PNEG: Prompt-based Negative Response Generation for Robust Response Selection Model

Anonymous ACL submission

Abstract

Dialogue response selection models typically predict an appropriate response relying on the context-response content similarity. However, the selection model with over-reliance only on superficial features is vulnerable to adversarial responses that are semantically similar but irrelevant to dialogue context. Recent studies have shown that leveraging these adversarial responses as negative training samples is useful for improving the robustness of the selection model. Nevertheless, existing methods often require further fine-tuning for data creation or have limited scalability. To overcome these limitations, this paper proposes a simple but effective method for generating adversarial negative responses leveraging a large-scale language model. Our method can generate realistic negative responses only with a few human-written examples and a prompt designed to optimize generation quality. Experimental results on the dialogue selection task show that our method outperforms existing methods for creating negative responses. Synthetic quality analyses and ablation studies prove that our method is scalable and can generate high-quality negative responses. These results suggest that our method can be an effective alternative to human annotators in generating adversarial responses.

1 Introduction

Dialogue response selection models aim to predict the most appropriate response among multiple candidates for a given dialogue context (Zhou et al., 2018; Wu et al., 2019). The selection model is usually trained with the dialogue dataset consisting of a relevant response (positive) and randomly selected irrelevant responses (negatives), but such a model generally poses the following problems. First, randomly selected negatives are often too easy to distinguish because they are totally irrelevant to the dialogue context (Li et al., 2019; Lin et al., 2020). In this case, the model is more likely to predict the response only by relying on the superficial content

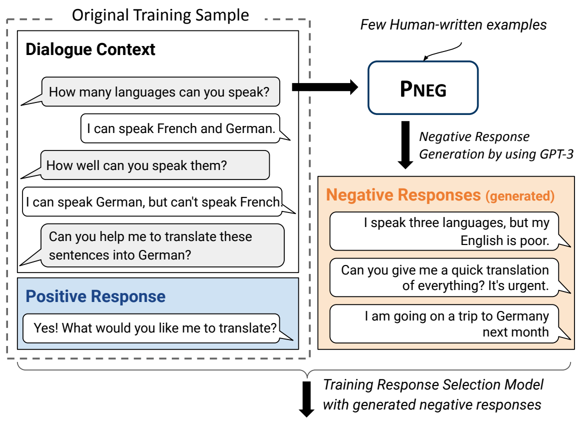

Figure 1: A conceptual pipeline of prompt-based negative response generation.

similarity of the context-response pairs (Yuan et al., 2019; Sai et al., 2020; Whang et al., 2021). These models are vulnerable to irrelevant responses which have high content similarity to the dialogue context in real-world scenarios. Second, random sampling can select a relevant response to a given dialogue context as a negative (Gupta et al., 2021). Such a false negative inherent in the training dataset prevents the correct prediction of the selection model, causing performance degradation.

To mitigate this problem, recent studies have proposed various methods to create and leverage adversarial negative training samples so that the selection model can learn features beyond content similarity (Srivastava et al., 2020; Kaushik et al., 2021). In particular, synthesizing methods can improve the robustness and generalization of the model by collecting synthetic samples besides the prepared dataset (Ebrahimi et al., 2018; Alzantot et al., 2018; Zhang et al., 2019; Qiu et al., 2021; Gupta et al., 2021). However, the existing methods can still synthesize negative responses that are grammatically incorrect or easily distinguished from the positive responses. They also usually require additional fine-tuning for generating negative responses. The most reliable approach is to collect human-written adversarial negatives (Sai et al., 2020), but it is costly, time-consuming, and difficult to create a large-scale dataset. To overcome such limitations, we can consider large-scale language models as an efficient alternative to human annotators. For example, GPT-3 (Brown et al., 2020) can effectively augment fluent text data in multiple NLP tasks without fine-tuning using prompt-based in-context learning (Yoo et al., 2021; Schick and Schütze, 2021b). This method requires a natural language-based prompt consisting of a task description and a few examples, and the prompt should be designed sensitively to ensure the quality of the generated samples (Reynolds and McDonell, 2021).

In this paper, we propose PNEG, a **P**rompt-based **NE**gative response **G**eneration method leveraging a large-scale language model (Figure 1). Since the negative response generation task contradicts the more general dialogue task of generating relevant responses, we need to elaborately design the prompts to ensure the quality of generated negative responses. As a result, our method can effectively generate adversarial negative responses using a few human-written samples and an optimized prompt. Experimental results on the dialogue response selection task show that negative responses generated by PNEG are more effective in training robust selection models than responses generated by other methods. We then conduct quality evaluation and ablation studies to analyze the validity of PNEG. Our method can efficiently produce high-quality negative responses with only a few human-written samples. Our contributions are as follows:

- We propose PNEG, a **P**rompt-based **NE**gative response **G**eneration method for robust dialogue response selection models.
- Our method can generate adversarial negative responses only with a few human-written examples and well-designed prompt.
- We show that our method outperforms strong baselines across multiple datasets and model architectures on the response selection task.

2 Related Work

Negative Response Creation Recently, several studies including the dialogue domain have proposed various negative sample creation methods for training robust and better retrieval model. Li et al. (2019) proposed an adaptive negative sampling method that selects a negative response based on similarity with a positive response. Gupta et al. (2021) introduced synthesizing methods based on masked language modeling or keyword-based generation to automatically create negative responses that have high contents similarity with a dialogue context. Similarly, Qiu et al. (2021) employed DialoGPT (Zhang et al., 2020b) to construct more challenging negative responses by providing garbled context. Sai et al. (2020) proposed a large-scale dialogue dataset including multiple positive and adversarial negative responses written by human annotators. Such human-written samples are the most reliable, but due to their lack of scalability to large-scale data, various synthesizing methods can be a scalable alternative. In this respect, we present an efficient synthesizing method utilizing human-written examples and the linguistic capabilities of large-scale language models.

Large-scale Language Models There has been grown attention in a prompt-based in-context learning paradigm with pre-trained language models (Shin et al., 2020; Jiang et al., 2020; Schick and Schütze, 2021a,c; Zhao et al., 2021). These studies have shown that the prompts written in natural language can be used to guide models to better understand a target task. In particular, mega-scale language models such as GPT-3 (Brown et al., 2020) achieve superior performance on zero- and few-shot tasks by in-context learning, even without parameter updates through fine-tuning. Yoo et al. (2021) proposed a data augmentation method that leverages GPT-3 to create realistic training samples for six sentence classification tasks. We extend this method to the dialogue domain by generating negative responses that are utilized for robust training of the dialogue response selection model.

3 PNEG: Prompt-based NEgative response Generation

Large-scale language models such as GPT-3 (Brown et al., 2020) can augment fluent text samples by using natural language prompts and in-context examples (Yoo et al., 2021; Schick and Schütze, 2021b). By extending these studies to the dialogue domain, we propose PNEG, a **P**rompt-based **NE**gative response **G**eneration method for robust response selection models.

Our method consists of three steps: (1) selecting examples from dialogue dataset, (2) constructing a prompt containing selected examples and target dialogue context, and (3) generating adversarial negative responses with a constructed prompt. The generated negative responses are used for the training of response selection models.

3.1 Example Selection

We first sample a total k dialogue examples from the dialogue dataset D_e to construct a prompt for in-context learning of GPT-3 with k-shot setting (A in Figure 2). The dialogue dataset consists of a dialogue context, an positive response, and multiple human-written negative responses. We uniformly sampled examples from the dialogue dataset. The context and human-written negative responses are used in the following prompt construction step.

3.2 Prompt Construction

Inspired by related studies (Reynolds and McDonell, 2021; Zhao et al., 2021; Yoo et al., 2021), we propose a prompt P that is designed to perform our target task. Our prompt is based on a template, k number of examples, and the target dialogue context c_t that we aim to generate multiple negative responses. The template consists of three components to clarify the role of each example and target context: (1) a task instruction I written in natural language, (2) an enumerator to receive each utterance from examples and the target context, and (3) a separator to separate each example or dialogue context in the prompt. The details of each component in the prompt template are as follows:

1. **Task instruction**: The task instruction I is used to explicitly guide GPT-3 to generate synthetic negative responses. The task instruction is located between the dialogue context and the negative responses of each example, and is located after the target context.

2. **Enumerator**: The enumerator indicates the location of each utterance in sampled examples and the target context on the prompt template. Specifically, The context enumerator for utterances from dialogue contexts is the repetition of two speaker information (*A:* and *B:*). The response enumerator for the negative responses starts with *1.* and increases by one to indicate each response. Besides indicating the utterances from examples and the target context, the enumerator also plays a role in constraining the generation of GPT-3 to suit the task goal (Reynolds and McDonell, 2021). For instance, PNEG can generate the desired number of negative responses at once by using the response enumerator to count the negative responses of sampled examples in the prompt.

3. **Separator**: The role of the separator is to separate different components in the prompt. Two types of separators are used: (1) The start separator to specify the beginning of every dialogue, and (2) The context separator to specify the start and end of each dialogue context.

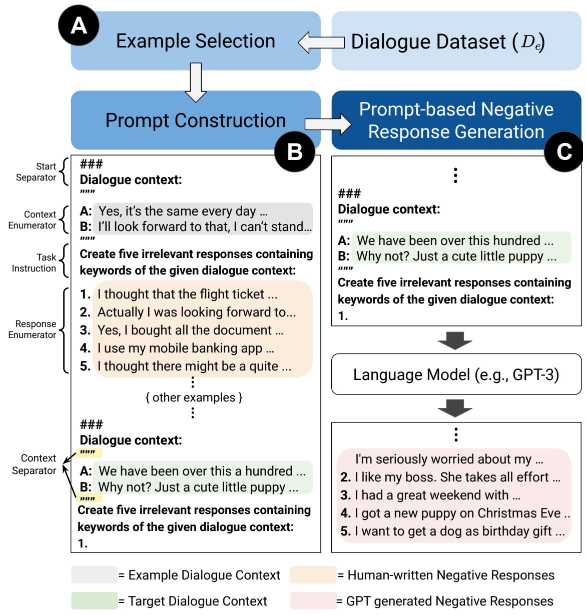

Figure 2: Overall pipeline of PNEG.

The examples and target contexts are added to the designated location of the template for prompt construction. Note that the target context is located at the last enumeration in the template. The constructed prompt is given as an input of GPT-3.

3.3 Prompt-based Negative Response Generation

GPT-3 gets our prompt as an input and generates new negative responses following the input prompt (C in Figure 2). At this time, the examples within the prompt encourage the language model to generate negative responses of similar patterns to the human-written negative responses, which indirectly explain the task. The task instruction directly guides the model to understand the target task and the relationship between a dialogue context and corresponding negative responses in the examples. The generated responses are used for the training of response selection models.

4 Prompt Optimization

This section introduces our prompt optimization process. The response selection models trained with randomly sampled negatives tend to predict high scores to responses with high content similarity with the context, and often ignore other features required to determine the appropriateness and consistency of responses (Gupta et al., 2021). Thus, we aim to generate adversarial negative responses that have high content similarities, but are irrelevant to the dialogue context. Since large-scale language models are generally pre-trained to generate relevant responses to a given context, they are familiar with generating relevant responses. In order to generate accurate negative responses, the prompt should be carefully designed to minimize the generation of relevant responses considered as false negatives. Accordingly, we set an optimization goal of prompt design in PNEG as follows: (1) minimizing false-negative generation and (2) maximizing adversarial negative generation.

We conduct iterative preliminary studies to determine the specifications of the prompt template which enables PNEG to achieve our optimization goal. As a result, we notice that the quality of generated negative responses is sensitive to the specification of task instruction or the number of examples in the prompt. To confirm our hypothesis, we analyze the performance of the downstream task according to prompt changes of PNEG (§6.3) and then determine the optimized specification of the prompt as follows.

Sufficient Examples Unlike other NLP tasks that show potential in zero-shot settings (e.g., neural machine translation), the negative response generation task increases the frequency of false-negative generation if in-context examples are not sufficiently provided. Although the contamination effect (e.g., word overlap) by examples may hinder the diversity of generated sentences (§6.3.2), in-context examples can be effectively used to achieve goals especially in non-typical tasks such as a negative response generation task. Depending on the analysis results for the number of examples (§6.3.2), we use two examples ($k = 2$) for the best performance.

Direct Task Instruction We observe that the generation quality is seriously affected by the type or abstraction level of the task instruction. Inspired by related works (Gupta et al., 2021; Reynolds and McDonell, 2021), we assume that providing a positive response or having a high abstraction level will affect the quality and diversity of generated negatives, respectively. Thus, we design and evaluate several types of task instruction (§6.3.3). According to the results, the direct task instruction is generally effective, while the instructions that may be ambiguous or misinterpreted are vulnerable to a false-negative generation. The optimized instruction is: *"Create five irrelevant responses containing keywords of the given dialogue context:"*.

5 Experimental Setup

5.1 Dialogue Response Selection Task

We evaluate our method and baselines on the dialogue response selection task. For the experiments, we train the selection model with candidate responses that have 1 positive response and 10 different negative responses per context. Five negative responses are randomly sampled responses and the other five negative responses are created by different methods that are described in §5.3. We report the R@1 and mean reciprocal rank (MRR) score of each selection model. The random and adversarial test datasets are used for evaluation, and the total number of candidate responses for each context is fixed to 6 in test datasets.

5.2 Datasets

DailyDialog++ We use the *DailyDialog++* (Sai et al., 2020) dataset for our overall experiments. This dataset consists of 16900, 1028, and 1142 dialogue contexts in training, validation, and test datasets, respectively. Since only the subset of 9259 contexts in the training dataset contains adversarial responses, we use them as our training dataset. Each context has five adversarially curated negative responses written by human annotators. Especially, the responses are created to have a high content similarity with the context. The dataset contains random and human-written adversarial test datasets with different negative response types, and both datasets contain a positive response and five negative responses for each context. In our experiment, we add a PNEG test dataset that contains PNEG generated negative responses.

PersonaChat We also use the *PersonaChat* dataset (Zhang et al., 2018) on the response selection task. The *PersonaChat* dataset consists of 8938, 1000, and 968 dialogue conversations in training, validation, and test datasets, respectively. We use 8938 contexts for training, and concatenate the persona sentences in front of the context. Since there are no human-written adversarial negative responses in this dataset, we create an adversarial test dataset by sampling one response from the context and including it in the candidate responses following Gupta et al. (2021) and Whang et al. (2021).

5.3 Baselines

We compare our approach with the following baselines that create negative responses. Every generated negative response from each method is used to train response selection models in §5.4.

Random Randomly sampled responses from other dialogue context.

BM25 (Karpukhin et al., 2020) Retrieved responses from BM25 (Robertson and Zaragoza, 2009) algorithm based on similarity with a dialogue context. In this work, we use the retrieved responses released by Gupta et al. (2021).

Semi-hard (Li et al., 2019) Retrieved responses from training dataset based on their similarity between positive response with a marin of α. We perform a static sampling using sentence-BERT (Reimers and Gurevych, 2019) with α as 0.07 following Gupta et al. (2021).

Mask-and-fill (Gupta et al., 2021) This method first randomly masks the words in a answer response, and infill them using masked language modeling conditioned on a random context.

Key-sem (Gupta et al., 2021) This method generates new responses conditioned on keywords in the context using GPT-2 (Radford et al., 2019).

Human (Sai et al., 2020) Human-written adversarial responses in *DailyDialog++* dataset.

PNEG (Ours) GPT-3 generated adversarial negative responses by using our method, PNEG.

5.4 Models

We train dialogue selection models with different negative responses described in Section 5.3. The models are based on cross-encoder architecture, and three different pre-trained language models are used in experiments: 1) BERT (Devlin et al., 2019), 2) RoBERTa (Liu et al., 2019), and 3) ELECTRA (Clark et al., 2019). For training of selection models, we predict the score of each context-response pair for every responses in a candidate responses and use cross entropy loss to maximize the score of the context-positive response pair.

5.5 Implementation Details

The inference on GPT-3 was carried out via the Open AI API Beta Access program. We used the largest GPT-3 model, *davinci*. Using the model, generating negative responses for the 9259 dialogue dataset takes an average of \$360 and 11 hours. The inference time can be shortened through parallel processing. Each inference consumes an average of 600 tokens. For the balance between diversity and quality of synthetic samples from our method, PNEG, we set the temperature to 0.8 and both frequency penalty and presence penalty to 0.4.

We use the pre-trained language models[1] released by Wolf et al. (2018) for experiments. We use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate as 2e-5, and set the maximum epochs to 3. We use the validation loss after each epoch to select the best model. The random seed is fixed, and the batch size is set to 16 per GPU on machines with 2 Nvidia TITAN RTX GPUs.

6 Experiments

In this section, we compare the performance of PNEG with the baselines on the dialogue response selection task (§ 6.1). Then we conduct quality evaluation and ablation studies (§ 6.2 and 6.3) on the *DailyDialog++*.

6.1 Performance on Response Selection Task

Response Selection Task on DailyDialog++ We compare the performance of our method with the baselines for the dialogue response selection task (Table 1). We first notice that PNEG shows the highest performance among dialogue response selection models trained with synthetic negative responses in the adversarial test datasets. This tendency is consistent in three different pre-trained language models. Although Semi-hard or other baselines often perform better than PNEG in the random test dataset, PNEG shows similar performance to human baseline in the mean of random and adversarial test datasets. These results suggest that our method can be an effective alternative to human annotators for collecting adversarial negative samples. As we

[1] `bert-base-uncased`, `roberta-base` and `google/electra-base-discriminator` are used.

Model	Approach	Test Set						Mean			
		Random		Adversarial		PNEG		Rand + Adv.		All	
		R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
RoBERTa	Random	0.879	0.932	0.658	0.797	0.599	0.749	0.768	0.865	0.712	0.826
	BM25	0.879	0.932	0.865	0.920	0.807	0.884	0.872	0.926	0.850	0.912
	Semi-hard	**0.892**	**0.937**	0.660	0.797	0.592	0.747	0.776	0.867	0.715	0.827
	Key-sem	0.889	**0.937**	0.868	0.924	0.775	0.865	0.879	0.931	0.844	0.909
	Mask-and-fill	0.873	0.927	0.868	0.922	0.806	0.884	0.871	0.925	0.849	0.911
	PNEG (Ours)	0.882	0.933	**0.942**	**0.967**	**0.907**	**0.947**	**0.912**	**0.950**	**0.911**	**0.949**
	Human	0.891	0.938	0.955	0.975	0.830	0.900	0.923	0.956	0.892	0.938
ELECTRA	Random	0.893	0.941	0.705	0.823	0.623	0.764	0.799	0.882	0.740	0.842
	BM25	0.853	0.916	0.900	0.940	0.839	0.904	0.877	0.928	0.864	0.920
	Semi-hard	**0.908**	**0.949**	0.730	0.840	0.632	0.772	0.819	0.894	0.757	0.853
	Key-sem	0.895	0.940	0.869	0.929	0.787	0.876	0.882	0.935	0.850	0.915
	Mask-and-fill	0.895	0.941	0.877	0.923	0.819	0.885	0.886	0.932	0.863	0.916
	PNEG (Ours)	0.873	0.928	**0.951**	**0.972**	**0.898**	**0.942**	**0.912**	**0.950**	**0.907**	**0.947**
	Human	0.896	0.941	0.967	0.982	0.851	0.914	0.931	0.961	0.905	0.946
BERT	Random	0.865	0.923	0.674	0.806	0.612	0.760	0.770	0.865	0.717	0.830
	BM25	0.845	0.911	0.857	0.915	0.795	0.877	0.851	0.913	0.833	0.901
	Semi-hard	**0.881**	**0.934**	0.672	0.804	0.607	0.757	0.777	0.869	0.720	0.832
	Key-sem	0.864	0.923	0.842	0.909	0.762	0.857	0.853	0.916	0.822	0.897
	Mask-and-fill	0.869	0.926	0.856	0.916	0.776	0.867	0.862	0.921	0.834	0.903
	PNEG (Ours)	0.867	0.924	**0.937**	**0.964**	**0.892**	**0.938**	**0.902**	**0.944**	**0.899**	**0.942**
	Human	0.870	0.926	0.954	0.974	0.823	0.897	0.912	0.950	0.882	0.932

Table 1: Performance in the dialogue response selection task on Random, Adversarial, and PNEG test sets based on the *DailyDialog++* dataset. We also report mean performance (Mean) of multiple test sets. We repeated the experiments three times with different random seeds and report the average performance. Among the methods except for human baseline, the best result is shown in **bold**, and the second-highest result is underlined.

Approach	Test Set		Mean
	Random	Adversarial	Rand + Adv.
Random	**0.815**	0.316	0.566
Semi-hard	0.814	0.338	0.576
BM25	0.718	0.637	0.678
PNEG (Ours)	0.774	**0.684**	**0.729**

Table 2: Performance of BERT models in the dialogue response selection task on the *PersonaChat* dataset. We repeated the experiments three times with different random seeds and report the average performance.

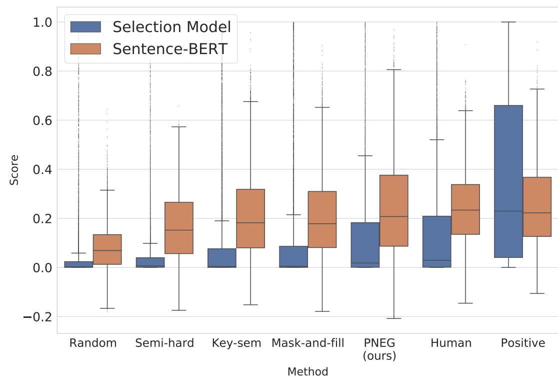

Figure 3: Box plot of prediction scores (blue) and similarity score (orange) for each type of response. The prediction scores are linearly normalized into the [0,1].

mentioned, the robustness of the models to the adversarial test dataset does not always lead to the random test dataset. We speculate that these results are due to data distribution shifts according to different negative response sampling strategies (Penha and Hauff, 2021). The examples of negative responses generated by each method are provided in Table 3.

Response Selection Task on PersonaChat We also compare our method with the baselines for the response selection task on the *PersonaChat* (Table 2). Although PNEG generates negative responses using human-written examples from *DailyDialog++*, it shows better performance than other baselines in the adversarial test dataset. Such results prove the scalability of PNEG across multiple dialogue datasets. We can expect higher performance of our method by collecting more suitable adversarial negative responses for each dataset.

6.2 Synthetic Dataset Quality

We conduct an automatic and human evaluation to verify the quality of generated responses.

6.2.1 Automatic Evaluation

We first automatically evaluate the quality of negative responses with predictive scores of the re-

Context	**A**: I am sorry to tell you that you failed in the job interview in our company. **B**: It is tough to accept it, I mean. But can you tell me why? **A**: It is hard to tell. Maybe it is because you're not resourceful enough to be a manager.
Mask-and-fill	**1.** No, that is not quite what i was thinking. What can you tell me? **2.** Can you tell me why do you want to get a taste for that? **3.** It is hard to tell. Get used to this kind of weather.
PNEG (Ours)	**1.** If you spend most of your time on the phone, then you are certainly not unemployed. **2.** It is hard to tell the difference between our company's products and those of my friend's company. **3.** I have tried to keep track of cross-cultural communication skills through the internet.
Human	**1.** Are you sorry for yourself for not being resourceful? **2.** It is tough to accept that I should attend the meeting without having lunch in the afternoon. **3.** Tomorrow there is a job interview in Titan company.

Table 3: Examples of negative responses of three methods. The dialogue context is from *DailyDialog++* dataset. The full results by each method are available at Appendix C.

Approach	Rand neg	Hard neg	False neg
Mask-and-fill	56.6%	41.0%	2.4%
PNEG (Ours)	43.6%	52.2%	4.2%
Human	47.4%	51.2%	1.4%

Table 4: Human evaluation results to verify the quality of synthetic adversarial negative responses.

Sub. (%)	$\lvert D_e \rvert$	Test Set		Mean
		Random	Adv	Rand + Adv
0.1 +REUSE	9+α	0.852	0.899	0.876
0.1	9	0.843	0.938	0.891
1	93	0.845	0.936	0.891
10	926	0.852	0.936	0.894
100 (PNEG)	9259	0.877	0.941	0.909

Table 5: Ablation study on the size of dataset D_e containing examples used to construct prompts of our method. We compare the 10%, 1%, 0.1%, and 0.1%+REUSE of the D_e.

sponse selection model and context-response similarity model. We assume that the higher the prediction score of the neural dialogue model for the adversarial negative response, the more effective it is for the robust training of the selection model. To this end, we first divide the training dataset of *DailyDialog++* dataset by 8:2 and use it as a training and test dataset, respectively. Then we train the response selection model using BERT with randomly sampled negatives. For context-response similarity model, We use a pre-trained Sentence-BERT. The evaluation results are shown in Figure 3, and the statistics of the scores are provided in Appendix F. In both models, the prediction score for negative responses generated by PNEG is higher on average than for negative responses from other methods. In particular, the difference is more evident in the selection model, suggesting that PNEG can produce more effective adversarial responses that are confused with the relevant response. Although Semi-hard samples negative responses using similarity scores from Sentence-BERT, the negative responses have lower scores than other methods because the sampling pool is limited.

6.2.2 Human Evaluation

In this experiment, we evaluate adversarial negative responses from the human point of view. We randomly sampled 100 data consisting of a dialogue context and 5 negative responses from three different method (Mask-and-fill, PNEG, and Human) which are selected according to the performance on the dialogue response selection tasks. Each response is evaluated by three human annotators[2]. Human annotators classify the type of each negative response as random, hard, and false negative according to the review criteria described in the *DailyDialog++*. The evaluation results are reported by a majority vote[3] on the three annotators. Table 4 shows the human evaluation results. Our PNEG has a slightly higher false negative ratio than Mask-and-fill, but shows the highest hard negative ratio. We also notice that our method cannot fully control the false negative generation. In future work, we may consider soft labeling (Wu et al., 2018; Chen et al., 2020) or label smoothing (Müller et al., 2019) techniques to alleviate this problem.

6.3 Ablation Studies

In this section, we conduct ablation studies and in-depth analysis of PNEG. The examples generated in each experiment are provided in Appendix D.

6.3.1 Size of Example Dataset (D_e)

We study the effect of the size of the dataset D_e containing examples used in prompts configuration

[2]We recruited a total of 9 human annotators (6 males and 3 females) for the human evaluation. The evaluation takes up to an hour and a half.

[3]The type of each data is basically determined by a majority, and if the evaluation result is a tie, such data is determined to be a random negative type.

k	Test Set		Mean
	Random	Adv	Rand + Adv
0	0.799	0.841	0.820
1	0.856	0.893	0.875
2 (PNEG)	0.859	0.928	0.894

Table 6: Ablation study on the number of examples k in the prompts of our method. (k = 0, 1, and 2)

k	position	Jaccard	Length Correlation	
	(pos/k)	Similarity	Pearson	Spearman
1	1/1	**0.046**	**0.376**	**0.351**
2	1/2	0.031	0.154	0.174
2	2/2	0.035	0.339	0.293
2	all	0.041	0.342	0.324

Table 7: Correlation of generated negative responses in our method with the few-shot examples (k>0). We measure the Jacquard similarity and length correlation between the example and the generated response.

Type	k	Test Set		Mean
		Random	Adv	Rand + Adv
I_dir	2	0.877	0.941	0.909
I_pos	2	0.857	0.940	0.898
I_imp	0	0.788	0.800	0.796

Table 8: Ablation studies on task instruction changes in the prompt of PNEG. The I_{pos} and the I_{pneg} are follows 2-shot setting, and the I_{imp} follows zero-shot setting.

on the performance of the selection model. As shown in Table 5, even if the size of D_e becomes extremely small (e.g., 0.1%), the performance of the adversarial test dataset hardly decreases. We conclude that our method can generate high-quality negative responses by collecting only a small number of real examples. To increase the diversity of examples, we further try +REUSE, which continuously adds the negative responses generated by PNEG to D_e. However, the 0.1%+REUSE has a significant performance drop in the adversarial test dataset. These results support our finding that example quality is more important than diversity to optimize the quality of the generated negative responses. In actual examples in Appendix D, the PNEG seems to generate adversarial negative responses well even when the size of D_e decreases.

6.3.2 Number of Examples (k)

We analyze the effect of the number of examples k in the prompts of our method on the response selection model. The results are in Table 6. Our method has the highest performance when using two examples (k=2), but using one example (k=1) also can be a reasonable alternative. The performance of prompts without examples (k=0) is rapidly degraded due to frequent occurrence or false-negative generation. These results show that it is important to provide an adequate number of examples to minimize the occurrence of false-negative responses.

We also measured the Jaccard similarity and length correlation between generated responses and each example in the prompt to qualitatively analyze the effect of the example on the generated responses. As shown in Table 7, the Jaccard similarity and length correlation coefficient are measured higher when k=1 than when k=2, and the generated responses are more affected by the closer example. Such contamination effect can increase the effectiveness of the in-context example as guidance of the task, but it can also limit the diversity.

6.3.3 Task Instruction Type (I)

We compare the performance of PNEG according to the change in specifications of the task instruction. We design the following three task instruction types: (1) direct task instruction (I_{dir}), (2) direct task instruction with a relevant response (I_{pos}), and (3) implicit task instruction (I_{imp}). We expect that I_{pos} can generate more challenging negatives by referring to the relevant response, and I_{imp} can generate diverse responses due to the reduced constraints in the prompt. As shown in Table 8, I_{pos} show lower performance than I_{dir} in the random test dataset, and we infer that the relevant response may negatively affect the quality of generated responses. Because I_{imp} is vulnerable to false-negative generation, it has the lowest performance in both random and adversarial test datasets.

7 Conclusion

This paper proposed PNEG, a prompt-based adversarial negative response generation method for training more robust dialogue response selection models. Our extensive experiments on dialogue response selection tasks show that negative responses generated by PNEG can improve the robustness of the selection models. Our method performs surprisingly well even when only a few human-written samples are available, suggesting that our method can be an efficient alternative to human annotators for generating adversarial negative responses. In future work, we are planning to extend our method to other open-domain dialogue tasks, such as dialogue context or relevant response augmentation.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling matters! an empirical study of negative sampling strategies for learning of matching models in retrieval-based dialogue systems. 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 1291–1296, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. 2020. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Gustavo Penha and Claudia Hauff. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170.

Yao Qiu, Jinchao Zhang, Huiying Ren, and Jie Zhou. 2021. Challenging instances are worth learning: Generating valuable negative samples for response selection training. *arXiv preprint arXiv:2109.06538*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm*. Association for Computing Machinery, New York, NY, USA.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Eliciting knowledge from language models using automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14041–14049.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. Transfertransfo: A transfer learning approach for neural network based conversational agents. In *NeurIPS Workshop on Conversational AI*.

Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 420–425, Melbourne, Australia. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019.

Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. 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 111–120, Hong Kong, China. Association for Computational Linguistics.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2020. Multi-referenced training for dialogue response generation. *arXiv preprint arXiv:2009.07117*.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

A Prompts used in PNEG

The optimized PNEG prompt is as follows:

```
###
Dialogue context:
"""
A: How about taking the damaged portion at a lower price?
B: What kind of price did you want?
A: I was thinking of 30% off.
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1.  I have not completed the portions of the children, ...
2.  Shall I inquire about the price of the plane tickets ...
3.  I have been thinking up new ways of supplying money ...
4.  My car roof was not damaged in the accident.
5.  I purchased a different kind of dress in the shopping mall ...
###
Dialogue context:
"""
A: No, but that was a random change of subject.
B: It may have been random, but have you?
A: I haven't lately.
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1.  Yeah, Our society is annoying.  They keep on changing ...
2.  I am not sure which subject I am going to take.  Lately, ...
3.  I don't know that day Prof.  Lesley was randomly picking up ...
4.  Today In college some random guy came and started talking ...
5.  Have you seen Tina lately?  I am feeling weird as ...
###
Dialogue context:
"""
A: Paul, a company called me for an interview.
B: That's great!  You need to prepare for it.
A: How?
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1.
```

Then, GPT-3 generate following negative responses:

```
I don't know how to handle an interview.
2.  How could I prepare for the interview?
3.  To be honest man, I don't know what to say.
4.  You are probably thinking that I could not do well in the interview, right?
5.  Did you remember to register for the interview?
```

This prompt is particularly sensitive to the task instruction. For example, GPT-3 generates emotionally negative responses when the word *"irrelevant"* in the task description is changed to *"negative"*. In addition, if we remove *"containing keywords of -"* from the task instruction, the generation quality of GPT-3 is degraded.

Context	**A**: I am sorry to tell you that you failed in the job interview in our company. **B**: It is tough to accept it, I mean. But can you tell me why? **A**: It is hard to tell. Maybe it is because you're not resourceful enough to be a manager.
Positive	**1.** Well what can I say! I am speechless, anyways thanks for calling. **2.** The interview was quite good, I answered everything I was asked for, and so what seems to be ... **3.** Well I was confident! It is really shocking that I couldn't clear it. **4.** But why! What is the matter? My performance was good. **5.** Anyways never mind, I will keep on trying, thank you for your information.
Random	**1.** Yogurt can be incredibly healthy. Unfortunately, most yogurts found in the grocery store are ... **2.** Well! I was really scared; maybe I need few more lessons. **3.** I too don't have any idea what they mean. **4.** Okay, you can have a day off. **5.** Well! Again it depends on family to family as I said to you earlier.
BM25	**1.** It is tough to accept it, i mean. But can you tell me why? **2.** I am indeed. What can you tell me? **3.** I am sorry, sir. Can you tell me your room number? **4.** Can you tell me why you replied to our advertisement? **5.** What's the matter? Please tell me. Maybe i can help you.
Key-sem	**1.** It's hard to say, but it sure was. You did a good job as a probationary employee, didn't you? **2.** No problem. The company will give me the job interview. **3.** No, that's better. I think we'd better do that at the least convenient time. **4.** I don't know exactly. I've never had any difficulty in starting a business. Why did you ask me to come in to talk about your business? **5.** Sorry, i'm not very resourceful enough to go there.", "I'm sorry. I'm too disappointed.
Mask-and-fill	**1.** No, that is not quite what i was thinking. What can you tell me? **2.** Can you tell me why do you want to get a taste for that? **3.** It is hard to tell. Get used to this kind of weather. **4.** First, i am looking for a shop to fix my bathroom. It's only the second time my water bill has been paying the bills recently. **5.** Anyways never mind giving up. I'm on my way out now, thank you for your information.
PNEG (Ours)	**1.** If you spend most of your time on the phone, then you are certainly not unemployed. **2.** It is hard to tell the difference between our company's products and those of my friend's company. **3.** I have tried to keep track of cross-cultural communication skills through the internet. **4.** You are so resourceful that you can work with people from all over the world. **5.** My partner has no resources to go over the financial reports. I am confused why she wants to be a part of this company again.
Human	**1.** Are you sorry for yourself for not being resourceful? **2.** It is tough to accept that I should attend the meeting without having lunch in the afternoon. **3.** Tomorrow there is a job interview in Titan company. **4.** She marched into the office and demanded to speak to the manager but she failed to talk. **5.** They were resourceful and talented people in our school.

Table 9: Negative response examples from various methods for a dialogue context with its positive responses. The dialogue context and its multiple positive responses are from *DailyDialog++* dataset.

B Task Instruction Type (I_{dir}, I_{pos}, and I_{imp})

The following task instructions are the detailed specification of the task instruction types used in Section 6.3.

I_{dir}=`Create five irrelevant responses containing keywords of the given dialogue context:`
`1.`

I_{pos}= `Relevant response: You need to know something about the company.`
`Create five irrelevant responses containing keywords of the given dialogue context:`
`1.`

I_{imp}=`Suddenly, B(or A) makes an awkward response. The response appears to be okay at first glance, but it's irrelevant to the conversation.`
`B (or A):`

C Negative Response Sample from PNEG and other baselines

Table 9 shows sample negative responses from various methods and 5 human-written relevant responses for a single dialogue context.

Context	**A**: I am sorry to tell you that you failed in the job interview in our company. **B**: It is tough to accept it, I mean. But can you tell me why? **A**: It is hard to tell. Maybe it is because you're not resourceful enough to be a manager.
PNEG (OURS) ***with optimal configuration***	
D_e=100% k=2 I=I_{pneg}	**1.** If you spend most of your time on the phone, then you are certainly not unemployed. **2.** It is hard to tell the difference between our company's products and those of my friend's company. **3.** I have tried to keep track of cross-cultural communication skills through the internet. **4.** You are so resourceful that you can work with people from all over the world. **5.** My partner has no resources to go over the financial reports. I am confused why she wants to be a part of this company again.
Varying D_e with {10%, 1%, 0.1%, and 0.1%+*REUSE*}, k=2, I=I_{pneg}	
D_e=10%	**1.** We must consider the ways to reduce our expenditure. **2.** We simply have to make more of an effort if we are going to succeed in the future. **3.** I already know that I want that job - what is your final offer? **4.** I am scared that our luck has run out and we will never win again. **5.** This is my personal opinion ."
D_e=1%	**1.** I don't know how to thank you for the kind of lesson I received from you. **2.** My sister would like to study in your school next year, but she doesn't want to leave the place. **3.** Our new teacher is rather young and pretty. What a pity that she's married. **4.** We will go on a trip this Saturday as my parent promised me. **5.** I am going on a date with my boyfriend tonight, but it doesn't seem too good as he told me that he would ask for a break up tomorrow night.
D_e=0.1%	**1.** I am the best and I know it. You can't beat me. **2.** I got a very good score in my English class. What's the big deal? **3.** I will be back to take revenge. **4.** The interview was held in the manager's room, who is currently out of town for a business trip. **5.** Tough to accept it, but I'm here to accept it with an open heart and welcome the next round
D_e=0.1% +REUSE	**1.** Please give me a raise in salary, please. **2.** I think you're an interesting manager. **3.** Here is the money I owe you. **4.** I really appreciate your kindness. **5.** Your answer is totally off my expectation! That's so amazing.
Varying k with {0,1}, D_e=100%, I=I_{pneg}	
k=1	**1.** Would you like to drink a cup of coffee together? **2.** I learned the recipe for a new dish yesterday and would like to try making it. **3.** Would you please give me some money now? **4.** You look tired; I'll take your job and do it myself then. **5.** Do you think you could go shopping with me later? I want to know what kind of houses are ...
k=0	**1.** I don't know why I failed to win the race. Maybe it is because my legs were too short. """, **2.** I am sorry to tell you that you failed in the job interview in our company. **3.** Well! I am stunned. That's it? Thanks for telling me at least a reason. **4.** Sir, I am sorry to tell you that you failed in the job interview in our company. **5.** It is hard to tell . Maybe it is because you're not resourceful enough to be a manager.
Varying (I,k) with {(I_{pos}, 2), (I_{imp}, 0)}, D_e=100%	
k=2 I_{pos}	**1.** Thank God, otherwise it would have been worse. **2.** What? I am not able to listen because the person next to me is shouting. **3.** Oh no! can I reduce my cell phone bill by getting to hear this news? **4.** I need the electric blanket in my room to survive this winter season. **5.** MHe's probably referring to the lack of managerial skills that I have established till now.
k=0 I_{imp}	**1.** I want you to think again. **2.** I already have the new pair of glasses. **3.** I used to be a soccer player when I was in high school. **4.** What a coincidence! How many first-class stamps do you have on hand? **5.** I have never been encouraged enough to have a good start.

Table 10: Example of negative responses generated by PNEG with varying the components. D_e, k, and I indicates the size of example dataset, number of examples, and task instruction type, respectively. The optimal configurations that are used in PNEG are D_e=100%, k=2, and $I = I_{pneg}$.

D Negative Response Sample from PNEG with Changing Prompts

Table 10 shows sample negative responses from PNEG with varying size of example dataset (D_e), number of examples in a context (k), and the task instruction type (I), following our ablation studies. Note that dialogue context in Table 10 is same with Table 9.

Approach	Pred. Score	Similarity
Random	$-2.749_{2.48}$	$0.078_{0.09}$
Semi-hard	$-2.051_{2.83}$	$0.161_{0.13}$
Mask-and-fill	$-1.925_{3.26}$	$0.207_{0.17}$
Key-sem	$-1.956_{3.34}$	$0.212_{0.17}$
PNEG (Ours)	$-0.598_{3.53}$	$0.241_{0.20}$
Human	$-0.279_{3.13}$	$0.242_{0.15}$
Positive	$2.779_{2.25}$	$0.256_{0.17}$

Table 11: Automatic evaluation results for response quality. **Pred. Score** and **Similarity** indicate the predicted score of each response by selection model and the similarity score between each response and the context measured by Sentence-BERT, respectively. The mean and standard deviation of each score are reported in the $mean_{std.}$ format.

Approach	Pearson	Spearman
Reference-based metrics		
BLEU1	0.189	0.081*
BLEU2	0.229	0.091
ROUGE-L	0.214	0.136
METEOR	0.220	0.090
Embedding Avg.	0.080*	0.095
BERTScore-recall	0.192	0.114
BERTScore-precision	0.269	0.235
BERT w. different negative samples		
Random	0.274	0.264
BM25	0.297	0.302
Semi-hard	0.298	0.294
Mask-and-fill	0.302	0.311
Key-sem	**0.345**	**0.349**
PNEG (Ours)	<u>0.340</u>	<u>0.348</u>
Human	0.316	0.323

Table 12: Correlation of our method and baselines with the human score in the dialogue evaluation task. Trainable metrics are based on BERT architecture. All results with p-value > 0.001 are marked with *. We repeated the experiments three times with different random seeds and report the average performance. The highest and second highest scores in each metric are highlighted in **bold** and <u>underline</u>, respectively.

E Results on Automatic Evaluation in Section 6.2.1

Table 11 shows statistics on the scores of each model for automatic evaluation in Section 6.2.1. Among the negative responses, human-written responses and our responses usually get the high predictive score than other negative responses. In terms of similarity score, our negative responses show high similarity with dialogue contexts. We speculate that the higher similarity of our responses with the dialogue contexts can improve the robustness of response selection models models by encouraging them to learn the features beyond superficial context-response similarity.

F Dialogue Response Evaluation Task

We also evaluate our method and baselines in Section 5.3 on the dialogue response evaluation task. The evaluation task aims to accurately assess the quality of each response. The performance of an evaluation model is measured by the correlation between human-annotated quality score and model prediction for responses to be evaluated. Pearson correlation (r) and Spearman's rank correlation coefficient (ρ) were used to measure the correlation.

As a source of human-annotated quality score, we leverage the three different datasets following Gupta et al. (2021): (1) 700 human scores from Zhao et al. (2020), where six different generation model with different decoding strategies. (2) 600 human scores from Zhao and Kawahara (2020), where hierarchical recurrent encoder-decoder models are used to generate responses. (3) 187 human score from Gupta et al. (2021), where the quality of human-written answer responses and retrieved responses from the dialogue corpus are annotated into binary score. All human scores are normalized from 0 to 1, and total 1487 human scores are used as an evaluation dataset for response evaluation task.

Baselines For dialogue evaluation task, a BERT-based selection model is trained with different type of negative response that are described in Section 5.3 for comparison. Besides, we also include the following reference-based metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), Embedding Average (Liu et al., 2016), and BERTScore (Zhang et al., 2020a). These metrics measure the similarity between human responses and the generated response to compute the response quality.

Results The results of evaluation task are in Table 12. The reference-based metrics usually show lower performance than BERT-based evaluation models. The Key-sem model shows the highest correlation, and our model shows competitive but slightly lower performance than Key-sem model. Unlike the tendency of previous experiments, the human baseline has a relatively low correlation. For this reason, we suspect that the evaluation set used in this experiment contains more randomness of the synthetic samples by the generative model than the regularity of the human-written samples (Gupta et al., 2021).

Aug.	Dataset num	Test Set		Mean
		Random	Adv	Rand + Adv
PNEG	9259	0.877	0.941	0.909
+ 5000	14259	**0.889**	0.946	0.917
+ 10000	19259	0.886	**0.950**	**0.918**
+ 15000	24259	0.871	0.937	0.904
+ 20000	29259	0.877	0.947	0.912

Table 13: Performance on our method with data augmentation techniques on additional 5,000, 10,000, 15,000, and 20,000 augmented dataset in the dialogue response selection task.

G Data Augmentation

We conduct data augmentation experiments by synthesizing adversarial negative responses to the additional datasets. For the experiment, we use the dialogue contexts in the original *DailyDialog* dataset that has no duplication with the contexts in *DailyDialog*++. The results are shown in Table 13. Data augmentation using our method generally leads to improved performance. However, if the training dataset is already large enough, the model can properly generalize it (Wei and Zou, 2019). In our experiments, the performance of the selection model is saturated, if the dataset is augmented by more than 10,000 (about 100%).

H Frequent Error Types in GPT-3 Generation

During our experiments, we often observed the weird generation results of GPT-3. The frequent error types in generated results of GPT-3 can be roughly categorized as follows: (1) n-gram or word repetition, (2) containing too many “__” or “_ _”, (3) out of numbering rules. We generate negative responses with GPT-3 for the given context until there is no error response that is aforementioned. Note that false negative is a semantic error type that needs to be evaluated by humans.