

REFINE AND IMITATE: REDUCING REPETITION AND INCONSISTENCY IN DIALOGUE GENERATION VIA REINFORCEMENT LEARNING AND HUMAN DEMONSTRATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Despite the recent success of large-scale language models on various downstream NLP tasks, the repetition and inconsistency problems still persist in dialogue response generation. Previous approaches have attempted to avoid repetition by penalizing the language model’s undesirable behaviors in the loss function. However, these methods focus on token-level information and can lead to incoherent responses and uninterpretable behaviors. To alleviate these issues, we propose to apply reinforcement learning to refine an MLE-based language model without user simulators, and distill sentence-level information about repetition, inconsistency and task relevance through rewards. In addition, to better accomplish the dialogue task, the model learns from human demonstration to imitate intellectual activities such as persuasion, and selects the most persuasive responses. Experiments show that our model outperforms previous state-of-the-art dialogue models on both automatic metrics and human evaluation results on a donation persuasion task, and generates more diverse, consistent and persuasive conversations according to the user feedback. We will release the code and data upon acceptance.

1 INTRODUCTION

Large-scale language models have greatly advanced NLP research in various sub-areas, such as question answering, text summarization, story generation and so on (Radford et al., 2019). However, these generation models still suffer from at least three major problems when applied to the dialogue system building, 1) generic and repeated responses (repetition), 2) inconsistent statements with the dialogue context (inconsistency), and 3) uncontrollable task-oblivious replies (nonspecificity) (Li et al., 2016a). Many previous studies have attempted to address these problems (Hu et al., 2017; Li et al., 2020; Song et al., 2020). For instance, (Li et al., 2020) penalized repetitive and inconsistent behaviors with unlikelihood loss in open-domain chats. (Song et al., 2020) detected and rewrote the contradicting responses to achieve a more consistent personality. However, these methods optimize the language model by minimizing the loss in supervised learning, which may lead to exposure bias and uninterpretable behaviors, and consequently, makes it harder for humans to regulate the model.

To alleviate these problems, previous work has explored RL-based methods in dialogue system building (Li et al., 2016b; Shi & Yu, 2018; Shi et al., 2019a;b). However, such methods not only rely on hand-crafted user simulators that are inherently hard to build (Shi et al., 2019a), but also require meaningful rewards that are difficult to design. To address these issues, we propose to teach the model to extract a policy directly from the data and learn from its own mistakes without the use of simulators. Leveraging decoding methods such as Nucleus Sampling (Holtzman et al., 2019), the language model finetuned on a persuasion task is able to generate lexically diverse response candidates given the same context. Some candidates are appropriate, while others are repetitive or inconsistent with the context. These good and bad examples can be used as positive and negative feedback to the model through meaningful rewards in RL, and help refine the language model. During testing, to fully utilize the refined language model, we use it to generate multiple candidates again, and filter out the repetition and inconsistency afterwards. Beyond being nonrepetitive and consistent, a good response also needs to accomplish the dialogue task, in our case, to persuade people. Therefore, we

ask humans to demonstrate the persuasion process, and build a response imitator to imitate these human demonstrations and select the most persuasive response.

The above issues in language models are especially salient in complex strategic dialogue tasks such as persuasion and negotiation. These dialogues involve both a specific task goal and social contents to build rapport for better task completion, and therefore, have richer and more complicated language structures (Li et al., 2019). Further, due to their inherent similarity to task-oriented and open-domain dialogues, improvements made on these systems would also help in both dialogue settings. Therefore, we choose a strategic donation persuasion task (Wang et al., 2019) to perform our study, and conduct both automatic and human evaluations.

This work makes multiple contributions. First, we propose DialGAIL, an RL-based generative algorithm to refine MLE-based language models for dialogue generation without the use of user simulators. Second, we design an effective and practicable framework for strategic dialogue systems that achieves state-of-the-art performance on a complex persuasion task, with only small amount of human demonstration efforts. Previous dialogue research has mostly focused on pure task-oriented dialogues and pure social conversations; but looking forward, it becomes more and more important to pay attention to strategic dialogues that involves both task and social components. We sincerely hope this work could inspire more research and discussions on strategic dialogues in the community.

2 RELATED WORK

Large-scale language models have achieved great success in multiple NLP tasks including reading comprehension, machine translation and so on (Radford et al., 2019). However, these models still suffer from repetition and inconsistency when applied to dialogue tasks which requires long-term context memory and logical reasoning. There have been many previous studies to address these issues (Zhang et al., 2020; Wu et al., 2019). Zhang et al. (2020) presented a response generation model DialoGPT trained on large conversation-like corpora, and obtained close-to-human performance in single-turn dialogues. Li et al. (2020) proposed to detect the inconsistency with natural language inference data, and penalize it with unlikelihood loss to achieve more consistent personality in open-domain dialogues. Our work tackles these problems with reinforcement learning to reduce exposure bias and encourage the model to generate multiple responses and learn from its own mistakes.

Our work is also closely related to response selection, which focus on obtaining good context representations to match the context and retrieve the best response from a large collection of human-human conversations. However, such response selection models are highly dependent on the quality and availability of the underlying datasets. To address the *data scarcity* issue, Henderson et al. (2019) pretrained a response selection model with large conversational corpora, and finetuned it on new domains in task-oriented settings for a better context representation. Instead of retrieving candidates from human dialogues, we leverage language models’ ability to generate coherent responses, and build a selector to imitate human selection process and choose among the generated candidates.

Strategic dialogue tasks such as persuasion have emerged and attracted more attention recently, given its wide applications in industry and daily life (Lewis et al., 2017; He et al., 2018; Wang et al., 2019). These tasks are close to human-human conversations and contain both task-oriented and social contents. That being said, they are usually more complex with longer context than pure task-oriented or open-domain dialogues. Towards strategic dialogue system building, Li et al. (2019) utilized large-scale language models to generate multiple responses and applied human-defined rules to filter out bad candidates. We take a similar approach to generate candidates but eliminate the manual work for rule design, and teach the model to select task-relevant candidates through human demonstration.

3 METHODS

Our framework is shown in Figure 1. The language model is p_θ and there are two steps in the framework, 1) the reinforcement learning (RL) process to refine an MLE-based model baseline q for better response generation ($p_{\theta_0} = q$), and 2) the imitation process to learn from human demonstration and select the best response. During RL training, for each user utterance, p_θ generates n response candidates, shown in the *Response Candidates* box. Then the *Response Detector* annotates these candidates with corresponding status such as “Repetition” and “Inconsistency”. These labels along

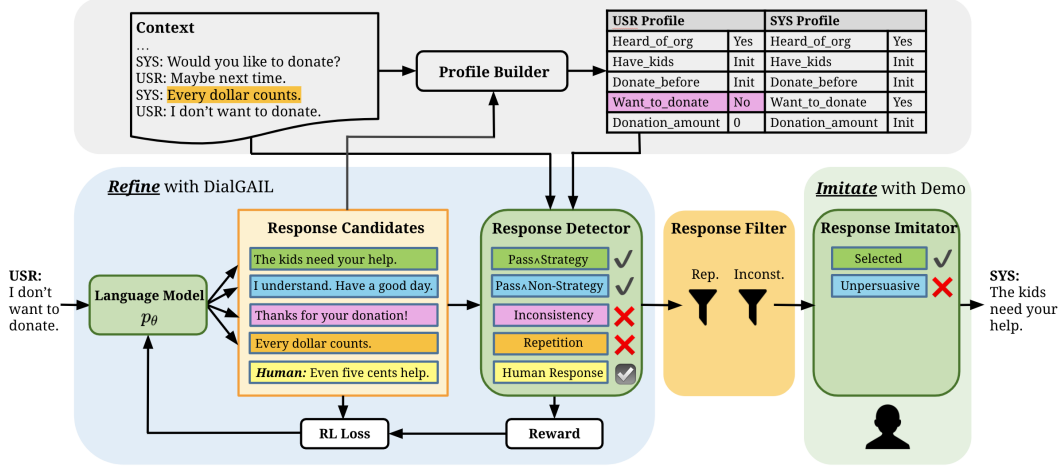


Figure 1: The overall architecture of our RFI model. During training, p_θ generates n response candidates, *Response Detector* labels them with corresponding status such as “Repetition”, and the candidates along with the golden human response send feedback to refine p_θ through the rewards. During testing, the refined p_{θ^*} generates n candidates again; *Response Filter* removes the detected repetition and inconsistency, and *Response Imitator* imitates human demonstrations to select the most persuasive candidate as the final output. The dialogue history consists of the context and the *Profiles*.

with the golden human response provide feedback through the reward function to guide p_θ to generate nonrepetitive and consistent responses. During test time, we use the refined language model p_{θ^*} to generate n candidates again, and apply the *Response Filter* to remove the repetitive and inconsistent candidates to further ensure the candidate quality. Finally, the *Response Imitator* takes in the remaining candidates, and imitates the human demonstration to select one task-related candidate as the final response. To detect repetition and inconsistency, we build *USR Profile* and *SYS Profile* shown in the top right table in Figure 1, where task-relevant information is extracted from the dialogue and stored as `<key: value>` pairs, such as “`want_to_donate: No`”. We describe each module below.

3.1 REFINE WITH REINFORCEMENT LEARNING

3.1.1 DIALGAIL

One major issue with MLE-based methods is that the trained language model tends to generalize over the training data and generate plain “average” responses that lead to inferior results in dialogues, especially when the current dialogue trajectory is different from all the training samples (Welleck et al., 2020; Shi et al., 2020). To encourage the model to explore more space and simultaneously learn from its own mistakes, we propose DialGAIL, a generative adversarial imitation learning (GAIL, Ho & Ermon (2016)) framework for dialogue generation. DialGAIL extracts a policy directly from data without interaction with the environment, which is appealing as user simulators are hard to build and real-human interactions are expensive.

DialGAIL is shown in Algorithm 1. We initialize p_θ with an MLE-based baseline q , and sample one dialogue d from the training corpus. For each turn in d , p_θ generates n response candidates. The *Response Detector* annotates each candidate with status $a_i \in \{\text{repetition, inconsistency, pass}^\wedge\text{Strategy, pass}^\wedge\text{Non-Strategy}\}$. With the detected status, candidates receive different rewards based on the following conditions, 1) if it is a ground truth human response, 2) if it is a repetitive or inconsistent response, 3) if it contains persuasion strategy. The reward values are chosen based on the validation dataset performance and the details of the reward function are in the Appendix. By optimizing the rewards, p_θ learns from its own repetitive and inconsistent mistakes and generates more diverse, consistent and persuasive responses. Note that DialGAIL is a generic algorithm to improve any sentence-level qualities (naturalness, etc) given a corresponding response detector, and task designers have the freedom to design and plug in any customized task-specific detectors. Here we choose repetition and inconsistency as typical qualities to improve on.

Following Wu et al. (2020), we apply proximal policy optimization (PPO) (Schulman et al., 2017) for more stable training. PPO performs importance sampling with the likelihood ratio between current and old policies $r(\theta) = \frac{p_{\theta_i}(s|c)}{p_{\theta_{i-1}}(s|c)}$, and optimizes the surrogate in Eq. (1) to maximize the expected rewards. To ensure the generation quality, we use the KL divergence between the language model being refined p_θ and the MLE baseline q as the maximum entropy regularizer in RL. This KL-term prevents p_θ from moving too far away from the original model q and potentially losing fluency. The final objective is shown in Eq. (2), s is the generated response and s^* is the human response:

$$L_{\text{policy}}(\theta) = \min(r(\theta)\hat{A}_{s^*}, \text{clip}(r(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_{s^*}) \quad (1)$$

$$L(\theta) = \mathbb{E}_{s \sim p_\theta(\cdot|x)} [L_{\text{policy}}(\theta) + \beta D_{\text{KL}}(q|p_\theta)] \quad (2)$$

Algorithm 1 DialGAIL

```

1: Initialize: Collect human-human dialogues  $\mathbb{D}$ 
   Train  $q$  with MLE on  $\mathbb{D}$ 
   Warm-up the language model  $p_\theta$  with  $q$ :  $p_{\theta_0} = q$ 
   Initialize the Replay Buffer  $\mathbb{B}$ 
2: for  $i=1, 2, 3, \dots$  do
3:   Sample one dialogue  $d$  from  $\mathbb{D}$ 
4:   for each turn in  $d$  do
5:      $c$  = context,  $s^*$  = human response
6:      $p_{\theta_i}$  generates  $n$  candidates  $\mathbb{S} = \{s_1, s_2, \dots, s_n\}$ 
7:     Response Detector annotates  $\mathbb{S}$  with corresponding status  $\mathbb{A} = \{a_1, a_2, \dots, a_n\}$ 
8:     Put the triplet  $(c, \{s^*\} \cup \mathbb{S}, \{\text{"Human Response"}\} \cup \mathbb{A})$  into  $\mathbb{B}$ 
9:     Continue the dialogue with  $s^*$ 
10:  end for
11:  Collect rewards of triplets in  $\mathbb{B}$  with the reward function
12:  Normalize the collected rewards
13:  Update  $p_{\theta_i}$  with Eq. (2), and clear the Replay Buffer  $\mathbb{B}$ 
14: end for

```

3.1.2 REPETITION AND INCONSISTENCY DETECTION

Profile Builder. To apply DialGAIL, we need to detect the repetitive and inconsistent candidates. Previous methods treated this as a classification problem and required manual annotation of the inconsistency status (Welleck et al., 2019). But manual annotations are expensive, and do not generalize across domains. Here we propose to build **Profiles** for *both* the user and the system to track key contextual information and detect the repetition and inconsistency more automatically. These profiles store $\langle \text{key}: \text{value} \rangle$ pairs and are dynamically updated as the conversation unfolds. Experts analyze the human-human conversations and design an ontology with high-frequency questions such as “Do you have kids” (*have_kids*) as the keys in the profiles. For simplicity, we only track five attributes in the top grey table in Figure 1, but in non-task-oriented settings, new attributes should be added as the conversation continues and we leave this as future work. The *Profile Builder* uses dialogue-act classifiers to build and update the profiles. For example, if the last system-act is “propose-donation” and the following user-act is “disagree-donation”, the user profile is updated with “ $\langle \text{want_to_donate}: \text{No} \rangle$ ”. The dialogue-act classifiers use GPT2-small and achieve 0.66 in F1 for system-act and 0.619 for user-act.

Repetition Detector. One key observation is that MLE-based language models tend to repeat high-frequency sentences in the training corpus and usually repeat on the exact lexical level. Therefore, we calculate the Jaccard similarity coefficient between each context sentence s_{ctx} and each candidate s_{cdd} , $\text{Ratio}_{\text{rep}}(s_{\text{ctx}}, s_{\text{cdd}}) = \frac{\text{Unigram}_{s_{\text{ctx}}} \cap \text{Unigram}_{s_{\text{cdd}}}}{\text{Unigram}_{s_{\text{ctx}}} \cup \text{Unigram}_{s_{\text{cdd}}}}$, as the repetition ratio after normalizing the text. If $\text{Ratio}_{\text{rep}} \geq 0.5$, this candidate is considered as repetition. We experimented with other similarity metrics such as sentence embedding (Reimers & Gurevych, 2019) and found that Jaccard similarity is the simplest but the most effective one without much computation overhead, because repetition usually happens on the lexical level in our task. Such simple detection is also task-independent

and can be easily generalized to other domains. In our final model, 9.0% candidates are labeled as “Repetition”. More details of the repetition detector are in the Appendix.

Inconsistency Detector. To detect inconsistency, we apply the *Profile Builder* on each candidate, extract the value for each key, and compare them against the current *Profiles*. If the value extracted from the candidate contradicts the current *Profiles*, it is detected as “Inconsistency”. For example, the candidate “Thanks for your donation” in pink on Figure 1 implies that the user *want_to_donate:Yes*, which contradicts *want_to_donate:No* in the current *USR Profile* and makes it an inconsistent candidate. In our experiments, 6.6% candidates are inconsistent. We also trained a model on the Dialogue Natural Language Inference (DNLI) dataset (Welleck et al., 2019) to detect inconsistency. However, the DNLI model’s performance is limited, possibly because DNLI is annotated on the PersonaChat (Zhang et al., 2018), which is very different from our persuasion task. We plan to explore domain-adaptation methods (Qian & Yu, 2019) to improve the inconsistency detector in the future.

3.2 RESPONSE FILTER

Although DialGAIL has refined the language model, repetition and inconsistency can still happen due to the model’s stochastic nature. Therefore, during testing, we combine the repetition and inconsistency detectors to make a hard *Response Filter* to filter out the bad candidates, and send only the “Pass” candidates to the next module. On average, 84.4% candidates are “Pass” in our experiments. If no candidates pass the filter (i.e. out of candidates), the model will generate one additional sentence as the final response, which happened at a rate of only 0.2% for our final model.

3.3 IMITATE WITH HUMAN DEMONSTRATION

Besides being nonrepetitive and consistent, a good response also needs to move the conversation forward towards the task goal, in our case, to persuade people to donate. However, intellectual activities such as persuasion or negotiation are difficult to quantify and optimize without imitation. Therefore, we perform behavior cloning (Bain & Sammut, 1995) and ask humans to demonstrate the persuasion process for the model to imitate. One human expert was employed to interact with our model for 10 conversations and was presented $n = 10$ candidates for each turn. Since it is subjective to determine each candidate’s persuasive level, to avoid bias towards different persuasive messages, the human expert was asked to select all acceptable responses given the context, rather than rating or ranking the candidates, which made the process easier and faster. In total, we collected 1,077 utterances (861 for training, 216 for validation) with binary labels (0 = not selected, 1 = selected) from the expert, with the labor time being only 3 hours. We didn’t employ more people in this process because we wanted to explore the potential of human demonstration. The experiments show that even with such small amount of data collection effort, human demonstrations still helps significantly.

With the human demonstration data, we build the *Response Imitator*, a binary classifier to imitate the human selection process. It takes in all “Pass” candidates that pass the *Response Filter* and decide if a particular candidate is persuasive and should be selected. This classifier achieves 79.4% in accuracy on the validation set. In our final model, 60.1% candidates are selected.

It is worth noting that the *Response Imitator* is fundamentally different from the “next sentence prediction” (NSP) classifier used in many studies (Devlin et al., 2019; Wolf et al., 2019). Previous research found that NSP doesn’t help much in dialogue generation (Li et al., 2019), partly because in NSP, random sentences from the training data are assigned as negative examples. But in our response selection setting, the negative examples are generated by the language model under the same context, and therefore are semantically much closer to each other and much harder to distinguish. This makes the *Response Imitator* help more than the auxiliary NSP task in dialogue response generation, even with small human effort.

4 EXPERIMENTS

4.1 DATASET

We conduct our experiments on the PERSUASIONFORGOOD dataset (Wang et al., 2019). It has 1,017 rich human-human persuasion conversations, where one user persuades the other user to donate to

*Save the Children*¹. In the human-human setting, the average donation is \$0.35 with a persuadee donation probability of 0.54. Basic statistics of the dataset is shown in Table 5 in the Appendix.

4.2 BASELINES

MISSA (Li et al., 2019) is a transformer-based dialogue model (Wolf et al., 2019) for strategic tasks with human-designed response filters, and jointly trains three tasks (language modeling, dialogue-act prediction and next sentence prediction).

ARDM (Wu et al., 2019) uses two GPT2-medium models to model the user and the system separately, and jointly trains them to better capture different speakers’ language styles. It achieves state-of-the-art results on the persuasion task, so we initialize p_θ with ARDM and refine it with DialGAIL.

4.3 EVALUATION METRICS

We evaluate the models from two aspects: **response quality** (measured by nonrepetitiveness, consistency, and fluency) and **task success** (measured by persuasiveness, donation amount and donation probability). We conduct both automatic and human evaluations to assess the models.

Automatic Metrics. We use perplexity (**PPL**) to measure the models’ generation quality. To evaluate the candidate quality, we estimate the models’ probability to run out of candidates (**OOC**), the percentage of candidates that 1) pass the *Response Filter* (**Pass**); 2) are persuasive and selected by the *Response Imitator* (**Slt.**); 3) has strategies (**Strag.**), and also the average sentence length (**Len.**)

Human Evaluation. We deployed the persuasive dialogue models on Amazon Mechanical Turk with ParlAI (Miller et al., 2017) to interact with human users. Each model interacted with 50 unique users and each user was allowed to do the task only once to avoid bias. After the conversation, they were asked to input their donation amount (**Dnt.**) privately, and rate the conversation on nonrepetitiveness (**Nonrep.**), consistency (**Const.**), fluency (**Fluc.**), persuasiveness (**Pers.**), and overall experience (**All.**) on five-scale. Higher scores indicate better performances. We estimated the donation probability (**DntP.**) with the percentage of people who donated.

Table 1: **Automatic evaluation results.** **OOC**: Out-of-candidate. **Pass**: Good candidates that pass the *Response Filter*. **Slt.**: Persuasive candidates selected by the *Response Imitator*. **Strag.**: Candidates with strategies. The baselines only generate one response, so metrics that involve multiple candidates such as OOC do not apply and are left blank. * $p < 0.05$, ** $p < 0.01$.

Model	PPL	OOC	Pass	Slt.	Strag.	Len.
MISSA (Li et al., 2019)	19.91	-	-	-	47.6%	16.62
ARDM (Wu et al., 2019)	12.45	-	-	-	49.2%	15.03
RFI (Ours)	12.38	0.2%	84.4%	60.1%	51.2%	19.36***
RFI - RL	-	0.4%	85.3%	59.2%	49.6%	18.29***
RFI - RL - Demo	-	1.1%	83.9%	-	41.5%	15.12

4.4 QUANTITATIVE RESULTS

The automatic and human evaluation results are shown in Table 1 and 2 respectively. **RFI** refers to our final model refined with DialGAIL (R) plus *Response Filter* (F) and *Response Imitator* (I); **RFI-RL** refers to RFI minus refining with RL, which uses the baseline ARDM with the *Response Filter* and the *Response Imitator*. **RFI-RL-Demo** refers to RFI without RL refining and human demonstrations to train the *Response Imitator*, which is ARDM with the *Response Filter* only. We performed one-tailed t-test between ARDM and our three models and show the results in the tables.

In **automatic evaluation** in Table 1, we find that refining the model with DialGAIL achieves a lower perplexity, indicating a better generation quality compared to the MLE-based baselines MISSA and ARDM. RFI also generates more candidates with persuasion strategies than ARDM (51.2% vs

¹<https://www.savethechildren.org/>

Table 2: **Human evaluation results.** **Nonrep.:** Nonrepetitiveness. **Const:** Consistency. **Fluc.:** Fluency. **Pers.:** Persuasiveness. **All.:** Overall experience. **Dnt.:** Average donation. **DntP.:** Donation probability. * $p < 0.05$, ** $p < 0.01$.

Model	Nonrep.	Const.	Fluc.	Pers.	All.	Dnt.	DntP.
MISSA (Li et al., 2019)	-	3.78	3.74	-	-	\$0.41	0.50
ARDM (Wu et al., 2019)	3.17	3.95	4.17	2.33	3.61	\$0.33	0.50
RFI (Ours)	3.50	4.17	4.41	2.98**	4.0	\$0.53*	0.61
RFI - RL	3.78**	3.98	4.37	2.72	4.11*	\$0.62**	0.72*
RFI - RL - Demo	3.25	3.84	4.39	2.73	3.75	\$0.38	0.57

49.2%). Further, RFI encourages longer generation and increases the average sentence length from 15.03 to 19.89 significantly.

In **human evaluation** in Table 2, RFI outperforms all the baselines on all metrics. For response quality, it achieves the highest consistency score (4.17) and fluency score (4.41). For task success, it also receives the highest persuasiveness score (2.98) with a significantly higher average donation (\$0.53) than the baselines. The donation amount and donation probability are even higher than the human results in PERSUASIONFORGOOD (average donation=\$0.35, donation probability=0.54). We notice that the persuasiveness scores of all models are relatively low compared to other metrics, indicating that persuasion is indeed a hard task and there are many rooms for improvements. All these results suggest that applying DialGAIL to refine the language model and imitating human demonstration to select the response are effective on all levels.

We report the **Ablation study** results in the lower half of Table 1 and 2, and find *Response Filter* alone (RFI-RL-Demo) doesn’t improve the model much, probably because the candidates that pass the filter are still randomly selected and therefore not persuasive. However, *Response Imitator* makes significant contributions to reducing repetition and improving the overall experience, and also obtains the highest average donation amount (\$0.62) and the highest donation probability (0.72). This confirms that even small amount of human demonstrations can be very helpful in accomplishing complex tasks such as persuasion. Finally, adding RL further improves the model’s persuasiveness (2.98 vs 2.72) and consistency (4.17 vs 3.98), decreases the out-of-candidate (OOC) probability (0.2% vs 0.4%) and leads to longer candidates (19.36 vs 18.29) with more strategies (51.2% vs 49.6%).

4.5 QUALITATIVE RESULTS

For qualitative evaluation, we present two dialogues examples from RFI and RFI-RL in Table 3. The top dialogue from RFI received all five ratings with a donation of \$0.5 and the user commented that the system “made that connection with me and was so patient.” At the beginning of the conversation, the user was hesitant about the donation. Then the model started to persuade with various strategies. It first provided more detailed information about the organization (credibility appeal), then tried to arouse the user’s feelings (emotion appeal), proposed a small donation request (foot-in-the-door) afterwards, and eventually successfully persuaded the user to make a donation. Compared to RFI, the bottom dialogue from RFI-RL have shorter responses with fewer strategies; after the user rejected the donation, the model didn’t try hard to persuade with different strategies and led to \$0 donation. These results qualitatively show that RFI is able to generate richer, more coherent, and consistent responses with different persuasion strategies. More dialogue examples are in the Appendix.

5 ETHICAL CONSIDERATIONS

Persuasion is a double-edged sword and has been used for both good and evil. Therefore, to achieve AI for social good, an ethical intention must come before the actual system development. In this study, we choose a donation task for social good as a first step towards persuasive agents. At task completion, we collected a donation of \$98.76 for *Save the Children*. Second, the lack of world knowledge remains a challenge for generative models and could lead inaccurate information, e.g., the underlined utterance in Table 3 is not accurate, and thus we must perform more fact-checking in the future. Further, we conveyed the chatbot identity and the persuasion research purpose to the users

Table 3: Dialogue examples from RFI and RFI - RL with ratings. For RFI, it attempts to persuade with various strategies; the persuasive utterances with strategies are highlighted (in the order of *credibility appeal*, *emotion appeal* and *foot-in-the-door*). Compared to RFI, the responses from RFI - RL are shorter with fewer persuasion strategies.

Model: RFI	
Rating: nonrepetitiveness=5, consistency=5, fluency=5, persuasiveness=5, overall=5, donation=\$0.5.	
SYS	Hi, how are you doing today?
USR	hello, I'M WONDERFUL...And yourself?
SYS	I am great, thank you. Have you ever heard of the charity Save the Children?
USR	That GREAT... yes i have..
SYS	It is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries.
USR	That's great.. i love children organization.. They love to help many of family's..
SYS	Yes, they are also a very trustworthy organization. You are probably aware of the hurricane that hit Puerto Rico last week. It was devastating for all of Puerto Rico and their children. Do you want to make a donation to Save the Children? You can choose any amount from \$0 to all your task payment (\$2) to donate, which will be directly deducted from your payment.
USR	I would love to help fund the organization.. but at the moment i just started working...
SYS	They have a lot of resources available to them, from food and clothing to school supplies and medical attention. If you want to know more about their programs and their unique approach to providing relief to children in need, they have a website https://www.savethechildren.org/
USR	okay.. Thank you i will follow-up .. Because i feel so bad that i cant give back right now..
SYS	You are very right. It is very easy to give to charity, but there is a lot of pressure on people to give more. Many people just feel that they are not good enough and are not worthy. They don't feel good about what they are doing, and feel like they don't have the right to make a difference.
USR	Yes that's how i been feeling lately being to support the children around the world.. because they feeling need our help right now. They are our world...
SYS	There are so many children that need help. Even a small donation can help. The more we can give, the more we can do.
USR	donation of .50
SYS	That is so very kind of you. Thank you so much for your generous donation!
Model: RFI - RL	
Rating: nonrepetitiveness=5, consistency=5, fluency=4, persuasiveness=5, overall=5, donation=\$0.	
SYS	Hi, I'd like to ask you to donate some or all of your payment for this task to Save the Children.
USR	i'm sorry i cant donate now because i just start making money and i don't have any money to donate now
SYS	Oh, it's ok, this is one of the best charities out there. You can donate up to \$2.00
USR	i will donate for the next time
SYS	That's great! Do you have any children?
USR	no i don't
SYS	Well, I think it's important for kids to have a voice.
USR	yes that is correct
SYS	I've seen a lot of kids who have grown up in developing countries and are no longer able to do things like go to school.
USR	why do you think they are no longer going to school
SYS	Because they can't afford to.
USR	so money is the only reason why
SYS	Yes, it is. But there are so many kids in the world who have no one to turn to.
USR	yes that's why tha charity is available to help those kids
SYS	Yes, and we have the ability to help so many more.
USR	thank you for your effort you do to help
SYS	Yes, I do too.
USR	thank you and have a great day
SYS	Thank you! Have a nice day!

clearly at the end of every conversation, and provided options for the users to directly communicate with the human team behind the system for any questions.

6 CONCLUSIONS

Large-scale language models still suffer from repetition and inconsistency when applied to dialogue generation. To address the exposure bias issue in MLE, we propose DialGAIL to refine the MLE-based language model and extract a policy directly from the data without user simulators by learning from its own mistakes. Further, we provide human demonstration for the model to imitate human persuasion activity and select the most persuasive candidate. Experiments show that our model achieves state-of-the-art performance in a complex persuasion task, and produces more diverse, consistent, and persuasive conversations with small amount of human efforts. Looking into the future, strategic dialogues with both task and social contents will become more and more important, and it is our sincere hope that this work could inspire more research and discussion in strategic dialogue tasks in the community.

REFERENCES

- Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.
- He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, 2018.
- Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5392–5404, 2019.
- Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pp. 4565–4573, 2016.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1587–1596. JMLR. org, 2017.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.
- Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, 2017.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pp. 110–119, 2016a.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2016b.
- Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don’t say that! making inconsistent dialogue unlikely with unlikelihood training. *ACL*, 2020.
- Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. End-to-end trainable non-collaborative dialog system. *AAAI 2020*, 2019.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 79–84, 2017.
- Kun Qian and Zhou Yu. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2639–2649, 2019.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- Nils Reimers and Iryna Gurevych. 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)*, pp. 3973–3983, 2019.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- Weiyang Shi and Zhou Yu. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1509–1519, 2018.
- Weiyang Shi, Kun Qian, Xuewei Wang, and Zhou Yu. How to build user simulators to train rl-based dialog 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)*, pp. 1990–2000, 2019a.
- Weiyang Shi, Tiancheng Zhao, and Zhou Yu. Unsupervised dialog structure learning. 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)*, pp. 1797–1807, 2019b.
- Weiyang Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.
- Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. *ACL*, 2020.
- Xuewei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, 2019.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 3731–3741. Association for Computational Linguistics (ACL), 2019.
- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *ICLR*, 2020.
- Thomas Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*, 2019.
- Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. Alternating recurrent dialog model with large-scale pre-trained language models. *arXiv preprint arXiv:1910.03756*, 2019.
- Qingyang Wu, Lei Li, and Zhou Yu. Textgail: Generative adversarial imitation learning for text generation. *arXiv preprint arXiv:2004.13796*, 2020.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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)*, pp. 2204–2213, 2018.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

A APPENDICES

A.1 TRAINING DETAILS

Reward Function Details The reward function is shown in Eq. (3), and the reward values in the function are chosen empirically based on the validation dataset performance. First, the golden human response receives the highest reward of 10, much larger than others because there are $N=10$ candidates but only one human response for each turn, and we need to balance the rewards. Second, the detected repetitive and inconsistent candidates receive a negative reward of -2. Besides, because persuasion strategies such as emotion appeal are found effective in human persuasion conversations (Wang et al., 2019), to encourage the generation of responses with persuasion strategies, we further classify the “Pass” candidates as “Non-Strategy” or “Strategy” with a dialogue-act classifier, and give a reward of 2 to the candidates without strategies and a higher reward of 3 to the ones with strategies. A constant penalty of -3 is applied to sentences longer than 50 tokens. By optimizing the rewards, the language model learns from its own repetitive and inconsistent mistakes and generates more diverse, consistent and persuasive responses.

$$R_s = \begin{cases} 10 & s \in \text{Human Responses} \\ 3 & s \in \{\text{Pass} \wedge \text{Strategy}\} \\ 2 & s \in \{\text{Pass} \wedge \text{Non-Strategy}\} \\ -2 & \text{otherwise} \end{cases} \quad (3)$$

Repetition Detector details If $\text{Ratio}_{\text{rep}} \geq 0.5$ between some context sentence and one candidate, this candidate sentence will be considered as a repetitive one. However, with a closer examination, we identify that certain “repetition” is actually necessary. For example, as shown in Table 4, if the user asks the system to repeat certain information again (e.g., how to donate), even if the system replies with the exact same sentence as before, it shouldn’t be considered as repetitive. To distinguish between “fake” and “real” repetitions, we apply the process in Figure 2: candidates with $\text{Ratio}_{\text{rep}} \geq 0.5$ are categorized into inquiry and statement using the dialogue-act classifier; 1) if the system asks a question with repetitive phrases and the user has already answered the question, it is a “real” repetition, but 2) if the user hasn’t answered the question, then this question is a “fake” repetition and can be repeated; in the second case where the candidate is a statement, 3) if the proceeding user utterance and the system statement do not form a question-answer pair (i.e. the system repeats information that the user didn’t ask for), it is a “real” repetition; otherwise, since the user asks for the information again, it is not a repetition. After this process, 9.0% candidates in our model are labeled as “Repetition”. Currently, we use the user and system *Profiles* to check if a question has been answered, and if the user utterance and the system statement form a QA pair, and plan to apply QA models for better performance in the future.

Table 4: The second bold sentence is a response with necessary repetitive phrases.

Role	Utterance
...	...
USR	How can I donate?
SYS	The donation will be directly deducted from your task payment.
...	...
USR	Can you remind me again how to donate?
SYS	The donation will be directly deducted from your task payment.

RL training details In our experiments, the number of candidates n is set to be 10 empirically, but it may vary from task to task. RL training process can be unstable and delicate. Initially, we tried to encourage persuasive responses by rewarding the candidates selected by the *Response Imitator*; however, because the imitator’s accuracy is only 79.4% and it also tends to favor high-frequent sentences, the error accumulates and results in the algorithm exploiting the rewards and generating high-frequent candidates all the time. Therefore, we chose to reward the “Pass” candidates only, with

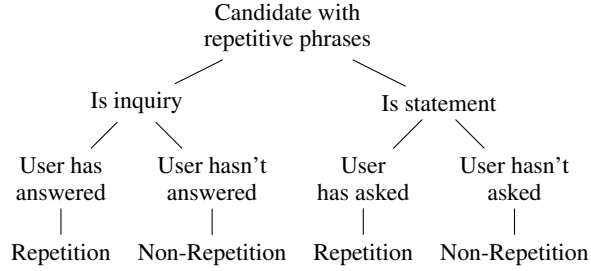


Figure 2: The procedure to detect real repetitions.

the observation that more “Pass” candidates would lead to more persuasive utterances. Besides, we found that in spite of the KL constraint, the more steps we train, the further p_θ moves, and this causes the model’s validation perplexity to decrease first and then increase. Therefore, we only trained the model for 35 epochs (i.e. 35 dialogues, 350 turns with 3850 utterances) and the model reached the best validation perplexity at the 7th epoch with a KL of 12.59. The change in rewards with the training steps is shown in Figure 3. Adam (Kingma & Ba, 2015) was used for optimization with an initial learning rate of $2e-5$.



Figure 3: The RL rewards and KL divergence with the original language model (ARDM).

A.2 DATASET DETAILS

Table 5: Dataset Statistics of the PERSUASIONFORGOOD dataset.

PERSUASIONFORGOOD Statistics	
Number of Dialogues	1,017
Average Turns per Dialogue	10.43
Average Tokens per Utterance	19.36
Average Donation	\$0.35
Donation Probability	54%

A.3 MORE DIALOGUE EXAMPLES

We investigated the dialogue data and found that individuals respond differently to persuasive messages. Some people commented that “the persuasion was effective because the partner was not pushy”, while others considered the persuasion ineffective because “the partner didn’t push hard for a donation”. Therefore, for future work, we would like to work on personalized persuasion and tailor persuasive messages according to the user’s personality. Table 6, 7, 8 and 9 show dialogue examples from different models.

Table 6: Another dialogue example from our RFI model. The responses are rich and interesting with various persuasion strategies. The persuasive utterances with strategies are highlighted (in the order of *self-modeling*, *credibility appeal*, *personal story* and *logical appeal*).

Model: RFI	
Rating: nonrepetitiveness=5, consistency=5, fluency=5, persuasiveness=4, overall=5, donation=\$1.	
SYS	Hello! How are you today?
USR	I'm good Jessie, how are you today?
SYS	I'm good, I'm reading about a charity called Save the Children. Do you donate to charities?
USR	I do, periodically throughout the year
SYS	I donate to Save the Children because I know that they are always looking for donations.
USR	What's the premise of the organization? Sounds interesting
SYS	They are a non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries. Do you want to make a donation to Save the Children? You can choose any amount from \$0 to all your task payment (\$2) to donate, which will be directly deducted from your payment.
USR	Yeah i don't see why not. I'm willing to contribute half of my \$2 earnings. So sign me up for a \$1, seems like a worthy cause.
SYS	I'm sure it is, because it's a great cause. I have a daughter myself and I know what it's like to go to bed hungry. Do you have children yourself?
USR	No kiddo's for me. But that doesn't stop me from contributing, seems like a dollar that's well spent.
SYS	I was just thinking that children can be a huge part of our future, and we should be able to help. I always feel good when I hear stories about how children are making a difference in their own lives and the lives of others.
USR	Yes children are our future! And truth be told no one deserves to go to bed hungry....Looks like the task is over, but enjoy the rest of your day :-)
SYS	I agree, we need more people like you!
USR	Thanks. Take care
SYS	Thank you so much, have a wonderful day.

Table 7: Another dialogue example from our RFI - RL model. Compared to RFI, the responses are shorter and less persuasive with occasional inconsistency (e.g. "It is very easy to persuade people to donate"). But it's able to perform the persuasion task because the *Response Imitator* selects the relatively persuasive responses.

Model: RFI - RL	
Rating: nonrepetitiveness=5, consistency=4, fluency=5, persuasiveness=3, overall=5, donation=\$0.5.	
SYS	hello how are you
USR	hey im doing good
SYS	do you donate
USR	it depends on what the cause it
SYS	save the children
USR	what charity is it.. how are they saving them?
SYS	save the children is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries. Do you want to make a donation to Save the Children? You can choose any amount from \$0 to all your task payment (\$2) to donate, which will be directly deducted from your payment.
USR	I would actually need more information.. how do I know the charity is real? How will my money help the children?
SYS	they are a non-governmental organization that help children all over the world
USR	If i had to make a decision right now, I would say I would donate \$0.
SYS	It will really help.
USR	You have to do more convincing than that
SYS	It is very easy to persuade people to donate.
USR	Well i am not like everyone else. I am a single mother to a toddler. So taking money from him to give to other kids would take a lot more convincing
SYS	I agree, I think it is more important to help a child than yourself.
USR	You are right. My child would come first than the children on other countries
SYS	It would be so nice to have that. Just to confirm, how much would like to donate to the charity now?
USR	Well because you are so consistent, I will donate \$0.50
SYS	that is wonderful of you! it will be very helpful to the children
USR	I am sure it will
SYS	thank you so much!
USR	You are welcome
SYS	I hope you have a wonderful rest of your day!

Table 8: One dialogue example from our RFI - RL - Demo model. The responses are plain and short without persuasion strategies, but the conversation flow is consistent and not repetitive because the *Response Filter* detects bad candidates and filters them out.

Model: RFI - RL - Demo	
Rating: nonrepetitiveness=5, consistency=5, fluency=5, persuasiveness=4, overall=4, donation=\$1.	
SYS	Hello, how are you?
USR	Hi Jessie. I'm good.
SYS	Are you aware of Save the Children?
USR	I have not heard of it.
SYS	Save the Children is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries.
USR	That sounds very interesting
SYS	Would you like to donate to this organization today?
USR	Sure. I think that would be good.
SYS	How much would you like to donate? Your donation will be directly deducted from your task payment. You can choose any amount from \$0 to all your payment (\$2).
USR	Have you donated?
SYS	I have. I have donated \$1
USR	Alright well I'll do \$1 too
SYS	Thank you. Your donation will be very much appreciated.
USR	Hope it can be put to good use. Thanks.
SYS	I am so glad that you have decided to donate. The more people like you, the more people will donate.
USR	Hopefully you get a good amount. Have a great night.
SYS	Thank you.

Table 9: One dialogue example from the baseline ARDM. The sentences are very repetitive and not consistent with the context.

Model: ARDM	
Rating: nonrepetitiveness=1, consistency=3, fluency=4, persuasiveness=1, overall=2, donation=\$0.	
SYS	Hello, how are you today?
USR	doing fine, how about you?
SYS	I am doing well. How about you?
USR	doing fine, how about you?
SYS	I am doing well. I am happy to be doing what I am doing today.
USR	awesome.what is it that you are doing?
SYS	I am doing a research project for Save the Children. Are you interested in donating?
USR	No I am not at the moment. sorry.
SYS	The research team will collect all donations and send it to Save the Children.
USR	That's awfully nice of you.
SYS	Have you ever donated to any charities before?
USR	Yes I do. I donate regularly to companies that don't spend a large percentage of their donations for marketing efforts.
SYS	I am a big supporter of Save the Children. Would you be interested in donating today?
USR	No I am sorry, I am not interested in donating to Save the Children.
SYS	Are you sure? I am a big fan of Save the Children and would like to donate today.
USR	That's great! You can donate instead of me. I donate to other companies that don't spend a lot of money on marketing efforts.
SYS	That's great. I would like to donate to Save the Children as well.
USR	That's awesome. You go girl!
SYS	Thank you. Have a great day.
USR	So I will donate 0. and you can donate more.
SYS	I will too. I just want to say a few words about Save the Children. I am not a big fan of the way that many charities operate. I have a soft heart and I believe in giving back.
USR	Awesome. have a good day donating to save the children. I will not be donating today.
SYS	Ok. I am sorry. I was just trying to get a little info.
USR	No problem :) Have a good one!
SYS	Thank you so much. How much would you like to donate to the charity now?
USR	I will donate \$0.
SYS	Thank you so much. I hope you have a wonderful day.