Deep Reinforcement Learning with Hierarchical Action Exploration for Dialogue Generation

Itsugun Cho[1], **Ryota Takahashi**[1], **Yusaku Yanase**[1], **Hiroaki Saito**[1]
Keio University, Japan[1]
{choitsugun, ryota.0226.tokky, y.y32851107}@keio.jp
hxs@ics.keio.ac.jp

Abstract

Traditionally, approximate dynamic programming is employed in dialogue generation with greedy policy improvement through action sampling, as the natural language action space is vast. However, this practice is inefficient for reinforcement learning (RL) due to the sparsity of eligible responses with high action values, which leads to weak improvement sustained by random sampling. This paper presents theoretical analysis and experiments that reveal the performance of the dialogue policy is positively correlated with the sampling size. To overcome this limitation, we introduce a novel dual-granularity Q-function that explores the most promising response category to intervene in the sampling process. Our approach extracts actions based on a grained hierarchy, thereby achieving the optimum with fewer policy iterations. Additionally, we use offline RL and learn from multiple reward functions designed to capture emotional nuances in human interactions. Empirical studies demonstrate that our algorithm outperforms baselines across automatic metrics and human evaluations. Further testing reveals that our algorithm exhibits both explainability and controllability, as well as generates responses with higher expected rewards.

Keywords: Dialogue Generation, Reinforcement Learning, Dual-granularity Q-function

1. Introduction

To ensure a satisfactory user experience, an intelligent dialogue agent is required to respond fluently and naturally while being endowed with a “forward-looking” capacity in the dialogue. A predominant approach to training agents is to optimize the maximum likelihood estimation (MLE) objective for the probability distribution of responses. However, this supervised technique is insufficient to learn a long-term behavior since the corpus often contains suboptimal dialogues, and MLE cannot model the future direction of the conversation. Instead, if we view the open-domain dialogue as a control problem, RL could enable agents to automatically adjust policy concerning the pre-defined appraisal functions via a trial-and-error process.

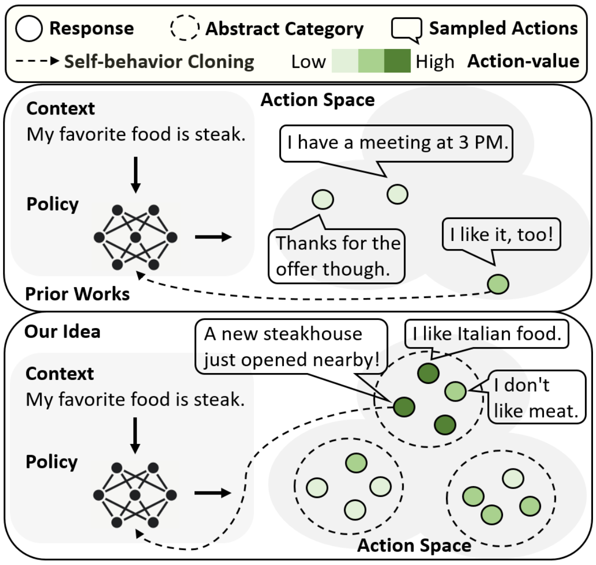


Figure 1: Comparison of our idea with prior work.

Recent work in RL for dialogue generation is well summarized by Lone et al. (2022). Most prior studies build dialogue agents using the actor-critic framework, optimizing the policy with the support of N generated possible responses from the agent to maximize the action-value function (i.e., Q-function). Although this self-behavior cloning avoids reliance on policy gradients, thereby preventing divergence from human language, it often suffers from slow improvement and falls into a trivial local optimum due to its indiscriminate treatment of the action space. We argue that the actions that can make the Q-function produce a higher value (i.e., Q-value) in a given state will be similar at an elevated abstraction rank. For example, in a conversation about favorite food, responses about restaurants are more likely to get higher cumulative expected rewards than those about business. If we apprehend which abstract category of actions can achieve a higher Q-value, generating responses from that category for greedy improvement will make the training policy more efficient. A simple illustration depicting this view is provided in Figure 1. To this end, we propose a dual-granularity Q-function to evaluate the Q-value associated with an action under different

LREC-COLING 2024, pages 4566–4579
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

levels of abstraction. Specifically, our algorithm includes a coarse-grained Q-function based on category-represented responses, which aims to lock in the optimal category, and a fine-grained Q-function based on token-represented responses, which strives to extract the optimal action. In this manner, the infinite action space is divided into several blocks at the high-level abstraction, enabling exploration of the entire action space to adapt policy on the fly. Since RL requires numerous costly interactions with the environment (i.e., real users), we applied offline RL to our algorithm, which leverages the previously collected English dataset DailyDialog (Li et al., 2017) for policy training. Moreover, inspired by the psychology of human conversation, four reward functions are devised to improve the agent's ability to engage in natural dialogue. Experimental results demonstrate that by training with our algorithm, four state-of-the-art dialogue models achieve a significant performance improvement. The controllability and effectiveness of our approach are clarified in the discussion. Our main contributions in this study are two-fold:

(1) To the best of our knowledge, this is the first attempt to implement offline RL using different-grained representations of natural language. This approach provides a unified algorithmic template suitable for tasks with large action spaces.

(2) The quantitative and qualitative empirical verifications have established that our approach exhibits a high level of trustworthiness.

2. Methodology

2.1. Preliminaries

We begin with a Markov decision process represented by a tuple $M = (S, A, T, R, \gamma)$, where S is the state space, A is the action space, T is the state transition function, R is the reward function, and $\gamma \in (0, 1)$ is a discount factor. In the dialogue setting, the agent observes a context s, executes its policy π, by generating a response a, according to $\pi(a|s)$, transitions to a new context s', and receives a reward $r = R(s, a)$. The goal is to learn π to maximize cumulative reward from a dataset $\mathcal{D}$, which consists of multiple (s, a, r, s') pairs produced under a potential behavior π_β. Therefore, prior works typically rely on the actor-critic style that alternates between fitting Q-function by the policy evaluation based on approximate dynamic programming (i.e., iterating the Bellman operator via minimizing the temporal difference error) and improving π by updating it toward responses that maximize the expected Q-value.

$$\begin{gathered}\text{Evaluation} : Q \leftarrow \arg\min_{Q} \mathbb{E}_{(s,a,r,s') \backsim \mathcal{D}} \\ [(r + \gamma \mathbb{E}_{a' \backsim \pi(a'|s')}[Q(s', a')] - Q(s, a))^2].\end{gathered} \tag{1}$$

$$\begin{gathered}\text{Improvement} : \pi \leftarrow \arg\max_{\pi} \mathbb{E}_{s \backsim \mathcal{D}, a \backsim \pi(a|s)} \\ [Q(s, a)].\end{gathered} \tag{2}$$

A challenge in implementing offline RL is that static $\mathcal{D}$ has limited coverage of S and A, whereby π may be biased toward out-of-distribution (OOD) actions for π_β with erroneously high Q-value (Fujimoto et al., 2019; Kumar et al., 2020; Kostrikov et al., 2021). Hence, we follow Jang et al. (2022) to employ the one-step algorithm (Brandfonbrener et al., 2021) for on-policy evaluation, as shown below, which can iterate in a rather stable manner since actions are always in $\mathcal{D}$ to avoid the OOD due to distribution deviations between π and π_β.

$$\begin{gathered}\text{Evaluation} : Q \leftarrow \arg\min_{Q} \mathbb{E}_{(s,a,r,s',a') \backsim \mathcal{D}} \\ [(r + \gamma Q(s', a') - Q(s, a))^2].\end{gathered} \tag{3}$$

2.2. Dual-granularity Q-function

Traditionally, to implement the $arg\,max$ operator in Eq.(2), a set of responses is sampled by $\pi(a|s)$, and π is updated based on the one that yields the highest action value according to the Q-function. We can show that the renewed policy by more responses has a higher state value (i.e., a better performance). When the sampling size is large enough to cover the entire action space, π can theoretically iterate to the optimum. We also show that the renewed policy by responses with a higher Q-value has a higher state value. We formalize the results in Theorem 1 and 2, respectively. The detailed proofs are presented in Appendix A.

Theorem 1. Given a policy π and the number of sampled actions L, if we update the new policy by

$$\forall s,\ \pi'_L = \mathop{\arg\max}_{a \in \{a_i\}_{i=1}^{L} \backsim \pi(a|s)} Q^\pi(s, a),$$

then for any N, M, such that $N \geq M \geq 1$, $\forall s$, $V^{\pi'_N}(s) \geq V^{\pi'_M}(s)$ always holds.

Theorem 2. Given the policy π_α, π_β, and π, s.t. $\mathbb{E}_{a \backsim \pi_\alpha(a|s)}[Q^\pi(s, a)] \geq \mathbb{E}_{a \backsim \pi_\beta(a|s)}[Q^\pi(s, a)]$, if the number of sampled actions is L, and we update the new policy by

$$\begin{aligned}\forall s,\ \pi'_1 &= \mathop{\arg\max}_{a \in \{a_i\}_{i=1}^{L} \backsim \pi_\alpha(a|s)} Q^\pi(s, a),\\ \forall s,\ \pi'_2 &= \mathop{\arg\max}_{a \in \{a_i\}_{i=1}^{L} \backsim \pi_\beta(a|s)} Q^\pi(s, a),\end{aligned}$$

then $\forall s$, $V^{\pi'_1}(s) \geq V^{\pi'_2}(s)$ always holds.

Since it is impractical to exhaust all possible responses, we focus on constructing the sampling process in a more organized manner instead of randomly in order to yield responses with a higher Q-value and learn an agent with better performance with the same sample size. We call response a the fine-grained action and its category representation

Algorithm 1 Dual-granularity Q-function

```
Input:
    The dataset $\mathcal{D} = \{\mathcal{D}_i = (s, a, r, s', a')\}_{i=1}^M$, the classifier $F$ with action category set $\{\bar{a}_i\}_{i=1}^N$.
Output:
    The agent with policy $\pi_\mu$.
Initialization:
    Build the dataset $\mathcal{D}^c = \{\mathcal{D}_i^c = (s, \bar{a}, r, s', \bar{a}')\}_{i=1}^M$ base on the dataset $\mathcal{D}$ using the classifier $F$.
    Initialize the critic and target network parameters $\phi$, $\hat{\phi}$, $\theta$, $\hat{\theta}$, control generator $\pi_\psi$, and agent $\pi_\mu$.
    Fine-tuning the control generator $\pi_\psi$ using $(s, a, \bar{a})$ triples, where $s$ and $a$ are from dataset $\mathcal{D}$.
for $i = 1$ to until $Q_\phi$ and $Q_\theta$ converge do
    # The iteration stops for the first converged Q-function, while the rest continue until convergence.
    # Policy Evaluation on the dual-granularity Q-function.
    $\phi \leftarrow \arg\min_{\phi}(r + \gamma Q_{\hat{\phi}}(s', \bar{a}') - Q_\phi(s, \bar{a}))^2 \quad (s, \bar{a}, r, s', \bar{a}') = \mathcal{D}_i^c$
    $\theta \leftarrow \arg\min_{\theta}(r + \gamma Q_{\hat{\theta}}(s', a') - Q_\theta(s, a))^2 \quad (s, a, r, s', a') = \mathcal{D}_i$
    Every $n$ step $\hat{\phi} \leftarrow \phi$, $\hat{\theta} \leftarrow \theta$.
end for
for $i = 1, s \in \mathcal{D}_i$ to until $\pi_\mu$ converge do
    # Policy Improvement for the agent.
    $\bar{a}^* = \arg\max_{\bar{a}} Q_\phi(s, \bar{a}) \quad \bar{a} \in \{\bar{a}_i\}_{i=1}^N$
    $\{a_i\}_{i=1}^L = \pi_\psi(a|s, \bar{a}^*)$
    $a^* = \arg\max_{a} Q_\theta(s, a) \quad a \in \{a_i\}_{i=1}^L$
    $\mu \leftarrow \arg\min_{\mu} -\log \pi_\mu(a^*|s)$
end for
```

$\bar{a}$ the coarse-grained action, where $\bar{a}$ belongs to a finite set of categories $\{\bar{a}_i\}_{i=1}^N$ and $\bar{a}$ can obtain by $\arg\max_{\bar{a}_i} F(\bar{a}_i|a)$, where F is a classifier. The coarse-grained Q-function searches the category $\bar{a}^*$ with the highest Q-value from $\{\bar{a}_i\}_{i=1}^N$, where policy evaluation is given by

$$\begin{aligned} \phi \leftarrow & \arg\min_{\phi} \mathbb{E}_{(s,\bar{a},r,s',\bar{a}') \backsim \mathcal{D}^c} \\ & [(r + \gamma Q_{\hat{\phi}}(s', \bar{a}') - Q_\phi(s, \bar{a}))^2], \end{aligned} \tag{4}$$

where ϕ and $\hat{\phi}$ are the parameters of the critic and target networks (Mnih et al., 2015), and the same is true for θ and $\hat{\theta}$ in Eq.(5). $\mathcal{D}^c$ is a new dataset built on $\mathcal{D}$ that replaces fine-grained actions with coarse-grained actions. Then, a fine-tuned control generator with policy π_ψ generates a set of responses $\{a_i\}_{i=1}^L$ under the specified category according to $\pi_\psi(a|s, \bar{a}^*)$. The fine-grained Q-function selects the response a^* with the highest Q-value from $\{a_i\}_{i=1}^L$, where policy evaluation is as follows.

$$\begin{aligned} \theta \leftarrow & \arg\min_{\theta} \mathbb{E}_{(s,a,r,s',a') \backsim \mathcal{D}} \\ & [(r + \gamma Q_{\hat{\theta}}(s', a') - Q_\theta(s, a))^2]. \end{aligned} \tag{5}$$

Finally, the agent with policy π_μ is optimized by

$$\mu \leftarrow \arg\min_{\mu} \mathbb{E}_{s \backsim \mathcal{D}}[-\log \pi_\mu(a^*|s)]. \tag{6}$$

Our pseudocode is presented in Algorithm 1.

2.3. Rewards

Our goal is to develop an agent that is sufficiently intrinsically motivated to enrich the interactive content by capturing affective cues in human reactions. To achieve this, we devised four reward functions that assess how empathetic the response is to the conversation. (1) The average cosine similarity between the agent's response and dull responses. An expression that lacks emotional engagement may limit the development of dialogue. (2) The outpouring of the emotion of surprise. It benefits to build trust and hold the partner's attention throughout the conversation (Shum et al., 2018). (3) The length of response (i.e., the number of tokens in a response). It is a critical signal of engagement in conversation (Zhou et al., 2020). (4) Asking questions. It is an active listening skill that links to conversation management and responsiveness (Bodie et al., 2012). The total of the above rewards was used as r, and more details on the scoring design are presented in Appendix B.

3. Experiments

3.1. Corpus

We evaluated our approach on the DailyDialog dataset, which was crawled from websites that serve English dialogue in daily life. This dataset is human-rewritten and manually labeled with communication intention and emotion. We referred to its labels of action and emotion for assigning re-

Agent	Training Method	Dataset-based				Simulator-based			
		CS ↓	SE	LR	AQ	CS ↓	SE	LR	AQ
GPT-2	MLE	0.712	0.082	10.396	0.308	0.685	0.146	11.276	0.390
	Standard	0.645	0.126	13.020	0.550	0.644	0.206	13.778	0.526
	Ours	**0.596**	**0.191**	**14.463**	**0.555**	**0.597**	**0.238**	**15.636**	**0.566**
DialoGPT	MLE	0.714	0.069	9.761	0.345	0.687	0.142	10.838	0.492
	Standard	0.645	0.142	12.182	0.579	0.654	0.206	13.772	0.538
	Ours	**0.598**	**0.171**	**13.055**	**0.586**	**0.588**	**0.240**	**14.466**	**0.604**
T5	MLE	0.720	0.063	9.704	0.316	0.651	0.088	10.242	0.396
	Standard	0.621	0.147	13.291	0.532	0.605	0.224	13.676	0.510
	Ours	**0.567**	**0.202**	**14.834**	**0.565**	**0.553**	**0.268**	**15.134**	**0.552**
GODEL	MLE	0.718	0.064	9.507	0.318	0.689	0.112	10.132	0.414
	Standard	0.625	0.165	13.553	0.529	0.615	0.235	13.108	0.614
	Ours	**0.571**	**0.232**	**15.272**	**0.557**	**0.571**	**0.258**	**14.608**	**0.628**

Table 1: Automatic evaluation results. For the standard offline RL algorithm and our approach, we use $L = 5$ for the number of candidate responses $\{a_i\}_{i=1}^{L}$. For the simulator-based evaluation, we conducted 1000 dialogues of 5 consecutive turns between the simulator and each method. Each metric is measured per response, and the best score in each metric is in bold. The statistical test revealed that the differences are significant, with a p-value < 0.05.

wards (2) and (4) designed in Section 2.3 to each response. This dataset contains 11,118 / 1,000 / 1,000 multi-turn dialogues for train / test / dev, respectively. We used the set of train and dev for Q-function training and fine-tuning agents and the set of test for evaluation and discussion.

3.2. Agents

The following four state-of-the-art generative methods were considered as agents in our experiments. **GPT-2** proposed by Radford et al. (2019), is an unsupervised autoregressive language model for textual generation. **DialoGPT** is a pre-trained dialogue model proposed by Zhang et al. (2020). This model is based on GPT-2, using the Reddit comments dataset. **T5** is a unified framework proposed by Raffel et al. (2020) that converts all text-based language tasks into a text-to-text format via the transfer learning technique. **GODEL** is a pre-trained dialogue model proposed by Peng et al. (2022). This model is based on T5, using the Reddit discussion dataset. All the agents used the base version of the corresponding pre-trained model.

3.3. Implementation

The critic and target networks are the BERT models (Kenton and Toutanova, 2019) with a fully connected head on top. The classifier[1] is a RoBERTa model fine-tuned on the TweetTopic dataset (Antypas et al., 2022), which divides the responses into 19 pre-defined topics as its action category. Further, the control generator is initialized by the corresponding agent model. To drive the control generator to respond for the specified category, we append the category representation at the beginning of the input for GPT-2 and DialoGPT during the learning and inference, the injection scheme for which followed Cho et al. (2022). In addition, for T5 and GODEL, we added the category representation into the task prefix of the T5 framework during the learning and inference. The task prefix was set as “Instruction: given a dialog context, you need to respond related to <category>.” Our implementation was based on PyTorch (Paszke et al., 2019) and HuggingFace libraries (Wolf et al., 2019).

All agents and the control generator were fine-tuned before executing RL. The patience for early stopping was set to 5. The batch size was fixed at 32. The Adam algorithm (Kingma and Ba, 2015) was utilized for optimization, with a learning rate of 2.6e-5 and a warmup step of 6000. The control generator constructs the actions using multinomial sampling with a temperature setting of 1.5 to collect diverse responses. In addition, the update rate of the target network is set as 2.4e-5. The synchronized interval for the target networks was 30 steps. The discount factor was set as 0.9. We considered the target network to have converged and terminated the iteration when the change in the loss for 10 consecutive epochs is less than 0.01.

---

[1] It is available in the official repository of Cardiff NLP : `https://huggingface.co/cardiffnlp`
Note that many other ways for the category decision are also feasible. Since TweetTopic and DailyDialog were both crawled from social networking, we consider this classifier is more appropriate for our task.

Agent	Training Method	Quality	Informativeness	Empathy	Engagingness
GPT-2	MLE	1.4	1.3	1.2	1.1
	Standard	**1.7**	1.3	1.4	1.4
	Ours	1.5	**1.5**	**1.5**	**1.6**
DialoGPT	MLE	1.3	1.1	0.7	0.7
	Standard	**1.5**	1.4	1.2	1.2
	Ours	1.4	**1.5**	**1.6**	**1.6**
T5	MLE	1.2	0.9	0.5	0.6
	Standard	1.1	0.8	0.6	0.7
	Ours	**1.4**	**1.4**	**1.4**	**1.3**
GODEL	MLE	1.5	1.3	0.8	1.0
	Standard	1.6	1.2	1.1	1.1
	Ours	**1.7**	**1.6**	**1.7**	**1.6**

Table 2: Human evaluation results. The final scores for each metric were calculated by taking the average of the annotator ratings. Each metric is measured per dialogue, and the best score in each metric is presented in bold. The Fleiss' kappa (Fleiss, 1971) score with human judges was approximately 0.29, which can be regarded as "fair agreement."

3.4. Evaluation

3.4.1. Automatic Metrics

We apply the reward perspective designed in Section 2.3 to automatic evaluation. In particular, for the view of reward (2), we count the number of generated responses that contain a word that expresses surprise (i.e., Aha, Oh, Wow, Whoa, Gee, Really?, Amazing) by conservative string-matching heuristics. For the view of reward (4), we count the number of generated responses that contain a question word or a question mark. We denote CS, SE, LR, and AQ as cosine similarity, surprise emotion, response length, and asking questions, respectively. In our automatic evaluation and discussion, each metric is measured per response, and its score is obtained by taking the average across all samples.

3.4.2. Human Metrics

Ten native speakers were recruited to evaluate all agents trained using different methods. We asked the annotators to engage in a conversation with all agents regarding daily life topics (e.g., hobbies and interests) for at least five consecutive turns and rate their overall experience based on the following metrics. The scale of these metrics is [0, 1, 2].

Quality measures the coherence and grammatical accuracy of the agents' responses. *Score 0*: Most responses are incoherent or contain grammatical errors, thereby preventing the dialogue from proceeding. *Score 1*: Although certain responses are incoherent or contain grammatical errors, the dialogue can continue. *Score 2*: Only a few (or no) incoherent or grammatical errors in the responses, and the overall dialogue flows fluently.

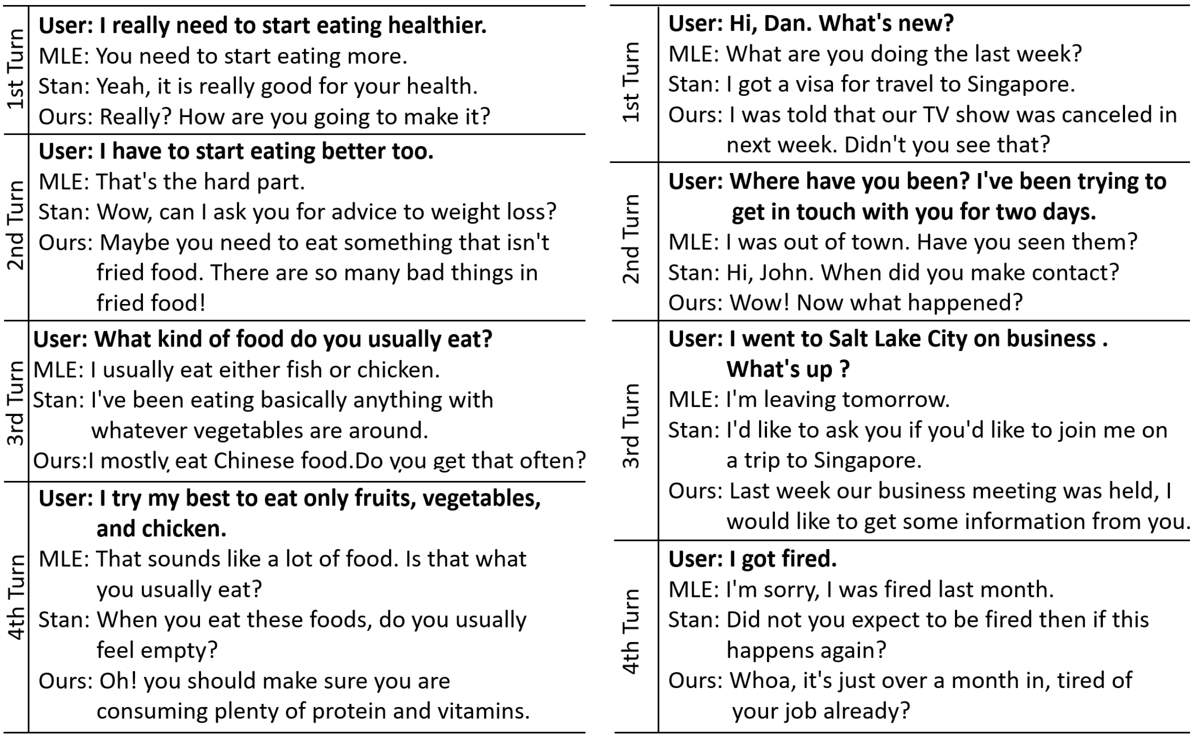
Informativeness measures the diversity and hallucination of the agents' responses. *Score 0*: Most responses simply repeat information from the context or are generic. *Score 1*: The information conflicts with common sense or contradicts the previous statement. *Score 2*: Most responses have the appropriate information.

Empathy measures the degree to which agents respond with concern or affectivity. *Score 0*: Most responses were short or showed little concern for the users in the dialogue. *Score 1*: Although not very coherent, certain responses convey an emotional tone or ask a question. *Score 2*: Certain responses are both coherent and show care for or emotional attachment to the user.

Engagingness measures the desire to engage the agents in a long conversation. *Score 0*: The replies are lackluster, thereby making it difficult to sustain the dialogue. *Score 1*: The responses are not particularly engaging, but they are fair for continuing the dialogue. *Score 2*: The responses are engaging and have the potential to further the dialogue.

3.4.3. Results

We assessed the performance of our approach using dataset-based evaluation and compared it with baseline methods, which include a standard offline RL algorithm (i.e., Eq.(3) and Eq.(2), where Eq.(3) is equivalent to our fine-grained Q-function, and it is referred to as the standard method in the following account) and MLE without RL. We also conducted a simulator-based evaluation by interacting with the user simulator Blenderbot (Roller et al., 2021) to assess the performance of different methods in a long-term dialogue. Table 1 reports the automatic



1st Turn	**User: I really need to start eating healthier.** MLE: You need to start eating more. Stan: Yeah, it is really good for your health. Ours: Really? How are you going to make it?
2nd Turn	**User: I have to start eating better too.** MLE: That's the hard part. Stan: Wow, can I ask you for advice to weight loss? Ours: Maybe you need to eat something that isn't fried food. There are so many bad things in fried food!
3rd Turn	**User: What kind of food do you usually eat?** MLE: I usually eat either fish or chicken. Stan: I've been eating basically anything with whatever vegetables are around. Ours:I mostly eat Chinese food.Do you get that often?
4th Turn	**User: I try my best to eat only fruits, vegetables, and chicken.** MLE: That sounds like a lot of food. Is that what you usually eat? Stan: When you eat these foods, do you usually feel empty? Ours: Oh! you should make sure you are consuming plenty of protein and vitamins.

(a) DialoGPT

1st Turn	**User: Hi, Dan. What's new?** MLE: What are you doing the last week? Stan: I got a visa for travel to Singapore. Ours: I was told that our TV show was canceled in next week. Didn't you see that?
2nd Turn	**User: Where have you been? I've been trying to get in touch with you for two days.** MLE: I was out of town. Have you seen them? Stan: Hi, John. When did you make contact? Ours: Wow! Now what happened?
3rd Turn	**User: I went to Salt Lake City on business . What's up ?** MLE: I'm leaving tomorrow. Stan: I'd like to ask you if you'd like to join me on a trip to Singapore. Ours: Last week our business meeting was held, I would like to get some information from you.
4th Turn	**User: I got fired.** MLE: I'm sorry, I was fired last month. Stan: Did not you expect to be fired then if this happens again? Ours: Whoa, it's just over a month in, tired of your job already?

(b) T5

Figure 2: Case study. DialoGPT and T5 are used as the agents in the case study. For brevity, the standard method is abbreviated as stan. The ground truth for responding to each user utterance is the next user utterance in the dialogue.

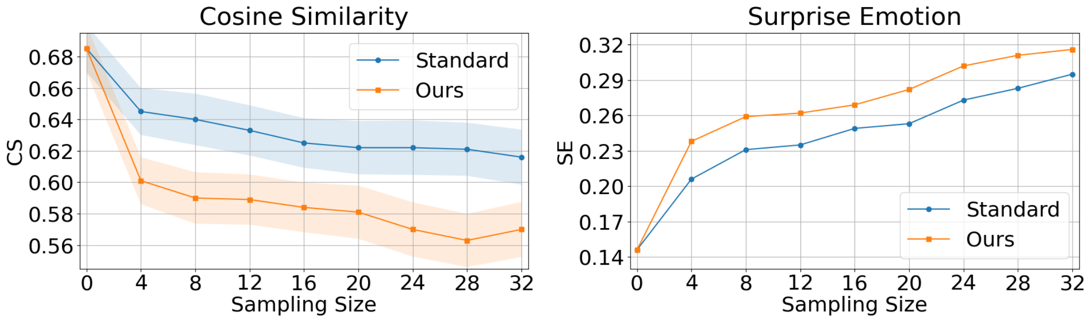
evaluation results. The standard method shows better performance than MLE, which can be credited to the policy improvement introduced by RL. In contrast, our approach achieved substantial gains in all metrics, thereby demonstrating the effectiveness of the dual-granularity Q-function.

We used a messenger APP as our platform, creating a specialized account for annotators to interact with agents and provide ratings via their smart devices. Table 2 summarizes the results of human evaluation. Despite the varied strengths and weaknesses of each agent according to individual human ratings, our approach exhibited markedly better results compared to baseline methods. Furthermore, RL-based agents displayed better proficiency than MLE-based agents in empathy and engagement by utilizing knowledge of the rewards outlined in Section 2.3. In terms of quality, most agents scored higher than other metrics because of the capacity of large-scale models to generate responses that are similar to human language and the stable one-step policy improvement of offline RL, which prevents the divergence of responses from human language. With regard to informativeness, upon analyzing instances of failure, we identified several agents that provided unrealistic information. Nevertheless, our approach generated more diverse responses, thereby resulting in a more favorable outcome than other methods. The interface and platform details for human evaluation are presented in Appendix C.

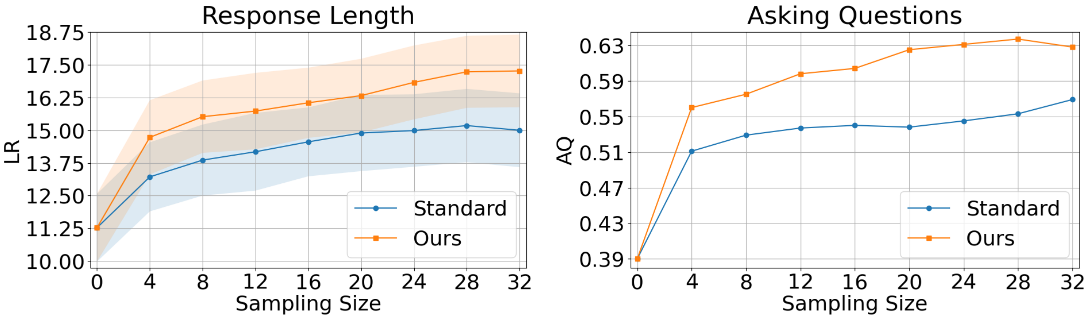
4. Discussion

4.1. Case Study

To conduct a comprehensive qualitative comparison between our approach and the baseline methods, we randomly selected four dialogues with varying topics from the testing set and shortened each dialogue into four consecutive utterances, thereby yielding four contexts per dialogue. We then instructed all agents to generate a response for each context based on their respective training methods. Figure 2 presents the varied responses generated by different methods across two agents. We found that our algorithm and the standard method generated longer responses compared to MLE, thereby indicating that RL-trained agents have better conversational engagement. Both the standard method and ours tend to ask questions, but our algorithm produces a more expressive tone of voice that conveys surprise, such as “Oh!,” “Wow!,” “Really?” and so on. Although our approach and the standard method received similar scores in the quality evaluation, a closer examination of the generated in-



(a) CS (b) SE



(c) LR (d) AQ

Figure 3: The evolution of the agent's performance for each metric with the increased sampling size. The scale for the X-axis is a multiple of 4, and 0 represents the MLE without RL. Bands indicate half a standard deviation.

stances reveals that the responses produced by our method are slightly more coherent than those of the standard one. We consider that this may be attributed to the coarse-grained Q-function, which tends to determine the category of context-related actions, as we will explain later.

4.2. Further Verification

First, to validate Theorem 1 and illustrate the impact of sampling size on policy performance, we conducted an additional trial. Figure 3 presents a comparison of the performance between the GPT-2 agents trained using the standard method and our approach, which was under the simulator-based setting with varying numbers of response candidates. As suggested by our theoretical derivation, increasing the sample size generally led to better performance. Our algorithm outperformed the standard method even when using the same number of actions for policy improvement, thereby indicating its efficiency in iterating policy. This result emphasizes the significance of sample size as a constraint on policy performance and highlights the efficacy of our approach in addressing this issue.

Next, we sought to verify whether our approach satisfies the hypothesis $\mathbb{E}_{a \backsim \pi_\alpha(a|s)}[Q^\pi(s,a)] \geq \mathbb{E}_{a \backsim \pi_\beta(a|s)}[Q^\pi(s,a)]$ in Theorem 2. With this intention, our control generator, which relies on the coarse-grained Q-function to provide the optimal category, can be represented as π_α, the agent that learned by the standard method can be considered as π_β, and Q^π is the fine-grained Q-function. The expected value was approximated by averaging the Q-value of each sample. The results reveal that for the GPT-2, DialoGPT, T5, and GODEL agents, $\mathbb{E}_{a \backsim \pi_\alpha(a|s)}[Q^\pi(s,a)]$ increases by 8.76%, 8.71%, 10.14%, and 9.61%, respectively, compared to $\mathbb{E}_{a \backsim \pi_\beta(a|s)}[Q^\pi(s,a)]$. This indicates that the categories selected by the coarse-grained Q-function can produce responses with a higher Q-value, thereby supporting the hypothesis in Theorem 2. Overall, these findings emphasize the potential of employing a coarse-to-fine-grained approach to narrow the scope of action for policy improvement to enhance dialogue agent performance.

Then, we examined the behavior of the coarse-grained Q-function in selecting action categories. We extracted 3000 contexts from the testing set and obtained the corresponding optimal action category by $\bar{a}^* = \arg\max_{\bar{a}} Q_\phi(s,\bar{a}) \quad \bar{a} \in \{\bar{a}_i\}_{i=1}^{N=19}$. The classifier was used to assign topics to these contexts. We then tallied the number of action categories selected by the coarse-grained Q-function for each context under each topic and this is visualized in Figure 4. It was revealed that the selections made by the coarse-grained Q-function largely align with human intuition. For example, when the context pertains to the "film tv and video topic," the coarse-grained Q-function often selects categories

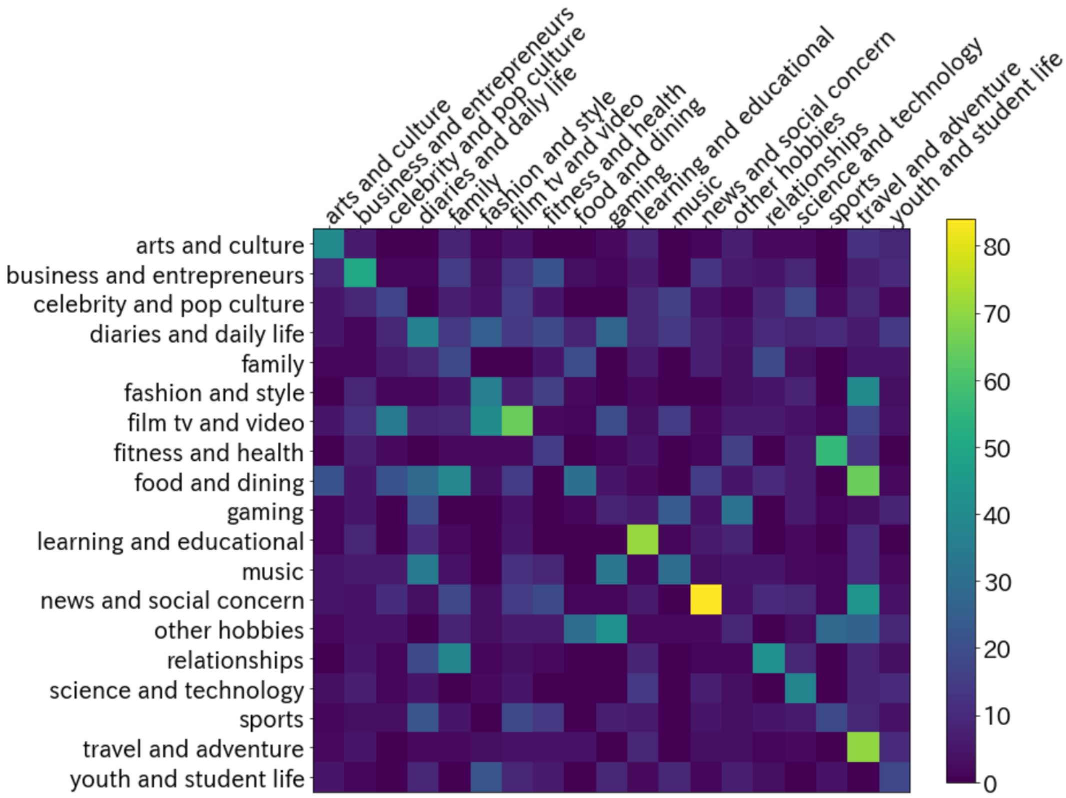


Figure 4: Behavioral analysis of coarse-grained Q-function. The label on the Y-axis represents the topic of each context, whereas the label on the X-axis represents the selected action category. The color bar represents the number of occurrences of the corresponding selection.

related to "fashion and style" and "celebrity and pop culture." Similarly, when the context relates to the "other hobbies" topic, it tends to select categories related to "travel and adventure," "gaming," "music," and "sports." We also observed a significant proportion of choices concentrated along the matrix's main diagonal, thereby indicating its propensity to select action categories similar to the context topic. It is worth noting that "travel and adventure" also constitute a considerable part of the selected categories. After closely analyzing the dialogues in the corpus, we observed that discussions on travel are typically lengthier and require higher participation from both parties. This may have led the coarse-grained Q-function to learn the extensive relevance of this category in dialogues.

Finally, we wanted to check if our control generator can generate responses for the specified category. For each $\bar{a} \in \{\bar{a}_i\}_{i=1}^{N=19}$, we used $\pi_\psi(a|s, \bar{a})$ to obtain a response and then used the classifier to determine if this response belongs to $\bar{a}$. The empirical results reveal that the percentage of accurate responses to a given category for the control generator performed by GPT-2, DialoGPT, T5, and GODEL are 28.16%, 29.89%, 35.42%, and 36.37%, respectively. This ratio is significantly higher than the correct percentage obtained by randomly generating responses over the 19 categories of the classifier (i,e., $1/19 \approx 5.26\%$).

5. Related Work

Since the RL for dialogue requires multiple steps of expensive human interaction, several prior studies have updated the agent's policy by the self-play method or the interaction with simulators (Li et al., 2016; Shah et al., 2018; Peng et al., 2018; Liu et al., 2020b). However, these online RL methods suffer from the issue of diverging from human language (Lewis et al., 2017; Zhao et al., 2019; Jang et al., 2020). On the other hand, offline RL (Fujimoto et al., 2019; Kumar et al., 2020; Brandfonbrener et al., 2021; Kostrikov et al., 2021) eliminates all need for environmental interaction or user simulators, instead of operating purely on static datasets of prior human interaction.

There are many closely related works (Jaques et al., 2019, 2020; Snell et al., 2022; Cohen et al., 2022; Verma et al., 2022; Jang et al., 2022) based on offline RL that lead to policy improvement via behavior cloning of self-generated utterances, which inherits the ability of pre-trained language models to generate human-like responses. Nevertheless, unlike RL tasks in which the actions are finite, such as Atari games (Mnih et al., 2015), the dialogue setting is hard to explore all probability space. Therefore, the policy obtained through the aforementioned methods is suboptimal. Some studies (Tiwari et al., 2022; Liu et al., 2020a; Rohmatillah and Chien, 2023; Liao et al., 2020; Saha et al., 2020) apply hierarchical RL to task-oriented dialogue, which consists of sub-tasks with a clear goal. The learning problem of the master policy is formulated as a Semi-Markov decision process (Sutton et al., 1999), requiring multiple-step environment interactions for each sub-task, thereby making offline learning challenging. Our dual-granularity Q-function focuses on the more structured action choices to implement policy improvement effectively for open-domain dialogue, which is suitable not only for offline but can also be extended to online learning.

6. Conclusion and Future Research

This paper presented a dual-granularity Q-function for mitigating suboptimal policy improvement due to the hard-to-traverse action space in RL, and we applied our method to the dialogue generation task. Theoretical and experimental results demonstrate the reliability of our algorithm, which significantly enhances the performance of the dialogue agent. Moving forward, we intend to design additional abstract categories for actions, such as those based on sentence embedding, to allow the coarse-grained Q-function to account for not only content but also utterance structure and expression. We will also investigate the affinity between the number of action categories and policy improvement. Ultimately, we plan to test our algorithm on other online RL tasks in NLP to confirm its broad applicability.

7. Ethics Statement

Ethical implications are not raised in this paper. The data and other relevant resources utilized in this study are openly available and widely employed by numerous existing research endeavors. Furthermore, the essence of this paper lies in proposing an efficient training approach rather than specific artifacts.

8. Acknowledgement

This work was supported by the JST SPRING (grant number: JPMJSP2123).

9. Bibliographical References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vitor Silva, and Francesco Barbieri. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Graham D Bodie, Kellie St. Cyr, Michelle Pence, Michael Rold, and James Honeycutt. 2012. Listening competence in initial interactions i: Distinguishing between what listening is and what listeners do. *International Journal of Listening*, 26(1):1–28.

David Brandfonbrener, William F Whitney, Rajesh Ranganath, and Joan Bruna. 2021. Offline rl without off-policy evaluation. In *35th Conference on Neural Information Processing Systems, NeurIPS 2021*, pages 4933–4946. Neural information processing systems foundation.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. 2022. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 367–377.

Deborah Cohen, Moonkyung Ryu, Yinlam Chow, Orgad Keller, Ido Greenberg, Avinatan Hassidim, Michael Fink, Yossi Matias, Idan Szpektor, Craig Boutilier, et al. 2022. Dynamic planning in open-ended dialogue using reinforcement learning. *arXiv preprint arXiv:2208.02294*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2020. Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7994–8001.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2022. Gpt-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *International Conference on Learning Representations*.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard.

2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3985–4003.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, pages 1–15.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1179–1191.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *arXiv preprint arXiv:2004.14254*.

Jianfeng Liu, Feiyang Pan, and Ling Luo. 2020a. Gochat: Goal-oriented chatbots with hierarchical reinforcement learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1793–1796.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020b. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427.

Mohsin Bashir Lone, Nahida Nazir, Navneet Kaur, Dinil Pradeep, Ammad Ul Ashraf, Pir Asrar Ul Haq, Nouman Barkat Dar, Abid Sarwar, Manik Rakhra, and Omdev Dahiya. 2022. Self-learning chatbots using reinforcement learning. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 802–808. IEEE.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Mahdin Rohmatillah and Jen-Tzung Chien. 2023. Hierarchical reinforcement learning with guidance for multi-domain dialogue policy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:748–761.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards integrated dialogue policy learning for multiple domains and intents using hierarchical deep reinforcement learning. *Expert Systems with Applications*, 162:113650.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gökhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From ELIZA to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology Electronic Engineering*, 19(1):10–26.

Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. 2022. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*.

Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Abhisek Tiwari, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A knowledge infused context driven dialogue agent for disease diagnosis using hierarchical reinforcement learning. *Knowledge-Based Systems*, 242:108292.

Siddharth Verma, Justin Fu, Sherry Yang, and Sergey Levine. 2022. Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4471–4491.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv e-prints*, pages arXiv–1910.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. 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.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. 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 1208–1218.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A. Proofs of Theorem

See Preliminary Derivation, Proof of Theorem 1, and Proof of Theorem 2 on the next page.

B. Details regarding Rewards

(1) Cosine Similarity: We manually created a list of dull responses consisting of utterances such as "I don't know," etc., which are short and frequently occur in the corpora. We penalize the cosine similarity between the agent's response and the dull responses to avoid the generation of dull responses by the agent. The response representations are computed by leveraging a state-of-the-art sentence embedding model (Conneau et al., 2017), and the score of this reward ranges from 0 to 1. Although there are more ways to generate a dull response, similar expressions are likely to fall into an adjacent vector space. The user "keeps away" from the utterances in the list, thereby also keeping away from other similar dull responses.
(2) Surprise Emotion: Since each utterance in the DailyDialog dataset was annotated by one of six universal emotions in human beings, we used the emotional label of "Surprise" to allocate a reward

Preliminary Derivation :

$$
\begin{aligned}
&\mathbb{E}_{\pi'}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s,a)] \\
&= \mathbb{E}_{\pi'}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_t,a)) + \gamma R(s_{t+1},a_{t+1}) + \gamma^2 R(s_{t+2},a_{t+2}) + ...|s_t = s, \{a_n \backsim \pi(a|s_n)\}_{n=t+1}^{T}] \\
&= \mathbb{E}_{\pi'}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_t,a)) + \gamma V^{\pi}(s_{t+1})|s_t = s, a \backsim \pi'(a|s_t)]
\end{aligned}
$$

Proof of Theorem 1.

$$
\begin{aligned}
&\text{Premise}: \quad \forall s,\ \pi'_L(\cdot|s) = \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}\backsim\pi(a|s)} Q^{\pi}(s,a) \\
&\text{Lemma}: \quad \text{if } N \geq M \geq 1,\ \text{then } \mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{N}} Q^{\pi}(s,a)] \geq \mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{M}} Q^{\pi}(s,a)]
\end{aligned}
$$

$$
\begin{aligned}
&V^{\pi'_N}(s) \\
&= \mathbb{E}_{\pi'_N}[R(s_t, a_t \backsim \pi'_N(a|s_t)) + \gamma R(s_{t+1}, a_{t+1} \backsim \pi'_N(a|s_{t+1})) + ...|s_t = s] \\
&= \mathbb{E}_{\pi}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{N}} Q^{\pi}(s_t,a)) + \gamma R(s_{t+1}, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{N}} Q^{\pi}(s_{t+1},a)) + ...|s_t = s] \\
&= \mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{N}} Q^{\pi}(s,a) - \gamma V^{\pi}(s_{t+1})|s_t = s] + \gamma\mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{N}} Q^{\pi}(s_{t+1},a) - \gamma V^{\pi}(s_{t+2})|s_t = s] + ... \\
&\geq \mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{M}} Q^{\pi}(s,a) - \gamma V^{\pi}(s_{t+1})|s_t = s] + \gamma\mathbb{E}_{\pi}[\max_{a\in\{a_i\}_{i=1}^{M}} Q^{\pi}(s_{t+1},a) - \gamma V^{\pi}(s_{t+2})|s_t = s] + ... \\
&= \mathbb{E}_{\pi}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{M}} Q^{\pi}(s_t,a)) + \gamma R(s_{t+1}, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{M}} Q^{\pi}(s_{t+1},a)) + ...|s_t = s] \\
&= \mathbb{E}_{\pi'_M}[R(s_t, a_t \backsim \pi'_M(a|s_t)) + \gamma R(s_{t+1}, a_{t+1} \backsim \pi'_M(a|s_{t+1})) + ...|s_t = s] = V^{\pi'_M}(s)
\end{aligned}
$$

Proof of Theorem 2.

$$
\begin{aligned}
&\text{Premise 1}: \quad \forall s,\ \pi'_1(\cdot|s) = \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}\backsim\pi_\alpha(a|s)} Q^{\pi}(s,a),\ \ \pi'_2(\cdot|s) = \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}\backsim\pi_\beta(a|s)} Q^{\pi}(s,a) \\
&\text{Premise 2}: \quad \mathbb{E}_{a\backsim\pi_\alpha(a|s)}[Q^{\pi}(s,a)] \geq \mathbb{E}_{a\backsim\pi_\beta(a|s)}[Q^{\pi}(s,a)],\ \ \sigma^2_{a\backsim\pi_\alpha(a|s)}[Q^{\pi}(s,a)] \approx \sigma^2_{a\backsim\pi_\beta(a|s)}[Q^{\pi}(s,a)] \\
&\text{Lemma} \quad : \quad \because\ \text{Premise 2}\ \ \therefore\ \mathbb{E}_{\pi_\alpha}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s,a)] \geq \mathbb{E}_{\pi_\beta}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s,a)]
\end{aligned}
$$

$$
\begin{aligned}
&V^{\pi'_1}(s) \\
&= \mathbb{E}_{\pi'_1}[R(s_t, a_t \backsim \pi'_1(a|s_t)) + \gamma R(s_{t+1}, a_{t+1} \backsim \pi'_1(a|s_{t+1})) + ...|s_t = s] \\
&= \mathbb{E}_{\pi_\alpha}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_t,a)) + \gamma R(s_{t+1}, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_{t+1},a)) + ...|s_t = s] \\
&= \mathbb{E}_{\pi_\alpha}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s,a) - \gamma V^{\pi}(s_{t+1})|s_t = s] + \gamma\mathbb{E}_{\pi_\alpha}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_{t+1},a) - \gamma V^{\pi}(s_{t+2})|s_t = s] + ... \\
&\geq \mathbb{E}_{\pi_\beta}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s,a) - \gamma V^{\pi}(s_{t+1})|s_t = s] + \gamma\mathbb{E}_{\pi_\beta}[\max_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_{t+1},a) - \gamma V^{\pi}(s_{t+2})|s_t = s] + ... \\
&= \mathbb{E}_{\pi_\beta}[R(s_t, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_t,a)) + \gamma R(s_{t+1}, \mathop{\arg\max}_{a\in\{a_i\}_{i=1}^{L}} Q^{\pi}(s_{t+1},a)) + ...|s_t = s] \\
&= \mathbb{E}_{\pi'_2}[R(s_t, a_t \backsim \pi'_2(a|s_t)) + \gamma R(s_{t+1}, a_{t+1} \backsim \pi'_2(a|s_{t+1})) + ...|s_t = s] = V^{\pi'_2}(s)
\end{aligned}
$$

to each sentence, for which the scale is [0, 1].

(3) Response Length: From an empathetic standpoint, we prefer that the agents generate responses that are more elaborate and longer. The reward is defined in the following manner.

```
if   the number of generated tokens < 5:
         reward = -0.2
elif the number of generated tokens < 10:
         reward = 0
elif the number of generated tokens < 15:
         reward = 0.2
else:
         reward = 0.5
```

(4) Asking Questions: Each utterance in the DailyDialog dataset was also labeled as one of four dialogue act classes. We used the act label of "Questions" to allocate a reward to each sentence. The scale of this reward is [0, 1].

C. Details regarding Interactive

We utilized LINE[2] as our platform and set up a specialized account for human evaluation, which evaluators accessed through their smart device to interact with each agent and provide ratings. The server was equipped with an NVIDIA A6000 (48G) graphics card, and the program was developed using the LINE Messaging API SDK and ran continuously in the background, ready to receive requests at any time. Each agent extended an abstract class that defined key methods for conversation generation and was registered to a dictionary via a decorator. To ensure a randomized order of appearance of agents for annotators during the evaluation process, we implemented a randomized selection of dictionary indices. Furthermore, due to the substantial startup times of the agents, all agents were kept in memory at all times in the background process. The current configuration was able to support hundreds of simultaneous users and concurrently host more than 20 agents. Figures 5 (a) and (b) depict the conversation interface utilized by annotators for interacting with the agents during the human evaluation process. This interface enabled participants to engage in a dialogue of at least five turns before initiating the rating phase by entering the word "end." In contrast, Figures 5 (c), (d), (e), and (f) exhibit the interface used for rating the agents after having a conversation of at least five turns with them as part of the human evaluation.

D. Details regarding Instructions

For the human evaluation, the full text of instructions given to annotators is as follows:

Thank you for participating in our experiment. You will sequentially have conversations with 12 different chatbots and have at least 5 exchanges with each chatbot. After each conversation with a chatbot, you will rate the experience based on the four criteria. Detailed instructions will be provided throughout the experiment. The experiment will take approximately 2 hours, but you are free to take breaks as needed. ※Please note the following:

(1) Please try to cooperate with the chatbot and engage in role-playing to ensure a smooth conversation. (if necessary)

(2) Please keep the conversation topics to daily life such as hobbies and interests, and avoid political, religious, and racial topics.

(3) Please avoid disclosing any personal information during the conversation.

(4) You can only rate each chatbot once, and after evaluation, you will move on to the next chatbot for conversation.

(5) Please wait for the chatbot to reply before sending the next message.

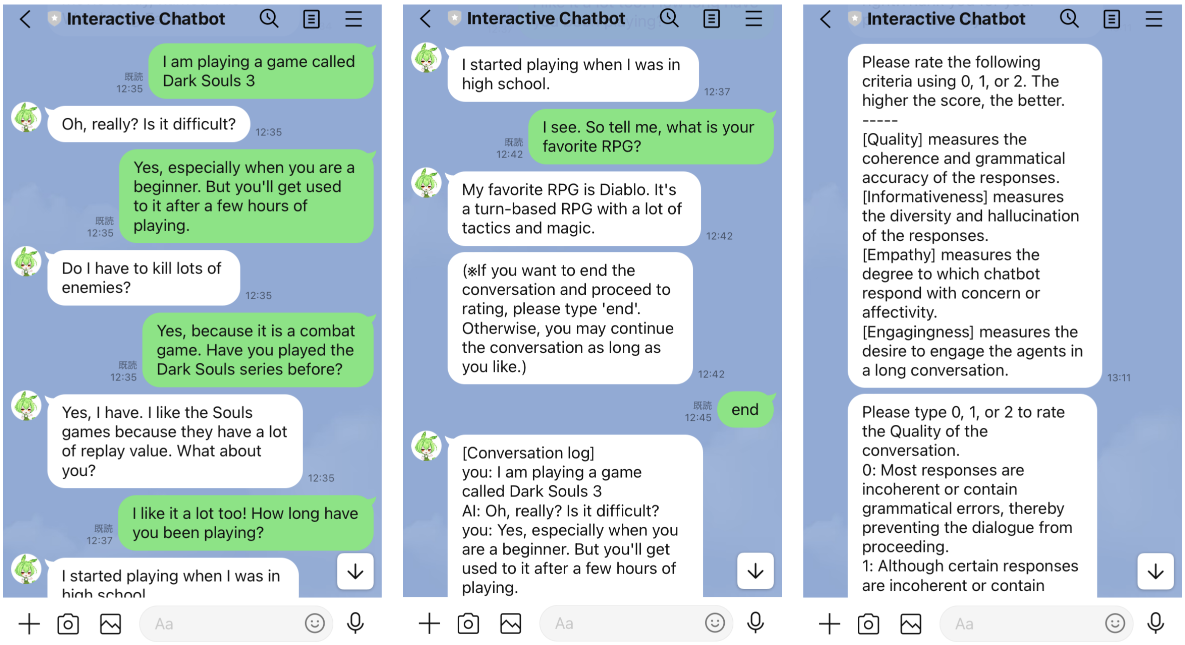
(6) Please avoid sending messages that contain anything other than English and commonly used punctuation marks, such as emojis, stamps, and special characters.

E. Details regarding Annotators

All the annotators are graduate students unrelated to this study recruited through an internal university-based recruitment process. For the human evaluation, annotators were required to engage in a dialogue for at least 5 turns with 12 agents, sending over 60 messages while understanding the context. Additionally, the annotators need advanced expertise in native-level English proficiency and natural language processing, as they will score the user experience for each agent based on four criteria. Considering the complexity of the task and the guidelines regarding the school's public funding, the hourly rate was set at 2,500 yen. Since the estimated duration of this task is 2 hours per person, the remuneration for the work would be approximately 5,000 yen (2,500 yen/hour × 2 hours). Considering that the minimum hourly wage for annotators residing in the city is 1,072 yen, we believe this payment is adequate.

---

[2] https://line.me/en/



I am playing a game called Dark Souls 3

Oh, really? Is it difficult?

Yes, especially when you are a beginner. But you'll get used to it after a few hours of playing.

Do I have to kill lots of enemies?

Yes, because it is a combat game. Have you played the Dark Souls series before?

Yes, I have. I like the Souls games because they have a lot of replay value. What about you?

I like it a lot too! How long have you been playing?

I started playing when I was in high school

I started playing when I was in high school.

I see. So tell me, what is your favorite RPG?

My favorite RPG is Diablo. It's a turn-based RPG with a lot of tactics and magic.

(※If you want to end the conversation and proceed to rating, please type 'end'. Otherwise, you may continue the conversation as long as you like.)

end

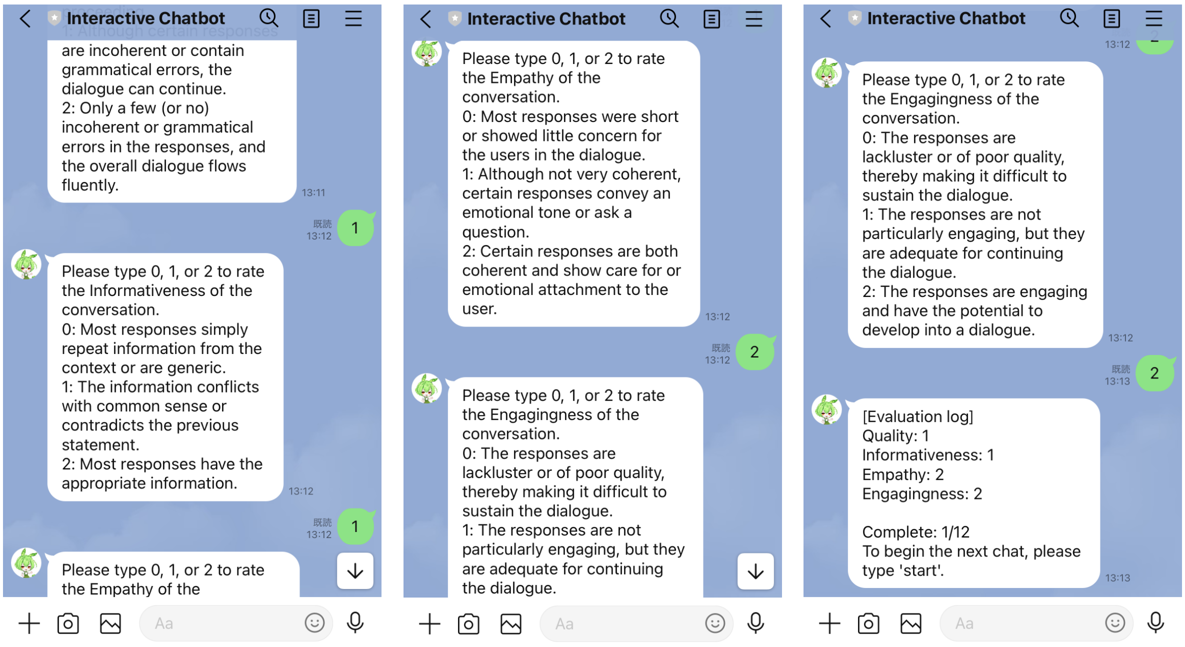
[Conversation log]
you: I am playing a game called Dark Souls 3
AI: Oh, really? Is it difficult?
you: Yes, especially when you are a beginner. But you'll get used to it after a few hours of playing.

Please rate the following criteria using 0, 1, or 2. The higher the score, the better.

[Quality] measures the coherence and grammatical accuracy of the responses.
[Informativeness] measures the diversity and hallucination of the responses.
[Empathy] measures the degree to which chatbot respond with concern or affectivity.
[Engagingness] measures the desire to engage the agents in a long conversation.

Please type 0, 1, or 2 to rate the Quality of the conversation.
0: Most responses are incoherent or contain grammatical errors, thereby preventing the dialogue from proceeding.
1: Although certain responses are incoherent or contain

(a) Chat Interface 1 (b) Chat Interface 2 (c) Ratings Interface 1



1: Although certain responses are incoherent or contain grammatical errors, the dialogue can continue.
2: Only a few (or no) incoherent or grammatical errors in the responses, and the overall dialogue flows fluently.

1

Please type 0, 1, or 2 to rate the Informativeness of the conversation.
0: Most responses simply repeat information from the context or are generic.
1: The information conflicts with common sense or contradicts the previous statement.
2: Most responses have the appropriate information.

1

Please type 0, 1, or 2 to rate the Empathy of the

Please type 0, 1, or 2 to rate the Empathy of the conversation.
0: Most responses were short or showed little concern for the users in the dialogue.
1: Although not very coherent, certain responses convey an emotional tone or ask a question.
2: Certain responses are both coherent and show care for or emotional attachment to the user.

2

Please type 0, 1, or 2 to rate the Engagingness of the conversation.
0: The responses are lackluster or of poor quality, thereby making it difficult to sustain the dialogue.
1: The responses are not particularly engaging, but they are adequate for continuing the dialogue.

Please type 0, 1, or 2 to rate the Engagingness of the conversation.
0: The responses are lackluster or of poor quality, thereby making it difficult to sustain the dialogue.
1: The responses are not particularly engaging, but they are adequate for continuing the dialogue.
2: The responses are engaging and have the potential to develop into a dialogue.

2

[Evaluation log]
Quality: 1
Informativeness: 1
Empathy: 2
Engagingness: 2

Complete: 1/12
To begin the next chat, please type 'start'.

(d) Ratings Interface 2 (e) Ratings Interface 3 (f) Ratings Interface 4

Figure 5: The chat and rating interfaces used for human evaluation.