MDPO: Customized Direct Preference Optimization with a Metric-based Sampler for Question and Answer Generation

Yihang Wang[1*], **Bowen Tian**[2*], **Yueyang Su**[3,4†], **Yixing Fan**[3,4†], **Jiafeng Guo**[3,4]

[1]Beijing University of Posts and Telecommunications, Beijing
[2]Hong Kong University of Science and Technology (Guangzhou), Guangzhou
[3]CAS Key Lab of Network Data Science and Technology, ICT, CAS, Beijing
[4]University of Chinese Academy of Sciences, Beijing

yihangwang1020@gmail.com;bowentian@hkust-gz.edu.cn;{suyueyang,fanyixing,guojiafeng}@ict.ac.cn

Abstract

With the extensive use of large language models, automatically generating QA datasets for domain-specific fine-tuning has become crucial. However, considering the multifaceted demands for readability, diversity, and comprehensiveness of QA data, current methodologies fall short in producing high-quality QA datasets. Moreover, the dependence of existing evaluation metrics on ground truth labels further exacerbates the challenges associated with the selection of QA data. In this paper, we introduce a novel method for QA data generation, denoted as MDPO. We proposes a set of unsupervised evaluation metrics for QA data, enabling multidimensional assessment based on the relationships among context,question and answer. Furthermore, leveraging these metrics, we implement a customized direct preference optimization process that guides large language models to produce high-quality and domain-specific QA pairs. Empirical results on public datasets indicate that MDPO's performance substantially surpasses that of state-of-the-art methods.

1 Introduction

With the advancement of large language models (LLMs), QA-based services have increasingly been adopted across various sectors, including finance (Passali et al., 2021), education (Yuhao Dan, 2024), and healthcare (Singhal et al., 2023). To improve the performance of general-purpose large models in specific domains, it is common to infuse domain-specific knowledge through Supervised Fine-Tuning (SFT) techniques (Mecklenburg et al., 2024), which necessitates the use of large-scale labeled datasets. However, the acquisition of large-scale labeled datasets through manual annotation is laborious and time-consuming. Consequently, the study of automatic generation methods for QA pairs, known as Question Generation (QG), has emerged as a topic of significant interest in both the academic and industrial communities.

Existing approaches for QG can be broadly classified into two categories: rule-based and model-based methods. Rule-based methods employ expert-designed templates to formulate questions with content extracted from contexts. Notable techniques within this category include those utilizing dependency syntax trees, implementing knowledge graphs, and applying semantic role labeling to systematically generate questions. However, these methods are constrained by the prior knowledge of domain experts, resulting in the low quality of QA pairs. Additionally, predefined templates are inadequate for handling diverse and complex application scenarios, demonstrating a lack of flexibility and scalability.

With the continuous advancement of deep learning, artificial neural networks, due to their exceptional nonlinear modeling capabilities, have demonstrated powerful semantic extraction potential. Model-based approaches have emerged and been widely applied to address the QG task. For instance, Chan and Fan (2019) employed BERT as both the Encoder and Decoder for question generation, while Lopez et al. (2021) used a fine-tuned GPT2 method to generate questions, the latest research, Ushio et al. (2022) builds on previous work in QG. It establishes a complete process for generating and evaluating paragraph or sentence-level QG based on fine-tuned language models. These methods typically take a part of the context as the answer and then generate questions using specially designed models, which limits diversity and fails to generate comprehensive questions that require cross-paragraph analysis. Recent advancements Ushio et al. (2023) introduce a new problem definition, discarding the assumption that the answer is a part of the context. Instead, they posit that the answer should also be generated, thereby aligning

[*]Equal contribution.
[†]Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10660–10671
January 19–24, 2025.

the approach more closely with real-world scenarios. However, since both the questions and answers need to be generated by the model using a black-box mechanism, the quality and reliability of the QA pairs cannot be guaranteed, severely limiting its application.

In this paper, we focus on the problem definition in Lee et al. (2020), *i.e.*, generating QA pairs given a context. It is challenging to solve, primarily in two aspects: (i) **Low Quality:** The generation of QA pairs requires the model to fully understand the semantic information of the context, which is inherently challenging. Additionally, ensuring the readability and diversity of the generated questions, as well as providing comprehensive and accurate answers, imposes further demands on the model. These challenges collectively hinder the generation of QA pairs, frequently resulting in low-quality outcomes. (ii) **Lack of Comprehensive Metric:** The quality of QA pairs directly impacts the fine-tuning effectiveness of the model. To filter high-quality QA pairs and enhance the reliability of the generated data, it is essential to design robust metrics for QA pair quality assessment. However, evaluating these pairs involves multiple dimensions, such as the relevance of the question to the context, the accuracy and completeness of the answer, and the logical consistency between the question and the answer, making comprehensive assessment challenging. In the absence of ground truth answers, aligning with human evaluation standards is impossible, further complicating the assessment.

In response, We propose a novel method for generating QA pairs. For the issue of low-quality, we employ a large language model as the foundational architecture, which has been widely validated for its robust performance across various tasks. To mitigate the generation of unanswerable questions and unfaithful answers, we fine-tune the model on a reading comprehension dataset to enhance its ability to uncover deep semantic meanings in the text. Furthermore, to guide the model to generate high-quality QA pairs, we propose a metric-based direct preference optimization mechanism that increases the logarithmic probability of preferred samples while simultaneously decreasing the logarithmic probability of non-preferred responses. For the issue of lacking metrics, we introduce a set of unsupervised evaluation metrics that analyze the correlations among pairs within triplets composed of context, question, and answer, assessing the quality of QA pairs from three distinct dimensions.

Overall, our contributions can be summarized as follows:

- We introduce a novel method for QA generation that involves the strategic fine-tuning of a large language model to adapt to the specific demands of QA tasks. Importantly, we incorporate a metric-based DPO mechanism that systematically guides the model to generate high-quality QA pairs.
- We propose a set of unsupervised evaluation metrics, designed to facilitate the selection of preferred samples during the optimization of our model. Additionally, these metrics are versatile enough to assess QA pair quality in various contexts such as validation set construction and QA system analysis.
- We conduct extensive experiments across multiple datasets to validate our model's performance. Our findings demonstrate its superiority in QA generation compared to state-of-the-art models.

2 Related Work

2.1 Question Generation

Traditional question generation mainly focuses on the task of generating a question given an input consisting of a passage chunk and an answer (Mitkov and Ha, 2003), a task commonly referred to as answer-aware QG. Early research on problem generation relied on rule-based methods (Rus et al., 2010), which generated fixed format problems using templates pre-designed by human experts. The rule-based methods mainly include methods based on dependency syntax trees and semantic role annotation (Dhole and Manning, 2020), as well as knowledge graph-based methods (Guo et al., 2022).

However, neural approaches quickly gained prominence, generating questions directly from text in an end-to-end manner (Du et al., 2017). After the emergence of masked pre-trained language models, such as BERT (Devlin et al., 2018) and DeBERTa (He et al., 2021b), significant advancements have been made in the fields of machine reading comprehension (QA, Question Answering) and machine question generation. This has led to the creation of high-quality datasets like SQuAD (Rajpurkar et al., 2016), which address complementary challenges in natural language processing, research on question generation based on neural

network models includes the use of the BERT architecture as both Encoder and Decoder structures to train a question generator, as described by Chan and Fan (2019). Lopez et al. (2021) fine-tuned GPT2 to accomplish this task. Murakhovs'ka et al. (2022) fine-tuned the T5 and BART models for question generation across multiple tasks, achieving state-of-the-art results on several different types of datasets.

Recently, Ushio et al. (2022) have proposed a paragraph-level automatic question generation approach. This method involves fine-tuning a sequence-to-sequence generative language model to generate questions directly from passage chunks and ultimately extends to question-answer generation (Ushio et al., 2023). However, this method heavily relies on ground-truth data and does not fully exploit the linguistic capabilities obtained by LLMs during pre-training, resulting in insufficient expressive power.

2.2 Reinforcement Learning from Human Feedback

In recent years, Reinforcement Learning with Human Feedback (RLHF) (Stiennon et al., 2020) has emerged as a significant approach in enhancing the performance of language models, particularly in tasks requiring nuanced understanding and generation capabilities. RLHF leverages human feedback to fine-tune models, guiding them to produce more accurate and contextually appropriate responses.

RLHF integrates human evaluations directly into the training loop of machine learning models, enabling them to adapt based on real-time feedback. This approach typically involves converting human feedback into numerical rewards that the model can optimize. Reward modeling is a crucial step where human feedback, either explicit (e.g., thumbs up/down) or implicit (e.g., click-through rates), is quantified. Algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) are then used to adjust the model's policy based on these rewards. By incorporating human evaluators who provide ongoing feedback, models continuously refine their outputs to better align with human expectations and preferences.

This methodology has proven particularly effective in domains where precise and contextually nuanced responses are crucial. Applications include conversational AI, where chatbots and virtual assistants benefit from RLHF by generating more accurate and contextually relevant responses, and personalized recommendation systems, which align suggestions more closely with user preferences and feedback.

Another emerging approach is Direct Preference Optimization (DPO) (Rafailov et al., 2023), which aims to learn strategies directly from data by understanding and optimizing preferences without the explicit need for a reward model. DPO works by collecting a dataset of human preferences, typically involving pairs of outputs where one is preferred over the other. Using machine learning algorithms, these preferences are modeled directly, often through techniques such as pairwise ranking or more complex neural network architectures. Once the preference model is trained, it guides the generation of new outputs by predicting which ones will be preferred based on learned patterns.

Previous studies have demonstrated the efficacy of RLHF in mitigating issues such as model hallucination (Ouyang et al., 2024), where the model generates responses that are not supported by the provided context. By incorporating human feedback, models can learn to avoid generating questions that the given context cannot answer, thereby improving the overall quality and reliability of the generated outputs.

3 Comprehensive Metric for QAG

Our goal is to generate diverse and consistent question-answer pairs (QA pairs) to provide training data for large language models fine-tuning in QA. Formally, given a context C containing N words, our objective is to generate (Q, A) pairs, where Q is the question and A is the answer derived from the context for Q.

To systematically introduce our work, we first present the proposed unsupervised evaluation metrics. Specifically, we concentrate on the intrinsic connections between the elements within (C, Q, A), segmenting the evaluation into three subtasks: assessing if the context contains information to answer the question, evaluating if the answer suitably responds to the question, and assessing if the answer remains faithful to the context. Correspondingly, we introduce three metrics: MRCScore, NLIScore, and SIMScore, as shown in Figure 1.

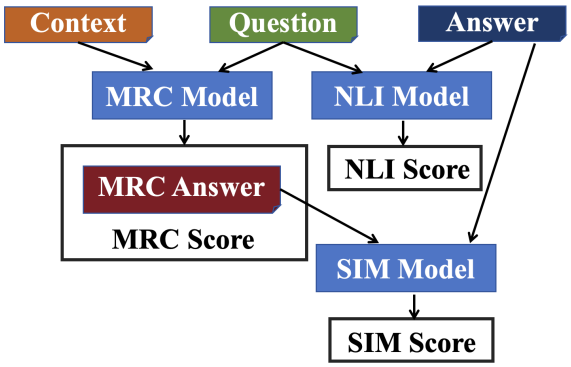

Figure 1: This is a diagram illustrating the calculation of a comprehensive metric using (C, Q, A) triples.

3.1 MRCScore

To evaluate whether the context contains the information needed to answer a question, the model must deeply analyze the underlying semantic information covered in the context, and dissect characteristics of the answer (including the inherent assumptions of the question and the topics covered by the answer) . To achieve this goal, we utilize a Transformer-based model (mdeberta-v3-base-squad2 (He et al., 2021a) (He et al., 2021b)) as the foundational framework and incorporate machine reading comprehension (MRC) as the pre-training task to bolster the model's capabilities in semantic analysis and logical reasoning.

Specifically, we utilize the Transformer-based model to extract potential answer intervals from contexts based on given questions. The model also provides a confidence value for each extracted answer interval. A low confidence level indicates a lower probability that the extracted interval correctly answers the question, thus serving as a measure of the correlation between questions and contexts. We refer to this confidence level as the MRCScore, which can be expressed in the following form:

$$(A_{ref}, S_{MRC}) = M_{MRC}(Q, C)$$

where Q represents the given question,and C represents the given context, A_{ref} represents the model's predicted answer, S_{MRC} is the confidence score for the answer.

3.2 NLIScore

To assess whether the answer appropriately responds to the question, we model it as a logical relationship discrimination problem, determining whether the answer logically supports or contradicts the content of the question. To achieve this goal, we utilize the same model as in section § 3.1, especially employing natural language inference (NLI) (Parikh et al., 2016) for pre-training to enhance the model's ability to discern the subtle connections between the question and the answer.

Follow the traditional NLI task setting (Parikh et al., 2016), we construct the premise by amalgamating the question and its corresponding answer into a sentence, and formulate the hypothesis:"this sentence constitutes a question-answer pair". Subsequently, the model assesses the semantic similarity and logical relationship between the hypothesis and the premise. It then generates a classification label that reflects the confidence level of the hypothesis given the premise. A lower confidence level suggests a lower probability that the question and answer form a logically coherent pair. Therefore, this confidence level can be used to measure the degree of logical matching between the question and answer. In this article, we refer to this measure as NLIScore (deberta-v3-large-tasksource-nli (Sileo, 2023)), which can be expressed in the following form:

$$S_{NLI} = M_{NLI}(Q, A)$$

where S_{NLI} represents the matching score between Q and A.

3.3 SIMScore

To assess whether the answer remains faithful to the context, we utilize the Transformer-based model in section § 3.1, which has been fine-tuned on MRC task, endowed with capabilities for semantic analysis and logical reasoning. Specifically, we employ this model to extract a reference answer A_{ref} that can be found in the context for the current generated question and compare it with the generated answer of the model. This calculation is based on the assumption that the answer should come from the original context, which can be expressed in the following format:

$$S_{SIM} = M_{SIM}(A_{ref}, A)$$

Where S_{SIM} represents the matching score between the generated answer and the context, M_{SIM} is generally a model that can calculate the semantic similarity between sentences, A represents the generated answer of the model, and A_{ref} represents

the most likely reference answer extracted from the context through the MRC model(§ 3.1).

4 LLM for Question Answer Generation

In this section, we will introduce the specific process of MDPO, as illustrated in Figure 2. This approach includes constructing negative samples, SFT for QAG, using RLHF to reduce hallucinations, and metric-based RLHF.

4.1 Constructing Negative Samples

We manually selected questions for which answers could not be found in the context and fed these context-question pairs into the closed-source model, forcing it to generate an answer.

4.2 SFT for QAG

The encoder-decoder architecture model lacks strong expressive capabilities. Dong et al. (2021) have demonstrated that the attention matrix can suffer from decreased expressive power due to the low-rank problem. In contrast, the attention matrix of decoder-only models is a lower triangular matrix, a full-rank lower triangular matrix represents stronger expressive power. Therefore, we choose to use a decoder-only model for the generation task.

We extracted (C, Q, A) from the dataset and formatted them as Figure 3. In this way, we constructed the SFT data needed for the QA generator.

To fully utilize the Encoder-only model's In-Context Learning (ICL) ability, we compared two types of prompts. One explicitly instructs the model to generate questions whose answers can be found in the reference contexts, while the other imposes no such restriction. We found that explicitly imposing restrictions on the model in the prompt significantly reduces the ratio of incorrect questions generated by the model. So the Prompts we design in the future all contain statements that explicitly limit the questions generated by the model must can be answered using reference contexts.

Through SFT, when given a specified instruction, the model can output in a designated format. We refer to this as acquiring instruction-following capability. Using this specified format, we can preliminarily generate a large number of QA pairs and parse them into the required format.

4.3 Using RLHF to Reduce Hallucinations

However, the model initially fine-tuned with SFT still generates some questions whose answers cannot be found in the context, instead relying on its own prior knowledge to respond. Therefore, we need to restrict the model from generating such questions.

RLHF is particularly relevant for the QAG task, where the goal is to generate coherent and contextually grounded questions and answers. Figure 4 illustrates our RLHF data template. Each sample consists of a query, which includes both the instruction and context. The response consists of a pair of question and answer that are contextually appropriate and coherent. Additionally, we provide a rejected response, which is another question-answer pair that is deemed inappropriate or incoherent for the given context. Therefore, we used negative samples constructed by a proprietary large model and found positive samples with matching contexts. We then constructed (query, response, rejected_response) samples for DPO training (Rafailov et al., 2023). The goal is to minimize the model's generation of unanswerable questions.

Using the above methods, the occurrence of hallucinations in the model-generated QA pairs has been significantly reduced, both from a human perspective and according to the MRCScore evaluation. A typical case can be found in Figure 5.

4.4 Metric-based RLHF

The generation model obtained through the aforementioned methods has already surpassed the previous generation in terms of capability. However, we can still leverage the proposed metrics to further automate the enhancement of the model's abilities.

Now, instead of relying on human feedback, we use model-based metric feedback to guide the fine-tuning process. This approach leverages an encoder-only model to provide more consistent and objective feedback, thus avoiding the problem of overestimating the model's capabilities (Hasselt et al., 2016).

We sampled contexts from the dataset and tasked the models obtained in § 4.2 and § 4.3 with generating QA pairs. These pairs were then scored using our proposed metrics. Samples with significant score differences were selected to construct DPO data. We performed another round of DPO training on the model from § 4.3. Evaluation results on the dataset showed an improvement in the metrics.

This automated approach ensures a systematic generation of training samples, further refining the model's output(§ 3) by penalizing unanswerable questions. The combination of using model-

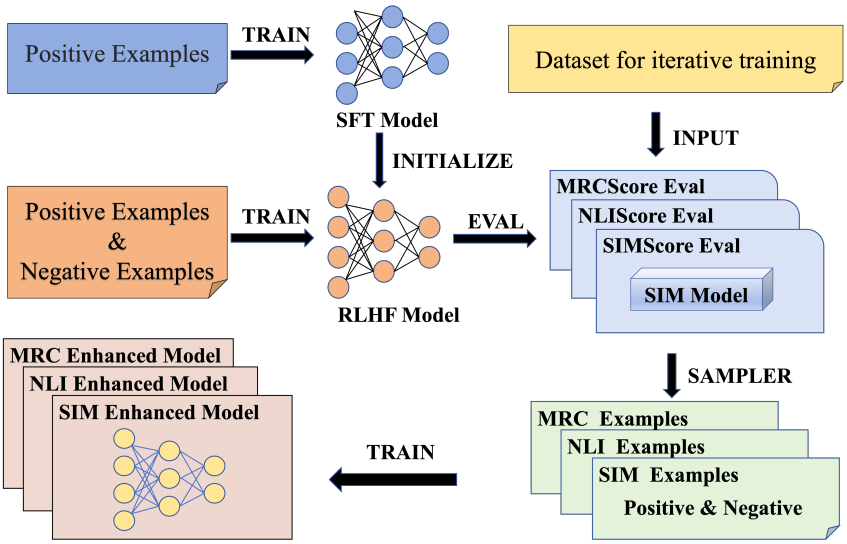

Figure 2: The overview of MDPO.

```
query:" [instruction] [context] "
response:" question: [question], answer: [answer] "
```

Figure 3: This illustrative diagram presents our SFT data template.

```
query:" [instruction] [context] "
response:" question: [question], answer: [answer] "
reject response:"question: [question], answer: [answer]"
```

Figure 4: This illustrative diagram presents our RLHF data template.

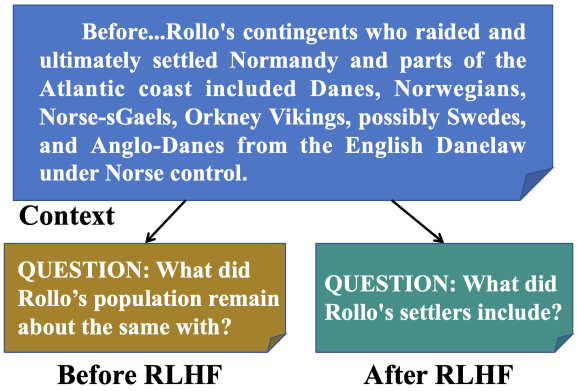

Figure 5: Different questions generated before and after RLHF training for the same context.

based feedback and DPO enhances the model's performance, aligning its outputs more closely with contextual expectations and reducing instances of hallucination. This dual approach leverages the strengths of both supervised fine-tuning and reinforcement learning to achieve a more robust and reliable QAG system.

5 Experiments

5.1 Dataset Description

To comprehensively evaluate the performance of our model against others, we conducted a QA generation capability assessment on three datasets: SQuAD, SQuAD_v2 (Rajpurkar et al., 2016), and Tweetqa (Xiong et al., 2019). SQuAD is a reading comprehension dataset consisting of questions posed by crowdsourced workers on a set of Wikipedia articles. This dataset includes reference contexts, questions, and answers extracted from those contexts. SQuAD_v2 expands on the original SQuAD dataset by adding manually annotated negative samples (questions that cannot be answered using reference contexts). Tweetqa, on the other hand, collects short tweets from social media, with human annotators writing questions and answers specific to each tweet. Unlike SQuAD, the answers in Tweetqa are not direct excerpts from the tweets but are freely formulated texts. These three datasets provide a comprehensive evaluation of the model's generation capabilities.

To validate the effectiveness of our evaluation metrics, we selected three commonly used question-answering datasets: NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and MultiRC (Khashabi et al., 2018). We constructed negative samples by randomly replacing questions, reference contexts, and answers to assess the metrics' ability to classify positive and negative cases.

5.2 Experiment Setting

Due to the lack of preset answer content for negative samples in the dataset, it is difficult to use them for reinforcement learning directly. Therefore, we called the DeepSeek API (DeepSeek-AI, 2024) to generate corresponding answers for the questions of these negative samples. As the questions of negative samples are not related to the context, the generated answers cannot be found either. We selected two sizes: gemma-2b-instruct and gemma-7b-instruct. For model fine-tuning and inference tools, we chose swift (Team, 2024), with all fine-tuning done using LoRA (Low-Rank Adaptation) (Hu et al., 2022).To comprehensively compare with the lmqg method (Ushio et al., 2023), we selected the results of lmqg method (Ushio et al., 2023) training on multiple base models and datasets, including fine-tuning models based on t5-small, t5-base, t5-large (Raffel et al., 2020), Bart-base, Bart-large (Lewis et al., 2019) on Tweetqa, as well as fine-tuning models based on flan-t5-small, flan-t5-base, and flan-t5-large (Chung et al., 2022) on SQuAD_v2.

5.3 Feasibility of Metrics

In Table 1, we tested the scores obtained by ARES (Saad-Falcon et al., 2023) and our metrics on various manually curated datasets, as well as the classification accuracy for positive and negative samples. ARES provides discrete values of 0 or 1, whereas our metrics yield continuous values. Both theoretically and experimentally, Our metrics is 21.17%, 30.67%, and 21.67% higher than that of ARES in terms of the accuracy of distinguishing positive and negative samples on three commonly used datasets NQ, HotpotQA, and MultiRC, respectively.

5.4 Models Performance on Comprehensive Metrics

In Table 2, we tested the comprehensive metrics scores of a series of models trained by lmqg and our models, both before and after RLHF. The NLIScore and SIMScore of our model on the SQuAD test set are 13.94% and 33.14% higher than the highest scores of the lmqg series models, respectively. On the SQuAD_v2 test set, these two scores are 14.51% and 34.59% higher, respectively. On the tweetqa test set, the highest NLIScore belongs to bart-large-tweetqa of the lmqg series, while the highest SIMScore of our model is 28.94% higher than the highest score of the lmqg series.We conducted the following analysis:

dataset		metric	score	accuracy
NQ	ARES	Context Relevance	78.74	
		Answer Faithfulness	72.00	65.50%
		Answer Relevance	99.96	
		MRCScore	34.60	
	Ours	NLIScore	51.80	**86.67%**
		SIMScore	75.66	
HotpotQA	ARES	Context Relevance	96.02	
		Answer Faithfulness	69.00	51.00%
		Answer Relevance	81.25	
		MRCScore	42.60	
	Ours	NLIScore	66.96	**81.67%**
		SIMScore	73.56	
MultiRC	ARES	Context Relevance	80.34	
		Answer Faithfulness	96.00	50.00%
		Answer Relevance	95.15	
		MRCScore	32.79	
	Ours	NLIScore	57.50	**71.67%**
		SIMScore	47.66	

Table 1: The evaluation results of ARES and our metrics on several commonly used datasets show the accuracy of positive and negative sample classification. Negative samples were obtained by randomly replacing elements in the (C, Q, A) triple. To increase the difficulty of classification, we selected replacement contexts that closely matched the theme of the correct context.

(i) Models with a larger number of parameters significantly outperformed smaller models on the NLIScore metric. This can be attributed to the fact that larger models develop stronger language capabilities during the pre-training phase.

(ii) Our models, both before and after RLHF, scored lower on the MRCScore compared to the encoder-decoder models trained by lmqg. We attribute this to the differences in the training methods of different models. Specifically, Decoder-only models rely on self-generated context to predict the next word, making them more prone to hallucinations, especially when earlier content deviates. In contrast, encoder-decoder models use an encoder to process input information, providing a stable context foundation that enhances consistency and accuracy during generation. This dual processing mechanism, combined with the ability to integrate external information, allows encoder-decoder models to perform better in tasks requiring precision and fact verification, thereby reducing the likelihood of hallucinations. Nevertheless, our model still demonstrates greater practical value, as its performance on the MRCScore metric is comparable to other models, especially when considering its advantages on SIMScore and NLIScore.

	SQuAD			SQuAD_v2			tweetqa		
	MRCScore	NLIScore	SIMScore	MRCScore	NLIScore	SIMScore	MRCScore	NLIScore	SIMScore
t5-small-tweetqa	40.04	46.00	39.92	39.16	47.46	40.16	43.04	40.68	40.32
t5-base-tweetqa	44.27	46.35	43.92	43.60	47.39	44.03	54.75	43.16	44.90
t5-large-tweetqa	48.18	46.36	43.42	46.43	48.30	44.05	59.45	46.32	46.96
bart-base-tweetqa	51.83	57.31	49.55	48.57	56.99	48.54	62.26	52.31	49.91
bart-large-tweetqa	50.62	55.29	50.07	48.44	55.61	49.41	63.49	**54.16**	52.13
flan-t5-small-squad	54.32	32.77	35.44	54.11	33.40	35.21	52.40	22.44	26.07
flan-t5-base-squad	62.10	37.63	41.00	60.38	36.68	39.73	57.86	32.81	41.84
flan-t5-large-squad	**65.30**	40.14	44.76	**62.83**	39.67	43.86	**64.61**	39.47	50.11
gemma-2b-sft(MDPO)	51.09	61.63	78.51	49.20	63.70	79.50	46.68	46.13	71.86
gemma-7b-sft(MDPO)	58.05	65.83	**83.21**	56.00	66.50	**84.00**	60.11	50.65	**81.07**
gemma-2b-rlhf(MDPO)	53.69	64.77	72.22	51.80	67.30	72.50	54.65	50.08	70.70
gemma-7b-rlhf(MDPO)	60.66	**71.25**	78.03	61.10	**71.50**	78.00	62.22	51.77	78.74

Table 2: The table compares the scores of a series of models trained by lmqg and our models on comprehensive metrics, tested on SQuAD, SQuAD_v2, and tweetqa datasets.

Model	MRCScore	NLIScore	SIMScore
gemma-2b-fewshot	52.60	64.30	**80.10**
gemma-2b-sft	49.20	63.70	79.50
gemma-2b-rlhf	51.80	67.30	72.50
gemma-2b-mrc-iter	54.40	62.97	57.77
gemma-2b-nli-iter	51.82	**74.15**	63.93
gemma-2b-sim-iter	**56.30**	70.39	71.00
gemma-7b-fewshot	57.60	65.20	**85.00**
gemma-7b-sft	56.00	66.50	84.00
gemma-7b-rlhf	**61.10**	**71.50**	78.00

Table 3: In the ablation experiment, we evaluated our model's performance under three conditions: trained solely with SFT, trained with a combination of SFT and RLHF, and trained with SFT combined with RLHF and further enhanced with Few-shot techniques. For the 2b model, we iteratively conducted RLHF training by selecting and pairing positive and negative examples based on samples where the SFT and RLHF models had significant scoring differences across various metrics. This approach yielded evaluation results for three iteratively trained models based on different metrics.

5.5 Evaluation of the Metric-Based RLHF

Using our metric-based RLHF, the model shows significant improvements in both MRCScore and NLIScore in Table 3. Human experts have also observed an enhancement in the generation quality. However, there is a noticeable decline in SIMScore. Our analysis suggests that SIMScore computation relies on the MRC metrics. When SIMScore is used directly for feedback, the model does not receive optimization from the MRCScore perspective, resulting in a deterioration of the intermediate variable A_{ref}(§ 3.1), which ultimately leads to a drop in the SIMScore.

6 Conclusion

In this study, we have demonstrated the effectiveness of our metrics. Specifically, MRCScore and NLIScore exhibited strong performance on human-annotated datasets, effectively distinguishing between positive and negative samples. This capability is crucial for filtering generated data and facilitating the automatic iteration of training datasets.

Our metric-based RLHF has shown a marked improvement in the quality of data generated by the model. Notably, when using MRCScore and NLIScore as feedback signals in RLHF training, we observed significant improvements in both metrics individually. Furthermore, when SIMScore was used as the feedback signal, we observed concurrent improvements in MRCScore and NLIScore. We attribute this phenomenon to the coupling nature of SIMScore, which inherently integrates aspects of both MRC and NLI evaluations.

These findings underscore the potential of using specialized metrics to drive the training process, leading to substantial enhancements in model performance. The iterative nature of our approach ensures continuous refinement and improvement, making it a valuable strategy for developing more robust and accurate machine reading comprehension models. In conclusion, our work highlights the importance of carefully chosen metrics in guiding the RLHF process and demonstrates the significant impact of such metrics on model performance. Future work will focus on further refining these metrics and exploring their integration with multi-task learning approaches to achieve even greater improvements across multiple evaluation criteria.

Limitations

While our approach for RLHF has demonstrated significant improvements, several limitations must be addressed to enhance the overall effectiveness and robustness of our methodology.

Firstly, our proposed evaluation metrics, *i.e.* MRCScore, NLIScore, and SIMScore, are all model-based and the scores largely depend on the model's performance, which in turn is heavily influenced by the quality of the training data. This dependency means that even true positive samples can receive low scores if the model's training data does not adequately cover those scenarios. Addressing this limitation requires the use of model alignment techniques to better calibrate the scoring models.

Secondly, while we selected models trained on tasks like Machine Reading Comprehension (MRC) to serve as scoring models based on the required capabilities, there is an inherent gap between the training tasks of these models and our specific application. Although the training tasks are somewhat similar, they do not perfectly align with our needs. To better address and explain our method, extensive research from a meta-learning perspective is necessary. This involves developing models that can generalize across various tasks more effectively.

Thirdly, our metric-based approach has significantly enhanced the model's capabilities. However, it is challenging to achieve simultaneous improvements across multiple metrics. This difficulty arises because the fine-tuning data sources for each metric differ, leading to varied impacts on model performance. To overcome this, we propose leveraging multi-task fine-tuning methods in future work. By utilizing diverse data sources, we aim to endow the model with multiple capabilities, thereby improving performance across all key metrics in our evaluations.

In conclusion, while our current approach has shown promising results, addressing these limitations is crucial. Future work should focus on better alignment techniques, exploring meta-learning strategies, and employing multi-task learning to ensure balanced improvements across all evaluation metrics. These steps will be vital for further advancing the performance and applicability of our machine reading comprehension models.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62372431, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102 and the Youth Innovation Promotion Association CAS under Grants No. 2021100.

References

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kaustubh D. Dhole and Christopher D. Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. In *Annual Meeting of the Association for Computational Linguistics*.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *ArXiv*, abs/2103.03404.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Shasha Guo, Jing Zhang, Yanling Wang, Qianyi Zhang, Cuiping Li, and Hong Chen. 2022. DSM: Question generation over knowledge base via modeling

diverse subgraphs with meta-learner. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4194–4207, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2094–2100. AAAI Press.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2021. Simplifying paragraph-level question generation via transformer language models. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II*, page 323–334, Berlin, Heidelberg. Springer-Verlag.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Leonardo Barros Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *ArXiv*, abs/2404.00213.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 17–22.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with mixed answer types. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1486–1497, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Christian Moldovan. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

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

Damien Sileo. 2023. tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

K. Singhal, S. Azizi, T. Tu, S.S. Mahdavi, J. Wei, H.W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G.S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Erratum in: Nature. 2023 Jul 27;: PMID: 37438534; PMCID: PMC10396962. Epub 2023 Jul 12.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

The ModelScope Team. 2024. Swift:scalable lightweight infrastructure for fine-tuning. https://github.com/modelscope/swift.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An empirical comparison of lm-based question and answer generation methods. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics: Findings*, Toronto, Canada. Association for Computational Linguistics.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yiyang Gu Yong Li Jianghao Yin Jiaju Lin Linhao Ye Zhiyan Tie Yougen Zhou Yilei Wang Aimin Zhou Ze Zhou Qin Chen Jie Zhou Liang He Xipeng Qiu Yuhao Dan, Zhikai Lei. 2024. Educhat: A large-scale language model-based chatbot system for intelligent education. *CCKS 2024*.

A Details of the indicator validity experiment

In order to better evaluate the ability of our proposed indicator to separate positive and negative samples, we employed an SVM (Support Vector Machine) classifier for our experiments in Table 1. We compared the separability of our three metrics (MRC-Score, NLI-Score, SIM-Score) with the three metrics proposed by ARES, and while performing the separability comparison of the three metrics independently, we trained a linear classifier to observe the separability in the case of the combination of the three metrics through SVM, which can be represented as follows:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b$$

where $\mathbf{w}$ represents the weight vector, $\mathbf{x}$ is the input feature vector, and b is the bias term. During the training process, the parameters $\mathbf{w}$ and b are optimized to maximize the margin between the positive and negative samples, thereby enhancing the classifier's ability to generalize to unseen data. This optimization is typically achieved through the use of quadratic programming techniques, which

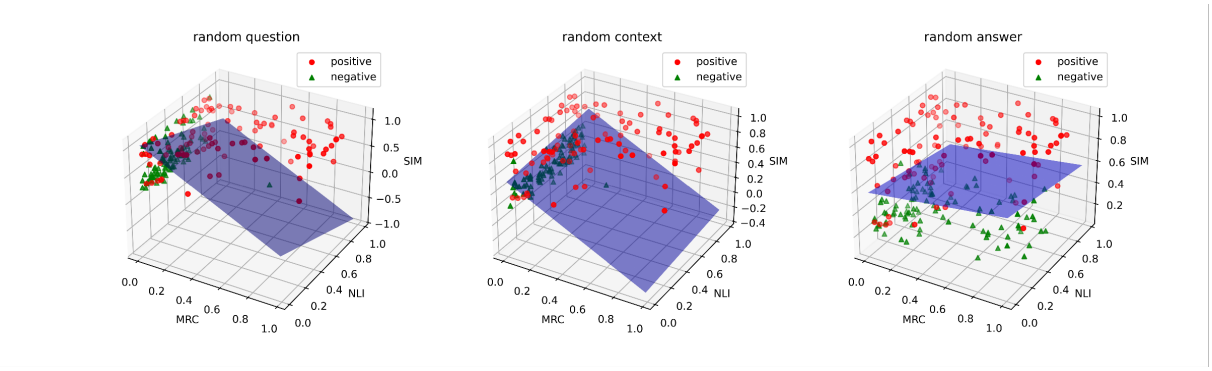

Figure 6: Visualizing the effect of SVM classifiers on our metrics.

aim to minimize the classification error while ensuring the largest possible margin between the support vectors.

The ability of our metric to separate positive and negative examples is visualized in Figure 6, where "random question" denotes a random choice of question Q between (C, Q, A) tuples, and "random context" as well as "random answer" denotes a random choice of C and A between different (C, Q, A) tuples, constructing the negative examples in this way without performing random transformed tuples represent positive examples