adapt from the implementation of PPO in Ouyang et al. (2022). Aside from

\(^2\)We choose PPO because it has shown successful results in other NLP tasks (Nakano et al., 2021; Stiennon et al., 2020). Our earlier experiments with REINFORCE did not show promising results.
the policy model, PPO additionally uses a *value model* (parameterized by $\phi$) to estimate the value function for states with incomplete decoded text, i.e. $V(s_t; \phi)$ for any $t$. PPO minimizes a joint loss,

$$L_{PPO}(\theta, \phi) = L_{Policy}(\theta) + \alpha \cdot L_{Value}(\phi), \quad (3)$$

where $L_{Policy}(\theta)$ is the loss on the policy model, $L_{Value}(\phi)$ is the loss on the value model, and $\alpha$ is a hyperparameter.

**Policy loss.** To obtain the policy loss, we first compute the *truncated estimated advantage function*,

$$\hat{A}_t = \sum_{t'=t}^{T-1} (\gamma\lambda)^{t'-t} \delta_{t'},$$

where $\delta_{t'} = r_{t'} + \gamma V(s_{t'+1}; \phi) - V(s_{t'}; \phi)$, where the value functions $V(\cdot)$ are estimated by the value model. PPO then maximizes the empirical expectation of a so-called *clipped surrogate objective term*,

$$cso(\hat{A}_t, \nu_t(\theta), \varepsilon) = \min \left( \nu_t(\theta)\hat{A}_t, \text{clip}(\nu_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t \right),$$

where $\nu_t(\theta) = \frac{p_K(k_t|q_t;\theta)}{p_K(k_t|q_t;\theta_{old})}$ is the ratio between the current policy $\theta$ and a lagging policy $\theta_{old}$. The lagging policy is updated to the current policy under a fixed interval of $s$ training steps, and is kept fixed otherwise. We adapt this to our use case, and define the policy loss as

$$L_{Policy}(\theta) = -\mathbb{E}\left[cso(\hat{A}_t, \nu_t(\theta), \varepsilon)\right]$$

where the expectation is taken over all instances in the training data ($x \sim D_{\text{train/seen}}$), the distribution of model-generated knowledge as determined by the current policy conditioning on the instance’s question ($k \sim p_K(k|q;\theta)$), and all tokens in the knowledge statement ($t \in [1, |k|]$).

**Value loss.** The value model is trained with MSE loss with respect to the target value, $V_t^{\text{targ}}$, which in turn is estimated with a lagging value model $\phi_{old}$:

$$L_{Value}(\phi) = \mathbb{E}\left[\left(V(s_t; \phi) - V_t^{\text{targ}}\right)^2\right],$$

where $V_t^{\text{targ}} = \sum_{t'=t}^{T-1} \gamma^{t'-t}r_{t'} + \gamma^{T-t}V(s_T; \phi_{old})$.

### 2.2.1 Reward Shaping

We define the reward function in reinforcement learning as the quantified effect of RAINIER’s knowledge on the QA model’s prediction. Suppose we already have a reasonably good QA model, which assigns a probability score $P_{QA}(a|q)$ to any candidate answer $a \in A$. Since we will use a sequence-to-sequence language model (i.e. UnifiedQA (Khashabi et al., 2020)) as the QA model, we define

$$P_{QA}(a|q) = \frac{\exp S_{QA}(a|q)}{\sum_{a' \in A} \exp S_{QA}(a'|q)},$$

where

$$S_{QA}(a|q) = \frac{1}{|a|} \sum_{i=1}^{|a|} -\log p_{QA}(a_i|q, a_{<i}; \psi_{QA}),$$

where $p_{QA}(a_i|q, a_{<i}; \psi_{QA})$ is the language modeling score received by $a_i$, the $i$-th token of $a$. The naive prediction would be the candidate answer that gets the highest $P_{QA}(a|q)$ (or equivalently, the highest $S_{QA}(a|q)$): $\hat{a} = \text{arg max}_{a \in A} P_{QA}(a|q)$.

We aim at maximizing $P_{QA}(a^*|q \circ k)$, the probability score received by the correct answer when the QA model is prompted with the knowledge $k$ generated by RAINIER, and $\circ$ denotes text concatenation. One naive definition of reward function may be

$$r(x, k) = P_{QA}(a^*|q \circ k) - P_{QA}(a^*|q).$$

However, this reward only captures the absolute change of score, but not whether the model prediction is changed or not. To remedy for this, we define the reward function as

$$r(x, k) = \frac{1}{2} \left[ \tanh\left(S_{QA}(a^*|q \circ k) - \max_{a' \in A, a' \neq a^*} S_{QA}(a'|q \circ k)\right) \right.$$

$$\left. - \tanh\left(S_{QA}(a^*|q) - \max_{a' \in A, a' \neq a^*} S_{QA}(a'|q)\right) \right].$$

Intuitively, this function would give a reward of near $+1$ if the naive prediction is incorrect (i.e. $S_{QA}(a^*|q \circ k) < \max_{a' \in A, a' \neq a^*} S_{QA}(a'|q)$), while the knowledge-prompted prediction is correct (i.e. $S_{QA}(a^*|q \circ k) > \max_{a' \in A, a' \neq a^*} S_{QA}(a'|q \circ k)$). Similarly, the reward would be near $-1$ if the naive prediction is correct but the knowledge-prompted prediction is incorrect. The hyperbolic tangent
serves as a smoothed sign function, and provides a soft interpolation between the two polarity of reward values by taking into account the margin of the correct answer.

We also experiment with some alternative definitions of the reward function. See Table 4.

**Reward normalization.** To stabilize training, we apply an affine transformation on the rewards so that initially they are normalized. Before starting Stage II training, we use the imitation model to generate a knowledge statement for each training instance, and estimate the population mean and standard deviation of rewards:

\[
\mathcal{R}_{\text{init}} = \{ r(x, k) : x \in D_{\text{train}}^\text{seen}, k \sim p_K(\cdot | q; \theta_{\text{init}}) \}, \\
\mu_0 = \mu(\mathcal{R}_{\text{init}}), \sigma_0 = \sigma(\mathcal{R}_{\text{init}}). 
\]

In Stage II training, each reward is normalized as:

\[
r(x, k) \leftarrow \frac{r(x, k) - \mu_0}{\sigma_0}.
\]

### 2.3 Inference: Knowledge Prompting and Aggregation

Following Liu et al. (2022), at inference time we use RAINIER to generate multiple knowledge per question, and prompt the QA model by individually concatenating each knowledge to the question. The knowledge are generated by RAINIER with nucleus sampling where \( p = 0.5 \) (Holtzman et al., 2020),

\[
K(q) = \{ \epsilon \} \cup \{ k_m : k_m \sim p_{K}^{p=0.5}(k | q; \theta), \quad m = 1 \ldots M \},
\]

where \( M \) is the number of knowledge per question, and \( \epsilon \) denotes empty string. We collect a set of outputs for prompting with each knowledge. The final prediction is the candidate answer that receives maximum confidence,

\[
\hat{a} = \arg \max_{a \in A} \max_{k \in K(q)} P_{QA}(a | q \circ k),
\]

and the prediction is supported by a single knowledge – the selected knowledge,

\[
\hat{k} = \arg \max_{k \in K(q)} \max_{a \in A} P_{QA}(a | q \circ k).
\]

**Training time model selection.** In Stage II training, we only generate one knowledge per question for the validation set.\(^3\) Predictions are made using the same knowledge prompting method as above, and the model checkpoint with the maximal accuracy on the union of all validation sets is selected.

---

\(^3\)This is for efficiency purposes. We use greedy decoding here because it is more stable than nucleus sampling when generating only one knowledge per question.

### 3 Experiments

**Seen datasets.** For both imitation learning and reinforcement learning, we use the 8 multiple-choice datasets that UnifiedQA\(_v2\) (Khashabi et al., 2022) uses for training: OpenBookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018), AI2Science (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), QASC (Khot et al., 2020), PhysicaliQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), and Winogrande (Sakaguchi et al., 2021).

**Unseen datasets.** We additionally evaluate our method on the following 4 multiple-choice QA datasets that our model was not trained on: NumerSense (Lin et al., 2020), RiddleSense (Lin et al., 2021), QuaRTz (Tafjord et al., 2019), and Hel-laSwag (Zellers et al., 2019).

**Models.** For Stage I training, we get silver knowledge from the GPT-3-Curie (13B) model (Brown et al., 2020). The knowledge introspector is initialized with T5-large (Raffel et al., 2019), which has 0.77B parameters. For Stage II training, we initialize the value model with T5-large, and replace the language modeling head with a value regression head, which is initialized from scratch; we use UnifiedQA-large (UQA-large) (Khashabi et al., 2020) as the QA model that provides reward, which means the text concatenation function is defined as \( q \circ k = \{ q \} \setminus \{ k \} \). We use the same question formatting as UnifiedQA. See Table 7 for hyperparameters.

**Baselines.** We mainly report performance improvements over the vanilla QA baseline (i.e. direct inference with the UnifiedQA-large model and without prompting RAINIER-generated knowledge). We also consider using knowledge from:

- **Few-shot GPT-3** (Liu et al., 2022), where knowledge statements are elicited from the GPT-3-Curie (13B) model – under the same prompts used for getting silver knowledge in Stage I training (§2.1), and the same hyperparameter setting for decoding (\( M = 10 \) knowledge per question, with nucleus sampling where \( p = 0.5 \)).

- **Self-talk** (Shwartz et al., 2020), where we generate \( M = 10 \) knowledge per question with GPT-3-Curie and a variety of templates.

\(^4\)We exclude MCTest and RACE because most questions in these reading comprehension datasets are too long to fit into our model’s input.
Table 1: Results on seen datasets. All experiments use UnifiedQA-large as the QA model, and optionally uses knowledge from one of the knowledge generation models. Skipped baselines are marked with “–”.

| Dataset → Method | CSQA | QASC | PIQA | SIQA | WG | Avg. |
|------------------|------|------|------|------|----|------|
| UQA-large (0.77B) | 61.43 | 53.00 | 43.09 | 45.65 | 63.66 | 65.50 | 53.84 | 57.21 | 53.35 | 54.67 | 55.07 | 55.21 |
| + Few-shot GPT-3-Curie (13B) | 66.34 | – | 53.24 | – | 64.25 | – | 58.29 | – | 55.56 | – | 59.54 | – |
| + Self-talk GPT-3-Curie (13B) | 63.31 | – | 49.89 | – | 65.23 | – | 51.89 | – | 52.96 | – | 56.66 | – |
| + DREAM (11B) | 64.54 | – | 49.46 | – | 64.74 | – | 51.59 | – | 56.12 | – | 57.29 | – |

+ RAINIER-large (0.77B) [ours] 67.24 | 60.18 | 54.97 | 54.13 | 65.67 | 67.09 | 57.01 | 59.01 | 56.91 | 57.39 | 60.36 | 59.56 |

Table 2: Results on unseen datasets.

| Dataset → Method | NS | RS | QuaRTz | HS | Avg. |
|------------------|----|----|--------|----|------|
| UQA-large (0.77B) | 26.50 | 19.61 | 28.11 | 38.34 | 68.75 | 67.60 | 35.00 | 34.30 | 39.59 | 41.85 |
| + Few-shot GPT-3-Curie (13B) | 38.00 | – | 35.65 | – | 69.01 | – | 37.33 | – | 45.00 | – |

+ RAINIER-large (0.77B) [ours] 30.00 | 21.81 | 30.07 | 41.22 | 70.31 | 68.24 | 35.73 | 34.80 | 41.53 | 43.76 |

4 Results

4.1 Main Results

Performance on seen datasets. Table 1 shows the performance of RAINIER-enhanced QA model on the seen datasets. On average, our method achieves more than 5% improvement over directly applying the QA model. The knowledge generated by RAINIER improves performance on five benchmarks: CommonsenseQA, QASC, PhysicalIQA, SocialIQA, and Winogrande, with the greatest improvement on CommonsenseQA (+6%) and QASC (+12%). As shown in Table 8, there is no performance gain on OpenBookQA, ARC, and AI2Science. We conjecture that this is because the QA model, UnifiedQA, is already trained on these three datasets, thus setting a strong baseline.

Comparison with other models. Compared to RAINIER, other knowledge generation models, including few-shot GPT-3, Self-talk, and DREAM, provide generally weaker improvements over the vanilla QA baseline. In particular, RAINIER outperforms GPT-3-based models while being 16x smaller in parameter size (0.77B vs. 13B).

Performance on unseen datasets. Table 2 shows that RAINIER’s knowledge substantially improves performance over the vanilla QA baseline on the four unseen datasets, demonstrating its generalization capability.

Choice of QA model for evaluation. To verify that our RAINIER model is not hacking into the rewards provided by the QA model we use during training, we evaluate the effect of RAINIER’s knowledge on different QA models. We choose three other UnifiedQA models with different sizes, as well as a different model known as Unicorn (Lourie et al., 2021). Results are shown in Figure 2. RAINIER consistently gives performance gains on top of all QA models, indicating that its knowledge are generally useful information rather than mere artifacts of model-specific reward hacking. We even observe performance gains with a QA model that is 4x as large as RAINIER, which means generating and prompting relevant knowledge can be a technique complementary to model scaling, and can be done meaningfully with smaller models. Finally, we see the largest improvement when the QA model itself has weak, but non-trivial, performance (UnifiedQA-small for seen datasets, and UnifiedQA-base for unseen datasets).

4.2 Ablations

Stage I and Stage II training. We experimented with omitting the Stage I (imitation) and/or Stage II (reinforcement) from the training pipeline. Results are shown in Table 3. Without Stage I training,
RAINIER does not improve the performance of the QA model (regardless of whether it is trained with Stage II or not), showing the indispensability of equipping the model with the basic functionality of knowledge generation. On the other hand, a model trained solely with Stage I gives weaker improvements than the fully trained RAINIER, stressing the importance of Stage II training as well.

Reward function. Table 4 shows the results for knowledge introspectors trained with different reward functions. Our reward shaping gives the best performance on unseen datasets, as well as one of the top performance on seen datasets. While the naive prob diff reward function gives slightly better performance on seen datasets, our reward shaping results in better generalization.

4.3 Analysis

To get a deeper understanding of the behavior and capability of RAINIER, we manually analyzed the generated knowledge along several quality and diversity aspects. We asked three NLP experts to annotate the selected knowledge (§2.3) for up to 100 questions per dataset among the validation sets of 8 benchmarks (5 seen, 3 unseen; see Figure 3). It was hidden from the annotators whether the knowledge rectifies or misleads QA model’s prediction, so potential bias is eliminated.

Quality. First, we follow Liu et al. (2022) by annotating the quality aspects — relevance, factuality, and helpfulness — of each knowledge with respect to the question. We find that RAINIER-generated knowledge are overwhelmingly related to the respective questions. 64% are factually correct, 25% are factually incorrect, and the remaining 11% have undetermined factuality due to various reasons (e.g. ambiguity, cultural sensitivity). 58% are seen by human as being helpful for reasoning about the question, whereas 24% are seen as harmful.

In our annotations, there are 420 knowledge that rectify UnifiedQA-large’s predictions (i.e. flipping from wrong to right), and 246 knowledge that mislead the predictions (i.e. flipping from right to wrong). Among the rectifying knowledge, 84% are deemed helpful by human; and among the misleading knowledge, 62% are deemed helpful by human; and among the misleading knowledge, 62% are deemed harmful. These results have similar trends as Liu et al. (2022), and show that RAINIER’s knowledge are of high quality and interpretability in helping QA models.

Diversity. Additionally, we analyze the diversity aspects by annotating each knowledge with the importance of both training stages.

Table 3: Ablations on the choice of reward function.

Table 4: Ablations on the choice of reward function.
and its syntactic property(s) (e.g., negation, comparison). See Figure 3 for complete list of options under each aspect. The knowledge’s domain distribution is strongly tied to the domain of the benchmark (e.g., scientific for QASC and QuaRTz, social for SocialIQA and Winogrande, numerical for NumerSense). The domain aspect is more diverse for benchmarks that test general commonsense, like CommonsenseQA and RiddleSense. For the relation aspect, there are many knowledge that express an “attribute” relation, while other relations are also substantially represented. As for syntax, a good proportion of the knowledge contain structures like comparison and negation. Therefore, RAINIER’s knowledge have good syntactic and semantic diversity while being able to adapt to the domain.

4.4 Qualitative Examples

We show some examples of good knowledge generated by RAINIER in Table 5.

5 Related Work

Explicit reasoning for commonsense QA. Commonsense question answering poses a significant challenge to modern neural models. To improve performance and interpretability, many work have proposed to do explicit reasoning for tasks in this area, that is, to verbalize the intermediate text artifacts that facilitate the reasoning process. Rajani et al. (2019) and Latcinnik and Berant (2020) use supervised learning to train models to generate text explanations, while Gu et al. (2022) and Bansal et al. (2021) use similar training regimes to obtain models that can generate scene elaborations and paths through a structured knowledge graph, respectively. Schwartz et al. (2020) and Paranjape et al. (2021) prompt pretrained models with pre-defined templates to generate question clarifications or contrastive explanations, which are in turn used to prompt the inference model. The above approaches all pose, implicitly or explicitly, certain constraints (e.g., domain, relation, syntax) on the model-generated text. In contrast, Wei et al. (2022) elicits full chain-of-reasoning from language models with in-context learning; Liu et al. (2022) uses few-shot demonstrations to elicit flexible, relevant knowledge statements from a language model, and Wang et al. (2022) distills this capability into smaller models using supervised learning. These methods provide more flexibility on the knowledge, yet they rely on accessing very large language models (e.g., GPT-3). Aside from methods that make reasoning explicit in a linear chain manner, another set of work produce recursive structures of reasoning, through either backward chaining (Dalvi et al., 2022; Jung et al., 2022) or forward chaining (Bostrom et al., 2022). Our work contributes to this line of research, yet we depart from prior work by presenting the first approach that learns to generate relevant knowledge without requiring human-labeled gold knowledge.

Reinforcement learning for NLP. Recently, reinforcement learning methods have been adopted for NLP tasks like question answering (Nakano et al., 2021), summarization (Stiennon et al., 2020; Paulus et al., 2018), machine translation (Shen et al., 2016; Wu et al., 2016), grounded text generation (Ammanabrolu et al., 2021, 2022), controlled text generation (Lu et al., 2022), and prompt generation (Guo et al., 2021; Deng et al., 2022). Our application of reinforcement learning on knowledge introspection is novel. The idea of reinforce-
| Task | Question / Knowledge | Domain | Relation | Syntax |
|------|----------------------|--------|----------|--------|
| CSQA | What would vinyl be an odd thing to replace? (A) pants (B) record albums (C) record store (D) cheese (E) wallpaper | scientific | member of | – |
| QASC | Some pelycosaurs gave rise to reptile ancestral to (A) lamphreys (B) angiosperm (C) mammals (D) paramecium (E) animals (F) protozoa (G) arachnids (H) backbones | scientific | member of | – |
| SIQA | Sydney rubbed Addison’s head because she had a horrible headache. What will happen to Sydney? (A) drift to sleep (B) receive thanks (C) be reprimanded | social | – | – |
| WG | Adam always spent all of the free time watching TV unlike Hunter who volunteered, due to (A) being lazy. (B) Adam (C) Hunter | social | attribute | comparison |
| RS | Causes bad breath and frightens blood-suckers (A) tuna (B) iron (C) trash (D) garlic (E) pubs | – | attribute | – |
| QuaRTz | If the mass of an object gets bigger what will happen to the amount of matter contained within it? (A) gets bigger (B) gets smaller | scientific | physical | – |

Table 5: Examples of good knowledge generated by RAINIER. Each of these knowledge rectifies UnifiedQA-large’s prediction, and is labeled by the annotator as relevant, factual, and helpful.

ment learning with model-provided feedback has been previously explored in Guo et al. (2021), Ammanabrolu et al. (2021), and Lu et al. (2022). The PPO algorithm has been previously employed to optimize rewards learned from human feedback (Nakano et al., 2021; Stiennon et al., 2020). In contrast, we use PPO to optimize reward purely derived from the decision-making neural models.

6 Conclusion

We introduced RAINIER, a neural model that can introspect for relevant knowledge on a broad range of commonsense question answering tasks. RAINIER is trained with a novel adaption of reinforcement learning, and does not need gold knowledge labels that are difficult to obtain. Knowledge generated by RAINIER can serve as useful prompts that improves the performance of QA models on both seen and unseen benchmarks, and outperform knowledge elicited from a few-shot GPT-3 which is 16x bigger. RAINIER generates knowledge in the form of natural language statements that are fluent, meaningful, high-quality, and diverse in terms of domain and relation; furthermore, the effect of these knowledge on the QA model is found to align well with human judgments.

Limitations

Despite the positive effect of our knowledge introspector RAINIER on commonsense QA tasks, its performance on non-commonsense applications is unknown and thus requires further investigation. Even for commonsense applications, there is still a large gap between model performance and human performance, so the resulting model is not ready for real-world applications. There is also a limit on the length of knowledge it generates in our experimental setting, and it has not been tested on generating long and coherent text. Furthermore, in some cases it may generate knowledge that express inappropriate social values (Table 10), are culture-specific (Table 11), or contain ethical risks (Table 12). See §B for examples. Extra care should be taken when applying our model in production environments, especially when making critical decisions or exposing its generated contents directly to human end users.

Acknowledgements

This work was funded in part by the DARPA MCS program through NIWC Pacific (N66001-19-2-4031), NSF IIS-2044660, and ONR N00014-18-1-2826. We thank OpenAI for offering access to the GPT-3 API.

We would like to thank Prithviraj Ammanabrolu, Alisa Liu and Weijia Shi for the discussion and feedback on early drafts of the paper. We also thank the anonymous reviewers for their valuable feedback.
References

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. arXiv preprint arXiv:2205.01975.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktaeschel, and Jason Weston. 2021. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 807–833. Online. Association for Computational Linguistics.

Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Neet Kaur, and Balaji Krishnamurthy. 2021. Cose-co: Text conditioned generative commonsense contextualizer.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, IAAI 2020, New York, NY, USA, February 7-12, 2020, pages 7432–7439. AAAI Press.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. arXiv preprint arXiv:2201.06028.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Khashabi, Ashish Sabharwal, Jack Urbanek, Margaret Li, Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457.

Mingkai Deng, Jiantao Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Riprompt: Optimizing discrete text prompts with reinforcement learning. arXiv preprint arXiv:2205.12548.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. DREAM: Improving situational QA by first elaborating the situation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Text generation with efficient (soft) q-learning. arXiv preprint arXiv:2106.07704.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeez Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. arXiv preprint arXiv:2205.11822.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQa-v2: Stronger generalization via broader cross-format training. arXiv preprint arXiv:2202.12359.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQAOA: Crossing format boundaries with a single QA system. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1896–1907. Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, IAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8082–8090. AAAI Press.

Veronica Lattinink and Jonathan Berant. 2020. Explaining question answering models through text generation. arXiv preprint arXiv:2004.05569.
Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. *QuaRTz: An open-domain dataset of qualitative relationship questions*. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

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.

Wenya Wang, Vivek Srikumar, Hanna Hajishirzi, and Noah A Smith. 2022. Elaboration-generating commonsense question answering at scale. *arXiv preprint arXiv:2209.01232*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. *Google’s neural machine translation system: Bridging the gap between human and machine translation*. *arXiv preprint arXiv:1609.08144*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. *HellaSwag: Can a machine really finish your sentence?* In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
A Additional Experimental Details

A.1 Hyperparameters
See Table 7.

A.2 Baselines

Self-talk. We generate $M = 10$ knowledge per question with GPT-3-Curie, using 10 pairs of question-answer templates adapted from Shwartz et al. (2020). We generate one knowledge from each template: first, query GPT-3 with the question template and using nucleus sampling ($p = 0.2$) to obtain a full question; next, query GPT-3 again with both the full question and the corresponding answer template, this time using nucleus sampling ($p = 0.5$), to obtain a full answer sentence. The answer sentence will be treated as the knowledge.

DREAM. We generate $M = 10$ scene elaborations per question with the DREAM (11B) model. Gu et al. (2022) proposes 4 types of scene elaborations: motivation ($M$), emotion ($E$), rule-of-thumb ($ROT$), and consequence ($Con$). Each is associated with a control code that guides the DREAM model. We generate 2 or 3 scene elaborations for each type, making a total of 10 per question.

B Additional Analysis

Table 9 through 12 show more analysis of knowledge generated by RAINIER. Table 9 shows semantically problematic knowledge. Table 10 shows knowledge that express some social value. Table 11 shows knowledge that are culture-specific. Table 12 shows knowledge that have potential ethical risks. All examples are taken from the validation set of the respective dataset.

C Prompts for Getting Silver Knowledge from GPT-3
See Table 13 through 20.

| Question Template          | Answer Template          |
|----------------------------|--------------------------|
| What is the definition of  | The definition of _ is   |
| What is the main purpose of| The purpose of _ is to   |
| What is the main function of| The main function of a _ is|
| What are the properties of a| The properties of a _ are that _ is|
| What is a                  | As a result of _         |
| What happened as a result of| The cause of _ was       |
| What might have caused     | A part of _ is           |
| What is a part of          | An example of _ is       |
| What is an example of      | One would _ by           |

Table 6: Templates used in the self-talk baseline.
| Symbol | Value | Description |
|--------|-------|-------------|
| $M$    | 20    | Number of knowledge statements to sample from GPT-3, per question. |
| $p$    | 0.5   | Parameter for nucleus sampling from GPT-3. |
| $L_{\text{output}}$ | 64 | Max length of output from GPT-3. |

### STAGE I: IMITATION LEARNING

| Symbol | Value | Description |
|--------|-------|-------------|
| $L_{\text{input}}$ | 256 | Max length of input to RAINIER (i.e. question plus choices). |
| $L_{\text{output}}$ | 64 | Max length of output from RAINIER (i.e. generated knowledge). |
| $B$    | 64    | Batch size for training. |
| $S$    | 50,000 | Total number of training steps. |
| $\eta$ | $1 \times 10^{-5}$ | Learning rate of Adam optimizer. |

### STAGE II: REINFORCEMENT LEARNING

| Symbol | Value | Description |
|--------|-------|-------------|
| $\alpha$ | 1.0 | Weight of value model loss in PPO. |
| $\beta$ | 0.2 | Weight of entropy bonus term in reward. |
| $\gamma$ | 1.0 | Discount factor for rewards. |
| $\lambda$ | 0.95 | Parameter for advantage estimation. |
| $\varepsilon$ | 0.2 | Clipping range for the clipped surrogate objective. |
| $L_{\text{input}}$ | 256 | Max length of input to RAINIER (i.e. question plus choices). |
| $L_{\text{output}}$ | 32 | Max length of output from RAINIER (i.e. generated knowledge). |
| $\tau$ | 0.7 | Temperature for knowledge sampling in PPO training. |
| $E$    | 1M    | Total number of training episodes. |
| $B$    | 64    | Batch size for training. |
| $S$    | 15,625 | Total number of training steps. |
| $s$    | 4     | Interval (in steps) for updating the lagging models (policy and value). |
| $\eta$ | $2 \times 10^{-5}$ | Learning rate of Adam optimizer (with a linear learning rate decay schedule). |

### INFEERENCE

| Symbol | Value | Description |
|--------|-------|-------------|
| $M$    | 10    | Number of knowledge statements to sample from RAINIER, per question. |
| $p$    | 0.5   | Parameter for nucleus sampling from RAINIER. |
| $L_{\text{input}}$ | 256 | Max length of input to RAINIER (i.e. question plus choices). |
| $L_{\text{output}}$ | 32 | Max length of output from RAINIER. |

Table 7: Hyperparameter settings.

### Table 8: Results on the other 3 seen datasets.

| Dataset                | Method                | OBQA | ARC easy | ARC hard | ARC elem | ARC mid |
|------------------------|-----------------------|------|----------|----------|----------|---------|
| UQA-large (0.77B)      |                       | 70.20| 69.12    | 55.85    | 69.11    | 64.80   |
| + Few-shot GPT-3-Curie (13B) |               | 68.80| 71.05    | 56.52    | 70.73    | 65.60   |
| + RAINIER-large (0.77B) [ours] |          | 69.60| 67.72    | 55.18    | 68.29    | 63.20   |

All experiments use UnifiedQA-large as the QA model, and optionally uses knowledge from one of the knowledge generation models. On these datasets, RAINIER-generated knowledge does not give an improvement over the vanilla QA baseline.
| Task | Question / Knowledge | Problem with the knowledge |
|------|-----------------------|-----------------------------|
| RS   | I am a fruit. I am tasty and provide lots of energy. You can also find me in a calendar. What am I? (A) shop (B) choppers (C) date (D) european (E) movie | Ambiguous The knowledge does not specify which fruit year, so it is not possible to verify its factuality. |
| QuaRTz | Carla is pushing two carts down a street. One cart is heavy, the other is light. The heavy cart, compared to the light one, will accelerate (A) slowly (B) quickly | Under-specified Missing the control factor that the forces exerted on the carts are the same. Though this flaw also exists in the question itself. |
| RS   | What has a head at night but not in the morning? (A) sleeping (B) pillow (C) neck (D) shampoo (E) beer | Over-specified It would be normal to say "animals have a head". It is weird to qualify with "sleeping" and "at night". |
| WG   | Because William developed a sore knee during his race against Kyle, _ won the race. (A) William (B) Kyle | Over-confident William might still be able to finish the race with a sore knee. |
| RS   | what is the begining of enternity the end of life the end of time and the end to every race (A) e (B) quick (C) finality (D) fix (E) habit | Tautological This knowledge does not give any meaningful information. |
| QuaRTz | Sharon is conducting an experiment on valence electrons and soon discovers that when they are closer to the nucleus, they are ______ easily removed from the atom. (A) more (B) less | Not applicable This knowledge implicitly compares the removability of valence vs. non-valence electrons. However, the question needs a comparison of valence electrons in atoms of different sizes, so the knowledge cannot be applied to answering this question. |

Table 9: Examples of knowledge generated by RAINIER which are semantically problematic.
| Task | Question / Knowledge | Problem with the knowledge |
|------|----------------------|---------------------------|
| SIQA | Riley broke loose from the house. He thought he would never get out of there. Why did Riley do this? (A) Stay in the house longer (B) think about his life (C) go home for a while. | Social value Break out of a bad habit is usually a bad idea. |
| SIQA | Tracy heard a faint buzzing noise and immediately ran for her life. How would you describe Tracy? (A) scared of bees (B) sad (C) not phased by bees. | Social value It is hard to decide whether this knowledge should be considered factual or not. |
| SIQA | Remy gave Skylar’s Netflix account password to one of Remy’s other friends. How would Skylar feel as a result? (A) like a bad friend (B) excited (C) used. | Social value One should not be scared of bees. |
| SIQA | Riley was the best of friends with the boy with cancer. What will Riley want to do next? (A) visit the hospital (B) shun the friend (C) become friends with the boy with cancer too. | Social value One should visit their sick friend. |
| SIQA | Carson tried to fight Robin last night because Robin hurt Carson a lot. What will Carson want to do next? (A) apologize (B) do nothing (C) hurt Robin. | Social value It is generally a kind thing to visit a sick friend. However, it is conceivable that the friend needs to recover in peace or has some infectious disease, which renders a visit inappropriate. |
| SIQA | Bailey told Alex to send the pdf because they didn’t want to do it themselves. How would Alex feel as a result? (A) lazy about work (B) happy (C) angry | Social value One should be willing to help others. |
| SIQA | Kendall wrapped a bandage around my neck after getting injured in a fight. What will Kendall want to do next? (A) harm them (B) punish them (C) protect them. | Social value One should help others in need. |

Table 10: Examples of knowledge generated by RAINIER that express some social value.
| Task | Question / Knowledge                                                                 | Problem with the knowledge                      |
|------|-------------------------------------------------------------------------------------|--------------------------------------------------|
| SIQA | Remy made hay getting home from school on Friday the 13th. Why did Remy do this?    | Culture-specific                                 |
|      | (A) go to school before this (B) had heard that on Friday the 13th, God would bless you if you just went home and hid (C) had heard that on Friday the 13th, that people pranked other people really bad People are more likely to be pranked on Friday the 13th. | This knowledge largely applies within western, especially Christian, culture. |
|      |                                                                                     |                                                  |
| WG   | Lindsey like to read graphic novels but Natalie liked classic literature to read. _ bought the new Frank Miller comic at the book store. (A) Lindsey (B) Natalie Graphic novels are popular among teens. | Culture-specific                                 |
|      |                                                                                     | This knowledge could be correct or not depending on the culture or geographical location. |
| CSQA | Where would a person live that isn’t in the metro area but still has good schools? (A) woods (B) bed (C) suburbs (D) rural (E) neighborhood A person who lives in the suburbs is likely to have good schools. | Culture-specific                                 |
|      |                                                                                     | This knowledge could be correct or not depending on the culture or geographical location, and it is not even always true within the U.S. |
| NS   | most people in the u.s. works <mask> consecutive hours in a business day. (A) no (B) one (C) two (D) three (E) four (F) five (G) six (H) seven (I) eight (J) nine (K) ten Most people work eight hours in a day. | Culture-specific                                 |
|      |                                                                                     | This knowledge is an over-generalization. In some other cultures, most people work more than eight hours, or less. |
| CSQA | What will god never do according to religion? (A) anything (B) judge people (C) work miracles (D) judge men (E) everywhere God is a judge of people. | Culture-specific                                 |
|      |                                                                                     | This is a topic specific to the religion. |

Table 11: Examples of knowledge generated by RAINIER that are culture-specific.

| Task | Question / Knowledge                                                                 | Problem with the knowledge                      |
|------|-------------------------------------------------------------------------------------|--------------------------------------------------|
| RS   | A woman screams. You go into a shop and find a man lying dead on the floor. There are three people in the room. A postman, a milkman and a doctor. Who killed him? (A) doctor (B) police station (C) liars (D) radio (E) water fountain The doctor is the one who killed the man. | Ethical risk                                     |
|      |                                                                                     | This knowledge might impress bias towards a certain profession. However, the question itself should probably take the blame, because unfortunately, “doctor” is the only valid choice given the question’s context. |
| PIQA | crossing out (A) ugly people from the bachelor is cool (B) ugly people from planet earth is cool Crossing out ugly people from planet earth is cool. | Ethical risk                                     |
|      |                                                                                     | This knowledge might impress bias towards a group with certain physical characteristic. |
| CSQA | With the card slot lit up he knew how to get started finding his balance with what? (A) slot machine (B) ticket machine (C) bank machine (D) telephone (E) automated teller A slot machine is a machine that takes cards and uses them to make money. | Ethical risk                                     |
|      |                                                                                     | This knowledge might advocate for gambling. |

Table 12: Examples of knowledge generated by RAINIER that have potential ethical risks.
Task	Prompt
OBQA	Input: The sun is responsible for (A) puppies learning new tricks (B) children growing up and getting old (C) flowers wilting in a vase (D) plants sprouting, blooming and wilting
Knowledge: **Natural light provides energy for photosynthesis.**

Input: Poison causes harm to which of the following? (A) a Tree (B) a robot (C) a house (D) a car
Knowledge: **Living organisms are susceptible to poisonous matter.**

Input: As a car approaches you in the night (A) the headlights become more intense (B) the headlights recede into the dark (C) the headlights remain at a constant (D) the headlights turn off
Knowledge: **The intensity of light increases when observed from a shorter distance.**

Input: When the weather changes as it does from Christmas to Easter, (A) the air may chill (B) the ground may freeze (C) the plants may die (D) the ground may warm
Knowledge: **Christmas is in winter and Easter is in spring.**

Input: Using mirrors to focus collected light from heavenly bodies allows (A) detailed observation (B) foregone conclusions (C) radiation experiments (D) celestial music
Knowledge: **Telescopes use mirrors to focus light from the stars.**

Input: [question]
Knowledge:

Table 13: Prompt for OpenBookQA.

---

Task	Prompt
ARC	Input: George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat? (A) dry palms (B) wet palms (C) palms covered with oil (D) palms covered with lotion
Knowledge: **Rubbing hands produces heat because of friction.**

Input: Which of the following is an example of a physical change? (A) lighting a match (B) breaking a glass (C) burning of gasoline (D) rusting of iron
Knowledge: **Physical changes must not involve chemical changes such as combustion and rusting.**

Input: On Earth, water can be a solid, a liquid, or a gas. Which energy source has the greatest influence on the state of matter of water? (A) the sun (B) the wind (C) ocean currents (D) the metal core
Knowledge: **Earth’s water circulation is mostly driven by heat radiated from the sun.**

Input: What do cells break down to produce energy? (A) food (B) water (C) chlorophyll (D) carbon dioxide
Knowledge: **Food contain calories.**

Input: What characteristic of DNA results in cell differentiation in developing embryos? (A) which genes are present (B) how many copies of each gene are present (C) which genes are active (D) what protein is produced by a gene
Knowledge: **Cell differentiation is caused by selective expression of genes.**

Input: [question]
Knowledge:

Table 14: Prompt for ARC.
**AI2Sci**

**Prompt**
Input: Which is a nonrenewable natural resource that is used to make electrical energy? 
(A) coal (B) wind (C) water (D) thermal

Knowledge: **Fossil fuel is nonrenewable natural resource.**

Input: Which adaptation will warn predators not to eat an animal? 
(A) bright colors (B) bulging eyes (C) geometric shapes (D) poisonous secretions

Knowledge: **Bright colors in animals are usually a sign of being poisonous.**

Input: An Italian scientist named Alessandro Volta invented the Voltaic pile in 1800. It was able to produce a steady electrical current. Based on this description, what is the modern equivalent of the Voltaic pile? 
(A) a wire (B) a battery (C) a resistor (D) a light bulb

Knowledge: **Batteries can produce steady electrical current.**

Input: What is the best measure to use in determining the effect of solar energy on Earth’s atmosphere? 
(A) the temperature of the air (B) the temperature of the ocean (C) the density of clouds in the sky (D) the amount of rainfall on a rainy day

Knowledge: **Solar radiation converts to heat in Earth’s atmosphere.**

Input: Which nongaseous compound can be made from two elements that are gases at room temperature? 
(A) water (B) table salt (C) iron oxide (D) carbon dioxide

Knowledge: **Water molecules are made of Hydrogen and Oxygen.**

**CSQA**

**Prompt**
Input: Google Maps and other highway and street GPS services have replaced what? 
(A) united states (B) mexico (C) countryside (D) atlas (E) oceans

Knowledge: **Electronic maps are the modern version of paper atlas.**

Input: The fox walked from the city into the forest, what was it looking for? 
(A) pretty flowers. (B) hen house (C) natural habitat (D) storybook (E) dense forest

Knowledge: **Natural habitats are usually away from cities.**

Input: You can share files with someone if you have a connection to a what? 
(A) freeway (B) radio (C) wires (D) computer network (E) electrical circuit

Knowledge: **Files can be shared over the Internet.**

Input: Too many people want exotic snakes. The demand is driving what to carry them? 
(A) ditch (B) shop (C) north america (D) pet shops (E) outdoors

Knowledge: **Some people raise snakes as pets.**

Input: The body guard was good at his duties, he made the person who hired him what? 
(A) better job (B) irritated (C) feel safe (D) save money (E) headache

Knowledge: **The job of body guards is to ensure the safety and security of the employer.**
**Task**

**Prompt**

QASC

Input: What type of water formation is formed by clouds? (A) pearls (B) streams (C) shells (D) diamonds (E) rain (F) beads (G) cooled (H) liquid

Knowledge: **Clouds are made of water vapor.**

Input: What can prevent food spoilage? (A) prolactin release (B) one celled organisms (C) hydrating food (D) cleaning food (E) airing out food (F) Electric generators (G) a hydraulic system (H) dehydrating food

Knowledge: **Dehydrating food is used for preserving food.**

Input: The process by which genes are passed is (A) Most plants (B) flow of electrons (C) mitosis (D) Summer (E) respiration (F) mutation (G) mechanical (H) reproduction

Knowledge: **Genes are passed from parent to offspring.**

Input: The stomach does what in the body? (A) decreases its bodily water (B) kills all germs (C) breaks food into nutrients (D) stores bile (E) heat is produced (F) extracts water from food (G) get chemical reactions started (H) cause people to become sick.

Knowledge: **The stomach is part of the digestive system.**

Input: What can cause rocks to break down? (A) Wind Barriers (B) Protective Barriers (C) Stone Sealers (D) wind (E) mines (F) Water (G) erosion (H) Gravity

Knowledge: **Mechanical weathering is when rocks are broken down by mechanical means.**

Input: [question]

Knowledge:

---

Table 17: Prompt for QASC.

---

PIQA

Input: how do you flood a room? (A) fill it with objects. (B) fill it with water.

Knowledge: **Too much water can cause flooding.**

Input: How can I get oil stains out of my driveway? (A) Douse each stain with a couple cans of beer. (B) Douse each stain with a couple cans of soda.

Knowledge: **Sodium carbonate solution can wash away oil stains.**

Input: Soothe a painful sunburn. (A) Wait until brewed tea bag is cool, then apply on burn. (B) Wait until brewed tea bag is hot, then apply on burn.

Knowledge: **Sunburn can be alleviated by applying cold material.**

Input: What can I use for fuel in an alcohol stove? (A) Use acetone. (B) Use vinegar.

Knowledge: **Acetone is flammable, while vinegar is not.**

Input: How can I cut the handles of metal cutlery? (A) Use a hand saw to cut the handles. (B) Use a hand drill to cut the handles.

Knowledge: **A hand saw is used for making cuts; a hand drill is used for making holes.**

Input: [question]

Knowledge:

---

Table 18: Prompt for PhysicalIQA.

8957
## Task Prompt

### SIQA

**Input:** What will Quinn want to do next? (A) Eat messy snacks (B) help out a friend (C) Pick up the dirty clothes

Quinn wanted to help me clean my room up because it was so messy.

Knowledge: A messly room likely contains dirty clothes.

**Input:** What will Aubrey want to do next? (A) help Aubrey go back home (B) keep on partying without the mom (C) going on with the mom

Aubrey took Sasha’s mom to the hospital.

Knowledge: One should attend to their sick family member.

**Input:** How would Jan feel afterwards? (A) scared of losing the cat (B) normal (C) relieved for fixing the problem

Their cat kept trying to escape out of the window, so Jan placed an obstacle in the way.

Knowledge: One usually has positive emotions after solving a problem.

**Input:** How would Sydney feel afterwards? (A) affected (B) like they released their tension (C) worse

Sydney had so much pent up emotion, they burst into tears at work.

Knowledge: Crying can be a catharsis.

**Input:** What does Sydney need to do before this? (A) be bad at her job (B) do a good job (C) be lazy

Sydney got a raise and a new promotion.

Knowledge: Pay raise and promotion are usually results of good job performance.

**Input:** {question}

Knowledge:

### Table 19: Prompt for SocialIQA.

| Task | Prompt |
|------|--------|
| SIQA | Input: The GPS and map helped me navigate home. I got lost when the _ got turned off. (A) GPS (B) map

Knowledge: A GPS device is electronic, while a map is paper-based. |

Input: I picked up a bag of peanuts and raisins for a snack. I wanted a sweeter snack out so I ate the _ for now. (A) raisins (B) peanuts

Knowledge: Peanuts contain a lot of fat. Raisins contain a lot of sugar. |

Input: The geese prefer to nest in the fields rather than the forests because in the _ predators are more hidden. (A) fields (B) forests

Knowledge: There are more trees in the forests than in the fields. |

Input: Once in Poland, Dennis enjoyed the trip more than Jason because _ had a shallow understanding of the Polish language. (A) Dennis (B) Jason

Knowledge: Those who know the native language would enjoy the trip better. |

Input: Adam put handwash only clothes in the washer but Aaron washed them by hand as _ was lazy. (A) Adam (B) Aaron

Knowledge: Washing clothes with washer takes less effort than by hand. |

Input: {question}

Knowledge: