LSTDial: Enhancing Dialogue Generation via Long- and Short-Term Measurement Feedback

Guanghui Ye[1], **Huan Zhao**[1*], **Zixing Zhang**[1], **Xupeng Zha**[1], **Zhihua Jiang**[2]
[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Department of Computer Science, Jinan University, China
{yghui,hzhao}@hnu.edu.cn

Abstract

Generating high-quality responses is a key challenge for any open domain dialogue systems. However, even though there exist a variety of quality dimensions especially designed for dialogue evaluation (e.g., coherence and diversity scores), current dialogue systems rarely utilize them to guide the response generation during training. To alleviate this issue, we propose LSTDial (**L**ong- and **S**hort-**T**erm **Dial**ogue), a novel two-stage framework which generates and utilizes conversation evaluation as explicit feedback during training. Specifically, we fine-tune pre-trained dialogue systems through using turn-level quality feedback in the first stage and further train ever-improving dialogue agents through using dialogue-level quality feedback in the second stage. By using our approach on dialogue systems, capable of enabling dialogue generation with both short-term capabilities (generating more fluent, relevant and varied responses at the turn-level) and long-term capabilities (generating more coherent, engaging and informative responses at the dialogue-level). We implement LSTDial on four strong baseline models and experiment with two open-domain dialogue datasets. Experimental results show that LSTDial achieves significant improvement, enabling to generate better dialogue responses in terms of both human and automatic evaluation.

1 Introduction

Generating high-quality responses is a key challenge for any open domain dialogue systems (Adiwardana et al., 2020; Roller et al., 2021; Kann et al., 2022; Ferron et al., 2023; Park et al., 2023). Current state-of-the-art systems do this by training agents with supervised learning on large amounts of labeled text data. However, these labeled data are usually reference responses or pre-defined task goals instead of multi-faceted quality dimensions designed for dialogue evaluation (e.g., relevance, coherence, and engagement, etc.), which leads to a great gap between dialogue generation and evaluation. Response generation systems are typically trained by optimizing the average likelihood of the training data (Maximum Likelihood Estimation, MLE), without a clear signal on the progression of the ongoing conversation (Sun et al., 2023).

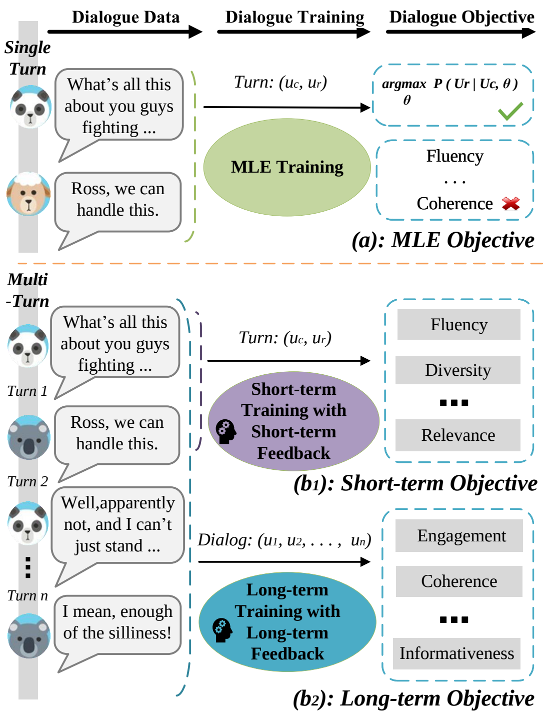
Dialogue Data
Dialogue Training
Dialogue Objective
Single Turn
What's all this about you guys fighting ...
Ross, we can handle this.
Turn: (u_c, u_r)
$\underset{\theta}{\text{argmax}}\ P(U_r \mid U_c, \theta)$ ✓
MLE Training
Fluency
...
Coherence ✗
(a): MLE Objective
Multi -Turn
What's all this about you guys fighting ...
Turn 1
Ross, we can handle this.
Turn 2
Well,apparently not, and I can't just stand ...
Turn n
I mean, enough of the silliness!
Turn: (u_c, u_r)
Short-term Training with Short-term Feedback
Fluency
Diversity
...
Relevance
(b1): Short-term Objective
Dialog: $(u_1, u_2, \ldots, u_n)$
Long-term Training with Long-term Feedback
Engagement
Coherence
...
Informativeness
(b2): Long-term Objective

Figure 1: A motivation example of comparing MLE and LSTDial (ours). (a) MLE does not aim at any dialogue quality. Conversely, LSTDial achieves (b_1) the short-term objective via training with short-term quality feedback and (b_2) the long-term objective via training with long-term quality feedback.

A possible solution to the above issue is to introduce quality annotations of the ongoing conversation during training. However, selecting relevant and meaningful quality dimensions and generating corresponding measurement annotations are crucial. In the dialogue evaluation task, an evaluation

[*]Corresponding authors.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5857–5871
June 16-21, 2024 ©2024 Association for Computational Linguistics

metric can be evaluated at the turn and dialogue level (Yeh et al., 2021; Mehri and Eskénazi, 2020). In a turn-level evaluation setting, the goal is, given prior dialogue history (frequently denoted as context) c and a response r, the metric will assign a score s of r conditioned on c while assuming a measured quality dimension q. Examples of turn-level quality dimensions include fluency, relevance, diversity, appropriateness, and safety, etc (Jiang et al., 2022). Conversely, in a dialogue-level evaluation setting, the goal is to evaluate the performance throughout the full dialogue. Examples of dialogue-level quality dimensions include coherence, engagement, topic depth, informativeness, and consistence, etc (Zhang et al., 2022, 2023). Because human evaluation is both expensive and time-consuming, automatic evaluation metric models that strongly correlate with human judgments have been proposed. Intuitively, if the mentioned-above quality annotations can be automatically generated and recast into explicit feedback, it is promising to study ever-improving dialogue agents that learn through interaction.

Feedback-based reinforcement learning (RL) has gained increasing attention in dialogue generation (Yu et al., 2023; Cai et al., 2023; Zhu et al., 2023). Because human-generated feedback is expensive to obtain, researchers have devised learned feedback generators while assuming one can train downstream models to utilize generated feedback (Zhou et al., 2023). Motivated by this, we design feedback functions which depict multi-sided qualities of conversation and utilize generated feedback to guide response generation in the RL-based setting. Considering the balance between efficiency and performance, we fine-tune pre-trained dialogue systems through using turn-level quality feedback (called *short-term feedback* in our method) while train ever-improving dialogue agents via continual learning through using dialogue-level quality feedback (called *long-term feedback* in our method). A motivation example is shown in Figure 1.

To this end, we propose the LSTDial (**L**ong- and **S**hort-**T**erm **Dial**ogue) framework. There are two stages involved in the learning process. In the first stage (STDial), we first extract single-turn dialogues from the training set of the experimental dataset and then calculate their turn-level evaluation scores. After this, we fine-tune dialogue systems via a multi-objective loss function including MLE and short-term feedback. In the second stage (LTDial), we further train the first-stage dialogue system but in the RL-based mode. Specifically, we first generate a response r via the training system conditioned on current history c. Then, we calculate dialogue-level evaluation scores of the conversation <c, r> as current reward or feedback of r. Particularly, to estimate future reward of r, we generate a future response f via user simulator (Peng et al., 2022; Hu et al., 2023) while assuming the conversation is ongoing. Next, we calculate turn-level evaluation scores of <r, f> as future reward or feedback of r. Finally, these two rewards are integrated and adapted in policy updating for RL. A novel feature of our methodology is generating and utilizing feedback that accesses relevant and meaningful conversation evaluation for better response generation.

The contributions of this paper are three folds:

- We propose LSTDial, a novel framework which utilizes turn- and dialogue-level conversation evaluation as explicit feedback during training for the first time, alleviating the gap between dialogue generation and evaluation.
- Our approach enables both short-term capabilities (generating more fluent, relevant and varied responses at the turn-level) and long-term capabilities (generating more coherent, engaging and informative responses at the dialogue -level) for dialogue generation.
- We conduct comprehensive experiments on two popular open-domain dialogue datasets and the results show that LSTDial achieves significant improvement in terms of both human and automatic evaluation.

2 Related Works

2.1 Dialogue Generation

In recent years, significant progress has been made in the field of open-domain response generation. GPT-2 (Radford et al., 2019) is a Transformer-based model which improves performance in a log-linear fashion across tasks. Then, DialoGPT (Zhang et al., 2020) is trained on 147M conversation-like exchanges extracted from Reddit comment chains, enabling to generate relevant, content-rich and context-consistent responses. In contrast with DialoGPT, GODEL (Peng et al., 2022) leverages a new phase of grounded pre-training that require information external to the current conversation to produce good responses. The most related work to ours is PLATO-2 (Bao et al., 2021) trained via curriculum learning. There are two stages involved in the learning process. However, the main difference between previous dialogue systems including PLATO-2 and ours is that we elaborately design the quality-driven training method while the others mainly optimize the average likelihood of the training data. For more advances on dialogue systems, we refer to (Ni et al., 2023; Deng et al., 2023) for interested readers.

2.2 Dialogue Evaluation

Automatic dialogue evaluation metrics can be divided into two categories: rule-based and model-based (Yeh et al., 2021). Rule-based metrics employ heuristic rules to evaluate responses while model-based metrics train deep networks or language models on specific dialogue data. Recently, combined model-based metrics have sprung up. USL-H (Phy et al., 2020) composites three quality groups (understandability, sensibleness, likability) in a linear hierarchy. IM^2 (Jiang et al., 2022) trains sub-metrics with each targeting a specific dimension and integrates them into meaningful categorical metrics. Some recent studies (Mendonça et al., 2023; Duan et al., 2023) propose ensemble models that take advantage of the strengths of current evaluation models with prompting Large Language Models, achieving state of the art results on the newest DSTC dialogue evaluation challenge. For more details of dialogue evaluation and DSTC challenges, we refer the readers to (Zhang et al., 2021; Rodríguez-Cantelar et al., 2023).

3 Methodology

The approach proposed in this paper belongs to the augmentation task of dialogue generation, and the specific objectives and motivations have been elaborated in Figure 1. Figure 2 shows the architecture of our method LSTDial. First, we crafted two sets of conversation evaluators to generate long- and short-term feedback (described in §3.1 and §3.2, respectively). Then, the two-stage training of LSTDial will be presented in §3.3 and §3.4.

3.1 Short-term Feedback

We integrated three turn-level dialogue evaluation models together to construct Turn-level Evaluator (i.e., TEval). TEval eventually generates a Short-term Feedback (i.e., ST-Feed) and participates in the training of short-term stage. In the figure 2(c), the Fluency, Relevance and Diversity evaluation models included in TEval are presented respectively. **Fluency**: We used VUP (Valid Utterance Prediction) proposed by USL-H (Phy et al., 2020) for Fluency. We run VUP with the original settings. **Relevance**: Similar to the Fluency metric, we proposed this metric to enhance the relevance prediction by using negative sampling. Specifically, the objective is to discern whether a given context-response pair is relevance or not. RoBERTa-base (Liu et al., 2019) is adopted as the text encoder. **Diversity**: We used a new automatic evaluation metric, Sem-Ent (Han et al., 2022), which can measure the semantic diversity based on the semantic distribution of generated responses.

$$Feed_{ST} = \alpha_1 s^{(\mathrm{Flu})} + \alpha_2 s^{(\mathrm{Rel})} + \alpha_3 s^{(\mathrm{Div})} \quad (1)$$

where we set $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$, which meaning that all three turn-level scores contribute the same to the final short-term feedback.

Notably, the design details of the evaluation model are in Appendix B. A description of all the qualities used to build the feedback is in the Appendix A. We follow DSTC 11 track 4[1] to select quality in feedback.

3.2 Long-term Feedback

In the figure 2(d), we also carefully constructed the Dialogue-level Evaluator (i.e., DEval) to generate the Long-term Feedback (i.e., LT-Feed). DEval is capable of evaluating the Engagement, Coherence, and Informativeness content of the entire conversation. **Engagement**: We trained a discriminatory model using the RED (Reddit-based Engagement Dataset) (Xu et al., 2022). The RED dataset is derived from Reddit and is carefully curated using a unique distant-supervision framework. In this framework, emotional, attentional, behavioral, and reply engagement are combined to form a single score called ENDEX. **Coherence**: We considered the use of the GRADE (Huang et al., 2020) as our model for coherence evaluation. We got k - 1 adjacency pairs for a dialogue containing k utterances and hence k - 1 coherence scores. The coherence score at the dialogue-level is calculated by averaging the k - 1 scores. **Informativeness**: We adopted a pretrained natural language inference (NLI) model[2] to calculate the topic depth of a

[1] https://chateval.org/dstc11

[2] https://huggingface.co/roberta-large-mnli

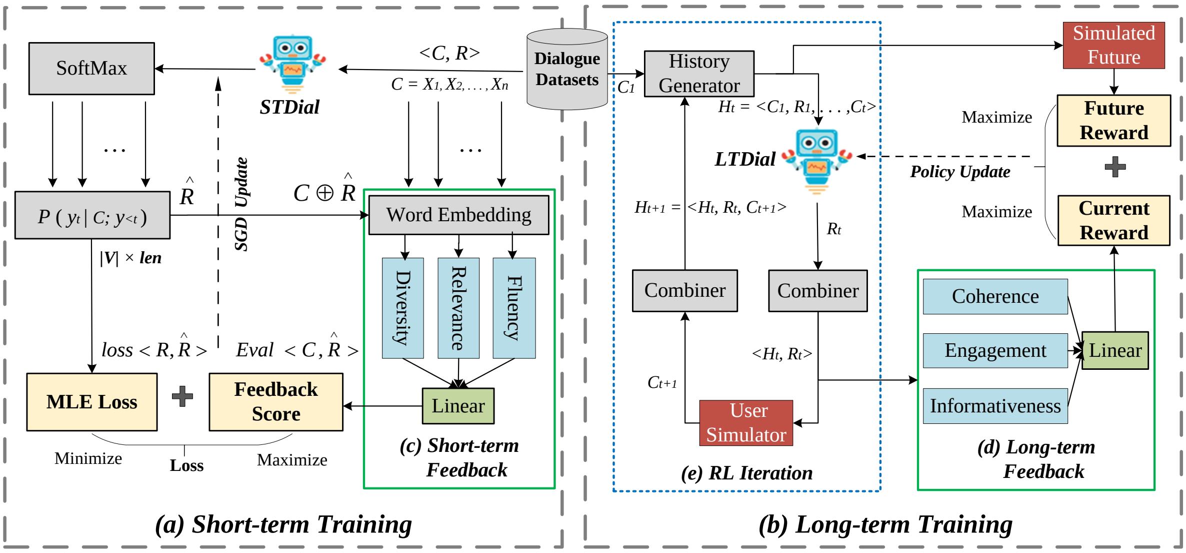

Figure 2: An architectural illustration of the introduced LSTDial method. LSTDial consists of (a) Short-term Training and (b) Long-term Training. In (a) we use (c) Short-term Feedback incorporated into regular MLE training to improve the short-term capabilities (relevance and diversity, etc.) of the dialogue. In (b) we build on the (e) RL Iteration algorithm and use (d) Long-term Feedback as a reward to improve the long-term capabilities of the dialogue (coherence and engagement, etc.).

conversation to represent the informativeness of a response.

$$Feed_{LT} = \beta_1 s^{(\text{Eng})} + \beta_2 s^{(\text{Coh})} + \beta_3 s^{(\text{Inf})} \quad (2)$$

Finally, we averaged the three dialogue-level scores, setting $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$ to get the LT-Feed. All the qualities used to construct the feedback are described in the Appendix A. Details of the design of the evaluation model are given in Appendix B.

3.3 Short-term Training

This section will introduce the short-term training of dialogue system, which we refer to as STDial. Figure 2(a) describes in detail the training process. In this strategy, we leverage the TEval introduced in §3.1 as a form of short-term feedback to enhance the short-term generation performance of the dialogue system. We introduce an additional feedback (eq. 4) term alongside the generative loss (eq. 3) within the overall loss equation (eq. 5), as demonstrated in Equation 3-5:

$$L_{MLE} = \sum_{t=1}^{len} p\left(y_t \mid C; y_{<t}\right) \log\left(q\left(\hat{y}_t \mid C; y_{<t}\right)\right) \quad (3)$$

$$F_{ST} {=} \| \, Eval\left(C, q\left(. \mid C; y_{<t}\right)\right) \| \quad (4)$$

$$L_{ST} = L_{MLE} - \lambda \cdot F_{ST} \quad (5)$$

where C is the context and $q \in R^{|V| \cdot len}$ in the Equation 3 is the softmax output produced by STDial (i.e., the predicted response matrix $\hat{R}$). The term $\hat{y}_t$ indicates the decoder's response at the t^{th} word it generates. The L_{MLE} of Equation 3 is the cross-entropy loss from STDial. In the Equation 4, the function *Eval* refers to TEval, which will generate the short-term feedback F_{ST}. The TEval module processes the input as a one-hot representation in a standalone assessment scenario. This means that the input consisting of *len* tokens first passes through an embedding lookup layer. This embedding lookup layer converts the input into a matrix of shape $\mathbb{R}^{D \times len}$ to rest of the network. D represents the dimension of the word embeddings. To enable the differentiability in the loss, we avoid using *argmax* to get the decoded token. Instead, we access the output of the softmax layer, which represents the likelihood distribution of output lengths across the entire vocabulary ($\mathbb{R}^{|V| \times len}$). In order to obtain the same input matrix $\mathbb{R}^{D \times len}$ for TEval network, we use this distribution to perform a weighted embedding lookup throughout the entire vocabulary. The input form of the transformed predicted responses $\hat{R}$ is:

$$\mathbb{R}^{D \times len} = \mathbb{R}^{D \times |V|} \times \mathbb{R}^{|V| \times len} \quad (6)$$

Then the context C and the predicted response $\hat{R}$ will get the same matrix form $\mathbb{R}^{D \times len}$, which is concatenated and fed into TEval to get the final feedback. We weighted the feedback by the hyperparameter λ in Equation 5. We set λ to 15 by grid search for optimising the final loss on the development set. The STDial is trained to minimize cross-entropy loss while maximizing ST-Feed. The TEval dialogue evaluation model is trained on the original corpus, following which the parameters are frozen. The updated loss is back-propagated to update STDial.

3.4 Long-term Training

Figure 2(b) describes in detail the training process and components of LTDial, which is mainly based on Reinforcement Learning (RL) iterations as well as the construction of effective feedback. In this section, RL iteration will be introduced first, followed by reward composition, which is involved in long-term training.

3.4.1 RL Iteration

History Generator: In the context of the t-th utterance, the dialogue state is derived from the dialogue history, denoted as H_t = <C_1, R_1, ..., C_t>. We define k as the iteration variable. When k = 1, we take the first utterance of the dialogue sample as the initialized dialogue history: $H_1 = C_1$. The purpose of History Generator is to update H_t to H_{t+1}. When moving to the next iteration $k = t + 1$, the new history H_k is assigned as <C_1, R_1, ..., C_t, R_t, C_{t+1}> such that the newly-generated R_t and C_{t+1} can be appended to the previous history H_t.

LTDial: LTDial takes the dialogue history H_t as input and generates the action R_t. We employed the stochastic policy gradient algorithm to optimize LTDial. In our method, the policy takes the form of the dialogue generation network (i.e., $P(R_t|H_t; \theta)$), and is defined by the parameters θ of LTDial. The objective to maximize is the expected future reward $\mathbb{E}[r_k]$:

$$L_{LT}(\theta) = \mathbb{E}[r_k] \quad (7)$$

According to the theorem of policy gradient, The gradient of the expected reward can be calculated given the parameter θ of LTDial as:

$$\nabla_\theta L_{LT}(\theta) = E\left[\nabla \log P\left(R_t \mid H_t; \theta\right) \cdot r_k\right] \quad (8)$$

where r_k, which involves long-term feedback, will be described in detail in the next section. We recommend that readers who are interested in the implementation of the policy gradient algorithm refer to (François-Lavet et al., 2018).

Combiner: Combiner acts as a connector to add the newly-generated utterance (e.g., action R_t) to the end of the previous history.

User Simulator: We introduced a User Simulator (Hu et al., 2023) to imitate human interaction with the dialogue system to obtain better training data for the online RL. Our method simulates a conversation between the User Simulator and our LTDial, and let them take turns talking with each other. We used the base version of GODEL[3] (Peng et al., 2022) as our User Simulator. The whole interaction is shown in Figure 2(e). Through extensive experiments, we chose iteration variable k = 5 to ensure moderate long-term training difficulty and better performance.

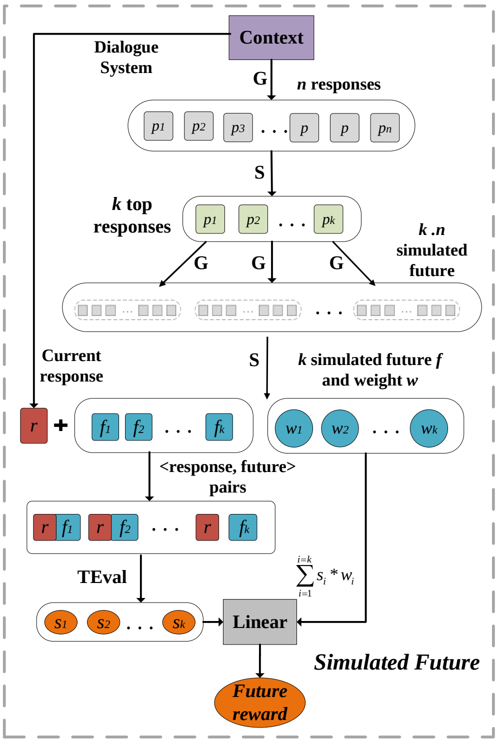

Figure 3: The calculation procedure for Future Reward.

3.4.2 Reward Composition

In our approach, the reward composition of t-th RL iteration consists of the current reward r_t^c and the future reward r_t^f. For the dialogue history H_t and response R_t, the current reward computes the score of <H_t, R_t>, which measures the importance of R_t to the overall dialogue history H_t. Conversely, the future reward computes the score of <R_t, F_t> where F_t is the content of the conversation that will probably be talked about in the following (i.e., Follow-up utterances), which measure the importance of R_t to the future utterances F_t.

Current Reward: We used long-term feedback as current reward r_t^c:

$$r_t^c = Feed_{LT}(H_t, R_t) \quad (9)$$

[3] https://github.com/microsoft/GODEL

Future Reward: The entire process of calculating future reward r_t^f is shown in Figure 3. We used the User Simulator as generator G and the top-k sampling to generate future dialogues, while using TEval as a scorer S to pick k-best future utterances. By using G and S twice, we finally get the set of future utterances $F_t = (f_1, f_2, \dots, f_k)$ about R_t and the importance $W_t = (w_1, w_2, \dots, w_k)$ about F_t. Then, we spliced R_t with all f_i to get k <R_t, f_i> pairs. Finally it is scored using TEval and weighted using w_i:

$$r_t^f = \sum_{i=1}^{k} w_i Feed_{ST}(R_t, f_i) \tag{10}$$

The final reward r_{final} consists of r_t^c and r_t^f:

$$r_{final} = \delta_1 r_t^c + \delta_2 r_t^f \tag{11}$$

where δ_1 and δ_2 are learnable weights for training.

4 Experiment Setup

4.1 Datasets

Following the works (Cai et al., 2020; Sun et al., 2023) on augmented dialog generation, we evaluated our method over two English open domain dialogue datasets: DailyDialog (Li et al., 2017) and Opensubtitles (Lison and Tiedemann, 2016). For the purpose of our study, we conducted pre-processing on these two datasets. We extracted $T-1$ single-turn dialogues [(u_1, u_2), (u_2, u_3), ..., (u_{T-1}, u_T)] from a multi-turn dialogue ($u_1, u_2, \dots, u_T$), where u represents an utterance. We gathered all dialogue pairs, reduced the repeat pairs, and divided them into training, validation and test sets. We splited the DailyDialog dataset to 54,894/6,000/5,700, and OpenSubtitles to 64,000/8,000/8,000.

4.2 Baselines

We compared our method on following classic generation baselines: (1) **Transformer** (Vaswani et al., 2017): an encoder-decoder architecture that relies solely on attention mechanisms. (2) **GPT-2** (Radford et al., 2019): a large-scale pre-trained language model, which is fine-tuned using the full training dataset. (3) **PLATO** (Bao et al., 2021): a pretrained dialogue generation model based on UniLM (Dong et al., 2019). We utilized the v1 version with the 132M parameters. (4) **BART** (Lewis et al., 2020): a pretrained sequence-to-sequence transformer model. We used the base version with the 110M parameters.

4.3 Implementation Details

Our implementation is based on the open-source toolkit Transformers (Wolf et al., 2020). All the experiments are conducted on 4 nvidia 3090 24GB GPUs. We adopted AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-5 and a batch size of 16 for training. The best performing model on the validation set was selected for testing. The choice of all the hyperparameters depends on how they perform on the development set. We used top-k sampling (Fan et al., 2018) during dialogue generation with the temperature as 0.7. The discount factor γ to 0.95 during the reward accumulation phase of RL. During the iteration of reinforcement learning, we can choose different Iteration variable k to balance the computation cost and the performance. We test $k \in \{3, 5, 7\}$ on both datasets and find k = 5 can get moderate long-term training difficulty and better performance.

4.4 Main Results

4.4.1 Automatic Evaluation

We instantiated our method on a set of mainstream dialogue generation models including Transformer, GPT-2, PLATO and BART. The automatic evaluation results are shown in Table 1. For each pair such as Plato+LST (i.e., using the long- and short-term training method in this paper) vs. Plato (i.e., using the original MLE method), as can be seen, our model consistently surpasses vanilla baselines across all automated metrics on both datasets, underscoring the effectiveness and versatility of our methodology. It can be observed that the results of GPT2 and Tranformer improve more significantly compared to models with larger number of parameters such as BART and PLATO. This is mainly due to the fact that large-scale models such as BART and PLATO already had a better dialogue generation capability.

4.4.2 Human Evaluation

Table 2 reports the result of human evaluation. We employed three annotators to assess the quality of the responses generated on the DailyDialog dataset. Totally, 100 randomly sampled responses generated by each model are rated by each annotator on six different dimensions. Ratings ranged from 1 to 3, indicating poor, normal and good (Assessment questions are consistent with the quality description in Appendix A). We conducted two groups of experiments. The first group

	Models	Dist-1 ↑	Dist-2	Dist-3	BLEU-1 ↑	BLEU-2	BLEU-3	BLEU-4	Len.	PPL ↓
(a)	Transformer	1.19	6.21	15.13	13.29	2.28	1.13	0.76	6.74	25.25
	Transformer (♠)	**4.55**	**16.64**	**32.14**	**33.76**	**27.91**	**26.55**	**18.78**	**12.24**	**18.07**
	GPT-2	2.16	7.44	16.15	15.27	2.84	1.66	0.78	7.01	15.91
	GPT-2 (♠)	**4.96**	**19.04**	**35.77**	**17.63**	**25.15**	**19.73**	**11.45**	**10.74**	**14.39**
	PLATO	6.20	28.41	49.93	31.81	24.85	20.42	16.12	8.31	12.73
	PLATO (♠)	**6.29**	**33.76**	**65.28**	**36.18**	**29.92**	**24.97**	**19.62**	**11.22**	**7.65**
	BART	7.64	24.42	40.99	9.97	8.01	6.69	5.33	6.35	11.59
	BART (♠)	**7.95**	**34.16**	**64.39**	**18.08**	**13.96**	**12.47**	**8.17**	**10.87**	**6.48**
(b)	Transformer	1.57	3.28	6.39	8.76	2.35	1.21	0.87	7.41	27.83
	Transformer (♠)	**3.81**	**9.45**	**23.67**	**38.13**	**30.77**	**26.38**	**22.69**	**11.02**	**25.76**
	GPT-2	3.12	4.32	7.29	10.97	3.30	2.15	1.15	8.47	26.47
	GPT-2 (♠)	**5.07**	**13.66**	**30.17**	**23.96**	**23.42**	**18.05**	**13.43**	**9.85**	**25.16**
	PLATO	6.06	25.55	47.82	35.81	28.85	24.14	19.26	10.83	24.25
	PLATO (♠)	**7.45**	**25.87**	**63.65**	**45.76**	**35.98**	**31.63**	**26.07**	**13.91**	**5.51**
	BART	7.17	23.92	39.30	10.24	8.37	7.06	5.65	7.51	21.84
	BART (♠)	**9.87**	**36.75**	**68.43**	**16.32**	**15.33**	**15.81**	**9.04**	**10.12**	**4.43**

Table 1: Automatic evaluation results (%) on (a) DailyDialog and (b) OpenSubtitles. The symbol "♠" indicates that the model is trained using our proposed method. Bold numbers mean that the enhancement to the best baseline is statistically signifcant (a two-tailed paired t-test with p-value <0.01).

	PLATO+LST	PLATO	STDial	LTDial
Flu.	**2.67**	2.14	**2.01**	1.79
Rel.	**2.55**	2.02	**1.95**	1.43
Div.	**2.19**	1.94	**1.74**	1.68
Eng.	**2.41**	1.58	1.04	**1.67**
Coh.	**2.20**	1.86	1.15	**1.73**
Inf.	**2.25**	2.11	1.35	**2.01**

Table 2: Human evaluation results in terms of turn-level quality (up) and dialogue-level quality (down). STDial and LTDial is instantiated on the naive Transformer.

(PLATO+LST vs. PLATO) proves the effectiveness of the training method proposed in this paper, and the PLATO model using our method signifcantly outperforms the origin PLATO in terms of all the evaluation aspects. In the second group (STDial vs. LTDial), STDial has better performance in turn-level quality, while LTDial has better performance in dialogue-level quality. This suggests that long-term and short-term training methods (LSTDial) can accomplish the long-term (Engagement/Coherence/Informativeness) and short-term(Fluency/Diversity/Relevance) objectives of dialog generation, respectively.

4.4.3 LLMs Evaluation

Since rule-based evaluation metrics are not effective in evaluating the performance of dialogue generation, we turn our attention to large language models. Specifically, we used GPT-4 (Bubeck et al., 2023) to conduct a dialogue evaluation of LSTDial and its comparison models on the same 100 well-collected samples as in the previous section. This process begins with an instruction that directs the Large Language Models (LLMs) to provide both reasons and ratings for a given dialogue. To enhance the comprehension of LLMs, we also supply evaluation criteria standards. See Appendix C for complete instruction. In Figure 4 the results of the GPT-4 Evaluation are presented. In the left figure, it is demonstrated that LSTDial improves the results across the board in six areas, while STDial and LTDial are only effective in the short- or long-term qualities. The right figure, on the other hand, demonstrates the significant enhancement of using LSTDial on different models.

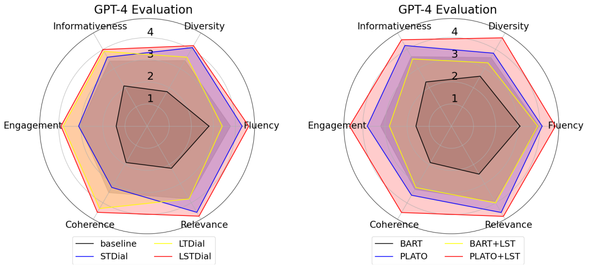

Figure 4: GPT4-Evaluation. Variant of LSTDial (left); Original model using the LSTDial method (right).

4.5 Ablation Study

We list the results of the ablation study in Table 3, aiming to investigate the influence of different modules in our proposed method. It can be seen that if we removed STDial or LTDial, the performance of all metrics drops. This suggests that both the short-term and long-term training modules proposed in this paper are important for enhanced dialogue generation. We also observed that after removing LTDial, the Distinct-1/2/3 are reduced by 41.32%

	Dist-1 ↑	Dist-2	Dist-3	BLEU-1 ↑	BLEU-2	BLEU-3	BLEU-4	Len.	PPL ↓
LSTDial	**4.55**	**16.64**	**32.14**	**33.76**	**27.91**	**26.55**	**18.78**	**12.24**	**18.07**
w/o STDial	3.75	14.28	30.07	9.48	2.04	1.28	0.66	10.48	19.31
w/o LTDial	2.67	10.38	15.74	31.91	24.89	23.73	16.77	8.97	20.42
STDial	**2.67**	**10.38**	**15.74**	**31.91**	**24.89**	**23.73**	**16.77**	**8.97**	**20.42**
w/o ST-Feed	1.19	6.21	15.13	34.33	25.43	24.59	16.76	6.74	25.25
LTDial	**3.75**	**14.28**	**30.07**	**9.48**	**2.04**	**1.28**	**0.66**	**10.48**	**19.31**
w/o LT-Feed	2.11	9.38	20.56	9.32	0.98	1.21	0.43	9.83	18.70
w/o Future Reward	3.04	12.76	25.73	8.45	1.47	1.09	0.65	10.21	19.25

Table 3: The ablation test of our model (%) on DailyDialog, which is implemented on the basic Transformer. Numbers in bold mean that the improvements to the ablation models are statistically signifcant.

/ 37.62% / 51.03%, respectively. However, the BLEU-1/2/3/4, only dropped an average of 9.39%. This indicates that LTDial, has a greater impact on diversity metric. An more obvious phenomenon is the dramatic reduction of BLEU-1/2/3/4 by 71.92% / 92.69% / 95.18% / 96.49% respectively after removing STDial. This is due to the fact that STDial incorporates MLE training, which is capable of generating dialogue responses with high BLEU values. When only the RL-based LTDial is left, the BLEU values naturally drops significantly since ground-truth is not involved in the training. Furthermore, most of the metrics decreased after removing the corresponding feedback in both STDial and LTDial, which are in line with our expectations.

Metric	Pearson Corr	Spearman Corr
Flu.	0.24	0.25
Rel.	0.48	0.45
Div.	0.35	0.33
Eng.	0.28	0.27
Coh.	0.41	0.43
Inf.	0.34	0.34

Table 4: Fine-grained correlation of short-term feedback (top) as well as long-term feedback (down) with human evaluation, respectively. All values are statistically significant to p < 0.05, unless in italic.

4.6 Discussion

We aim to address the following research questions (RQs) about LSTDial: (1) Does the two-type feedback work effectively enough? (2) How does the order of the two-stage training influence the method? (3) Can our method outperform other RL-based methods? (4) Can it be effectively implemented on LLMs? (5) What insights can be drawn from the case study?

RQ1: Feedback Validity

The previous section 4.4.2 presents the 100 samples scored by human. To verify the validity of the two types of feedback, we calculated the fine-grained correlation of short-term feedback as well as long-term feedback with human evaluation on there samples, respectively. From Table 4, it can be observed that the two types of feedback has significant correlation with human evaluation[4]. Considering the substantial individual differences in evaluating open-domain conversations, we observed that our evaluators (contained TEval and DEval) can be used to provide effective feedback for a human-chatbot conversation.

RQ2: Impact of Training Order

We empirically find that it is easier for dialogue system to converge when training at the long-term stage after it is already trained at the short-term stage. Thus, we only report the results of the designed order (short-term first and then long-term) in the paper. Here, we compare with the other order (long-term first and then short-term) and report the results at different RL steps, as shown in Table 5, using the GPT-4 evaluation guidelines given in Appendix C. The experiments employed the PLATO model trained by our LSTDial. The results show that the improvement in the long-term first is less efficient compared to the short-term first.

Steps	Short-term First			Long-term First		
	Eng.	Coh.	Inf.	Eng.	Coh.	Inf.
20000	2.68	2.94	2.48	1.83	2.07	1.80
40000	3.27	3.34	3.08	2.11	2.26	1.90
60000	3.62	3.81	3.56	2.14	2.20	1.92

Table 5: Results of dialogie-level evaluation of two order settings at different steps of long-term training.

RQ3: Comparison with RL-based methods

We conducted experiments to compare the effectiveness of LSTDial with other RL-based ap-

[4] Pearson correlation of 0.2-0.5 can be proved to be moderately to highly significant with human ratings in dialogue evaluation tasks (Yeh et al., 2021).

	Models	Dist-1 ↑	Dist-2	Dist-3	BLEU-1 ↑	BLEU-2	BLEU-3	BLEU-4	Len.	PPL ↓
(a)	VHRL (2020)	4.10	13.73	25.07	32.45	25.72	**26.68**	17.73	9.41	21.22
	HRLG (2023)	4.21	12.65	20.33	31.43	26.89	24.95	15.47	10.73	19.86
	LSTDial (ours)	**4.55**	**16.64**	**32.14**	**33.76**	**27.91**	26.55	**18.78**	**12.24**	**18.07**
(b)	InstructGPT	13.21	40.76	69.75	38.15	**30.68**	25.71	19.95	14.33	6.41
	InstructGPT (♠)	**15.10**	**41.13**	**69.76**	**39.14**	29.92	**26.40**	**20.48**	**15.02**	**6.38**

Table 6: Automatic evaluation results (%) on DailyDialog. (a) Comparison of our approach (instantiated on the naive Transformer) with two RL-based dialogue generation methods. (b) We implemented our approach on InstructGPT (denoted by "♠") to explore the applicability of LSTDial for Large Language Models (LLMs).

proaches applied to dialogue generation tasks. Table 6(a) shows the results of LSTDial as well as two open-source RL-based methods VHRL (Saleh et al., 2020) and HRLG (Rohmatillah and Chien, 2023) on DialyDialog. The above experimental results show that our LSTDial outperforms both of VHRL and HRLG in terms of all metrics except for BLEU-3 (where LSTDial is slightly inferior to VHRL). In particular, for Dist-3, BLEU-4, and PPL, LSTDial exhibits a significant advantage over other methods, showing that the model enhanced with our method can generate responses that are more diverse, fluent, and similar responses to the reference. In addition, the reasons why the RL algorithm should be used in the long-term stage rather than short-term are given in Appendix D.

RQ4: Implementation with LLMs

To investigate whether LSTDial is effective on LLMs, we implemented our approach on a small-sized LLM called InstructGPT (1.3B) [5] (Ouyang et al., 2022) and showed its comparative results on the DailyDialog dataset in Table 6(b). The above experimental results show that (1) InstructGPT outperforms PLATO and even PLATO(♠) in terms of all metrics, showing the powerful ability of LLMs for our task; (2) InstructGPT(♠) slightly outperforms InstructGPT in terms of all metrics except for BLEU-2, showing that our method is still effective for the implementation on LLMs.

RQ5: Case Study

We randomly sampled cases of dialogue responses from DailyDialog and then compared the MLE to our approach (implemented on the PLATO model). From Table 7(a), it can be seen that the model using LSTDial can generate a more diverse and interesting response. We have further observed responses generated by our model across different stages in Table 7(b). It can be observed that the baseline's response (i.e., the original PLATO without our trainings method) is simple and generic, while the STDial response is more informative and relevant. In particular, LSTDial tends to ask questions and show emotion to further deepen the dialogue. This response demonstrates long-term capabilities at the dialogue-level. See Appendix E for more case studies along with detailed explanations.

- So you have all the news on broadcast?		
(a)	**Ground-truth**:	Yes , it's so helpful.
	MLE:	Yep, got it.
	LSTDial:	Yes, I do. It keeps me informed!
- You need to know the type of shoes you want .		
(b)	**Baseline**:	I'm not sure.
	STDial:	Guess I'll decide on the shoes once I'm at the store.
	LSTDial:	Knowing the type of shoes I want is key, but have you ever had a pair of shoes that pleasantly surprised you?

Table 7: (a) Response generated by MLE and our method (the top case). (b) Case of response generation at different stages of the model (the bottom case).

5 Conclusion

In this study, we introduced LSTDial, a novel framework that utilizes turn- and dialogue-level conversation evaluation as explicit feedback during training. By bridging the gap between dialogue generation and evaluation, LSTDial enables dialogue systems to exhibit both short-term capabilities at the turn-level and long-term capabilities at the dialogue-level. Experimental results on two popular open-domain dialogue datasets demonstrate the significant improvement achieved by LSTDial in terms of both human and automatic evaluation metrics. These findings highlight the importance of incorporating conversation evaluation into dialogue system training for generating high-quality responses. Future research will explore more feedback mechanisms to better utilize feedback to enhance dialogue generation.

[5] We reproduced InstructGPT based on: `https://github.com/LanXiu0523/RLHF_instructGPT`

Limitation

Firstly, due to limited resources, this paper did not attempt to use larger evaluation models to generate feedback, as well as LLMs with a larger number of parameters to apply our approach. Secondly, in the §3.4.1 RL Iteration phase, we defaulted to taking the first utterance of the dialogue sample as the initialized dialogue history, without attempting to use a greater number of utterances. This may affect the difficulty and effectiveness of dialogue interaction, as a greater number of initial utterances would lead to richer background knowledge for the conversation. Finally, a key factor contributing to the successful performance of LSTDial is the sequence of short-term and long-term training, meaning that LSTDial needs to be trained strictly in order.

Ethics Statement

There are no ethical issues involved in this study. All datasets used in this paper are publicly available. Due to the limitations of rule-based metrics, we conducted a human evaluation of the response generation quality. We recruited three part-time postgraduate students to conduct the human evaluation with clearly defined evaluation rules. They were paid 5 CNY per sample for their work during the evaluation period.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62076092).

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2513–2525. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6334–6343. Association for Computational Linguistics.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2023. Generating better responses from user feedback via reinforcement learning and commonsense inference. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, pages 376–387.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6583–6591.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *CoRR*, abs/2310.13650.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2078–2100.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. 2018. An introduction to deep reinforcement learning. *CoRR*, abs/1811.12560.

Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7789–7796. AAAI Press.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 934–950. Association for Computational Linguistics.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 3953–3957. ACM.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9230–9240. Association for Computational Linguistics.

Zhihua Jiang, Guanghui Ye, Dongning Rao, Di Wang, and Xin Miao. 2022. Im2: an interpretable and multi-category integrated metric framework for automatic dialogue evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 11091–11103. Association for Computational Linguistics.

Katharina Kann, Abteen Ebrahimi, Joewie J. Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*, pages 148–165. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

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, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations, ICLR*.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235.

John Mendonça, Patrícia Pereira, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. *CoRR*, abs/2308.16797.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1439–1450. Association for Computational Linguistics.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: a systematic survey. *Artif. Intell. Rev.*, 56(4):3055–3155.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe.

2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

ChaeHun Park, Seungil Chad Lee, Daniel Rim, and Jaegul Choo. 2023. Density: Open-domain dialogue evaluation metric using density estimation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14222–14236.

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. *CoRR*, abs/2206.11309.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pages 4164–4178. International Committee on Computational Linguistics.

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

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander Rudnicky. 2023. Overview of robust and multilingual automatic evaluation metrics for open-domain dialogue systems at DSTC 11 track 4. *CoRR*, abs/2306.12794.

Mahdin Rohmatillah and Jen-Tzung Chien. 2023. Hierarchical reinforcement learning with guidance for multi-domain dialogue policy. *IEEE ACM Trans. Audio Speech Lang. Process.*, 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, and Jason Weston. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind W. Picard. 2020. Hierarchical reinforcement learning for open-domain dialog. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020*, pages 8741–8748. AAAI Press.

Shiki Sato, Reina Akama, Hiroki Ouchi, and Jun Suzuki and. 2020. Evaluating dialogue generation systems via response selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 593–599. Association for Computational Linguistics.

Katherine Stasaski and Marti A. Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 85–98. Association for Computational Linguistics.

Bin Sun, Yitong Li, Fei Mi, Weichao Wang, Yiwei Li, and Kan Li. 2023. Towards diverse, relevant and coherent open-domain dialogue generation via hybrid latent variables. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 13600–13608. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Reddy Chandra, and Nanyun Peng. 2022. Endex: Evaluation of dialogue engagingness at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4884–4893. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706.

Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. 2023. KRLS: improving end-to-end response generation in task oriented dialog with reinforced keywords learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12338–12358.

Chen Zhang, Luis F. D'Haro, Chengguang Tang, Ke Shi, Guohua Tang, and Haizhou Li. 2023. xdial-eval: A multilingual open-domain dialogue evaluation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5579–5601.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022. Finedeval: Fine-grained automatic dialogue-level evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3336–3355. Association for Computational Linguistics.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael E. Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *CoRR*, abs/2111.02110.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill 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, ACL 2020, Online, July 5-10, 2020*, pages 270–278.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12481–12490.

Ying Zhu, Bo Wang, Dongming Zhao, Kun Huang, Zhuoxuan Jiang, Ruifang He, and Yuexian Hou. 2023. Grafting fine-tuning and reinforcement learning for empathetic emotion elicitation in dialog generation. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, pages 3148–3155.

A Quality Description

The dialogue metrics used in this paper include the following two categories.

Turn-level:

Fluency: Responses are free of grammatical and semantic errors.

Relevance: Responses are on-topic with the immediate dialogue history.

Diversity: The response is informative, with long sentences including multiple entities and conceptual or emotional words.

Dialogue-level:

Engagement: Throughout the dialogue, the system displays a likeable personality.

Coherence: Throughout the dialogue, the system can maintain a good conversation flow.

Informativeness: Throughout the dialogue, the system provides unique and non-generic information.

B Feedback Details

B1: ST-Feed Details

Fluency: We used VUP (Valid Utterance Prediction) proposed by USL-H (Phy et al., 2020) for Fluency. The authors trained a model based on BERT-base to capture the Fluency of an utterance by classifying whether it is valid. For doing this, they applied many rules to get a negative sample, e.g., word reorder, word drop, and word repeat. We ran VUP via following the original setting. **Relevance**: Similar to the Fluency metric, We proposed this metric to enhance the relevance prediction by using negative sampling. Specifically, the objective is to discern whether a given context-response pair is relevance or not. To create positive samples (c_i, r_i^+) for this binary classification task, we leveraged two consecutive utterances (u_i, u_{i+1}) extracted from existing human-human dialogue corpora (Li et al., 2017), where u_i serves as the context and u_{i+1} represents the corresponding response. Similar to the work of Sato et al. (Sato et al., 2020), we selected the utterances in the dataset that are the most similar to the positive example as the negative samples (c_i, r_i^-). RoBERTa-base (Liu et al., 2019) is adopted as the text encoder. **Diversity**: According to previous studies (Stasaski and Hearst, 2022), these lexical-level evaluation metrics such as Distinct-n (Dist-*n*) (Li et al., 2016) often fail to capture semantic diversity. So we used a recent automatic evaluation metric, Sem-Ent (Han et al., 2022), which can measure the semantic diversity based on the semantic distribution of generated responses. Sem-Ent correlates with human judgments on response diversity more than other automatic diversity metrics.

B2: LT-Feed Details

Engagement: Engagement is widely acknowledged as a crucial evaluation criterion for assessing the quality of dialogue system (Ghazarian et al., 2020). As a result, we trained a discriminatory model using the RED (Reddit-based Engagement Dataset) (Xu et al., 2022). The RED dataset is derived from Reddit and is carefully curated using a unique distant-supervision framework. In this framework, emotional, attentional, behavioral, and reply engagement are combined to form a single score called ENDEX. Subsequently, a hyperparameter threshold is utilized to group posts into positive and negative samples based on this score. **Coherence**: Coherence is a dialogue-level metric that measures how well the dialogue flows, showing how the utterances are coordinated for a seamless interaction (Mesgar et al., 2020). We consider the use of the GRADE (Huang et al., 2020) as our model for coherence evaluation. GRADE is a graph-enhanced dialogue evaluation model that uses both utterance-level contextualized representations and topic-level graph representations to evaluate the response. We got k - 1 adjacency pairs for a dialogue containing k utterances and hence k - 1 coherence scores. The coherence score at the dialogue-level is calculated by averaging the k - 1 scores. **Informativeness**: During the human-human interaction, when the interlocutors deeply dive into a topic, they tend to convey a large amount of information (Zhang et al., 2022). We adopt a pre-trained natural language inference (NLI) model[6] to calculate the topic depth of a conversation to represent the informativeness of a response. More specifically, given a dialogue of k utterances, a pre-trained NLI model is used to provided entailment score to each utterance pair in the dialogue. In total, there are $\frac{k(k-1)}{2}$ entailment scores per dialogue. The dialogue-level entailment score is the average of all utterance-pair entailment scores in the dialogue.

C GPT-4 Evaluation Guideline

Figure 5 and Figure 6 elaborate the turn-level and dialogue-level GPT-4 evaluation guidelines, respectively. Steps to conduct the evaluation are:

Step1. Clarify task requirements by reading the instruction.

Step2. Read the dialogue, and the system response carefully.

Step3. Give some brief analysis from the aspects mentioned before.

Step4. Read the Definitions and Criteria provided above to help you with your in-depth analysis.

Step5. On a scale of 1-5, evaluate the above three aspects of the dialogue and produce the required output.

[6] https://huggingface.co/roberta-large-mnli

Instruction: Next is a sample of a turn-level dialogue, which is a combined pair containing a context and a response. Assuming that you are an expert in dialogue quality assessment, you are asked to assess the current dialogue response based on the context. First, please analyze the quality of the dialogue in terms of the following three turn-level dialogue dimensions:
1.Fluency.
2.Relevance.
3.Diversity.
Based on the above analysis, provide scores for each of the above three aspects from the set [1, 2, 3, 4, 5] and output the three scores as a list. *Output example*: [Fluency:4, Relevance:4, Diversity:3].
To help you better evaluate, here is the dialogue dimensions Definitions:
1.Fluency: Responses are free of grammatical and semantic errors.
2.Relevance: Responses are on-topic with the immediate dialogue history.
3.Diversity: The response is informative, with long sentences including multiple entities and conceptual or emotional words.
To help you better evaluate, here is the evaluation Criteria:
A score of 1 means very dissatisfied.
A score of 2 means dissatisfied.
A score of 3 means normal.
A score of 4 means satisfied.
A score of 5 very satisfied.
Steps to conduct the evaluation are:
1.Read the dialogue, and the response carefully;
2.Give some brief analysis from the aspects mentioned before;
3.Read the Definitions and Criteria provided above to help you with your in-depth analysis.
4.on a scale of 1-5, evaluate the above three turn-level aspects of the dialogue and produce the required output.

Figure 5: Turn-level GPT-4 Evaluation Guideline.

Instruction: Next is the entire human-system conversation, which contains multiple dialogue utterances. Assuming that you are an expert in dialogue quality assessment, you are asked to assess the performance of the system throughout the conversation. First, please analyze the quality of the dialogue in terms of the following three dialogue-level dialogue dimensions:
1.Engagement.
2.Coherence.
3.Informativeness.
Based on the above analysis, provide scores for each of the above three aspects from the set [1, 2, 3, 4, 5] and output the three scores as a list. *Output example*: [Engagement:2, Coherence:4, Informativeness:3].
To help you better evaluate, here is the dialogue dimension Definitions:
1.Engagement: Throughout the dialogue, the system displays a likeable personality.
2.Coherence: Throughout the dialogue, the system is maintaining a good conversation flow.
3.Informativeness: Throughout the dialogue, the system provides unique and non-generic information.
To help you better evaluate, here is the evaluation Criteria:
A score of 1 means very dissatisfied.
A score of 2 means dissatisfied.
A score of 3 means normal.
A score of 4 means satisfied.
A score of 5 very satisfied.
Steps to conduct the evaluation are:
1.Read the dialogue, and the system response carefully;
2.Give some brief analysis from the aspects mentioned before;
3.Read the Definitions and Criteria provided above to help you with your in-depth analysis.
4.on a scale of 1-5, evaluate the above three dialogue-level aspects of the dialogue and produce the required output.

Figure 6: Dialogue-level GPT-4 Evaluation Guideline.

D Why not Use RL at Short-term Stage

The RL-based training is more complicated and time-consuming than the multi-task learning. At the short-term stage, we leverage the multi-task learning to quickly train a high-quality dialogue system via turn-level feedback. After that, we move to the long-term stage and continuously train the resulting system via the complicated RL-based training to enhance replies considering the quality of the whole dialogue. However, if we utilize the RL-based training in the first stage, such design will make the structure of LSTDial very complicated, probably bringing negative influence to the model's reliability and generalization in practical scenarios.

E Detailed Case Analysis

Our model has the following three incremental variants as it is trained:

(a) **Baseline**: the original dialogue system without our trainings.

(b) **STDial**: baseline + short-term training.

(c) **LSTDial**: STDial + long-term training.

User: I'm considering going for a run.
Baseline: Oh, that's good.
STDial: Running is a great way to stay energized and enjoy the outdoors.
LSTDial: Running is fantastic! It's an excellent way to boost your mood and energy levels. Are you planning to run through a park or nearby streets?

Table 8: Case of response generation at different stages of our model (case 1).

User: Hi, what's the weather like today?
Baseline: I'm not sure.
STDial: Today's weather is sunny with a temperature of 25 degrees Celsius.
LSTDial: Today's weather is sunny with a temperature of 25 degrees Celsius. It's a great day to go outside and enjoy the sunshine. What are your plans for the day?

Table 9: Case of response generation at different stages of our model (case 2).

We analyze the change of generated responses across these models of different stages with the following two random samples obtained from user inputs to the above models.

Analysis1 (See table 8):

(1) The baseline's response is simple and generic, expressing a positive attitude towards the user's intention, but it lacks specific information or elements to further engagement in the conversation. (i.e., Lacking of short- and long-term capabilities)

(2) The STDial response is more informative and relevant due to mentioning the benefits of running for staying energized and enjoying the outdoors. However, it still lacks a proactive conversation guidance. (i.e., Having short-term capabilities but lacking of long-term capabilities)

(3) The LSTDial response goes a further step, not only affirming the benefits of running but also adding a subjective evaluation and asking a question to further engage in a deeper-in conversation. This kind of response not only facilitates a deeper conversation but also shows an interest in the user's preferences and plans, thereby enhancing user engagement. (i.e., Having both short-term and long-term capabilities)

Analysis2 (See table 9):

(1) The baseline model provides a poor response, displaying uncertainty about the weather.

(2) The STDial model offers a specific and informative response about the weather, showing the improved diversity and relevance. Compared to the baseline's response, the STDial response exhibits the short-term capabilities (i.e., fluency, relevance, and diversity).

(3) The LSTDial model enriches the response by adding a subjective evaluation of the weather and engaging the user in the conversation, increasing participation.

Compared to STDial, we observe that LSTDial tends to ask questions and express opinions actively, which is very friendly for users to participate in multiple rounds of conversations. Therefore, after the two-stage training, LSTDial can enhance long-term capabilities (i.e., coherence, engagement, and informativeness) beyond short-term capabilities.