RAINIER: Reinforced Knowledge Introspector for Commonsense Question Answering

Jiacheng Liu[♡] Skyler Hallinan[♡] Ximing Lu[♡♠] Pengfei He[◇]
Sean Welleck[♡♠] Hannaneh Hajishirzi[♡♠] Yejin Choi[♡♠]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Allen Institute for Artificial Intelligence [◇]Petuum, Inc.
liujc@cs.washington.edu

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 RAINIER[1], 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 in-domain benchmarks that are seen during reinforcement learning, as well as 4 out-of-domain benchmarks 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 knowledge elicited from GPT-3 for commonsense QA.

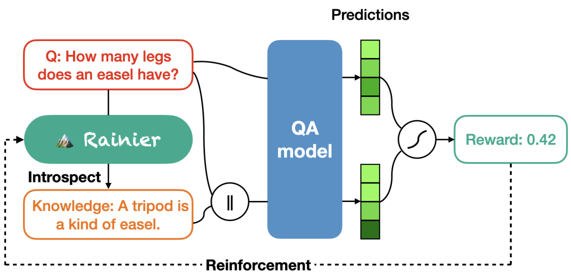

Figure 1: RAINIER can introspect for knowledge that underpin the reasoning of commonsense questions, and is trained via reinforcement learning where the reward is given by a fixed, generic QA model.

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., 2021; Liu et al., 2021), 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 given 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 domains without additional engineering effort (e.g. finetuning), (2) produces commonsense knowledge of 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 su-

[1]Code and model will be available at anonymized

pervision, because there are usually no gold knowledge labels on commonsense datasets. In order to ensure that RAINIER 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 on 8 (in-domain) commonsense QA datasets – encompassing general, scientific, physical, and social commonsense – so that the model can have better generalization to unseen domains (out-of-domain).

Experiments show that RAINIER substantially improves the performance of QA models on 9 commonsense benchmarks (5 in-domain and 4 out-of-domain), and gives larger and more consistent gains than a few-shot GPT-3 (Liu et al., 2021) 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 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.

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:

$q =$ `{question} (A) {choice_A} (B) {choice_B} ...`

Algorithm 1 Training RAINIER

Input initial policy model θ_0, initial value model ϕ_0, pretrained QA model ψ_{QA}

$\mathcal{D}_{\text{imit}} \leftarrow$ Get silver knowledge on $\mathcal{D}_{\text{ID}}$ from GPT-3.
$\theta_{\text{imit}} \leftarrow$ Optimize θ_0 with Eqn 2 from $\mathcal{D}_{\text{imit}}$. ▷ Section 2.1
$\theta_{\text{RAINIER}} \leftarrow$ REINFORCEDLEARNING($\mathcal{D}_{\text{ID}}, \theta_{\text{imit}}, \phi_0, \psi_{\text{QA}}$) ▷ Section 2.2

procedure REINFORCEDLEARNING($\mathcal{D}_{\text{ID}}, \theta, \phi, \psi_{\text{QA}}$)
  $\theta_{\text{old}} \leftarrow \theta, \phi_{\text{old}} \leftarrow \phi$
  for iterations = 1, 2, ... **do**
    for s minibatches from $\mathcal{D}_{\text{ID}}$ **do**
      Compute $\mathcal{L}_{\text{PPO}}$ on the minibatch with Eqn 3.
      Optimize θ and ϕ with $\mathcal{L}_{\text{PPO}}$ for one step.
    $\theta_{\text{old}} \leftarrow \theta, \phi_{\text{old}} \leftarrow \phi$
  return θ

Output θ_{RAINIER}

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 to train this model.

Training. Since we do not have gold knowledge to train RAINIER, we obtain this model by finetuning 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 some *in-domain* QA datasets, i.e. $\mathcal{D}_{\text{ID}} = \bigcup_{d=1}^{\Delta_{\text{ID}}} \mathcal{D}_d$, where $\mathcal{D}_d = \{(q_j, A_j, a_j^*)\}_{j=1}^{|\mathcal{D}_d|}$. The generic QA model we use may or may not have seen the in-domain datasets. The complete training process is outlined in Algorithm 1.

Inference. The effectiveness of RAINIER is evaluated against a set of *out-of-domain* QA datasets, $\mathcal{D}_{\text{OOD}}$, in addition to the in-domain datasets. Note that RAINIER is not trained on any OOD datasets, which means we neither get silver knowledge, nor do imitation learning or reinforcement learning on them. The generic QA model we use has not seen any OOD 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 dataset labeled with high-quality knowledge, but GPT-3 has been shown as a good generator for relevant knowledge (Liu et al., 2021). Therefore, we get silver knowledge from GPT-3 on our in-domain datasets. Following Liu et al. (2021), 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:

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

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

$$\mathcal{D}_{\text{imit}} = \Big\{(q, k) : (q, A, a^*) \in \mathcal{D}_{\text{ID}}, k \in K(q)\Big\}, \tag{1}$$

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

$$\mathcal{L}^{\text{train}}(\theta) \propto \sum_{(q,k)\in\mathcal{D}_{\text{imit}}^{\text{train}}} -\log p_K(k|q;\theta). \tag{2}$$

We denote the parameterization of the resulting model as θ_{imit}

2.2 Training Stage II: Reinforcement Learning

As we will see in the empirical results, the imitation model obtained in Stage I fails to provide 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, $p_K(k_t|q, k_{<t};\theta)$, 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 a KL penalty between the learned policy and the initial policy to the reward (Ouyang et al., 2022),

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

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

$$\begin{aligned} r_T &= R(x, k) \quad (\text{where } T = |k| \text{ and } k_T = \text{[EOS]}); \\ r_t &= 0 \quad (\text{where } 1 \le t < T). \end{aligned}$$

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 the policy model, PPO additionally uses a *value model* (parameterized by ϕ) 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,

$$\mathcal{L}_{\text{PPO}}(\theta, \phi) = \mathcal{L}_{\text{Policy}}(\theta) + \alpha \cdot \mathcal{L}_{\text{Value}}(\phi), \tag{3}$$

where $\mathcal{L}_{\text{Policy}}(\theta)$ is the loss on the policy model, $\mathcal{L}_{\text{Value}}(\phi)$ is the loss on the value model, and α is a hyperparameter.

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

$$\hat{A}_t = \sum_{t'=t}^{T-1} (\gamma\lambda)^{t'-t} \delta_{t'},$$
$$\text{where} \quad \delta_{t'} = r_{t'} + \gamma V(s_{t'+1};\phi) - V(s_{t'};\phi),$$

[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.

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,

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

where $\nu_t(\theta) = \frac{p_K(k_t|q;\theta)}{p_K(k_t|q;\theta_{\text{old}})}$ is the ratio between the current policy θ and a lagging policy θ_{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

$$\mathcal{L}_{\text{Policy}}(\theta) = -\hat{\mathbb{E}}\Big[\text{cso}(\hat{A}_t, \nu_t(\theta), \varepsilon)\Big]$$

where the expectation is taken over all instances in the training data ($x \sim \mathcal{D}_{\text{ID}}^{\text{train}}$), 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^{targ}, which in turn is estimated with a lagging value model ϕ_{old}:

$$\mathcal{L}_{\text{Value}}(\phi) = \hat{\mathbb{E}}\Big[\big(V(s_t;\phi) - V_t^{\text{targ}}\big)^2\Big],$$

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

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 $P_{\text{QA}}(a|q)$ to any candidate answer $a \in A$. Since we will use a sequence-to-sequence language model (i.e. UnifiedQA) as the QA model, we define

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

where

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

where $p_{\text{QA}}(a_i|q, a_{<i}; \psi_{\text{QA}})$ 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 $P_{\text{QA}}(a|q)$ (or equivalently, the highest $S_{\text{QA}}(a|q)$): $\hat{a} = \arg\max_{a \in A} P_{\text{QA}}(a|q)$.

Our goal is to maximize $P_{\text{QA}}(a^*|q \circ k)$, 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

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

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

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

Intuitively, this function would give a reward of near $+1$ if the naive prediction is incorrect (i.e. $S_{\text{QA}}(a^*|q) < \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'|q)$), while the knowledge-prompted prediction is correct (i.e. $S_{\text{QA}}(a^*|q \circ k) > \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'|q \circ k)$). 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 6.

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 generate a knowledge statement for each training instance, and estimate the population mean and standard deviation of rewards:

$$\mathcal{R}_{\text{init}} = \big\{r(x,k) : x \in \mathcal{D}_{\text{ID}}^{\text{train}}, k \sim p_K(\cdot|q;\theta_{\text{imit}})\big\}, \\ \mu_0 = \mu(\mathcal{R}_{\text{init}}), \sigma_0 = \sigma(\mathcal{R}_{\text{init}}). \quad (4)$$

In Stage II training, each reward is transformed:

$$r(x,k) \leftarrow \frac{r(x,k) - \mu_0}{\sigma_0}. \quad (5)$$

2.3 Inference: Knowledge Prompting and Aggregation

Following Liu et al. (2021), 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),

$$K(q) = \{\varepsilon\} \cup \big\{k_m : k_m \sim p_K^{p=0.5}(k \mid q; \theta)\big\}.$$

We collect a set of outputs for prompting with each knowledge. The final prediction is the candidate answer that receives maximum confidence,

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

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

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

Training time model selection. During Stage II training, we only generate one knowledge per question during validation.[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 Experiments

In-domain datasets. For both imitation learning and reinforcement learning, we use the multiple-choice datasets that UnifiedQA$_{\text{v2}}$ (Khashabi et al., 2022) uses for training (i.e. $\mathcal{D}_{\text{ID}}$): OpenBookQA, ARC, AI2Science, CommonsenseQA, QASC, PhysicalIQA, SocialIQA, and Winogrande.[4] See §A.1 for complete references on these datasets.

Out-of-domain datasets. We additionally evaluate on the following multiple-choice QA datasets (i.e. $\mathcal{D}_{\text{OOD}}$): NumerSense, RiddleSense, QuaRTz, and HellaSwag.

Models. For Stage I training, we get silver knowledge from the curie version of GPT-3 (13B) (Brown et al., 2020). The knowledge introspector is initialized with T5-large (Raffel et al., 2019). 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} \n {k}`. We use the same question formatting as UnifiedQA. See Table 8 for hyperparameters.

Baselines. We report performance improvements over direct inference with the same UnifiedQA-large model (i.e. without prompting RAINIER's knowledge). We also compare with knowledge from few-shot GPT-3 (Liu et al., 2021), where we generate knowledge with GPT-3 (curie, 13B) and 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 $p = 0.5$).

We could not directly compare with Self-talk (Shwartz et al., 2020) and DREAM (Gu et al., 2021) due to discrepancy in the base QA models, and we do not compare with chain-of-thought prompting (Wei et al., 2022) because it does not work well with smaller models. (See §A.3 for details.)

4 Results

4.1 Main Results

In-domain performance. Table 1 shows the performance of RAINIER-enhanced QA model on the in-domain 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%). There is no performance gain on OpenBookQA, ARC, and AI2Science. This is because the QA model, UnifiedQA, is already trained on these three datasets, thus setting a strong baseline.

Comparison with few-shot GPT-3. The few-shot GPT-3 knowledge generator is only able to provide small improvement on CommonsenseQA, QASC, and Winogrande, while hurting the performance on other benchmarks. This has been shown in Liu et al. (2021), as the 13B version of GPT-3 serves as a weaker knowledge generator. In contrast, RAINIER can provide larger and more consistent improvements, while being more than one

[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.

[4]We exclude MCTest and RACE because most questions in these reading comprehension datasets are too long to fit into our model's input.

Dataset → **Method ↓**	OBQA	ARC easy	ARC hard	AI2Science elem	AI2Science mid	CSQA	QASC	PIQA	SIQA	WG	**Avg.** (last 5)
UQA	70.20	69.12	55.85	69.11	64.80	61.43	43.09	63.66	53.84	53.35	55.07
RAINIER (T5-large) + UQA	69.60	67.72	55.18	68.29	63.20	**67.24**	**54.97**	**65.67**	**57.01**	**56.91**	**60.36**
Few-shot GPT-3 (13B) + UQA	68.80	71.05	56.52	70.73	65.60	66.34	53.24	64.25	58.29	55.56	59.54
Self-talk GPT-3 (13B) + UQA	–	–	–	–	–	63.31	49.89	65.23	51.89	52.96	56.66
DREAM (11B) + UQA	–	–	–	–	–	64.54	49.46	64.74	51.59	56.12	57.29

Table 1: Results on in-domain datasets. Accuracy on dev set is reported. Predictions made by UQA-large.

Dataset → **Method ↓**	CSQA	QASC	PIQA	SIQA	WG	**Avg.**
UQA	53.00	45.65	65.50	57.21	54.67	55.21
RAINIER (T5-large) + UQA	**60.18**	**54.13**	**67.09**	**59.01**	**57.39**	**60.00**

Table 2: Results on in-domain datasets. Accuracy on test set is reported. Predictions made by UQA-large.

Dataset → **Method ↓**	NS	RS	QuaRTz	HS	**Avg.**
UQA	26.50	28.11	68.75	35.00	39.59
RAINIER (T5-large) + UQA	**30.00**	**30.07**	**70.31**	**35.73**	**41.53**
Few-shot GPT-3 (13B) + UQA	38.00	35.65	69.01	37.33	45.00

Table 3: Results on out-of-domain datasets. Accuracy on dev set is reported. Predictions made by UnifiedQA-large.

QA Model → **Knowledge Gen. ↓**	UQA-small ID	UQA-small OOD	UQA-base ID	UQA-base OOD	UQA-large ID	UQA-large OOD	UQA-3b ID	UQA-3b OOD	Unicorn-large ID	Unicorn-large OOD
None	39.07	25.51	45.51	31.19	55.07	39.59	66.51	43.72	50.80	39.94
RAINIER (T5-large)	**48.60**	**27.99**	**54.77**	**37.21**	**60.36**	**41.53**	**67.85**	**45.45**	**59.86**	**41.90**
Δ	+9.53	+2.48	+9.26	+6.02	+5.29	+1.94	+1.34	+1.73	+9.06	+1.96
Few-shot GPT-3 (13B)	46.88	28.68	52.80	43.17	59.54	45.00	67.42	52.23	–	–

Table 4: Effectiveness of generated knowledge on different QA models. Average accuracy on dev set is reported.

order of magnitude smaller in size (770M).

Out-of-domain performance. Table 3 shows that RAINIER's knowledge substantially improves performance on the four out-of-domain benchmarks, 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, and show the results in Table 4. 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 in-domain benchmarks, and UnifiedQA-base for out-of-domain benchmarks).

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 5. 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 smaller improvements than the fully trained RAINIER, stressing the importance of Stage II training as well.

Reward function. Table 6 shows the results for knowledge introspectors trained with different reward functions. Our reward shaping gives the best out-of-domain performance as well as one of the top in-domain performance. The naive *prob diff* reward function gives slightly better in-domain performance, but our reward shaping results in better out-of-domain generalization.

4.3 Analysis

To get a deeper understanding of the behavior and capability of RAINIER, we manually analyzed the

QA Model → Knowledge Gen. ↓	UQA-large ID	OOD
None	55.07	39.59
RAINIER (T5-large)	**60.36**	**41.53**
– Stage I	53.68	36.83
– Stage II	57.00	40.70
– Stage I – Stage II	53.29	36.72

Table 5: Ablations on the importance of both training stages.

QA Model → Reward Func. ↓	Definition: $r(x, k) = \dots$	UQA-large ID	OOD
RAINIER's	$\frac{1}{2}\Big[\tanh\big(S_{\text{QA}}(a^*\|q \circ k) - \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'\|q \circ k)\big) - \tanh\big(S_{\text{QA}}(a^*\|q) - \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'\|q)\big)\Big]$	60.36	**41.53**
Prob only	$P_{\text{QA}}(a^*\|q \circ k)$	59.11	40.61
Prob diff	$P_{\text{QA}}(a^*\|q \circ k) - P_{\text{QA}}(a^*\|q)$	**60.69**	40.91
Score diff	$S_{\text{QA}}(a^*\|q \circ k) - S_{\text{QA}}(a^*\|q)$	58.26	39.86
Hard activation	$\frac{1}{2}\Big[\text{sgn}\big(S_{\text{QA}}(a^*\|q \circ k) - \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'\|q \circ k)\big) - \text{sgn}\big(S_{\text{QA}}(a^*\|q) - \max_{a' \in A, a' \neq a^*} S_{\text{QA}}(a'\|q)\big)\Big]$	58.32	41.16

Table 6: Ablations on the choice of reward function.

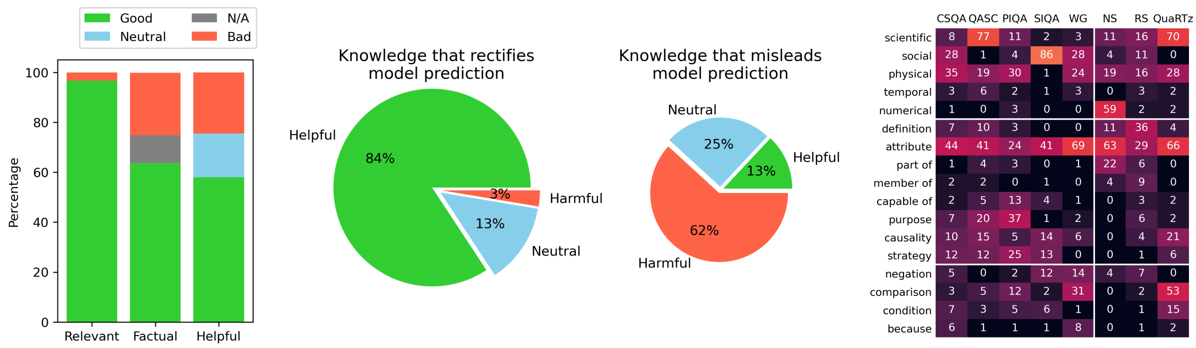

Figure 2: Human analysis of RAINIER-generated knowledge. **Left:** Percentage of good knowledge in each quality aspect. **Mid:** Agreement between human and machine on helpfulness of *selected knowledge*. **Right:** Percentage of RAINIER-generated knowledge categorized by domain, expressed relation, and syntax. The percentages do not add up to 100% because some knowledge have none of these characteristics, while some others may have multiple.

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 in-domain, 3 out-of-domain; see Figure 2). 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. (2021) 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 harmful. These results have similar trends as (Liu et al., 2021), 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 *domain*(s) it belongs to (e.g. scientific, social), the *relation*(s) it expresses (e.g. attribute, capable of), and its *syntactic* property(s) (e.g. negation, comparison). See Figure 2 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.

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 **Vinyl is a type of plastic.**	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 **Reptiles are the ancestors of all mammals.** **Lavender is a deterrent to house flies.**	scientific temporal scientific	attribute attribute	– –
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 **A good deed will be rewarded.**	social	–	–
WG	Adam always spent all of the free time watching Tv unlike Hunter who volunteered, due to _ being lazy. (A) Adam (B) Hunter **Hunter is more active than Adam.**	social	attribute	comparison
RS	Causes bad breath and frightens blood-suckers (A) tuna (B) iron (C) trash (D) garlic (E) pubs **Garlic is a strong-smelling food.**	–	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 **The mass of an object is proportional to the amount of matter it contains.**	scientific physical	–	–

Table 7: 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.

4.4 Qualitative Examples

We show some examples of good knowledge generated by RAINIER in Table 7.

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. (2021) 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. Shwartz et al. (2020) and Paranjape et al. (2021) prompt pretrained models with predefined 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, Liu et al. (2021) uses few-shot demonstrations to elicit flexible, relevant knowledge statements from a language model, whereas Wei et al. (2022) elicits full chain-of-reasoning from language models with in-context learning. These two methods rely on 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 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 tasks, 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; the effect of these knowledge on the QA model also aligns 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.

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 Rocktäschel, 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 LeBras, 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, EAAI 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79, Online. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems. *arXiv preprint arXiv:2204.13074*.

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

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2021. Dream: Uncovering mental models behind language models. *arXiv preprint arXiv:2112.08656*.

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, Faeze 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. UNIFIEDQA: 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, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515, Online. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Ximing Lu, Sean Welleck, Liwei Jiang, Jack Hessel, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: Controllable text generation with reinforced unlearning. *ArXiv*, abs/2205.13636.

Hugo Mercier and Dan Sperber. 2017. The enigma of reason.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

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 More Experimental Details

A.1 Datasets

The in-domain datasets we use are: 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). The out-of-domain datasets we use are: NumerSense (Lin et al., 2020), RiddleSense (Lin et al., 2021), QuaRTz (Tafjord et al., 2019), and HellaSwag (Zellers et al., 2019).

A.2 Hyperparameters

See Table 8.

A.3 Baselines

We set baselines for (1) direct inference with the QA model, and (2) prompting knowledge elicited from few-shot GPT-3 (Liu et al., 2021).

We could not directly compare with Self-talk (Shwartz et al., 2020) and DREAM (Gu et al., 2021) due to discrepancy in the base QA models. Specifically, Self-talk mainly uses GPT2-XL (1.6B) as inference model, and DREAM uses Macaw which has 11B parameters. Both are larger than UnifiedQA-large. Adapting these knowledge generators to our evaluation will require a lot of task-specific engineering work.

We do not compare with chain-of-thought prompting (Wei et al., 2022), because this method does not work well on commonsense tasks with smaller models.

B More 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.

Symbol	Value	Description
		GETTING SILVER KNOWLEDGE FROM FEW-SHOT GPT-3
M	20	Number of knowledge statements to sample from GPT-3, per question.
p	0.5	Parameter for nucleus sampling from GPT-3.
L_{output}	64	Max length of output from GPT-3.
		STAGE I: IMITATION LEARNING
L_{input}	256	Max length of input to RAINIER (i.e. question plus choices).
L_{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.
η	1×10^{-5}	Learning rate of Adam optimizer.
		STAGE II: REINFORCEMENT LEARNING
α	1.0	Weight of value model loss in PPO.
β	0.2	Weight of entropy bonus term in reward.
γ	1.0	Discount factor for rewards.
λ	0.95	Parameter for advantage estimation.
ε	0.2	Clipping range for the *clipped surrogate objective*.
L_{input}	256	Max length of input to RAINIER (i.e. question plus choices).
L_{output}	32	Max length of output from RAINIER (i.e. generated knowledge).
τ	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).
η	2×10^{-5}	Learning rate of Adam optimizer (with a linear learning rate decay schedule).
		INFERENCE
M	10	Number of knowledge statements to sample from RAINIER, per question.
p	0.5	Parameter for nucleus sampling from RAINIER.
L_{input}	256	Max length of input to RAINIER (i.e. question plus choices).
L_{output}	32	Max length of output from RAINIER.

Table 8: Hyperparameter settings.

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 **The fruit of the year is the date.**	**Ambiguous** The knowledge does not specify which 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 **The heavy cart will accelerate slower than the light one.**	**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 **Sleeping animals have a head at night.**	**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 against Kyle, _ won the race. (A) William (B) Kyle **William was not able to run the race.**	**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 **The end of every race is the end of every race.**	**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 **Valence electrons are more prone to being removed from the atom.**	**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 **Breaking out of a bad habit is usually a bad idea.**	**Social value** This knowledge is a generally true statement, so we labeled it as factual.
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 **One should not be scared of 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 **A friend can be used by a friend.**	**Social value** It is ambiguous whether *can* means *it is possible that* ..., or *ought to*. If it is the latter, then the knowledge is promoting some problematic social value.
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 **One should visit their sick friend.**	**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	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 **One should apologize when they hurt someone.**	**Social value** This knowledge is generally accepted. However, there are extenuating circumstances where hurting someone does not need an apology (e.g. hurting a violent criminal to protect oneself).
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 **One should be willing to help others.**	**Social value** This knowledge is generally accepted, but it is not a good fit to the question's context. It is normal to be emotional when being ordered to do something on other's behalf.
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 **One should help others in need.**	**Social value** This knowledge is generally accepted, and appropriate to the question's context.

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? (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.**	**Culture-specific** 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 then 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 gointo 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 \n (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? \n (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 \n (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, \n (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 \n (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? \n (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? \n (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? \n (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? \n (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? \n (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.

Task	Prompt
AI2Sci	Input: Which is a nonrenewable natural resource that is used to make electrical energy? \n (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? \n (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? \n (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? \n (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? \n (A) water (B) table salt (C) iron oxide (D) carbon dioxide Knowledge: **Water molecules are made of Hydrogen and Oxygen.** Input: {question} Knowledge:

Table 15: Prompt for AI2Science.

Task	Prompt
CSQA	Input: Google Maps and other highway and street GPS services have replaced what? \n (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? \n (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? \n (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? \n (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? \n (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.** Input: {question} Knowledge:

Table 16: Prompt for CommonsenseQA.

Task	Prompt
QASC	Input: What type of water formation is formed by clouds? \n (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? \n (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 \n (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? \n (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? \n (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.

Task	Prompt
PIQA	Input: how do you flood a room? \n (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? \n (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. \n (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? \n (A) Use acetone. (B) Use vinegar. Knowledge: **Acetone is flammable, while vinegar is not.** Input: How can I cut the handles of metal cutlery? \n (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.

Task	Prompt
SIQA	Input: What will Quinn want to do next? \n (A) Eat messy snacks (B) help out a friend (C) Pick up the dirty clothes \n Quinn wanted to help me clean my room up because it was so messy. Knowledge: **A messy room likely contains dirty clothes.** Input: What will Aubrey want to do next? \n (A) help Aubrey go back home (B) keep on partying without the mom (C) going on with the mom \n Sasha's mom passed out in the middle of the party. Aubrey took Sasha's mom to the hospital. Knowledge: **One should attend to their sick family member.** Input: How would Jan feel afterwards? \n (A) scared of losing the cat (B) normal (C) relieved for fixing the problem \n 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? \n (A) affected (B) like they released their tension (C) worse \n 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? \n (A) be bad at her job (B) do a good job (C) be lazy \n 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
WG	Input: The GPS and map helped me navigate home. I got lost when the _ got turned off. \n (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. \n (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. \n (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. \n (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. \n (A) Adam (B) Aaron Knowledge: **Washing clothes with washer takes less effort than by hand.** Input: {question} Knowledge:

Table 20: Prompt for Winogrande.