

# PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning

Siqi Bao\* Huang He\* Fan Wang\* Hua Wu\* Haifeng Wang  
Wenquan Wu Zhen Guo Zhibin Liu Xinchao Xu

Baidu Inc., China

{baosiqi, hehuang, wang.fan, wu.hua}@baidu.com

## Abstract

To build a high-quality open-domain chatbot, we introduce the effective training process of PLATO-2 via curriculum learning. There are two stages involved in the learning process. In the first stage, a coarse-grained generation model is trained to learn response generation under the simplified framework of one-to-one mapping. In the second stage, a fine-grained generative model augmented with latent variables and an evaluation model are further trained to generate diverse responses and to select the best response, respectively. PLATO-2 was trained on both Chinese and English data, whose effectiveness and superiority are verified through comprehensive evaluations, achieving new state-of-the-art results.

## 1 Introduction

Recently, task agnostic pre-training with large-scale transformer models has achieved great success in natural language processing (Devlin et al., 2019), especially open-domain dialogue generation. For instance, based on the general language model GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020) is further trained for response generation using Reddit comments. To obtain a human-like open-domain chatbot, Meena (Adiwardana et al., 2020) scales up the network parameters to 2.6B and employs more social media conversations in the training process, leading to significant improvement on response quality. To mitigate undesirable toxic or bias traits of large corpora, Blender (Roller et al., 2021) fine-tunes the pre-trained model with human annotated datasets and emphasizes desirable conversational skills of engagingness, knowledge, empathy and personality.

In addition to the attempts from model scale and data selection, PLATO (Bao et al., 2020) aims

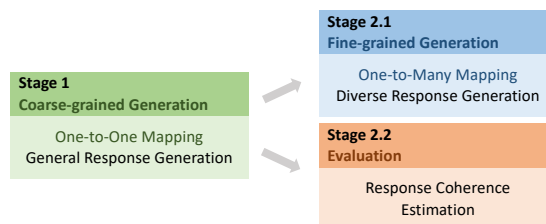


Figure 1: Curriculum learning process in PLATO-2.

to tackle the inherent one-to-many mapping problem to improve response quality. The one-to-many mapping refers to that one dialogue context might correspond to multiple appropriate responses. It is widely recognized that the capability of modeling one-to-many relationship is crucial for open-domain dialogue generation (Zhao et al., 2017; Chen et al., 2019). PLATO explicitly models this one-to-many relationship via discrete latent variables, aiming to boost the quality of dialogue generation. PLATO has a modest scale of 132M network parameters and trained with 8M samples, achieving relatively good performance among conversation models on a similar scale. However, scaling up PLATO directly encounters training instability and efficiency issues, which might result from the difficulty to capture the one-to-many semantic relationship from scratch.

In this work, we try to scale up PLATO to PLATO-2 and introduce an effective training schema via curriculum learning (Bengio et al., 2009). There are two stages involved in the whole learning process, as shown in Figure 1. In the first stage, under the simplified one-to-one mapping modeling, a **coarse-grained** generation model is trained for response generation under different conversation contexts. This model tends to capture typical patterns of diversified responses, sometimes resulting in general and dull responses during inference. Despite the problem of safe responses,

\*Equal contribution.

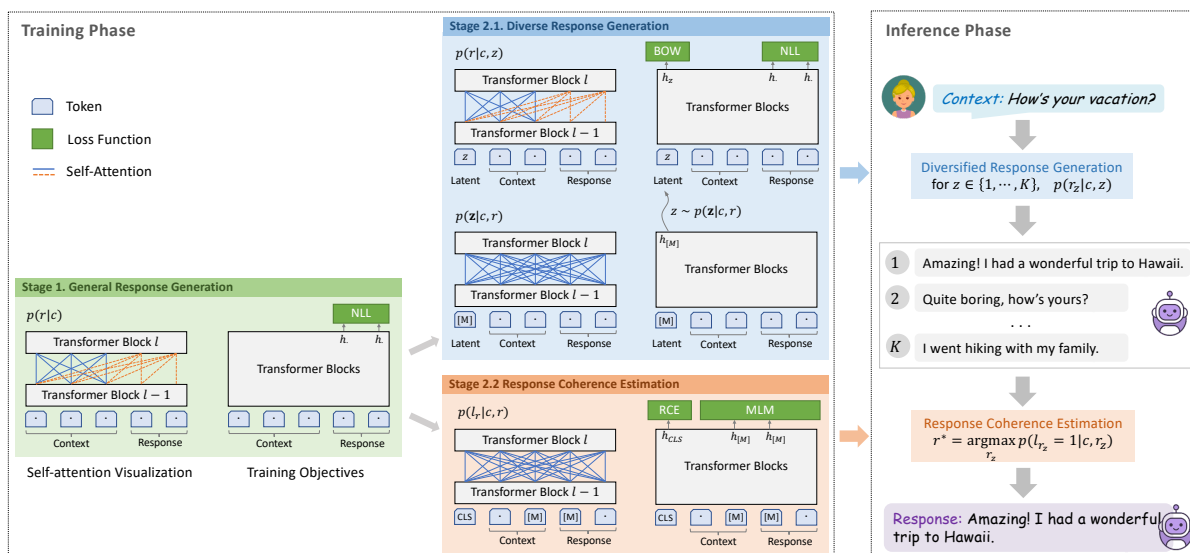


Figure 2: PLATO-2 illustration. Left: training phase via curriculum learning, model parameters in the second stage are warm started by those trained well in the first stage. Right: toy example to illustrate inference phase.

this coarse-grained model is still highly effective in learning general concepts of response generation.

The curriculum learning continues to the second stage, which contains the training of a **fine-grained** generation model and an evaluation model. The fine-grained generation model explicitly models the one-to-many mapping relationship via latent variables for diverse response generation. To select the most appropriate response, an evaluation model is trained to estimate the bi-directional coherence between the dialogue context and responses. Distinct with multi-task PLATO, the separate design of fine-grained generation and evaluation enables the model to concentrate more on its corresponding task, getting exempt from multi-task disturbance (Standley et al., 2020).

As compared with PLATO, PLATO-2 leverages curriculum learning to learn response generation gradually, from the general concept of one-to-one mapping to the complex concept of one-to-many mapping. With curriculum learning, we successfully scale the model up to billions of parameters, achieving new state-of-the-art results. Besides open-domain chitchat, the models learned in these two stages can also benefit task-oriented conversation and knowledge grounded dialogue respectively, whose effectiveness is verified thoroughly in DSTC9 (Gunasekara et al., 2020).

To sum up, we trained PLATO-2 with different model sizes: 1.6B, 314M and 93M parameters. In addition to the English models, we also trained Chinese models with massive social media conversa-

tions. Comprehensive experiments on both English and Chinese datasets demonstrate that PLATO-2 outperforms Meena, Blender and other state-of-the-art models. We have released our English models and source codes at GitHub, hoping to facilitate the research in open-domain dialogue generation.<sup>1</sup>

## 2 Methodology

The backbone of PLATO-2 is consisted of transformer blocks with pre-normalization (Radford et al., 2019). Distinct with conventional Seq2Seq, there are no separate encoder and decoder networks in our infrastructure. PLATO-2 keeps the unified network for bi-directional context encoding and uni-directional response generation through flexible attention mechanism (Dong et al., 2019).

### 2.1 Curriculum Learning

In this work, we carry out effective training of PLATO-2 via curriculum learning. As shown in Figure 2, there are two stages involved in the learning process: during stage 1, a coarse-grained baseline model is trained for general response generation under the simplified one-to-one mapping relationship; during stage 2, two models of fine-grained generation and evaluation are further trained for diverse response generation and response coherence estimation respectively.

<sup>1</sup><https://github.com/PaddlePaddle/Knover/tree/develop/projects/PLATO-2>

### 2.1.1 General Response Generation

It is well known that there exists a one-to-many relationship in open-domain conversations, where a piece of context may have multiple appropriate responses. Since conventional approaches try to fit the one-to-one mapping, they tend to generate generic and dull responses. Whereas, it is still an efficient way to capture the general characteristics of response generation. As such, we first train a coarse-grained baseline model to learn general response generation under the simplified relationship of one-to-one mapping. Given one training sample of context and response  $(c, r)$ , we need to minimize the following negative log-likelihood (NLL) loss:

$$\begin{aligned} \mathcal{L}_{NLL}^{\text{Baseline}} &= -\mathbb{E} \log p(r|c) \\ &= -\mathbb{E} \sum_{t=1}^T \log p(r_t|c, r_{<t}), \end{aligned} \quad (1)$$

where  $T$  is the length of the target response  $r$  and  $r_{<t}$  denotes previously generated words. Since the response generation is a uni-directional decoding process, each token in the response only attends to those before it, shown as dashed orange lines in Figure 2. As for the context, bi-directional attention is enabled for better natural language understanding, shown as blue lines in Figure 2.

### 2.1.2 Diverse Response Generation

Based upon the coarse-grained baseline model, diverse response generation is warm started and further trained under the relationship of one-to-many mapping. Following the previous work PLATO, the discrete latent variable  $z$  is introduced for the one-to-many relationship modeling.  $z$  is one  $K$ -way categorical variable, with each value corresponding to a particular latent speech act in the response. The model will first estimate the latent act distribution of the training sample  $p(\mathbf{z}|c, r)$  and then generate the response with the sampled latent variable  $p(r|c, z)$ . It is notable that these two tasks of response generation and latent act recognition are trained jointly within the shared network. The NLL loss of diverse response generation is defined as:

$$\begin{aligned} \mathcal{L}_{NLL}^{\text{Generation}} &= -\mathbb{E}_{z \sim p(\mathbf{z}|c, r)} \log p(r|c, z) \\ &= -\mathbb{E}_{z \sim p(\mathbf{z}|c, r)} \sum_{t=1}^T \log p(r_t|c, z, r_{<t}), \end{aligned} \quad (2)$$

where  $z$  is the latent act sampled from  $p(\mathbf{z}|c, r)$ . As sampling is not differentiable, we approximate it with Gumbel-Softmax (Jang et al., 2017). The

posterior distribution over latent values is estimated through the task of latent act recognition:

$$p(\mathbf{z}|c, r) = \text{softmax}(W_1 h_{[M]} + b_1) \in \mathbb{R}^K, \quad (3)$$

where  $h_{[M]} \in \mathbb{R}^D$  is the final hidden state of the special mask token [M],  $W_1 \in \mathbb{R}^{K \times D}$  and  $b_1 \in \mathbb{R}^K$  denote the weight matrices of one fully-connected layer.

To facilitate the training process of discrete latent variables, the bag-of-words (BOW) loss (Zhao et al., 2017) is also employed:

$$\begin{aligned} \mathcal{L}_{BOW}^{\text{Generation}} &= -\mathbb{E}_{z \sim p(\mathbf{z}|c, r)} \sum_{t=1}^T \log p(r_t|c, z) \\ &= -\mathbb{E}_{z \sim p(\mathbf{z}|c, r)} \sum_{t=1}^T \log \frac{e^{f_{r_t}}}{\sum_{v \in V} e^{f_v}}, \end{aligned} \quad (4)$$

where  $V$  refers to the whole vocabulary. The function  $f$  tries to predict the words within the target response in a non-autoregressive way:

$$f = W_2 h_z + b_2 \in \mathbb{R}^{|V|}, \quad (5)$$

where  $h_z$  is the final hidden state of the latent variable.  $f_{r_t}$  denotes the estimated probability of word  $r_t$ . As compared with NLL loss, the BOW loss discards word orders and forces the latent variable to capture the global information of target response.

To sum up, the objective of the fine-grained generation model is to minimize the following integrated loss:

$$\mathcal{L}^{\text{Generation}} = \mathcal{L}_{NLL}^{\text{Generation}} + \mathcal{L}_{BOW}^{\text{Generation}} \quad (6)$$

### 2.1.3 Response Coherence Estimation

By assigning distinct values to the latent variable, the fine-grained generation model is able to produce multiple high-quality and diverse responses. To select the most appropriate response from these candidates, one straightforward way is to rank them according to  $p(z|c)p(r|c, z)$ . However, it is widely recognized that the prior distribution  $p(\mathbf{z}|c)$  is difficult to estimate and the uniform distribution is not an effective approximation. To this end, we adopt an alternative approach to train an evaluation model in the second stage, estimating the coherence between each response and the given dialogue context. The loss of response coherence estimation (RCE) is defined as follows:

$$\begin{aligned} \mathcal{L}_{RCE}^{\text{Evaluation}} &= -\log p(l_r = 1|c, r) \\ &\quad -\log p(l_{r^-} = 0|c, r^-) \end{aligned} \quad (7)$$

The positive training samples come from the dialogue context and corresponding target response  $(c, r)$ , with coherence label  $l_r = 1$ . And the negative samples are created by randomly selecting responses from the corpus  $(c, r^-)$ , with coherence label  $l_{r^-} = 0$ .

In addition to our coherence evaluation function  $p(l_r|c, r)$ , there are two other functions widely used for response selection. One is the length-average log-likelihood (Adiwardana et al., 2020), which considers the forward response generation probability  $p(r|c)$ . The other one is the maximum mutual information (Zhang et al., 2020), which considers the backward context recovery probability  $p(c|r)$ . However, the forward score favors safe and generic responses due to the property of maximum likelihood, while the backward score tends to select the response with a high overlap with the context, resulting in repetitive conversations. By contrast, the discriminative function  $p(l_r|c, r)$  considers the bi-directional information flow between the dialogue context and response. Our coherence evaluation is able to ameliorate the aforementioned problems, whose effectiveness is verified in the experiments.

To maintain the capacity of distributed representation, the task of masked language model (MLM) (Devlin et al., 2019) is also included in the evaluation network. Within this task, 15% of the input tokens will be masked at random and the network needs to recover the masked ones. The MLM loss is defined as:

$$\mathcal{L}_{MLM}^{\text{Evaluation}} = -\mathbb{E} \sum_{m \in M} \log p(x_m | x_{\setminus M}), \quad (8)$$

where  $x$  refers to the input tokens of context and response.  $\{x_m\}_{m \in M}$  stands for masked tokens and  $x_{\setminus M}$  denotes the rest unmasked ones.

To sum up, the objective of the evaluation model is to minimize the following integrated loss:

$$\mathcal{L}^{\text{Evaluation}} = \mathcal{L}_{RCE}^{\text{Evaluation}} + \mathcal{L}_{MLM}^{\text{Evaluation}} \quad (9)$$

## 2.2 Inference

For open-domain chitchat, the inference is carried out with the second stage’s models as follows.

- 1) Diverse response generation. Conditioned on each latent value  $z \in \{1, \dots, K\}$ , its corresponding candidate response  $r_z$  is produced by the fine-grained generation model  $p(r_z|c, z)$ .
- 2) Response coherence estimation. The evaluation model will preform ranking and select the one with highest coherence value as the final response  $r^* = \operatorname{argmax}_{r_z} p(l_{r_z} = 1|c, r_z)$ .

## 3 Experiments

### 3.1 Training Data

PLATO-2 has English and Chinese models, with training data extracted from open-domain social media conversations. The English training data is extracted from Reddit comments, which are collected by a third party and made publicly available on pushshift.io (Baumgartner et al., 2020). To improve the generation quality, we carry out elaborate data cleaning, as discussed in the Appendix. After filtering, the data is split into training and validation sets in chronological order. The training set contains 684M (context, response) samples, ranging from December 2005 to July 2019. For the validation set, 0.2M samples are selected from the rest data after July 2019. The English vocabulary contains 8K BPE tokens (Sennrich et al., 2016), constructed with the SentencePiece library.

The Chinese training data is collected from public domain social medias. After filtering, there are 1.2B (context, response) samples in the training set, 0.1M samples in the validation set, and 0.1M samples in the test set. As for the Chinese vocabulary, it contains 30K BPE tokens.

### 3.2 Training Details

PLATO-2 has three model sizes: a standard version of 1.6B parameters, a small version of 314M parameters, and a tiny version of 93M parameters. Detailed network and training configurations are summarized in the Appendix. The main hyperparameters used in the training process are listed as follows. The maximum sequence lengths of context and response are all set to 128.  $K$  is set to 20 for the discrete latent variable (Bao et al., 2020; Chen et al., 2019). We use Adam (Kingma and Ba, 2015) as the optimizer, with a learning rate scheduler including a linear warmup and an invsqrt decay (Vaswani et al., 2017). To train the large-scale model with a relatively large batch size, we employ gradient checkpointing (Chen et al., 2016) to trade computation for memory. The training was carried out on 64 Nvidia Tesla V100 32G GPU cards. It takes about 3 weeks for 1.6B parameter model to accomplish curriculum learning process.

### 3.3 Evaluation Settings

#### 3.3.1 Compared Methods

The following methods have been compared in the experiments.

- PLATO (Bao et al., 2020) is trained on the basis of BERT<sub>BASE</sub> using 8.3M Twitter and Reddit conversations (Cho et al., 2014; Zhou et al., 2018; Galley et al., 2019). There are 132M network parameters in this model.
- DialoGPT (Zhang et al., 2020) is trained on the basis of GPT-2 (Radford et al., 2019) using Reddit comments. There are three model sizes: 117M, 345M and 762M. Since the 345M parameter model obtains the best performance in their evaluations, we compare with this version.
- Blender (Roller et al., 2021) is first trained using Reddit comments and then fine-tuned with human annotated conversations – BST (Smith et al., 2020), to help emphasize desirable conversational skills of engagingness, knowledge, empathy and personality. Blender has three model sizes: 90M, 2.7B and 9.4B. Since the 2.7B parameter model obtains the best performance in their evaluations, we compare with this version.
- Meena (Adiwardana et al., 2020) is an open-domain chatbot trained with social media conversations. There are 2.6B network parameters in Meena. Since Meena has not released the model or provided a service interface, it is difficult to perform comprehensive comparison. In the experiments, we include the provided samples in their paper for static evaluation.
- Microsoft XiaoIce (Zhou et al., 2020) is a popular social chatbot in Chinese. The official Weibo platform is used in the evaluation.

For the sake of comprehensive and fair comparisons, different versions of PLATO-2 are included in the experiments.

- PLATO-2 1.6B parameter model is the standard version in English, which is first trained using Reddit comments and then fine-tuned with BST conversations. To measure the effectiveness of PLATO-2, this model will be compared to the state-of-the-art open-domain chatbot Blender.
- PLATO-2 314M parameter model is a small version in English, which is trained with Reddit comments. This model will be compared to DialoGPT, as they have similar model scales.
- PLATO-2 93M parameter model is a tiny version in English, which is trained with Reddit comments. As it is difficult to scale up PLATO, we use this version to compare with PLATO.
- PLATO-2 336M parameter Chinese model<sup>2</sup> will be compared to XiaoIce in the experiments.

<sup>2</sup>This model has 24 transformer blocks and 16 attention

### 3.3.2 Evaluation Metrics

We carry out both automatic and human evaluations in the experiments. In automatic evaluation, to assess the model’s capacity on lexical diversity, we use the corpus-level metric of distinct-1/2 (Li et al., 2016a), which is defined as the number of distinct uni- or bi-grams divided by the total number of generated words.

In human evaluation, we employ four utterance-level and dialogue-level metrics, including coherence, informativeness, engagingness and humanness. Three crowd-sourcing workers are asked to score the response/dialogue quality on a scale of [0, 1, 2], with the final score determined through majority voting. The higher score, the better. These criteria are discussed as follows, with scoring details provided in the Appendix.

- Coherence is an utterance-level metric, measuring whether the response is relevant and consistent with the context.
- Informativeness is also an utterance-level metric, evaluating whether the response is informative or not given the context.
- Engagingness is a dialogue-level metric, assessing whether the annotator would like to talk with the speaker for a long conversation.
- Humanness is also a dialogue-level metric, judging whether the speaker is a human being or not.

## 3.4 Experimental Results

In the experiments, we include both static and interactive evaluations.

### 3.4.1 Self-Chat Evaluation

Self-chats have been widely used in the evaluation of dialogue systems (Li et al., 2016b; Bao et al., 2019; Roller et al., 2021), where a model plays the role of both partners in the conversation. As compared with human-bot conversations, self-chat logs can be collected efficiently at a cheaper price. As reported in Li et al. (2019), self-chat evaluations exhibit high agreement with the human-bot chat evaluations. In the experiments, we ask the bot to perform self-chats and then invite crowd-sourcing workers to evaluate the dialogue quality.

The way to start the interactive conversation needs special attention. As pointed out by Roller et al. (2021), if starting with ‘Hi!’, partners tend

heads, with the embedding dimension of 1024. As the Chinese vocabulary contains 30K BPE tokens, this model has 22.5M more parameters than the English small model.

Model	Human Evaluation				Automatic Evaluation		Average Length
	Coherence	Informativeness	Engagingness	Humanness	Distinct-1	Distinct-2	
PLATO	0.568	0.564	0.340	0.280	0.042	0.255	34.961
PLATO-2 93M	0.688	0.672	0.640	0.560	0.047	0.276	18.292
DialoGPT	0.720	0.712	0.340	0.100	0.150	0.508	9.335
PLATO-2 314M	1.572	1.620	1.300	1.160	0.065	0.435	22.732
Blender	1.856	1.816	1.820	1.540	0.117	0.385	16.873
PLATO-2 1.6B	<b>1.920</b>	<b>1.892</b>	<b>1.840</b>	<b>1.740</b>	<b>0.169</b>	<b>0.613</b>	15.736

Table 1: Self-chat evaluation results, with best value written in bold.

Model	Human Evaluation				Automatic Evaluation		Average Length
	Coherence	Informativeness	Engagingness	Humanness	Distinct-1	Distinct-2	
Microsoft XiaoIce	0.869	0.822	0.560	0.260	<b>0.289</b>	<b>0.764</b>	6.979
PLATO-2 336M Chinese	<b>1.737</b>	<b>1.683</b>	<b>1.600</b>	<b>1.480</b>	0.212	0.713	6.641

Table 2: Chinese interactive evaluation results, with best value written in bold.

to greet with each other and only cover some shallow topics in the short conversation. Therefore, to expose the model’s weaknesses and explore the model’s limits, we choose to start the interactive conversation with pre-selected topics. We use the classical 200 questions as the start topic (Vinyals and Le, 2015) and ask the bot to performance self-chats given the context. There are 10 utterances in each dialogue, including the input start utterance. We carry out automatic evaluation on the 200 self-chat logs and randomly select 50 conversations for human evaluation.

The compared models are divided into three groups. The first group includes PLATO 132M model and PLATO-2 93M model. Both of them have similar model scales. The second group includes DialoGPT 345M model and PLATO-2 310M model. Both of them are trained using Reddit comments and have similar model scales. The third group includes Blender 2.7B model and PLATO-2 1.6B model. Both of them are first trained using Reddit comments and further fine-tuned with BST conversations. In human evaluation, two self-chat logs, which are from the same group and have the same start topic, will be displayed to three annotators. One example is given in Figure 3. As suggested in ACUTE-Eval (Li et al., 2019), we ask crowd-sourcing workers to pay attention to only one speaker within a dialogue. In the evaluation, they need to give scores on coherence and informativeness for each P1’s utterance, and assess P1’s overall quality on engagingness and humanness.

The self-chat evaluation results are summarized in Table 1. These results indicate that PLATO-2 1.6B model obtains the best performance across human and automatic evaluations. In the first group, PLATO-2 achieves better performance than PLATO on a similar model scale, which might mainly result from the stable curriculum learning and large-scale conversation data. In the second group, DialoGPT tends to generate repetitive conversations due to the backward scoring function, resulting in poor performance in interactive evaluation. In the third group, PLATO-2 outperforms the state-of-the-art open-domain chatbot Blender. The gap of Blender and PLATO-2 on the corpus-level metric distinct-1/2 suggests that PLATO-2 has a better capacity on lexical diversity. In addition, the difference among these three groups suggests that enlarging model scales and exploiting human annotated conversations help improve the dialogue quality.

### 3.4.2 Human-Bot Chat Evaluation

In the Chinese evaluation, it is difficult to carry out self-chats for Microsoft XiaoIce, as there is no public available API. Therefore, we collect human-bot conversations through their official Weibo platform. The interactive conversation also starts with a pre-selected topic and continues for 7-14 rounds. 50 diverse topics are extracted from the high-frequency topics of a commercial chatbot, including travel, movie, hobby and so on. The collected human-bot conversations are distributed to crowd-sourcing

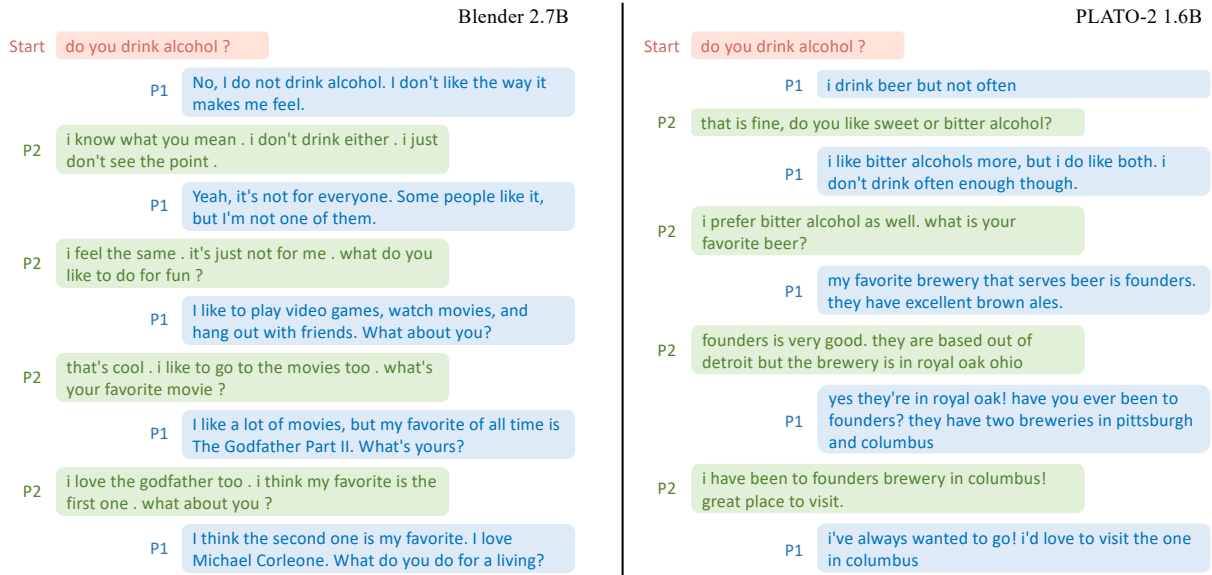


Figure 3: Self-chat examples by Blender and PLATO-2.

workers for evaluation. The human and automatic evaluation results are summarized in Table 2. XiaoIce obtains higher distinct values, which may use a retrieval-based strategy in response generation. The human evaluations indicate that our PLATO-2 model achieves significant improvements over XiaoIce across all the human evaluation metrics.

### 3.4.3 Static Evaluation

Besides the interactive evaluation, we also employ static evaluation to analyze the model’s performance. In static evaluation, each model will produce a response towards the given multi-turn context. Those powerful models are involved in the evaluation: Meena, Blender, DialoGPT and PLATO-2 1.6B. To compare with Meena, we include their provided 60 static samples in the Appendix of the paper and generate corresponding responses with other models. We also include 60 test samples about daily life from Daily Dialog (Li et al., 2017) and 60 test samples about in-depth discussion from Reddit. Given that the measurement of humanness usually needs multi-turn interaction, this metric is excluded from static evaluation. The evaluation results are summarized in Table 3. It can be observed that PLATO-2 is able to produce coherent, informative and engaging responses across different chat scenarios. The average Fleiss’s kappa (Fleiss, 1971) of human evaluation is 0.466, indicating annotators have reached moderate agreement.

Model	Meena Samples		
	Coherence	Informativeness	Engagingness
Meena	1.750	1.617	1.583
DialoGPT	1.233	1.067	1.017
Blender	1.800	1.767	1.683
PLATO-2 1.6B	<b>1.900</b>	<b>1.917</b>	<b>1.850</b>

Model	Daily Dialog Samples		
	Coherence	Informativeness	Engagingness
DialoGPT	1.117	1.033	0.917
Blender	1.767	1.617	1.633
PLATO-2 1.6B	<b>1.867</b>	<b>1.850</b>	<b>1.833</b>

Model	Reddit Samples		
	Coherence	Informativeness	Engagingness
DialoGPT	1.283	1.283	1.183
Blender	1.767	1.550	1.583
PLATO-2 1.6B	<b>1.900</b>	<b>1.900</b>	<b>1.883</b>

Table 3: Static evaluation results, with the best scores written in bold.

## 3.5 Discussions

### 3.5.1 Case Analysis

To further analyze the models’ features, two self-chat examples of Blender and PLATO-2 are provided in Figure 3. Although both models are able to produce high-quality engaging conversations, they exhibit distinct discourse styles. Blender tends to switch topics quickly in the short conversation, including alcohol, hobbies, movies and work. The emergence of this style might be related with BST

	Blender Win	Tie	PLATO-2 Win
In-depth Discussion	8	18	<b>24</b>

Table 4: In-depth discussion w.r.t. the start topic.

	PLATO-2 Stage-1 Win	Tie	PLATO-2 Stage-2 Win
Engagingness	3	31	<b>16</b>
Humanness	6	29	<b>15</b>

Table 5: Comparison of the models in PLATO-2.

fine-tuning data. For instance, persona chat in BST is about the exchange of personal information between two partners, where topics need to switch quickly to know more about each other. Due to the task settings of data collection, some human annotated conversations might be a little unnatural. Nevertheless, fine-tuning with BST conversations is essential to mitigate undesirable toxic traits of large corpora and emphasize desirable skills of human conversations.

Distinct with Blender, PLATO-2 can stick to the start topic and conduct in-depth discussions. The reasons might be two-fold. First, our model is able to generate diverse and informative responses with the accurate modeling of one-to-many relationship. Second, the evaluation model helps select the coherent response and stick to current topic. We asked crowd-sourcing workers to annotate which model’s in-depth discussion is better w.r.t. the start topic. The comparison result is shown in Table 4, which also verifies our above analysis on discourse styles.

### 3.5.2 Why PLATO-2 Performs Better?

Why PLATO-2 achieves better performance as compared with Meena, Blender and other state-of-the-art models? As analyzed above, major reasons might come from two aspects: fine-grained generation and evaluation. First, PLATO-2 employs discrete latent variable for the one-to-many relationship modeling, which is able to generate high-quality and diverse responses. Second, the evaluation model in PLATO-2 is effective at selecting the most appropriate response from the candidates.

In fact, these two aspects are associated with the curriculum learning in the second stage, modeling the one-to-many relationship for open-domain conversations. By contrast, Meena and Blender are learned under the one-to-one mapping relationship, similar to the first stage in PLATO-2. To dissect the effects of these two stage models, we further

ask crowd-sourcing workers to evaluate the models’ self-chat logs on the dialogue-level metrics. The comparison results are summarized in Table 5. These results verify the effectiveness of curriculum learning in PLATO-2.

### 3.5.3 Further Exploration of PLATO-2

In addition to open-domain chitchat, there are two other kinds of dialogues in conversational AI (Gao et al., 2018): knowledge grounded dialogue, and task-oriented conversation. Similar to open-domain conversation, the one-to-many mapping relationship also exists in knowledge grounded dialogue (Kim et al., 2020): given a dialogue context, multiple pieces of knowledge might be applicable for the response generation. Therefore, the one-to-many mapping models of the second stage can also be adapted for knowledge grounded dialogue. By expanding the network input with the knowledge segment, the background knowledge is encoded and grounded for response generation. Distinct from the open-domain conversation and knowledge grounded dialogue, task-oriented conversations usually need to accomplish a specific goal. Accordingly, the conversation flow would become less diverse and concentrated on task completion. Therefore, the one-to-one mapping generation model of the first stage can be used for the end-to-end task-oriented conversation.

For the exploration of PLATO-2 two-stage framework, we participated in several tasks of DSTC9 (Gunasekara et al., 2020), including interactive evaluation of open-domain conversation (Track3-task2), static evaluation of knowledge grounded dialogue (Track3-task1), and end-to-end task-oriented conversation (Track2-task1). PLATO-2 has achieved the first place in all three tasks (Bao et al., 2021). To sum up, the benefits brought by the two-stage curriculum learning in PLATO-2 are two-fold. Firstly, given the difficulties to scale up PLATO, the two-stage curriculum learning is an essential ingredient for the successful training of 1.6B parameter PLATO-2. Secondly, the two-stage PLATO-2 adapts well to multiple conversational tasks, indicating its potentials as a unified pre-training framework for conversational AI.

## 4 Related Work

Related works include large-scale language models and open-domain dialogue generation.

**Large-scale Language Models.** Pre-trained large-scale language models have brought many break-



throughs on various NLP tasks. GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) are representative uni-directional and bi-directional language models, trained on general text corpora. By introducing pre-normalization and modifying weight initialization, GPT-2 (Radford et al., 2019) successfully extends the model scale from 117M to 1.5B parameters. To cope with memory constraints, Megatron-LM (Shoeybi et al., 2019) exploits model parallelism to train an 8.3B parameter model on 512 GPUs. GPT-3 (Brown et al., 2020) further trains an 175B parameter autoregressive language model, demonstrating strong performance on many NLP tasks. The development of large-scale language models is also beneficial to the task of dialogue generation.

**Open-domain Dialogue Generation.** On the basis of GPT-2, DialoGPT (Zhang et al., 2020) is trained for response generation using Reddit comments. To obtain a human-like open-domain chatbot, Meena (Adiwardana et al., 2020) scales up the network parameters to 2.6B and utilizes more social media conversations in the training process. To emphasize desirable conversational skills of engagingness, knowledge, empathy and personality, Blender (Roller et al., 2021) further fine-tunes the pre-trained model with human annotated conversations. In addition to the attempts on model scale and data selection, PLATO introduces discrete latent variable to tackle the inherent one-to-many mapping problem to improve response quality. In this work, we explore the effective training of PLATO-2 via curriculum learning.

## 5 Conclusion

In this work, we discuss the effective training of open-domain chatbot PLATO-2 via curriculum learning, where two stages are involved. In the first stage, one coarse-grained model is trained for general response generation. In the second stage, two models of fine-grained generation and evaluation are trained for diverse response generation and response coherence estimation. Experimental results demonstrate that PLATO-2 achieves substantial improvements over the state-of-the-art methods in both Chinese and English evaluations.

## Acknowledgments

We would like to thank Jingzhou He, and Tingting Li for the help on resource coordination; Daxiang Dong, and Pingshuo Ma for the support on Pad-

dlePaddle; Yu Sun, Yukun Li, and Han Zhang for the assistance with infrastructure and implementation. This work was supported by the Natural Key Research and Development Project of China (No. 2018AAA0101900).

## 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](#). *arXiv preprint arXiv:2001.09977*.
- Siqi Bao, Bingjin Chen, Huang He, Xin Tian, Han Zhou, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, and Yingzhan Lin. 2021. [A unified pre-training framework for conversational ai](#). *arXiv preprint arXiv:2105.02482*.
- Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. [Know more about each other: Evolving dialogue strategy via compound assessment](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5382–5391.
- Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. [Plato: Pre-trained dialogue generation model with discrete latent variable](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, pages 1877–1901.
- Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. [Generating multiple diverse responses with multi-mapping and posterior mapping selection](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4918–4924.

- Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. [Training deep nets with sublinear memory cost](#). *arXiv preprint arXiv:1604.06174*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- 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*, pages 13063–13075.
- Joseph L Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). In *Psychological Bulletin*, pages 378–382.
- Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019. [Grounded response generation task at dstc7](#). In *AAAI Dialog System Technology Challenge Workshop*.
- Jianfeng Gao, Michel Galley, and Lihong Li. 2018. [Neural approaches to conversational AI](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7.
- Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. [Overview of the ninth dialog system technology challenge: Dstc9](#). *arXiv preprint arXiv:2011.06486*.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *International Conference on Learning Representations*.
- Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. [Sequential latent knowledge selection for knowledge-grounded dialogue](#). In *International Conference on Learning Representations*.
- Diederik P Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *International Conference on Learning Representations*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. [Deep reinforcement learning for dialogue generation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. [ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons](#). *arXiv preprint arXiv:1909.03087*.
- 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 8th International Joint Conference on Natural Language Processing*, pages 986–995.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). *Technical report, OpenAI*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *Technical report, OpenAI*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M 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*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. [Megatron-LM: Training multi-billion parameter language models using model parallelism](#). *arXiv preprint arXiv:1909.08053*.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. [Can you put it all together: Evaluating conversational agents’](#)

ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030.

Trevor Standley, Amir R Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. [Which tasks should be learned together in multi-task learning?](#) In *Proceedings of the 37th International Conference on Machine Learning*, pages 9120–9132.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need.](#) In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. [A neural conversational model.](#) *arXiv preprint arXiv:1506.05869*.

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*, pages 270–278.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. [Learning discourse-level diversity for neural dialog models using conditional variational autoencoders.](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 654–664.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. [Commonsense knowledge aware conversation generation with graph attention.](#) In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629.

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 Data Cleaning Process

PLATO-2 has English and Chinese models, with training data extracted from open-domain social media conversations. As the comments are formatted in message trees, any conversation path from the root to a tree node can be treated as one training sample, with the node as response and its former turns as context. To improve the generation quality, we carry out elaborate data cleaning. A message node and its sub-trees will be removed if any of the following conditions is met.

- 1) The number of BPE tokens is more than 128 or less than 2.

- 2) Any word has more than 30 characters or the message has more than 1024 characters.
- 3) The percentage of alphabetic characters is less than 70%.
- 4) The message contains URL.
- 5) The message contains special strings, such as `r/`, `u/`, `&amp;`.
- 6) The message has a high overlap with the parent’s text.
- 7) The message is repeated more than 100 times.
- 8) The message contains offensive words.
- 9) The subreddit is quarantined.
- 10) The author is a known bot.

After data cleaning, the English training data contains 684M (context, response) samples and the Chinese training data contains 1.2B (context, response) samples. Each English/Chinese sample has 2.78/2.82 utterances and each utterance has 26.29/22.20 tokens on average.

## B Training Configurations

PLATO-2 has three model sizes: a standard version of 1.6B parameters, a small version of 314M parameters, and a tiny version of 93M parameters. The 1.6B parameter model has 32 transformer blocks and 32 attention heads, with the embedding dimension of 2048. The 314M parameter model has 24 transformer blocks and 16 attention heads, with the embedding dimension of 1024. The 93M parameter model has 12 transformer blocks and 12 attention heads, with the embedding dimension of 768. The training configurations of PLATO-2 1.6B are provided in Table 6. The training was carried out on 64 Nvidia Tesla V100 32G GPU cards. It takes about 3 weeks for the 1.6B parameter model to accomplish the curriculum learning process.

## C Chinese Case Analysis

We also provide two human-bot chat examples of XiaoIce and PLATO-2 in Figure 4, with original interactive logs shown on the left and translated logs on the right. It can be observed that some responses produced by XiaoIce are not coherent with the contexts and there are some abrupt changes of topics. By contrast, the interaction with PLATO-2 is more coherent and engaging.

## D Scoring Criteria in Human Evaluation

The criteria used in human evaluation are provided in Table 7.

Model	Stage	Batch Size	Learning Rate	Warmup Steps	Training Steps
PLATO-2 1.6B Parameter	1. Coarse-grained Generation	524,288	5e-4	3125	20w
	2.1. Fine-grained Generation	65,536	1e-5	3125	10w
	2.2. Evaluation	524,288	5e-5	3125	10w

Table 6: Training configurations of PLATO-2.



Figure 4: Human-bot chat examples by Microsoft Xiaolce and PLATO-2.

Score	Coherence
0	<ul style="list-style-type: none"> <li>• The response is not related with the context.</li> <li>• The response simply repeats the context.</li> <li>• The response has obvious conflicts with the context.</li> <li>• There are serious logic conflicts within the response.</li> </ul>
1	<ul style="list-style-type: none"> <li>• The response has minor conflicts with the context.</li> <li>• There are some minor logic conflicts in the response.</li> </ul>
2	<ul style="list-style-type: none"> <li>• The response is coherent with the context.</li> </ul>

Score	Informativeness
0	<ul style="list-style-type: none"> <li>• The response doesn't contain any information.</li> <li>• This response just repeats the context and fails to bring any additional information.</li> <li>• The information is invalid, as the coherence score is 0.</li> </ul>
1	<ul style="list-style-type: none"> <li>• The information has conflicts with common sense.</li> <li>• There are factual errors in the response.</li> </ul>
2	<ul style="list-style-type: none"> <li>• The response has appropriate and correct information.</li> </ul>

Score	Engagingness
0	<ul style="list-style-type: none"> <li>• I don't want to talk with this speaker.</li> </ul>
1	<ul style="list-style-type: none"> <li>• It is kind of boring, but it is still ok to talk with this speaker.</li> </ul>
2	<ul style="list-style-type: none"> <li>• I would like to talk with this speaker for a long conversation.</li> </ul>

Score	Humanness
0	<ul style="list-style-type: none"> <li>• This speaker seems like a bot.</li> </ul>
1	<ul style="list-style-type: none"> <li>• This speaker gives unnatural responses occasionally and seems not that human-like.</li> </ul>
2	<ul style="list-style-type: none"> <li>• This speaker seems like a human being.</li> </ul>

Table 7: Score details of four metrics in human evaluation.

## E Response Selection Comparison

We carry out more experiments to compare the performance of distinct scoring functions in response selection. Firstly, one Chinese response selection dataset is constructed: 100 dialogue contexts are selected from the test set and 10 candidate responses are retrieved for each context with a commercial chatbot. Secondly, we ask crowd-sourcing workers to annotate the label whether the candidate response is coherent with the context. Thirdly, we train three 336M parameter models as the scoring function, including the forward response generation probability  $p(r|c)$ , the backward context

Score Function	MAP	MRR	P@1
$p(r c)$	0.705	0.790	0.700
$p(c r)$	0.672	0.737	0.610
$p(l_r c, r)$	<b>0.754</b>	<b>0.819</b>	<b>0.750</b>

Table 8: Comparison of different score functions in response selection, with best value written in bold.

recover probability  $p(c|r)$  and the bi-directional coherence probability  $p(l_r|c, r)$ . Their results on the annotated response selection dataset are summarized in Table 8. The metrics of mean average precision (MAP), mean reciprocal rank (MRR) and precision at position 1 (P@1) are employed. These results indicate that PLATO-2's evaluation model is better at selecting appropriate responses.