Enhancing Question Generation through Diversity-Seeking Reinforcement Learning with Bilevel Policy Decomposition: Technical Appendix

Tianyu Ren, Hui Wang*, Karen Rafferty

School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom
{tren01, h.wang, k.rafferty}@qub.ac.uk

Additional Details on BPD-DSRL

MDP of the Bilevel Policy for QG

Recall that we divide a typical QG policy $\pi_\theta(q|x,a)$ into a rationale policy $\pi_\theta(d|x,a)$ and an action policy $\pi_\theta(q|x,a,d)$ to accelerate policy learning. We here formalize the bilevel policy using an MDP to help readers better understand BPD. Initially, the agent observes a (x,a) pair and begins the autoregressive sampling of rationale tokens d_t ($1 \leq t \leq T_{\text{max}}$) from its vocabulary $\mathcal{V}$ governed by the rationale policy $\pi_\theta(d_t|d_{<t},x,a)$. Upon termination ($t$ exceeds T_{max} or the end-of-rationale token is sampled), the process reward model evaluates $d=(d_1,\dots,d_T)$ and assigns a singular reward r_d to it. Then, the action policy $\pi_\theta(q_s|q_{<s},d,x,a)$ starts to guide the agent to sample question tokens $q_s \in \mathcal{V}$ (for $1 \leq s \leq S_{\text{max}}$). The MDP concludes when the question sampling terminates, at which point the outcome reward model assigns a reward r_q to $q=(q_1,\dots,q_S)$.

Pseudo-code of BPD-DSRL training

We provide the general training process of our proposed BPD-DSRL for QG in Algorithm 1.

Human Evaluation

We follow similar procedures in (Xia et al. 2023; Gou et al. 2023) to conduct human evaluations. We enlist three annotators to critically evaluate the performance of the RLQG approaches alongside two supervised baselines using a set of 50 test samples from SQuAD 1.1 / 2. Our annotators focus on two primary criteria: the QG quality and the QG diversity.

QG Quality. For each test sample, each QG model produces five questions. The question with the highest predictive confidence by each model is selected for evaluation. In every evaluation round, annotators are provided with the reading passage, the associated answer and all model-generated questions. They need to determine whether the supplied answer can sufficiently address the questions posed and assign a score ranging from 1 (poor) to 5 (excellent) to reflect the quality of each question (see Fig 1 for details).

Algorithm 1: Training of BPD-DSRL

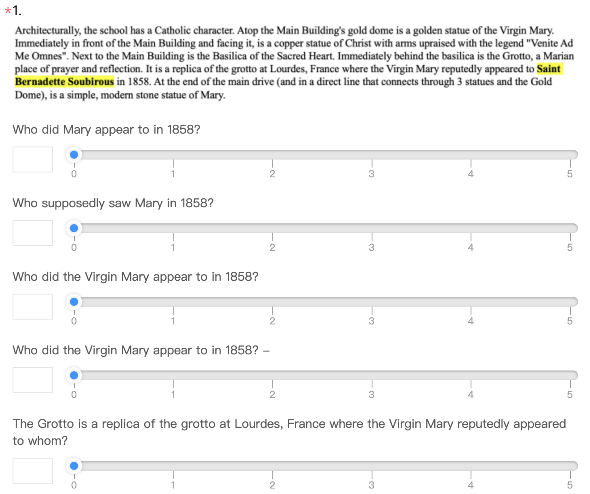
```
Input: Training dataset $D = \{(x_n, a_n, d_n, q_n)\}_{n=1}^N$, SFT bilevel policy $\pi_\theta(q, d|x, a)$, reward models $r_d(x, a, d)$ and $r_q(x, a, q)$, partition function estimator $Z_\mu(x, a)$.
Parameter: Trajectory batch size $M$, Time horizon $\mathcal{T}$.
Output: RL refined bilevel policy .
Let $t = 0$.
Initialize the replay buffer $\mathcal{B}$ by iterating $D$ and calculating the rewards of $q_n$ and $d_n$.
while $t < \mathcal{T}$ do
    Sample one $(x, a)$ pair from $D$.
    Randomly pick one behavior policy $\pi$ from $\pi_\theta$, tempered $\pi_\theta$ ($\pi_\theta^{\text{TEM}}$) and $\mathcal{B}$.
    if $\pi \in \{\pi_\theta, \pi_\theta^{\text{TEM}}\}$ then
        Sample trajectories $\{(d_m, q_m)\}_{m=1}^M$ from $\pi$.
        Calculate the reward score of $\{(d_m, q_m)\}_{m=1}^M$.
        Add $\{(d_m, q_m, r_d(x, a, d_m), r_q(x, a, q_m)\}_{m=1}^M$ to $\mathcal{B}$, with $(x, a)$ being the key.
    else
        Sample $M$ trajectories together with their rewards from $\pi$ based on the key $(x, a)$.
    end if
    Update $\theta$ with gradient $\mathbb{E}_{(d,q)\sim\pi}[\nabla_\theta \mathcal{L}_{\text{RL}}(d, q)|x, a]$.
    Update $\mu$ with gradient $\mathbb{E}_{(d,q)\sim\pi}[\nabla_\mu \mathcal{L}_{\text{RL}}(d, q)|x, a]$.
    $s = s + 1$.
end while
return $\pi_\theta^{\text{RL}}$.
```

QG Diversity. In this phase of evaluation, the annotators examine all five questions generated by each model for each test sample. The focus here is on determining the uniqueness of each question, i.e., how varied the questions are in terms of their underlying meanings. The annotators assign a diversity score from 1 (least diverse, high repetitiveness) to 5 (most diverse, no repetitiveness), indicating the number of representative questions (see Fig 2 for details).

The human evaluation results are reported in Fig. 3 based on the majority vote. These results align closely with previous automated metrics, revealing two key insights: (1) All RLQG methods consistently outperform the SFT baselines in generating high-quality questions, with our method achieving the most superior performance. (2) Although RL

*Corresponding author

typically compromises the diversity, our method maintains the highest level of diversity among all RLQG methods.



*1.

Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to **Saint Bernadette Soubirous** in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Who did Mary appear to in 1858?

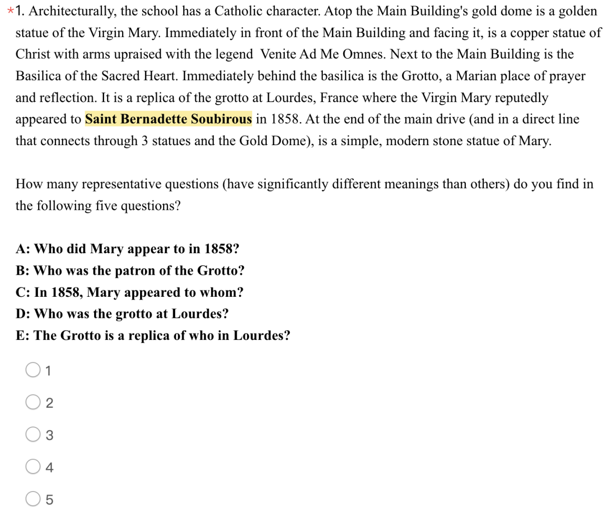
Who supposedly saw Mary in 1858?

Who did the Virgin Mary appear to in 1858?

Who did the Virgin Mary appear to in 1858? –

The Grotto is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to whom?

Figure 1: Questionnaire for question quality evaluation.



*1. Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend Venite Ad Me Omnes. Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to **Saint Bernadette Soubirous** in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

How many representative questions (have significantly different meanings than others) do you find in the following five questions?

A: Who did Mary appear to in 1858?
B: Who was the patron of the Grotto?
C: In 1858, Mary appeared to whom?
D: Who was the grotto at Lourdes?
E: The Grotto is a replica of who in Lourdes?

○ 1
○ 2
○ 3
○ 4
○ 5

Figure 2: Questionnaire for question quality evaluation.

Additional Details on Experiments

We conduct all of our experiments using PyTorch 2.1.0 on Ubuntu 22.04. More implementation details are as follows.

SFT Phrase of RLQG Methods

All RLQG methods (BPD-DSRL, PPO and REINFORCE) are fine-tuned using the same reward models from the same bilevel policy initialized by SFT.

Before the bilevel policy warm-up, we enrich the standard QG dataset with answer-centric summaries generated by an off-the-shelf question conversion model $f(q, a)$[1]. All processed QG datasets utilized in our experiments are included in the accompanying "Code & Data" zip file. Post data preprocessing, we train the bilevel policy using T5-large model[2] via SFT. We keep the checkpoint that achieves the best performance on the validation dataset. Detailed hyperparameter settings are presented in table 1.

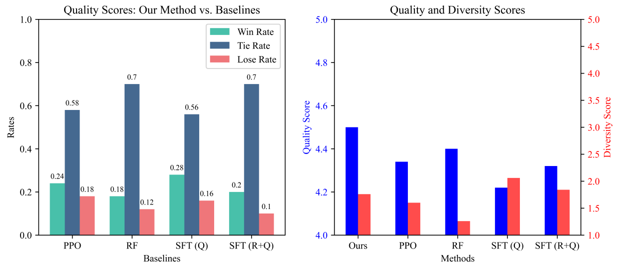


Figure 3: Human evaluations on consistency and diversity.

Similarly, we learn the outcome reward model $r_q(x, a, q)$ from T5-large by adopting the multi-task SFT objective outlined in the main paper. The checkpoint that has the best performance on the validation dataset is used as the final outcome reward model. Specific hyperparameter configurations are delineated in table 2.

RL Phrase of RLQG Methods

Training BPD-DSRL. To learn BPD-DSRL, we further fine-tune the supervised bilevel policy using DSRL with the reward models in a bandit environment. We here provide more details about the implementation.

- **Training dataset.** We sample a subset from the SFT QG dataset to present (x, a) pairs to the bilevel policy.
- **Process reward model.** We use the NLI model[3] from (Nie et al. 2020) as our process reward model.
- **Scheduler of reward temperature and learning rate.** We follow (Hu et al. 2024) to use two schedulers of reward temperature and learning rate to smooth training. The reward temperature scheduler is utilized to linearly decrease the reward temperature across the training horizon. Whereas the learning rate scheduler progressively increases the learning rate during the warm-up phase.
- **Behavior policy.** Recall that the trajectories used to estimate policy gradient come from three sources: (1) the current policy, (2) the tempered version of the current policy, and (3) a replay buffer. At each time step, we randomly pick one behavior policy from the above three sources with equal probability and sample (d, q). The tempered policy samples sequences using a random temperature in a pre-defined range and the replay buffer utilizes the reward-prioritized sampling to replay the high-reward experiences (Shen et al. 2023).

The hyperparameter settings for training BPD-DSRL in presented in table 3.

[1] https://huggingface.co/domenicrosati/question_converter-3b

[2] https://huggingface.co/google-t5/t5-large

[3] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

Training the REINFORCE baseline. The REINFORCE algorithm (Williams 1992) is widely used in previous RLQG work. In the BPD setting, the REINFORCE algorithm aims to maximize the following objective:

$$\mathcal{L} = \mathbb{E}_{(d,q)\sim\pi}[r_d(x,a,d)r_q(x,a,q)\log\pi_\theta(d,q|x,a)|x,a] \tag{1}$$

We use the same settings as BPD-DSRL to train the REINFORCE baseline, i.e., the only difference between BPD-DSRL training and the REINFORCE baseline training is the learning objective.

Training the PPO baseline. The PPO algorithm has demonstrated its superior performance in many natural language generation tasks. Formally, the PPO algorithm aims to maximize the following objective:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{(d,q)\sim\pi}[r_d(x,a,d)r_q(x,a,q) - \beta(\log\pi_\theta^{\text{RL}}(d,q|x,a) \\ - \log\pi_\theta^{\text{SFT}}(d,q|x,a))|x,a]\end{aligned} \tag{2}$$

where β is the KL coefficient which controls the the strength of the KL penalty.

We use the trl[4] library to implement the PPO baseline. The hyperparameter settings are provided in table 4.

Inference Settings of RLQG Methods

We use consistent inference settings for all RLQG methods, which are provided in table5.

LLM Baselines

We employ two open-sourced LLMs, LLaMA-8B-Instruct[5] and Mistral-7B-Instruct-V0.3[6], as our LLM baselines for QG. During inference, we use a relatively low temperature of 0.65 to make sure it can well follow our instructions. The prompt and demonstrations are shown in table 6.

Answerability Evaluation Using GPT-3.5

To balance the evaluation precision and the cost, we use GPT-3.5-Turbo-0125 as the QA model for measuring the answerability of generated questions. We use the zero-shot configuration with a temperature of 0.65 for evaluation. The relevant prompts can be found in table 7.

PLM	T5-Large
Epoch	5
Learning Rate	5e-5
Optimizer	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Warmup Ratio	0.1
Training Batch Size Per Device	2
Gradient Accumulation Steps	16
Evaluation Intervals	400
Evaluation Metric	Perplexity
Random Seed	1234
Max Input Length	384
Devices	2 × RTX 4090

Table 1: Hyperparameters for bilevel policy warm-up using supervised fine-tuning.

PLM	T5-Large
Epoch	5
Learning Rate	5e-5
Optimizer	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Warmup Ratio	0.1
Loss Coefficient	0.5
Training Batch Size Per Device	2
Gradient Accumulation Steps	16
Evaluation Intervals	400
Evaluation Metric	Perplexity
Random Seed	1234
Max Input Length	384
Devices	2 × RTX 4090

Table 2: Hyperparameters for training the outcome reward model $r_q(x,a,q)$.

References

Gou, Q.; Xia, Z.; Yu, B.; Yu, H.; Huang, F.; Li, Y.; and Cam-Tu, N. 2023. Diversify Question Generation with Retrieval-Augmented Style Transfer. In *EMNLP 2023*.

Hu, E. J.; Jain, M.; Elmoznino, E.; Kaddar, Y.; Lajoie, G.; Bengio, Y.; and Malkin, N. 2024. Amortizing intractable inference in large language models. In *ICLR 2024*.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL 2020*.

Shen, M. W.; Bengio, E.; Hajiramezanali, E.; Loukas, A.; Cho, K.; and Biancalani, T. 2023. Towards understanding and improving GFlowNet training. In *ICML 2023*.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.*, 8(3–4).

Xia, Z.; Gou, Q.; Yu, B.; Yu, H.; Huang, F.; Li, Y.; and Cam-Tu, N. 2023. Improving Question Generation with Multi-level Content Planning. In *EMNLP 2023 Findings*.

[4] https://pypi.org/project/trl

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[6] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Number of Training examples	800
Time Horizon	100
Training Batch Size Per Device	2
Gradient Accumulation Steps	4
Policy Learning Rate	1e-5
Estimator Learning Rate	1e-4
Optimizer	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Warmup Ratio	0.1
Maximum Input Length	384
Rationale Sequence Length Range	[5, 32]
Question Sequence Length Range	[5, 32]
Rationale Policy Temperature Range	[0.5, 1]
Action Policy Temperature Range	[0.5, 1]
Reward Temperature Range	[0.5, 1]
Buffer Size	4
Random Seed	0
Devices	1 × L20

Table 3: Hyperparameters for training BPD-DSRL and the REINFORCE baseline.

Number of Training examples	800
Time Horizon	100
Training Batch Size Per Device	8
Gradient Accumulation Steps	8
Policy Learning Rate	1e-4
Target KL Divergence	0.1
Initial KL Coefficient	0.2
Adaptive KL controller	True
Optimizer	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Maximum Input Length	384
Rationale Sequence Length Range	[5, 32]
Question Sequence Length Range	[5, 32]
Random Seed	0
Devices	1 × L20

Table 4: Hyperparameters for training the PPO baseline.

Rationale Policy Temperature	0.75
Action Policy Temperature	0.75
Action Policy Top-p	0.95
Random Seed	0
Rationale Sequence Length Range	[5, 32]
Question Sequence Length Range	[5, 32]
Devices	1 × L20

Table 5: Inference settings for all QG models during our experiments.

Instruction:
You are a helpful teacher who will ask questions (question only) from the reading passage and the given answer.

User:
Reading Passage: It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.

Answer: Saint Bernadette Soubirous

Question: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?

Reading Passage: Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend ”Venite Ad Me Omnes”.

Answer: a copper statue of Christ

Question: What is in front of the Notre Dame Main Building?

Reading Passage: Next to the Notre Dame Main Building is the Basilica of the Sacred Heart.

Answer: the Main Building

Question: The Basilica of the Sacred heart at Notre Dame is beside to which structure?

Reading Passage: {}

Answer: {}

Question: {}

Table 6: Prompt and demonstrations for LLaMA 3-8B-Instruct and Mistral-7B-Instruct-V0.3 during QG.

User:
Given a reading passage and a question, you need to judge whether this question is answerable or not. If the question is answerable, you need to respond with the answer to the question. Otherwise, simply reply me with NO.
The context is: {}
The question is: {}

Table 7: Prompt for GPT-3.5-Turbo-0125 for answerability evaluation.