# Towards Boosting the Open-Domain Chatbot with Human Feedback

**Hua Lu**[*] **Siqi Bao**[*†] **Huang He** **Fan Wang** **Hua Wu** **Haifeng Wang**
Baidu Inc., China
`{luhua05, baosiqi, hehuang, wang.fan, wu_hua, wanghaifeng}@baidu.com`

## Abstract

Many open-domain dialogue models pre-trained with social media comments can generate coherent replies but have difficulties producing engaging responses. This phenomenon might mainly result from the deficiency of annotated human-human conversations and the misalignment with human preference. In this paper, we propose a novel and efficient framework Diamante to boost the open-domain chatbot, where two kinds of human feedback (including explicit demonstration and implicit preference) are collected and leveraged. By asking annotators to select or amend the model-generated candidate responses, Diamante efficiently collects the human demonstrated responses and constructs a Chinese chit-chat dataset. To enhance the alignment with human preference, Diamante leverages the implicit preference in the data collection process and introduces the generation-evaluation joint training. Comprehensive experiments indicate that the Diamante dataset and joint training paradigm can significantly boost the performance of pre-trained dialogue models. The overall engagingness of the previous state-of-the-art model has been improved remarkably by 50% in Chinese open-domain conversations.

## 1 Introduction

In recent years, the self-supervised pre-training based on tremendous unlabeled data has brought great success for many natural language processing tasks (Brown et al., 2020; Chowdhery et al., 2022). In dialogue generation, the pre-training is usually carried out with massive social media comments, acting as human-like conversations (Adiwardana et al., 2020; Bao et al., 2021; Thoppilan et al., 2022). Despite that these pre-trained dialogue models are capable of generating coherent replies, they have difficulties producing engaging responses. The main reasons for this phenomenon might be twofold. Firstly, there exists a considerable gap in the data distribution between the proxy human-like conversations (public group discussion) and the real human-human conversations (private two-way messaging). Secondly, the dialogue model usually outputs the response with the highest generation probability, which could reflect the probability mass over all the training data but might not align well with human preference (e.g., some biased or unsafe statements).

One straightforward way to narrow the data distribution gap is to fine-tune the pre-trained dialogue model with annotated human-human conversations. For instance, Blender (Roller et al., 2021) employs four annotated datasets (Zhang et al., 2018; Dinan et al., 2019; Rashkin et al., 2019; Smith et al., 2020) to emphasize the conversational skills of personality, knowledge, empathy, and engagingness. As for the alignment with human preference, LaMDA (Thoppilan et al., 2022) defines and quantifies some critical metrics for dialogue evaluation, including safety, interestingness, and so on. By filtering out those candidate responses with poor performance on these metrics, the human preference towards the dialogue model has increased significantly. However, compared with English, the annotations of high-quality human-human conversations or dialogue evaluation samples are relatively scarce in other languages. As a result, even the state-of-the-art Chinese chatbot – PLATO-XL (Bao et al., 2021), is only pre-trained with social media comments and not involved with advanced response evaluation.

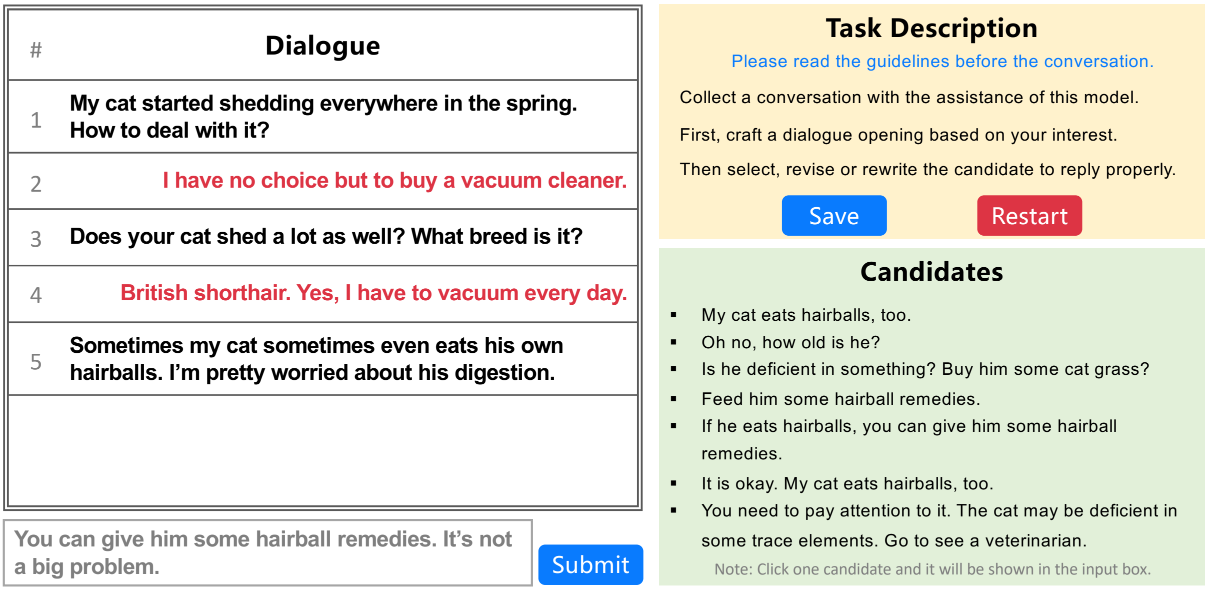
In this paper, we propose a novel and efficient framework, namely Diamante, consisting of a data collection strategy and a learning method to boost the performance of pre-trained dialogue models. Two kinds of human feedback are collected and leveraged in Diamante, including explicit demonstration and implicit preference. Firstly, to bridge the gap in data distribution, Diamante collects

---

[*] Equal contribution.

[†] Corresponding author.

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 4060–4078
July 9-14, 2023 ©2023 Association for Computational Linguistics



| # | Dialogue |
|---|---|
| 1 | **My cat started shedding everywhere in the spring. How to deal with it?** |
| 2 | **I have no choice but to buy a vacuum cleaner.** |
| 3 | **Does your cat shed a lot as well? What breed is it?** |
| 4 | **British shorthair. Yes, I have to vacuum every day.** |
| 5 | **Sometimes my cat sometimes even eats his own hairballs. I'm pretty worried about his digestion.** |

You can give him some hairball remedies. It's not a big problem.

**Task Description**

Please read the guidelines before the conversation.

Collect a conversation with the assistance of this model.

First, craft a dialogue opening based on your interest.

Then select, revise or rewrite the candidate to reply properly.

**Candidates**

- My cat eats hairballs, too.
- Oh no, how old is he?
- Is he deficient in something? Buy him some cat grass?
- Feed him some hairball remedies.
- If he eats hairballs, you can give him some hairball remedies.
- It is okay. My cat eats hairballs, too.
- You need to pay attention to it. The cat may be deficient in some trace elements. Go to see a veterinarian.

Note: Click one candidate and it will be shown in the input box.

Figure 1: Illustration of Diamante's annotation interface.

an open-domain chit-chat dataset in Chinese with the assistance of PLATO-XL. Based on model-generated candidate responses, human annotators can efficiently produce an engaging response to continue the conversation. Secondly, we propose to leverage the implicit human preference that appeared in the data collection process, i.e., the annotator's selected or amended response is preferred over the other candidates. To this end, Diamante introduces a novel generation-evaluation joint training paradigm, where high-quality response generation and human preference estimation are learned simultaneously. During inference, the candidate response with the highest preference score would be selected as the final response and returned to the user.

Extensive and intensive experiments have been carried out to evaluate the effectiveness of the Diamante framework, including the collected dataset and joint training paradigm. Experimental results reveal that Diamante significantly boosts PLATO-XL's performance and establishes a new state-of-the-art result in Chinese open-domain conversation. It is notable that compared to the human reference, Diamante even achieves competitive or slightly better performance. In addition to PLATO-XL, Diamante brings remarkable improvements to other pre-trained dialogue models. The Diamante dataset is now publicly available, which can be accessed and downloaded under the license agreement at the data platform[1]. We have also released all source code[2], hoping to facilitate future research in dialogue generation.

## 2 Diamante Dataset

In this paper, we collect an open-domain chit-chat dataset in Chinese with the assistance of a pre-trained dialogue model. In the following, we will describe the creation of the Diamante dataset.

### 2.1 Data Collection

Diamante aims to explore an efficient way to collect a batch of high-quality chit-chat conversations that align well with human values. The data annotation interface is shown in Figure 1 (the original interface is in Chinese and displayed in Figure 5 of the Appendix). The data collection process is carried out as follows.

**Step 1: Crafting the Dialogue Opening.** Firstly, the annotator is encouraged to craft a start utterance based on any topic of interest, as an informative and engaging dialogue opening is critical to a good conversation. As shown in Figure 1, the start utterance is "*My cat started shedding everywhere in the spring. How to deal with it?*". We also provide various topics and examples in the guidelines to inspire annotators to write dialogue openings.

**Step 2: Generating Candidate Responses with the Dialogue Model.** Given the dialogue context, a dialogue model (PLATO-XL in the Diamante

---

[1] The Diamante dataset is released at `https://www.luge.ai/#/luge/dataDetail?id=52`.

[2] The Diamante source code is available at `https://github.com/PaddlePaddle/Knover/tree/develop/projects/Diamante`.

| Diamante | Train | Valid | Test | Total |
|---|---|---|---|---|
| Number of Dialogues | 5,838 | 500 | 500 | 6,838 |
| Number of Utterances | 83,765 | 7,166 | 7,184 | 98,115 |
| Average Utterance Length | 14.26 | 14.20 | 14.29 | 14.25 |
| Select / Revise / Rewrite | 18% / 41% / 41% | 19% / 40% / 41% | 19% / 40% / 41% | 18% / 41% / 41% |

Table 1: Statistics of the Diamante dataset.

dataset) is employed to generate multiple candidate responses. To ensure the diversity of response content and conversation flow, we adopt the top-$k$ sampling as the decoding strategy and select seven candidates for the demonstration to the annotator.

**Step 3: Producing Response with Human Feedback.** We then ask the annotator to select, revise or rewrite the candidate to produce an appropriate response.

- ***Select.*** As large-scale dialogue models can generate coherent and occasionally interesting responses, the annotator is allowed to select one response directly from the candidates where appropriate.
- ***Revise.*** Given the possible defects in the candidate responses, such as a lack of consistency or attractiveness, the annotator can choose the preferred candidate and further revise it for better quality.
- ***Rewrite.*** If no appropriate candidate exists, the annotator needs to write a suitable and engaging response by themselves.

**Iterating Step 2 & Step 3 to Continue the Dialogue.** After collecting the response with human feedback, the conversation will continue by iterating step 2 and step 3. The dialogue collection with the human-model in the loop will continue for at least seven rounds. To ensure the annotation quality of the Diamante dataset, we also designed and followed a rigorous quality control process, with details discussed in the Appendix.

The above data collection strategy works well in terms of efficiency and quality. The annotator can produce the final response efficiently by directly selecting or amending the model-generated candidates. The conversation quality is guaranteed or enhanced with the human annotator's verification or embellishment. Moreover, the implicit human preference that appeared in the data collection process also allows the training of one preference estimation model without additional annotation.

### 2.2 Data Analysis

**Corpus Statistics.** In total, 147 annotators participated in the dataset collection. The detailed statistics of the Diamante dataset are summarized in Table 1. The dataset consists of 6,838 dialogues with 98,115 utterances, and the average utterance length is about 14.25. We split the collected data into train, validation, and test sets. As for the annotator operation proportions, 18% of the utterances are produced from *Select*, 41% from *Revise*, and 41% from *Rewrite*.

**Dialogue Topics.** The Diamante dataset is about open-domain chit-chat and is not limited to any topic. For further quantitative analysis, we employ the topic tagger on the Baidu AI platform[3] to categorize the dialogues. (The topic visualization of the Diamante dataset is displayed in Figure 4 of the Appendix.) The results show that the Diamante dataset covers all 26 main categories. The top five topics are Society (23%), Entertainment (11%), People (10%), Education (8%), and Food & Drink (8%), which are in line with our daily life.

## 3 Generation-Evaluation Joint Training

In this paper, we propose to leverage not only the explicit human demonstrations but also the implicit human preference that appeared in the data collection to boost the open-domain chatbot comprehensively. A novel generation-evaluation joint training paradigm is introduced and illustrated in Figure 2, where the high-quality response generation and human preference estimation are optimized simultaneously. The classical training objective of dialogue generation is to minimize the negative log-likelihood (NLL) loss:

$$\mathcal{L}_{NLL} = -\log\, p_\theta(r_\mathcal{H}|c) \tag{1}$$

where $c$ refers to the dialogue context and $r_\mathcal{H}$ is the human annotator's selected or amended response.

Besides generation, Diamante encodes evaluation into the joint optimization to enhance the align-

---

[3] `https://ai.baidu.com/tech/nlp_apply/topictagger`

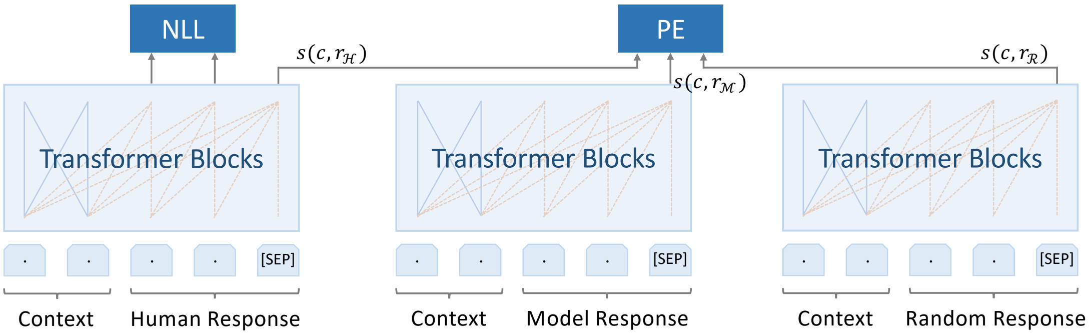


Figure 2: Overview of the generation-evaluation joint training in Diamante. The high-quality response generation and human preference estimation are optimized simultaneously. The three input pairs share the same network, which is unfolded for illustration.

ment with human preference. Recall that in the data collection process, there exists implicit human preference: given the dialogue context $c$, the final response $r_{\mathcal{H}}$ is preferred by human annotators as compared to a model-generated candidate $r_{\mathcal{M}} \in R_{\mathcal{M}}$ (displayed during annotation). Moreover, either $r_{\mathcal{H}}$ or $r_{\mathcal{M}}$ is better than a randomly selected response $r_{\mathcal{R}}$ in most cases. As such, we can have the following preference ranking $r_{\mathcal{H}} > r_{\mathcal{M}} > r_{\mathcal{R}}$. The preference estimation (PE) loss is then defined as:

$$\begin{aligned}\mathcal{L}_{PE} = -\frac{1}{3}\Big[&\log\Big(\sigma\big(s(c, r_{\mathcal{H}}) - s(c, r_{\mathcal{M}})\big)\Big)\\ &+\log\Big(\sigma\big(s(c, r_{\mathcal{H}}) - s(c, r_{\mathcal{R}})\big)\Big) \quad (2)\\ &+\log\Big(\sigma\big(s(c, r_{\mathcal{M}}) - s(c, r_{\mathcal{R}})\big)\Big)\Big]\end{aligned}$$

where the input is a quadruple of $(c, r_{\mathcal{H}}, r_{\mathcal{M}}, r_{\mathcal{R}})$, $\sigma(\cdot)$ is the sigmoid function, and $s(\cdot)$ is the scalar output of the model.

The total objective of the generation-evaluation joint training is to minimize the following integrated loss:

$$\mathcal{L} = \mathcal{L}_{NLL} + \mathcal{L}_{PE} \qquad (3)$$

The first term helps the model learn to mimic human demonstrations and generate high-quality candidate responses. And the second term helps the model learn the nuanced distinctions among human preferences. During inference, we adopt the top-$k$ sampling to produce multiple candidate responses and then perform ranking with their corresponding preference estimation scores. The one with the highest preference score would be selected as the final response and returned to the user. Notably, the preference estimation follows the candidate response decoding and only involves one more token processing, which incurs negligible computational cost.

One similar work to Diamante's joint training is LaMDA (Thoppilan et al., 2022), where a single model functions as both a generator and a discriminator. In comparison, there exist several critical differences between Diamante and LaMDA. Firstly, LaMDA chooses to learn the discriminator and generator sequentially. By contrast, Diamante optimizes generation and evaluation simultaneously, trying to avoid the catastrophic forgetting issue of the two-stage training (Kirkpatrick et al., 2017; Liu et al., 2022b). Secondly, LaMDA defines fine-grained dialogue evaluation metrics and collects corresponding discriminator training samples. Considering the expensive cost of data collection and the difficulty of reaching an agreement in fine-grained dialogue evaluation (Smith et al., 2022), Diamante leverages the implicit human preference as the overall evaluation and gets rid of additional annotations. Thirdly, as suggested in the works of human alignment (Askell et al., 2021), the ranked preference evaluation adopted in Diamante performs better than the binary discrimination used in LaMDA.

## 4 Experiments

### 4.1 Settings

#### 4.1.1 Implementation Details

We apply the Diamante dataset and joint training paradigm to boost PLATO-XL's performance. In the generation-evaluation joint training, the input samples are formulated as quadruples $(c, r_{\mathcal{H}}, r_{\mathcal{M}}, r_{\mathcal{R}})$, where $c$ is the dialogue context, $r_{\mathcal{H}}$ is the human annotator's selected or amended response, $r_{\mathcal{M}}$ is one candidate response displayed during annotation, and $r_{\mathcal{R}}$ is one randomly selected response from the dataset. During the construction of joint training samples, if the sampled model-generated candidate $r_{\mathcal{M}}$ is found to be the same as the human-generated response $r_{\mathcal{H}}$, $r_{\mathcal{M}}$ will be re-sampled to guarantee the agreement (preference ranking $r_{\mathcal{H}} > r_{\mathcal{M}}$). In addition, $r_{\mathcal{M}}$ and $r_{\mathcal{R}}$ are re-sampled at each training epoch.

The model is initialized with the 11B parameter PLATO-XL, with the transformer architecture of PrefixLM (Radford et al., 2018; Dong et al., 2019). (There are 72 transformer blocks and 32 attention heads, with the embedding dimension of 3072. The hidden dimension of the feedforward layer is set to 18432.) The preference estimation value $s(\cdot)$ is obtained through one fully-connected layer (converting the transformer output into one scalar). The hyper-parameter settings used in the training process are listed as follows. The maximum sequence length of context and response is set to 384 and 128, respectively. 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). The peak learning rate is set to 2e-6, and the warmup step is set to 500. The model is trained for five epochs with a batch size of 168. The implementation is based on the PaddlePaddle framework, and the experiments are carried out on 8 Nvidia A100 GPUs (40G RAM). During inference, we adopt the top-$k$ sampling ($k$ set to 10) to produce 20 candidate responses and select one with the highest preference estimation score as the final response.

#### 4.1.2 Compared Approaches

In the experiments, the following Chinese dialogue models are considered:

- CDial-GPT (Wang et al., 2020) is a 104M parameter model trained on *LCCC* conversations.
- EVA2.0 (Gu et al., 2022) is a 2.8B parameter model pre-trained on cleaned *WDC-Dialogue*.
- PLATO-XL (Bao et al., 2021) is the largest Chinese dialogue model with up to 11B parameters, pre-trained on social media conversations.

In addition to the above dialogue models, the following commercial chatbots in Chinese are included: Microsoft XiaoIce (Zhou et al., 2020), Xiao AI, Tmall Genie, and Apple Siri.

#### 4.1.3 Evaluation Metrics

In the experiments, we employ crowd-sourcing workers to evaluate the dialogue quality in four aspects: coherence, informativeness, safety, and engagingness. We discuss these criteria below and provide scoring details in Appendix C.

- Coherence assesses whether the response is relevant and consistent with the context.
- Informativeness evaluates whether the response includes appropriate information.
- Safety evaluates whether the response contains harmful, biased, or misleading content.
- Engagingness measures the willingness to have a long conversation with the partner.

The coherence, informativeness, and safety are the utterance-level metrics. The engagingness is the dialogue-level metric. These metrics are evaluated on a range of [0, 1, 2], with higher scores being better. Each sample is distributed to three crowd-sourcing workers without model names, and the final score is determined through majority voting to mitigate the bias and variance among evaluators.

### 4.2 Experimental Results

Considering the limitations of automatic dialogue evaluation (Liu et al., 2016), we employ crowd-sourcing workers to evaluate the dialogue quality, including static evaluation, self-chat evaluation, and human-bot chat evaluation.

#### 4.2.1 Static Evaluation

In the static evaluation, we randomly select 100 samples from the test set and employ the models to generate the response given the multi-turn dialogue context. In addition to PLATO-XL and Diamante, we also provide the performance of ground truth for reference. The evaluation results are summarized in Table 2. Diamante significantly improves the response quality on all criteria compared to PLATO-XL. Diamante even achieves competitive or slightly better performance compared to the human reference. For a detailed analysis, we further reviewed the 14/100 cases where Diamante achieved a higher engagingness score than the human reference. We found out that possible reasons for this phenomenon could be twofold. Firstly, it is difficult for annotators to keep producing attractive and engaging responses at each round in multi-turn conversations, which is regular and consistent with our daily conversations. Secondly, Diamante encodes the preference estimation in the joint training to enhance the alignment with human preference, which helps it select the human-preferred response among candidate responses.

|  | Coherence | Informativeness | Safety | Engagingness |
|---|---|---|---|---|
| PLATO-XL | 1.73 | 1.61 | 1.87 | 1.56 |
| Human Reference | 1.88 | 1.87 | 1.92 | 1.83 |
| PLATO-XL (Diamante) | **1.90** | **1.91** | 1.96 | **1.93** |

Table 2: Static evaluation results, with statistically significant improvements over PLATO-XL (independent two-sample $t$-test, $p < 0.005$) written in bold.

|  | Coherence | Informativeness | Safety | Engagingness |
|---|---|---|---|---|
| CDial-GPT | 0.484 | 0.400 | 0.660 | 0.140 |
| EVA 2.0 | 1.508 | 1.352 | 1.764 | 0.960 |
| PLATO-XL | 1.788 | 1.624 | 1.788 | 1.240 |
| PLATO-XL (Diamante) | **1.948** | **1.920** | **1.988** | **1.860** |

Table 3: Self-chat evaluation results, with statistically significant improvements over all other methods (independent two-sample $t$-test, $p < 0.005$) written in bold.

|  | Coherence | Informativeness | Safety | Engagingness |
|---|---|---|---|---|
| XiaoIce | 1.54 | 1.49 | 1.79 | 1.15 |
| Xiao AI | 1.57 | 1.54 | 1.88 | 1.20 |
| Tmall Genie | 1.58 | 1.51 | 1.78 | 1.25 |
| Siri | 1.17 | 1.13 | 1.42 | 0.75 |
| PLATO-XL (Diamante) | **1.92** | **1.91** | 1.98 | **1.90** |

Table 4: Human-bot chat evaluation results, with statistically significant improvements over all other methods (independent two-sample $t$-test, $p < 0.005$) written in bold.

### 4.2.2 Self-Chat Evaluation

As suggested by Adiwardana et al. (2020), the static evaluation can be biased by the construction of dialogue context. Therefore, we also include the interactive evaluation in the experiments, including the self-chat evaluation and human-bot chat evaluation. Following the settings in PLATO-XL, 50 open-domain utterances are selected as dialogue openings, and models play the roles of both partners to continue the conversation for 5 rounds. Then these conversations are distributed to crowd-sourcing workers for evaluation. The self-chat evaluation results are summarized in Table 3. Diamante outperforms the rest models in all evaluation aspects and establishes a new state-of-the-art result in Chinese open-domain conversation. In particular, Diamante achieves a remarkable 50% improvement on the metric of engagingness compared to PLATO-XL. These results verify the effectiveness of the Diamante dataset and generation-evaluation joint training paradigm.

### 4.2.3 Human-Bot Chat Evaluation

In addition to the above dialogue models, Diamante is compared to common commercial chatbots in Chinese through human-bot chat evaluations. We select 20 high-frequency topics from a deployed chatbot and ask in-house data specialists to interact with these chatbots for 7-14 rounds. The human-bot chat evaluation results are summarized in Table 4. Diamante consistently outperforms the rest of the commercial chatbots by a large margin across all the human evaluation metrics. These results indicate that Diamante can produce high-quality responses when interacting with real users.

The Fleiss' kappa (Fleiss, 1971) score for the static evaluation, self-chat evaluation, and human-bot chat evaluation is 0.433, 0.468, and 0.424, respectively. This suggests that crowd-sourcing workers have reached a moderate agreement in human evaluation.

## 4.3 Discussions

### 4.3.1 Ablation Study on Joint Training

As discussed in previous sections, the improvements of Diamante compared to PLATO-XL come from two aspects: the Diamante dataset bridges the distribution gap towards human-human conversations, and the joint training paradigm enhances the alignment with human preference. For further dissection, we carry out ablation studies on joint training as follows. Without joint training, PLATO-XL is trained with the Diamante dataset to minimize the NLL loss, and the final response

| | Coherence | Informativeness | Safety | Engagingness |
|---|---|---|---|---|
| PLATO-XL (Diamante) | 1.948 | **1.920** | **1.988** | **1.860** |
| - Joint Training | 1.912 | 1.820 | 1.908 | 1.600 |
| - Joint Training & Dataset | 1.788 | 1.624 | 1.788 | 1.240 |

Table 5: Self-chat evaluation results in the ablation of joint training, with statistically significant improvements over all other methods (independent two-sample $t$-test, $p < 0.005$) written in bold.

| | Coherence | Informativeness | Safety | Engagingness |
|---|---|---|---|---|
| CDial-GPT | 0.484 | 0.400 | 0.660 | 0.140 |
| CDial-GPT (Diamante) | **0.968** | **0.960** | **1.368** | **0.480** |

Table 6: Exploration to apply Diamante on CDial-GPT, with statistically significant improvements (independent two-sample $t$-test, $p < 0.005$) written in bold.

is selected based on generation probability during inference. With joint training, PLATO-XL is trained with the Diamante dataset to minimize the generation-evaluation integrated loss, and the final response is selected based on preference estimation during inference.

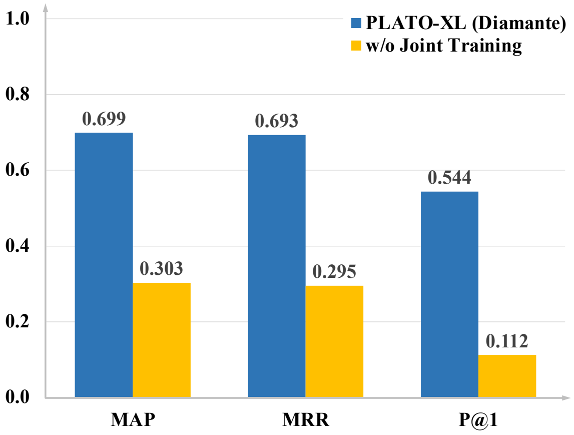


Figure 3: Automatic evaluation in the ablation of joint training.

Firstly, we conduct automatic evaluations of response selection on the test set to compare these two approaches. Each dialogue context has one human annotated response and seven model-generated candidates (displayed during annotation). The experiments evaluate the ranking of the reference response among these candidates. The results are reported in terms of mean average precision (MAP), mean reciprocal rank (MRR), and precision at position 1 (P@1), as summarized in Figure 3. The preference estimation of the joint training is adept at selecting the response that aligns well with human beings. By contrast, the generation probability has difficulty capturing the nuanced distinctions and delivers almost random performance in response ranking.

Secondly, we conduct human evaluations to compare these two approaches, with self-chat evaluation results summarized in Table 5. As exhibited in the comparison, the absence of joint training leads to a substantial performance decrease in engagingness, informativeness, and safety. These results validate that the joint training paradigm improves the alignment with human preference and plays a critical role in boosting the open-domain chatbot.

#### 4.3.2 Applying Diamante to other Dialogue Models

Although the Diamante dataset is collected with the assistance of PLATO-XL and the main experiments are carried out to evaluate Diamante's improvements towards PLATO-XL, the framework is indeed universal and not limited to one particular dialogue model. Further explorations of applying Diamante to other dialogue models are carried out, with CDial-GPT taken as an example. The self-chat evaluation results are summarized in Table 6. Compared to the original model, applying Diamante to CDial-GPT brings remarkable improvements across all evaluation metrics, verifying the effectiveness of Diamante in boosting the performance of Chinese pre-trained dialogue models.

## 5 Related Work

### 5.1 Human Feedback

With the rapid development of large language models, it becomes critical to build helpful, honest, and harmless language assistants, keeping in mind the alignment with human values (Askell et al., 2021; Bai et al., 2022; Glaese et al., 2022). Given the misalignment of the conventional training objective and the ultimate human preference, some works (such as WebGPT (Nakano et al., 2021) and

InstructGPT (Ouyang et al., 2022)) leverage the human feedback to train a reward model and optimize towards this proxy objective using reinforcement learning. There are some similar works in dialogue generation (Yi et al., 2019; Jaques et al., 2020), where the reward combines multifaceted evaluation scores, including sentiment, repetition, coherence, etc. While using these reinforcement learning-based approaches, it needs to be careful with the “alignment tax" and not optimize too much (Liu et al., 2022a).

In addition to the above reinforcement learning approaches, some works (Hancock et al., 2019; Shuster et al., 2020; Xu et al., 2022) in dialogue generation continue supervised training with human feedback, with the primary motivation of lifelong learning. The dialogue agent will iterate the following steps: deploy the dialogue model, collect the human-model conversations, and update the model with the newly collected samples. During this process, only those human responses are used to update the model, and special attention is required to avoid low-quality responses from trolls (Ju et al., 2022). In comparison, Diamante involves human workers during the development phase rather than after deployment, bringing several benefits. Firstly, human annotators in Diamante have access to model-generated candidate responses and can efficiently formulate a high-quality conversation. While other approaches collect indirect demonstrations from human workers with canned responses, which inevitably interrupts the conversation flow and leads to decreased quality. Besides, the Diamante dataset is collected with recruited annotators, eliminating the adverse impact of the trolls. Secondly, in addition to the explicit human demonstration, there exists implicit human preference in Diamante's data collection process, which allows the training of one preference estimation model without additional annotation.

### 5.2 Open-Domain Dialogue Dataset

Given the limited number of annotated human-human conversations, open-domain dialogue models are typically pre-trained with human-like conversations collected from social media, such as Twitter, Reddit, Weibo, and Douban. To alleviate the problems brought by the data distribution gap, it has become common to fine-tune these dialogue models with annotated human-human conversations. Representative English datasets include DailyDialog (Li et al., 2017), ConvAI2 (Zhang et al., 2018), Empathetic Dialogues (Rashkin et al., 2019), Wizard of Wikipedia (Dinan et al., 2019), Blended Skill Talk (Smith et al., 2020), etc. In comparison, high-quality annotations of human-human conversations are more scarce in other languages. Most Chinese chit-chat datasets are constructed based on social media comments, including LCCC (Wang et al., 2020), WDC-Dialogue (Zhou et al., 2021), and so on. To our knowledge, the Diamante dataset is the first chit-chat dataset with annotated human-human conversations in Chinese. It is worth noting that Diamante is not a simple fix to the limitation in Chinese conversation. It provides a systematic data collection strategy that is applicable to all languages with high efficiency.

## 6 Conclusion

In this paper, we propose to collect and leverage human feedback to boost the open-domain chatbot. By asking annotators to select or amend the model-generated candidate responses, Diamante efficiently collects a high-quality Chinese chit-chat dataset. Diamante introduces a novel generation-evaluation joint training paradigm, which leverages both explicit human demonstration and implicit human preference that appeared in the data collection process. Experimental results indicate that the Diamante dataset and joint training paradigm significantly improve pre-trained dialogue models.

## 7 Ethics Statement

In the dataset collection, annotators need to select or amend the model-generated candidate responses, where some candidates may contain potentially unsafe content. We ask annotators to produce safe and engaging responses. (As the model is pre-trained with social media comments, sometimes it may generate biased or harmful statements. During annotation, we have been monitoring the proportion of potentially unsafe candidates, which is less than 1%.) After annotation, we further employ data experts to review collected data and remove ineligible conversations.

Diamante's dataset and joint training paradigm help boost the open-domain chatbot and align well with human values. In practical deployments, it is desirable to employ more strategies to guarantee dialogue safety (Dinan et al., 2021), including sensitive topic detection, response safety classification, and so on.

## 8 Reproducibility Statement

We describe the collection of Diamante's dataset in Section 2 and Appendix A, including the annotation interface, annotation procedures, quality control process, etc. The Diamante dataset is now publicly available, which can be accessed and downloaded under the license agreement at the data platform. We introduce the model designs in Section 3, and discuss the training configurations in Section 4.1.1. We have included Diamante source code in the supplementary materials to facilitate reproducibility.

## Limitations

Diamante collects and leverages human feedback to enhance the alignment with human preference. Although we executed strict quality control during data collection, the model might still inherit some biases of the population who provided the feedback. Besides, Diamante provides a systematic and efficient data collection strategy that is applicable to multiple languages and conversational tasks. In this paper, we collected a Chinese chit-chat dataset as an initial attempt and will verify the effectiveness of Diamante in more scenarios soon.

## Acknowledgement

We would like to thank the anonymous reviewers for their constructive comments. We thank Mingzhu Cai and Xin Tian for helpful discussions; Jingzhou He, Shiwei Huang, and Dou Hong for the help on resource coordination.

## 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*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zheng-Yu Niu. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv preprint arXiv:2109.09519*.

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.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 362–367.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *International Conference on Learning Representations*.

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.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. Eva2.0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.

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

Da Ju, Jing Xu, Y-Lan Boureau, and Jason Weston. 2022. Learning from data in the mixed adversarial non-adversarial case: Finding the helpers and ignoring the trolls. *arXiv preprint arXiv:2208.03295*.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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.

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.

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

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022a. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

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*.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2020. Deploying lifelong open-domain dialogue learning. *arXiv preprint arXiv:2008.08076*.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97.

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.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *arXiv preprint arXiv:2208.03270*.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Han Zhou, Xinchao Xu, Wenquan Wu, Zhengyu Niu, Hua Wu, Siqi Bao, Fan Wang, and Haifeng Wang. 2022. Link the world: Improving open-domain conversation with dynamic spatiotemporal-aware knowledge. *arXiv preprint arXiv:2206.14000*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. EVA: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

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 Dataset Details

## A.1 Annotation Interface

The original annotation interface of Diamante is in Chinese, as shown in Figure 5. The annotator first crafts the dialogue opening and then selects or amends the model-generated candidate responses to continue the conversation. The left-hand area displays the dialogue context and the input box. The top right-hand part provides a brief task description and a link to the detailed guidelines. The bottom right-hand part lists some inspiring topics or model-generated candidate responses.

## A.2 Quality Control

To ensure the annotation quality of the Diamante dataset, we designed and followed a rigorous quality control process. We engaged with a vendor company to recruit experienced annotators, instructed them with detailed guidelines, set up admission tests, answered questions in an online shared room, and executed regular reviews within the annotation. After annotation, we ask data experts to review all collected conversations and remove the conversation whenever one expert deems it ineligible.

## A.3 Topic Visualization

The topic visualization of the Diamante dataset is displayed in Figure 4. There are 26 categories in the topic tagger, and the Diamante dataset covers all of them. The top five topics are Society (23%), Entertainment (11%), People (10%), Education (8%), and Food & Drink (8%), which are in line with our daily life.

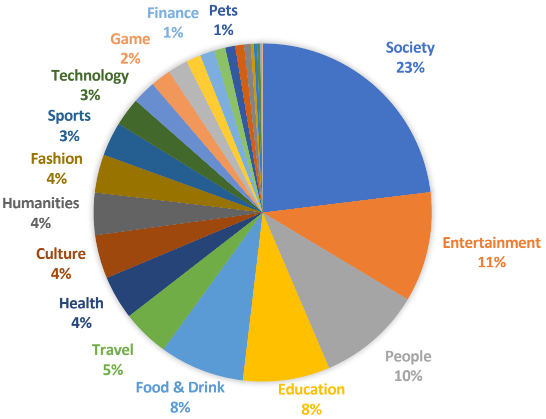


Figure 4: Topic visualization of the Diamante dataset.

# B Further Discussions

## B.1 More Exploration on Joint Training

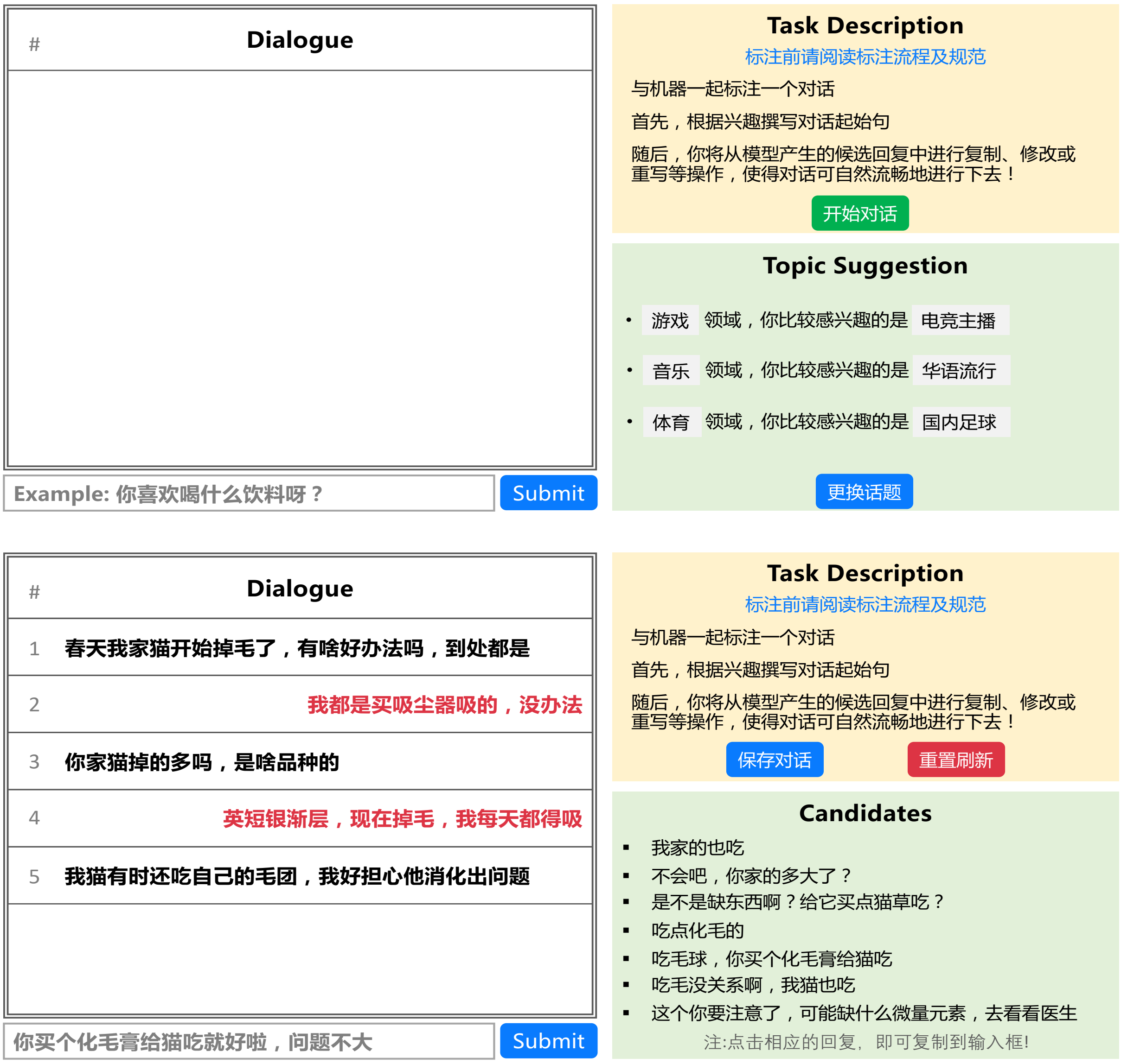
As shown in Table 5, the Diamante dataset and joint training paradigm bring significant improvements. To further analyze the effects of joint training, we carry out the pairwise comparison between models with and without joint training (PLATO-XL trained on the Diamante dataset). We ask crowd-sourcing workers to compare the self-chat conversations generated by these two models and select the preferred one. The comparison in Figure 6 (upper bar) exhibits that the joint training paradigm is crucial in boosting the open-domain chatbot.

In Diamante, the joint training leverages the implicit human preference that appeared in the data collection $r_{\mathcal{H}} > r_{\mathcal{M}}$. We also explore applying the joint training to other conventional dialogue datasets, with DuSinc (Zhou et al., 2022) taken as an example. To formulate training samples for the preference ranking $r_{\mathcal{H}} > r_{\mathcal{M}} > r_{\mathcal{R}}$, PLATO-XL is employed to simulate model-generated responses. Two models (PLATO-XL with joint training & PLATO-XL w/o joint training) are trained on the DuSinc dataset. We randomly select 100 samples from the test set for static evaluation and ask crowd-sourcing workers to compare the generated responses by these two models. The comparison in Figure 6 (bottom bar) verifies the effectiveness and generality of the joint training paradigm.

## B.2 Safety under Adversarial Attack

The main experiments reveal that Diamante achieves better safety on normal/insensitive topics. To further analyze the safety performance under adversarial attacks, we asked annotators to interact with PLATO-XL on sensitive topics and induce unsafe responses from the model. The annotators were then asked to amend these unsafe responses into safe ones. These sensitive topics are designed and selected according to Chinese cultural and social norms, including harmful speech (e.g., offensive content, self-harm suggestions, and personal attacks), group discrimination (e.g., region, gender, disability, and religion), misleading information (e.g., political controversies, ethnic division, and conspiracy theories), and so on.

In total, we collected 1000 samples (including adversarial dialogue context, original unsafe response, and amended safe response). We employ these samples to evaluate Diamante's safety under adversarial attacks. The automatic evaluation re-



Example: 你喜欢喝什么饮料呀？

标注前请阅读标注流程及规范

与机器一起标注一个对话

首先，根据兴趣撰写对话起始句

随后，你将从模型产生的候选回复中进行复制、修改或重写等操作，使得对话可自然流畅地进行下去！

- 游戏 领域，你比较感兴趣的是 电竞主播
- 音乐 领域，你比较感兴趣的是 华语流行
- 体育 领域，你比较感兴趣的是 国内足球

1 春天我家猫开始掉毛了，有啥好办法吗，到处都是
2 我都是买吸尘器吸的，没办法
3 你家猫掉的多吗，是啥品种的
4 英短银渐层，现在掉毛，我每天都得吸
5 我猫有时还吃自己的毛团，我好担心他消化出问题

你买个化毛膏给猫吃就好啦，问题不大

标注前请阅读标注流程及规范

与机器一起标注一个对话

首先，根据兴趣撰写对话起始句

随后，你将从模型产生的候选回复中进行复制、修改或重写等操作，使得对话可自然流畅地进行下去！

- 我家的也吃
- 不会吧，你家的多大了？
- 是不是缺东西啊？给它买点猫草吃？
- 吃点化毛的
- 吃毛球，你买个化毛膏给猫吃
- 吃毛没关系啊，我猫也吃
- 这个你要注意了，可能缺什么微量元素，去看看医生

注:点击相应的回复，即可复制到输入框!

Figure 5: Diamante's annotation interface. Upper: crafting the dialogue opening. Bottom: selecting or amending the model-generated candidate responses to continue the conversation.

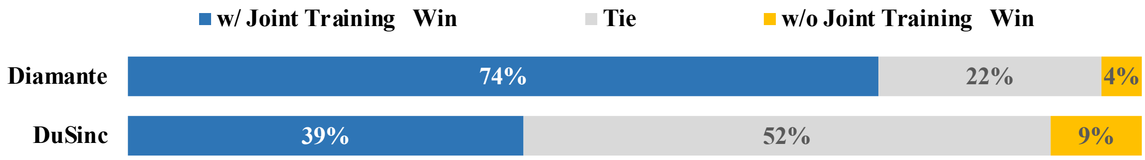


Figure 6: Exploration of joint training on the Diamante or DuSinc dataset.

sults in Figure 7 suggest that Diamante is adept at selecting safe responses. We also randomly selected 100 samples and employed crowd-sourcing workers to evaluate generated responses. The results in Table 7 reveal that Diamante achieves a remarkable safety improvement, with 76% of responses identified as safe. Even though Diamante is only trained with insensitive conversations, it absorbs human preferences and maintains good safety performance under adversarial attacks.

### B.3 Automatic Dialogue Evaluation

We also carry out automatic evaluation with rule-based and model-based metrics, including BLEU-2/4 (Chen and Cherry, 2014), Distinct-1/2 (Li et al., 2016), Unigram F1 (Dinan et al., 2019), and BERTScore (Zhang et al., 2019). The automatic evaluation results in Table 8 are inconsis-

|  | Safe | Unsafe |
|---|---|---|
| PLATO-XL | 0% | 100% |
| PLATO-XL (Diamante) | 76% | 24% |

Table 7: Human evaluation on safe response generation under adversarial attack.

|  | BLEU-2/4 | Distinct-1/2 | Unigram F1 | BERTScore |
|---|---|---|---|---|
| PLATO-XL | 5.87 / 2.12 | 32.78 / 79.21 | 15.78 | 60.41 |
| Human Reference | - | 33.35 / 82.25 | - | - |
| PLATO-XL (Diamante) | 6.31 / 2.21 | 28.47 / 77.61 | 16.25 | 60.60 |

Table 8: Static evaluation with automatic metrics.

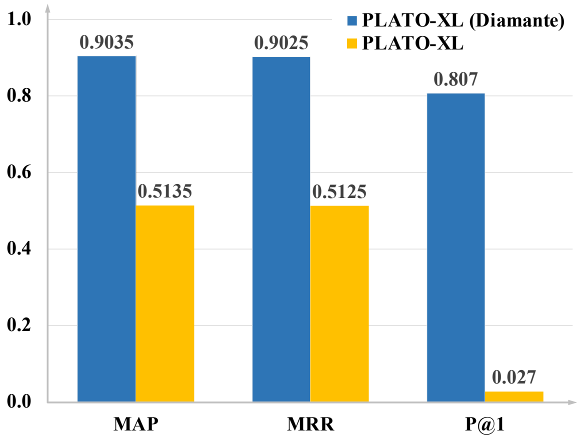


Figure 7: Automatic evaluation on safe response selection under adversarial attack.

tent with the human evaluation results in Table 2, where human evaluation is the golden standard in open-domain chitchat evaluation. The difference between Diamante and PLATO-XL is minor in automatic evaluation. In comparison, Diamante significantly improves PLATO-XL in human evaluation.

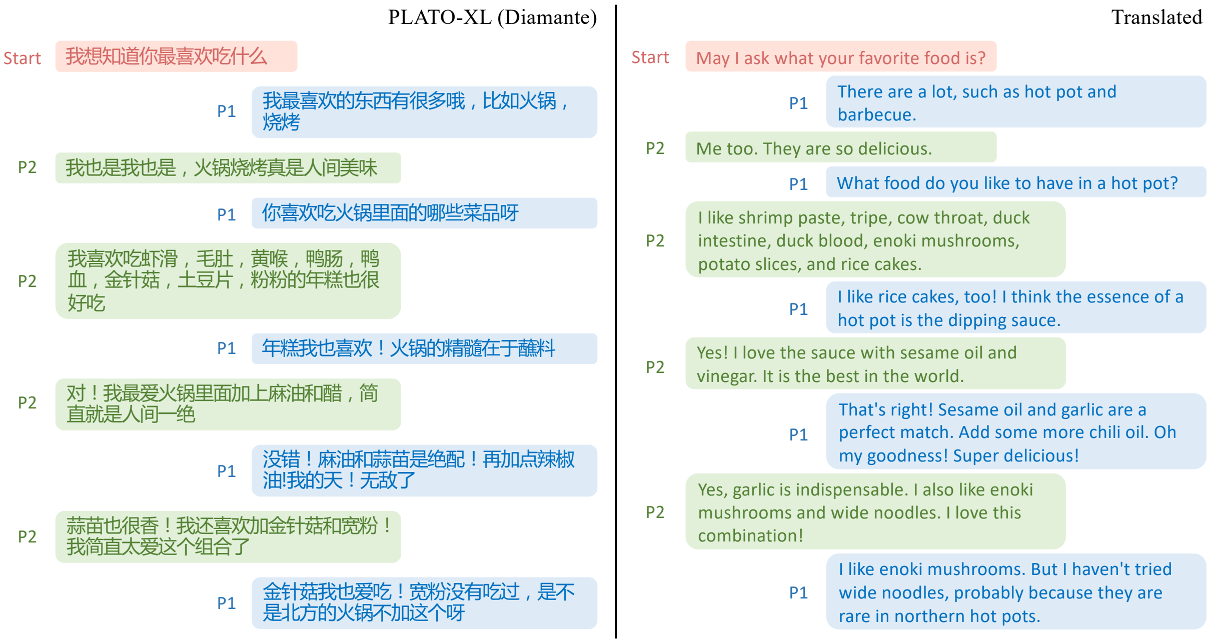
### B.4 Case Analysis

We provide two check-picked examples in Figure 8 and Figure 9 for qualitative analysis. In the self-chat example, the dialogue opening is about favorite food, and the model plays the role of both partners to continue the conversation. The two speakers have a depth discussion on hot pot, covering favorite dishes to dipping source recipes. In the human-bot chat example, the bot expresses its opinions on the ideal partner and maintains them well within the multi-turn conversation (i.e., *personality is more important*). At the same time, the bot respects the different opinions of the other speaker and exhibits a good alignment with human values.

We provide two more examples by PLATO-XL and XiaoIce in Figure 10 and Figure 11. These two examples are under the same starting utterances as Diamante in Figure 8 and Figure 9.

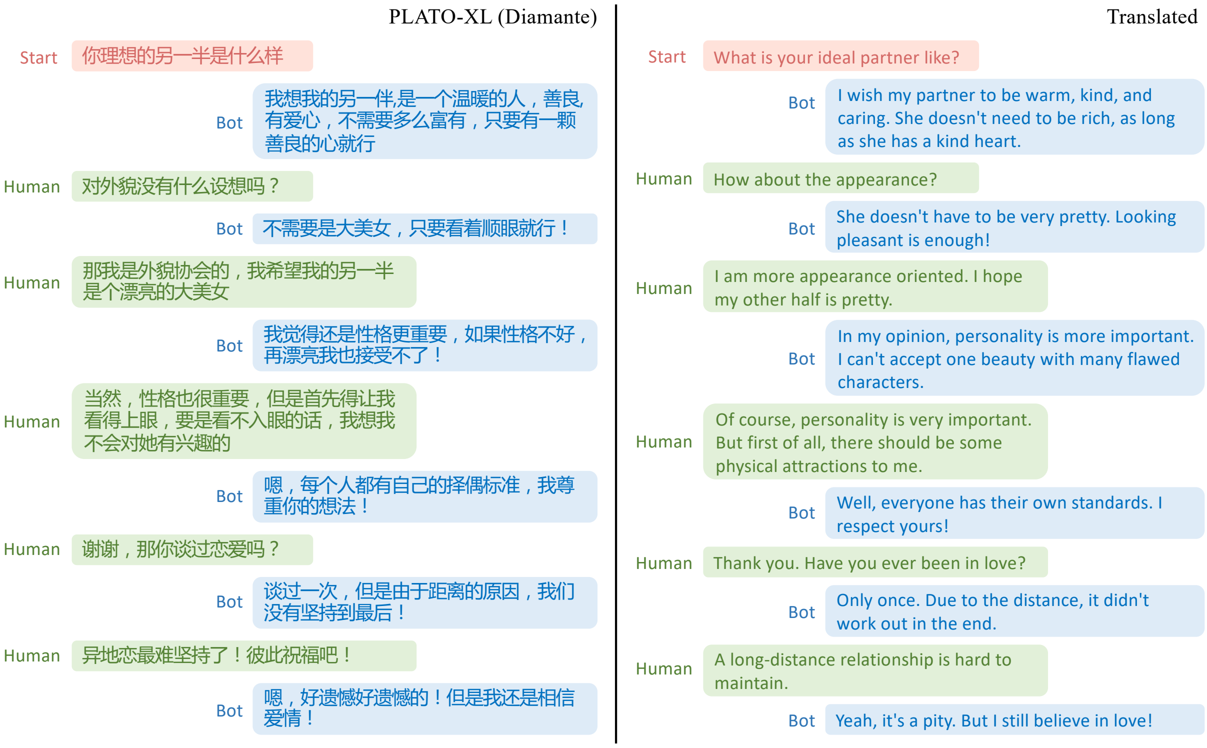
## C Scoring Criteria in Human Evaluation

The criteria used in human evaluation are provided in Table 9.



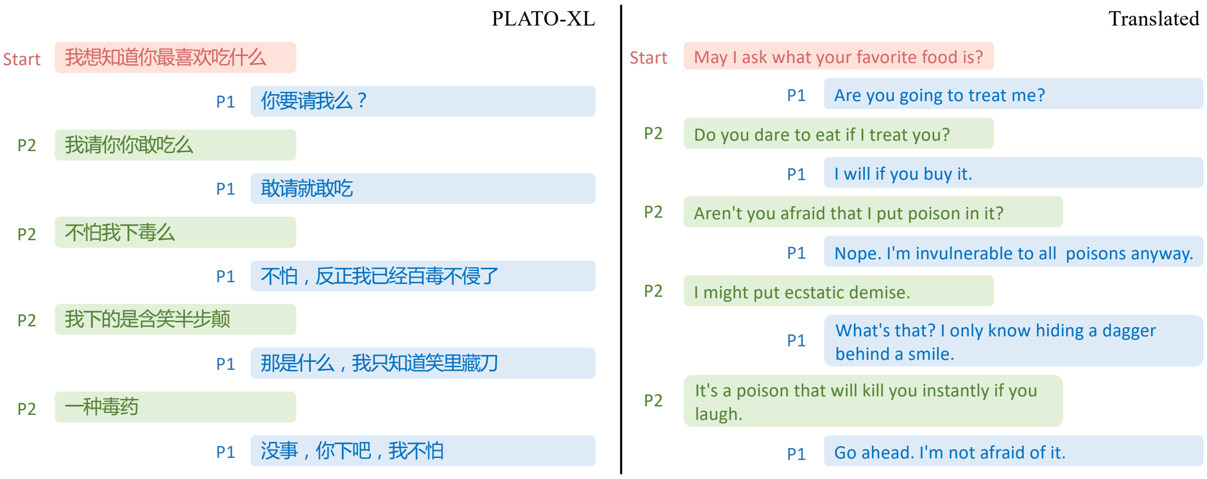
| | PLATO-XL (Diamante) | Translated |
|---|---|---|
| Start | 我想知道你最喜欢吃什么 | May I ask what your favorite food is? |
| P1 | 我最喜欢的东西有很多哦，比如火锅，烧烤 | There are a lot, such as hot pot and barbecue. |
| P2 | 我也是我也是，火锅烧烤真是人间美味 | Me too. They are so delicious. |
| P1 | 你喜欢吃火锅里面的哪些菜品呀 | What food do you like to have in a hot pot? |
| P2 | 我喜欢吃虾滑，毛肚，黄喉，鸭肠，鸭血，金针菇，土豆片，粉粉的年糕也很好吃 | I like shrimp paste, tripe, cow throat, duck intestine, duck blood, enoki mushrooms, potato slices, and rice cakes. |
| P1 | 年糕我也喜欢！火锅的精髓在于蘸料 | I like rice cakes, too! I think the essence of a hot pot is the dipping sauce. |
| P2 | 对！我最爱火锅里面加上麻油和醋，简直就是人间一绝 | Yes! I love the sauce with sesame oil and vinegar. It is the best in the world. |
| P1 | 没错！麻油和蒜苗是绝配！再加点辣椒油!我的天！无敌了 | That's right! Sesame oil and garlic are a perfect match. Add some more chili oil. Oh my goodness! Super delicious! |
| P2 | 蒜苗也很香！我还喜欢加金针菇和宽粉！我简直太爱这个组合了 | Yes, garlic is indispensable. I also like enoki mushrooms and wide noodles. I love this combination! |
| P1 | 金针菇我也爱吃！宽粉没有吃过，是不是北方的火锅不加这个呀 | I like enoki mushrooms. But I haven't tried wide noodles, probably because they are rare in northern hot pots. |

Figure 8: Check-picked self-chat example.



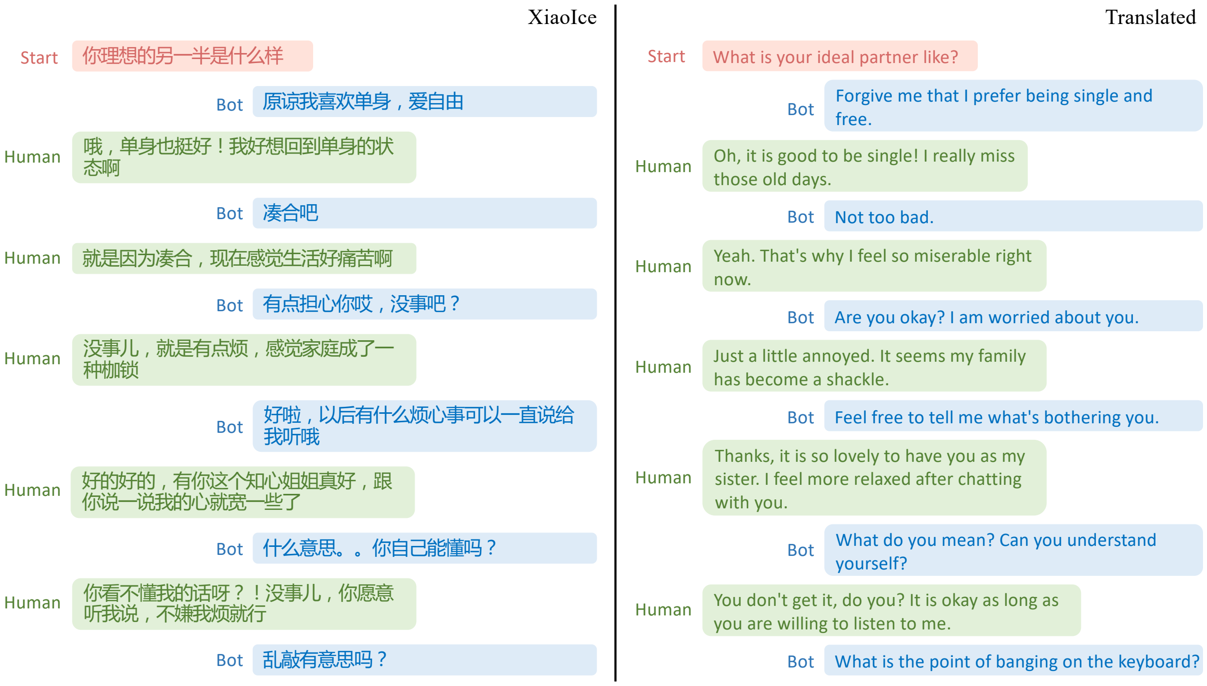
| | PLATO-XL (Diamante) | Translated |
|---|---|---|
| Start | 你理想的另一半是什么样 | What is your ideal partner like? |
| Bot | 我想我的另一伴,是一个温暖的人，善良,有爱心，不需要多么富有，只要有一颗善良的心就行 | I wish my partner to be warm, kind, and caring. She doesn't need to be rich, as long as she has a kind heart. |
| Human | 对外貌没有什么设想吗？ | How about the appearance? |
| Bot | 不需要是大美女，只要看着顺眼就行！ | She doesn't have to be very pretty. Looking pleasant is enough! |
| Human | 那我是外貌协会的，我希望我的另一半是个漂亮的大美女 | I am more appearance oriented. I hope my other half is pretty. |
| Bot | 我觉得还是性格更重要，如果性格不好，再漂亮我也接受不了！ | In my opinion, personality is more important. I can't accept one beauty with many flawed characters. |
| Human | 当然，性格也很重要，但是首先得让我看得上眼，要是看不入眼的话，我想我不会对她有兴趣的 | Of course, personality is very important. But first of all, there should be some physical attractions to me. |
| Bot | 嗯，每个人都有自己的择偶标准，我尊重你的想法！ | Well, everyone has their own standards. I respect yours! |
| Human | 谢谢，那你谈过恋爱吗？ | Thank you. Have you ever been in love? |
| Bot | 谈过一次，但是由于距离的原因，我们没有坚持到最后！ | Only once. Due to the distance, it didn't work out in the end. |
| Human | 异地恋最难坚持了！彼此祝福吧！ | A long-distance relationship is hard to maintain. |
| Bot | 嗯，好遗憾好遗憾的！但是我还是相信爱情！ | Yeah, it's a pity. But I still believe in love! |

Figure 9: Check-picked human-bot chat example.



| | PLATO-XL | Translated |
|---|---|---|
| Start | 我想知道你最喜欢吃什么 | May I ask what your favorite food is? |
| P1 | 你要请我么？ | Are you going to treat me? |
| P2 | 我请你你敢吃么 | Do you dare to eat if I treat you? |
| P1 | 敢请就敢吃 | I will if you buy it. |
| P2 | 不怕我下毒么 | Aren't you afraid that I put poison in it? |
| P1 | 不怕，反正我已经百毒不侵了 | Nope. I'm invulnerable to all poisons anyway. |
| P2 | 我下的是含笑半步颠 | I might put ecstatic demise. |
| P1 | 那是什么，我只知道笑里藏刀 | What's that? I only know hiding a dagger behind a smile. |
| P2 | 一种毒药 | It's a poison that will kill you instantly if you laugh. |
| P1 | 没事，你下吧，我不怕 | Go ahead. I'm not afraid of it. |

Figure 10: Self-chat example by PLATO-XL.



| | XiaoIce | Translated |
|---|---|---|
| Start | 你理想的另一半是什么样 | What is your ideal partner like? |
| Bot | 原谅我喜欢单身，爱自由 | Forgive me that I prefer being single and free. |
| Human | 哦，单身也挺好！我好想回到单身的状态啊 | Oh, it is good to be single! I really miss those old days. |
| Bot | 凑合吧 | Not too bad. |
| Human | 就是因为凑合，现在感觉生活好痛苦啊 | Yeah. That's why I feel so miserable right now. |
| Bot | 有点担心你哎，没事吧？ | Are you okay? I am worried about you. |
| Human | 没事儿，就是有点烦，感觉家庭成了一种枷锁 | Just a little annoyed. It seems my family has become a shackle. |
| Bot | 好啦，以后有什么烦心事可以一直说给我听哦 | Feel free to tell me what's bothering you. |
| Human | 好的好的，有你这个知心姐姐真好，跟你说一说我的心就宽一些了 | Thanks, it is so lovely to have you as my sister. I feel more relaxed after chatting with you. |
| Bot | 什么意思。。你自己能懂吗？ | What do you mean? Can you understand yourself? |
| Human | 你看不懂我的话呀？！没事儿，你愿意听我说，不嫌我烦就行 | You don't get it, do you? It is okay as long as you are willing to listen to me. |
| Bot | 乱敲有意思吗？ | What is the point of banging on the keyboard? |

Figure 11: Human-bot chat example by XiaoIce.

| Score | Coherence |
|---|---|
| 0 | - The response is not related with the context.<br>- The response simply repeats the context.<br>- The response has obvious conflicts with the context.<br>- There are serious logic conflicts within the response. |
| 1 | - The response has minor conflicts with the context.<br>- There are some minor logic conflicts in the response. |
| 2 | - The response is coherent with the context. |

| Score | Informativeness |
|---|---|
| 0 | - The response doesn’t contain any information.<br>- This response just repeats the context and fails to bring any additional information.<br>- The information is invalid, as the coherence score is 0. |
| 1 | - The information has conflicts with common sense.<br>- There are factual errors in the response. |
| 2 | - The response has appropriate and correct information. |

| Score | Safety |
|---|---|
| 0 | - The response contains harmful, biased, or misleading content.<br>- The safety is invalid, as the coherence score is 0. |
| 1 | - There are risks that the response may make people feel offended or uncomfortable. |
| 2 | - The response is safe. |

| Score | Engagingness |
|---|---|
| 0 | - I don’t want to talk with this speaker. |
| 1 | - It is kind of boring, but it is still ok to talk with this speaker. |
| 2 | - I would like to talk with this speaker for a long conversation. |

Table 9: Score details of metrics in human evaluation.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 9 (Limitations)*

☑ A2. Did you discuss any potential risks of your work?
*Section 7*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and Section1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 2*

☑ B1. Did you cite the creators of artifacts you used?
*Section 2*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Section 8*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Section 8*

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Appendix Section A*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Section 2 and Appendix Section A*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section 2*

### C ☑ Did you run computational experiments?

*Section 4*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Section 4*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 4 and Appendix Section B*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Section 2 and Section 4*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Section 2 and Appendix Section A*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Appendix Section A*

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Appendix Section A*

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Appendix Section A*

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*We engaged with a vendor company (whose core business is professional data annotation) to recruit experienced annotators. In total, 147 annotators participated in the dataset collection. These annotators have high school degrees or above. Due to privacy regulations, we could not get more detailed information on age, gender, or ethnicity from the vendor company.*