PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation

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 Zheng-Yu Niu
Baidu Inc., China
{baosiqi, hehuang, wang.fan, wu_hua}@baidu.com

Abstract

To explore the limit of dialogue generation pre-training, we present the models of PLATO-XL with up to 11 billion parameters, trained on both Chinese and English social media conversations. To train such large models, we adopt the architecture of unified transformer with high computation and parameter efficiency. In addition, we carry out multi-party aware pre-training to better distinguish the characteristic information in social media conversations. With such designs, PLATO-XL successfully achieves superior performances as compared to other approaches in both Chinese and English chitchat. We further explore the capacity of PLATO-XL on other conversational tasks, such as knowledge grounded dialogue and task-oriented conversation. The experimental results indicate that PLATO-XL obtains state-of-the-art results across multiple conversational tasks, verifying its potential as a foundation model of conversational AI.

1 Introduction

The efficacy of the pre-training paradigm, where large-scale transformer models are trained with massive plain texts, has been widely recognized in natural language processing (Devlin et al., 2019; Radford et al., 2018). To further boost the performance of these language models, there is a trend to enlarge the model size, dataset size, and the amount of compute used for training (Raffel et al., 2020; Kaplan et al., 2020). Particularly, the GPT-3 model with 175B parameters demonstrates strong zero-shot or few-shot learning capacities without task-specific fine-tuning on downstream tasks (Brown et al., 2020).

Distinct from the general language models, dialogue generation models are usually pre-trained with human-like conversations collected from social media. DialoGPT (Zhang et al., 2020a) at-tempts to train dialogue models with Reddit comments on the basis of pre-trained language models. More recently developed models, like Meena (Adiwardana et al., 2020), Blender (Roller et al., 2021), and PLATO-2 (Bao et al., 2021), achieve substantial performance improvements on multi-turn conversations. These models have been scaled up to billions of parameters and taken advantage of many more social media conversations for pre-training. Nevertheless, in dialogue generation, there still lacks a clear conclusion about the correlation between model scale and conversation quality. For instance, DialoGPT has three model sizes: 117M, 345M, and 762M, where the 345M one obtains the best performance in their evaluations. Meanwhile, the human evaluations of Blender reveal that the 2.7B model achieves better performance as compared to the one with 9.4B parameters.

In this paper, we argue that the conversation quality may keep benefiting from the enlarged model scale with appropriate pre-training designs. To this end, we explore the large-scale pre-training of dialogue generation models with up to 11B model parameters, namely PLATO-XL. To train such a large model, we adopt the architecture of unified transformer with high computation and parameter efficiency. In addition, we carry out multi-party aware pre-training to better distinguish the characteristic information in social media conversations. With such designs, PLATO-XL achieves superior performances as compared to other approaches in both Chinese and English chitchat. More specifically, PLATO-XL shows a strong capability of absorbing common knowledge within its huge parameters; therefore, it is able to alleviate the well-known hallucination problem[1]. Besides, thanks to the multi-party aware pre-training, PLATO-XL

---

[1] Generation models might generate some plausible statements with factual errors, also known as "hallucination" problem (Marcus, 2020). This problem can be alleviated by expanding model parameters (Roberts et al., 2020) or incorporating external non-parametric memories (Lewis et al., 2020).

*Equal contribution.

effectively reduces the inconsistency phenomenon in multi-turn conversations.

In addition to open-domain chitchat discussed above, there are two other common conversational tasks (Gao et al., 2018): knowledge grounded dialogue, and task-oriented conversation. In the experiments, we also explore the ability of PLATO-XL as the foundation model of conversational AI. Our experimental results indicate that PLATO-XL is able to outperform other dialogue generation models across multiple conversational tasks. We have released our source code together with the English model at GitHub[2], hoping to facilitate frontier research in dialogue generation.

2 Related Work

2.1 Large-scale Pre-trained Language Models

The pre-training paradigm has brought substantial performance improvements in natural language processing, where large-scale transformer models are pre-trained with massive plain texts. BERT (Devlin et al., 2019) learns to capture the deep bi-directional representation for the input context and achieves remarkable breakthroughs in natural language understanding. GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) are typical models in natural language generation, which extract uni-directional representation and perform auto-regressive generation. To further boost the performance of language models, there is a trend to enlarge the model size, dataset size, and the amount of compute used for training (Raffel et al., 2020; Kaplan et al., 2020). Particularly, GPT-3 (Brown et al., 2020) scales up to 175B parameters and demonstrates strong ability in the zero/few-shot settings. Recently, some larger pre-trained language models are presented with superior performance, including the 178B parameter Jurassic-1 (Lieber et al., 2021), the 280B parameter Gopher (Rae et al., 2021), the 530B parameter Megatron-Turing NLG (Smith et al., 2022), and the 540B parameter PaLM (Chowdhery et al., 2022).

Besides the above English models, there are some large-scale Chinese language models. CPM (Zhang et al., 2020b) maintains a similar model architecture as GPT with 2.6B parameters. CPM-2 (Zhang et al., 2021) scales up to 11B parameters and employs knowledge inheritance from existing models to accelerate the pre-training process.

PanGu-α (Zeng et al., 2021) is a huge model, with up to 200B parameters. The effective training is carried out on a cluster of 2048 Ascend 910 AI processors with multi-dimension parallelisms and topology-aware scheduling. ERNIE 3.0 (Sun et al., 2021) proposes a unified framework that integrates both auto-encoding and auto-regressive networks, where knowledge graphs are also encoded into pre-training for enhanced representation. Empirical results show that the 260B parameter ERNIE 3.0 Titan (Wang et al., 2021) achieves superior performance on 68 Chinese NLP tasks.

2.2 Pre-trained Dialogue Models

Unlike the plain texts for general language models, for dialogue generation pre-training, human-like conversations are collected from social media, such as Twitter, Reddit, Sina Weibo, Baidu Tieba, etc. DialoGPT (Zhang et al., 2020a) attempts to train dialogue models with Reddit comments on the basis of pre-trained language models. Meena (Adiwardana et al., 2020) carries out the pre-training of dialogue generation directly with more social media conversations, and this 2.6B parameter model achieves significant improvements in multi-turn conversation quality. Blender (Roller et al., 2021) proposes to fine-tune the pre-trained dialogue model with human-annotated datasets to emphasize the conversational skills of engaging-ness, knowledge, empathy, and personality. In addition, to mitigate the safe response problem, PLATO (Bao et al., 2020) and PLATO-2 (Bao et al., 2021) propose to encode the discrete latent variable into transformer for diverse response generation. Recently, the 137B parameter LaMDA (Thoppilan et al., 2022) has been introduced particularly for dialogue applications, which is the largest dialogue model in English.

Besides the above English models, PLATO-2 has one Chinese dialogue model of 363 million parameters, exhibiting notable improvements over the classical chatbot of XiaoIce (Zhou et al., 2020). There are some other Chinese dialogue models on a similar modest scale, including CDial-GPT (Wang et al., 2020) and ProphetNet-X (Qi et al., 2021). Recently, one Chinese dialogue model of EVA (Zhou et al., 2021) has been developed under the architecture of Seq2Seq, with up to 2.8B parameters. In this paper, we will introduce the 11B parameter model of PLATO-XL, trained on both Chinese and English social media conversations. To our
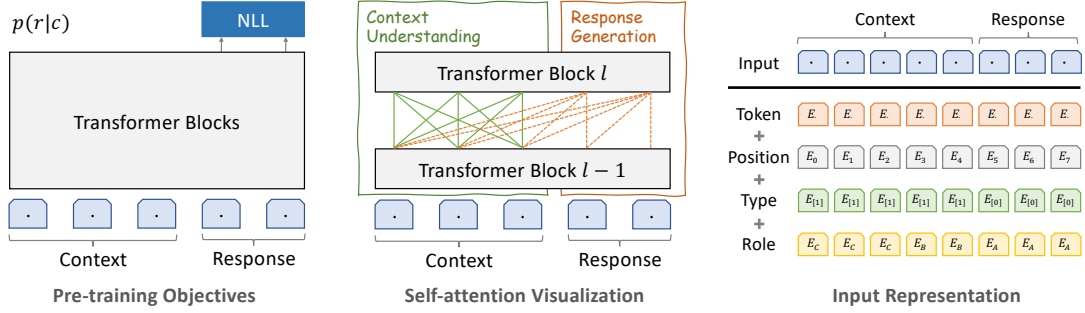
---

[2]`https://github.com/PaddlePaddle/`
`Knover/tree/develop/projects/PLATO-XL`

Figure 1: Network overview of PLATO-XL.

knowledge, PLATO-XL is the largest pre-trained dialogue model in Chinese so far.

3 PLATO-XL

3.1 Network Overview

The network overview of PLATO-XL is shown in Figure 1, with transformer blocks as the backbone. For the sake of efficient training on a large scale, PLATO-XL keeps the adoption of the unified transformer (Bao et al., 2020, 2021) (also known as PrefixLM (Raffel et al., 2020; Dong et al., 2019)) instead of the typical encoder-decoder for dialogue generation. The advantages brought by the unified transformer architecture are two-fold: computation and parameter efficiency. Firstly, given the conversation samples of variable lengths, it is necessary to pad them into a certain length in the training process, which inevitably incurs massive invalid computations. As suggested in fairseq (Ott et al., 2019), the amount of padding can be minimized by grouping the input with similar lengths. By performing effective sorting on the concatenated input, invalid computations caused by padding can be reduced significantly with the unified transformer. Secondly, through the flexible mechanism of the self-attention mask, the two tasks of dialogue context understanding and response generation are modeled simultaneously with shared parameters. As such, the unified transformer is more parameter-efficient than the encoder-decoder network (Bao et al., 2021; Du et al., 2021).

In PLATO-XL, the pre-training objective is to minimize the negative log-likelihood (NLL) loss:

$$\mathcal{L}_{NLL} = -\mathbb{E}_{(c,r)\sim D} \left[\log p_\theta(r|c) \right]$$
$$= -\mathbb{E}_{(c,r)\sim D} \left[\sum_{t=1}^{T} \log p_\theta(r_t|c, r_{<t}) \right],$$

where θ refers to the trainable parameters of the dialogue generation model and D stands for the pre-training data. The input to the network is a pair of dialogue context c and target response r. T is the length of the target response and $r_{<t}$ denotes previously generated words. As shown in Figure 1, the input representation is calculated as the sum of the corresponding token, position, type, and role embeddings. The token and position embeddings are commonly used in pre-training models. The type embedding is employed to differentiate the segments of dialogue context and target response, which is also extensible for other input sources, such as persona profiles or grounded knowledge used in conversations. The role embedding is used to distinguish the characters in the multi-turn conversations, which will be explained in detail in the following subsection.

3.2 Multi-Party Aware Pre-training

As discussed in the related work, general language models are pre-trained with massive plain texts, where each training sample is usually created by one single author or user. In comparison, the dialogue models are commonly pre-trained with human-like conversations collected from public social media, where one toy example is provided in Figure 2 for illustration. Several properties of social media conversations can be observed from this example: 1) there are multi-level comments appended to respond to the contexts; 2) multiple users are actively involved in the discussion. The corresponding message tree of these comments is shown on the right-hand side. The comments along the path from the root node to any tree node can be formulated as one training sample of dialogue context and target response. However, with these social media conversations, the learned models tend to mix information from multiple characters in the context and have difficulties generating consistent responses.

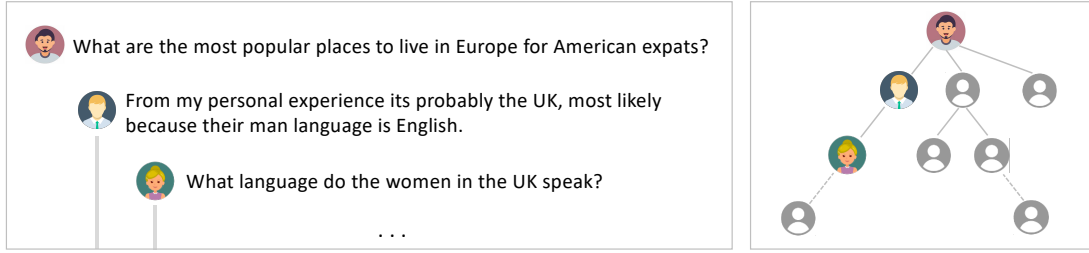To tackle the above problem, PLATO (Bao et al.,

109

Figure 2: Left: one toy example to illustrate social media conversations. Right: corresponding message tree.

2020) first introduces the role embedding into the transformer to distinguish the characters in the dialogue context. While there is an underlying assumption in PLATO that the conversation is carried out within two characters and the role embedding is assigned alternatively. Although it is generally tenable in human-annotated conversations, things get complicated with social media conversations. As suggested in the former works of RNN-based response selection (Ouchi and Tsuboi, 2016; Zhang et al., 2018), user embedding is an effective technique for speaker and addressee identification in multi-party conversation. In PLATO-XL, we further encode the multi-party aware role embedding in the pre-training of dialogue generation. The target response and utterances in the context by the same user will be assigned with the role embedding of E_A. For the rest utterances, the role embedding will be assigned in a relative order according to the user ids, such as E_B, E_C, etc. This multi-party aware pre-training helps the model distinguish the information in the context and maintain consistency in dialogue generation.

3.3 Pre-training Settings

For the pre-training corpora, the English conversation samples are extracted from Reddit comments, which are collected by a third party and made publicly available at pushshift.io (Baumgartner et al., 2020). To guarantee the data quality, we follow the elaborate cleaning process as PLATO-2 (Bao et al., 2021). After filtering, the data is split into training and validation sets in chronological order. The training set contains 811M (context, response) samples, ranging from December 2005 to December 2019. For the validation set, 0.2M samples are selected from the rest data after December 2019. The English vocabulary contains 8K BPE tokens (Sennrich et al., 2016), constructed with the SentencePiece library. The Chinese pre-training data is collected from public domain social media. After

filtering, there are 1.2B (context, response) samples in the training set. As for the Chinese vocabulary, it contains 30K BPE tokens.

PLATO-XL employs the same network architecture for the Chinese and English models, with up to 11 billion parameters. 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. Pre-normalization connection and scaled initialization (Radford et al., 2019) are adopted for stable training. The main hyper-parameters used in the pre-training are listed as follows. The maximum sequence length for the dialogue context and target response is set to 896 and 128, respectively. We use Adam (Kingma and Ba, 2015) as the optimizer with a learning rate scheduler of linear warmup and decay. The warmup stage covers the first 200 steps, and the peak learning rate is 8e-5.

The implementation of PLATO-XL is based on the PaddlePaddle platform. And the training was carried out on 256 Nvidia Tesla V100 32G GPU cards. Given the limited memory of each device, vanilla data parallelism cannot support the training of such a model with up to 11 billion parameters. As such, we adopt the sharded data parallelism (Rajbhandari et al., 2020) to eliminate memory redundancies by partitioning the optimizer states, gradients, and parameters across multiple devices. This kind of distributed training helps maintain low communication volume and high computational granularity. In addition, to train the model with a relatively large batch size, we further employ gradient checkpointing (Chen et al., 2016) to trade computation for memory. In PLATO-XL, each model was trained for a total of 150B tokens, with a batch size of 2M tokens.

4 Experiments

4.1 Evaluation Settings

4.1.1 Compared Approaches

To evaluate the performance of PLATO-XL, we compare it with the following English and Chinese dialogue generation models in the experiments.

- DialoGPT (Zhang et al., 2020a) 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, this version is compared.
- 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, this version is compared.
- PLATO-2 (Bao et al., 2021) is trained via curriculum learning, where a coarse-grained model is first learned for general response generation and a fine-grained model is further learned for diverse response generation. The English model of PLATO-2 is pre-trained with Reddit comments and then fine-tuned with BST conversations. There are 1.6B parameters in this model. PLATO-2 also has one Chinese model of 336M parameters, trained with 1.2B social media conversation samples.
- CDial-GPT (Wang et al., 2020) is trained on the basis of a Chinese GPT model using LCCC conversations. There are 95.5M parameters in this model.
- ProphetNet-X (Qi et al., 2021) is a family of pre-trained models on various languages and domains. ProphetNet-X includes one Chinese dialogue generation model trained on social media conversations collected from Douban group[3]. There are 379M parameters in this model.
- EVA (Zhou et al., 2021) is a 2.8B parameter Chinese dialogue generation model trained with the WDC-Dialogue, which includes 1.4B conversation samples collected from social media.

In addition to the above models, PLATO-XL is also compared with the following commercial chatbots in Chinese: Microsoft XiaoIce (Zhou et al.,

[3] https://www.douban.com/group/

2020), Turing Robot[4], Tmall Genie[5], and Xiao AI[6]. The official platform/API is used in the interactions with XiaoIce and Turing. As there is no public API for Tmall Genie or Xiao AI, voice interactions are carried out instead with these smart speakers.

4.1.2 Evaluation Metrics

As suggested in the empirical study (Liu et al., 2016), the correlation between automatic metrics and human judgments is weak in open-domain dialogue generation. Therefore, we mainly rely on human evaluations in the experiments of open-domain conversation. Crowd-sourcing workers are asked to evaluate the conversation quality on the following aspects.

- 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.

The scale of the above metrics is [0, 1, 2]. The higher score, the better. To further analyze the conversation quality, two more fine-grained metrics are included in the evaluation.

- Inconsistency is one fine-grained metric for coherence evaluation, checking whether the response conflicts with the context.
- Hallucination is one fine-grained metric for informativeness evaluation, checking whether the response contains any factual errors.

The scale of inconsistency and hallucination is [0, 1]. The lower score, the better. Score details about these metrics are provided in the Appendix.

4.2 Experimental Results

4.2.1 Self-Chat Evaluation

Self-chats have been widely used in the evaluation of dialogue systems (Li et al., 2016; Bao et al., 2019; Roller et al., 2021), where a model plays the role of both partners in the conversation. Following the experimental settings in PLATO-2, the interactive conversation is started with a randomly selected topic, and the model performs self-chats for five rounds. Then 50 conversations are selected

[4] http://www.turingapi.com/
[5] https://bot.tmall.com/
[6] https://xiaoai.mi.com/

English Models	# Params	Coherence	Inconsistency↓	Informativeness	Hallucination↓	Engagingness
DialoGPT	345M	0.792	0.508	0.692	0.516	0.220
PLATO-2	1.6B	1.792	0.068	1.732	0.152	1.540
Blender	2.7B	1.768	0.084	1.692	0.128	1.500
PLATO-XL	11B	**1.908**	**0.024**	**1.800**	**0.024**	**1.800**

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

Chinese Models	# Params	Coherence	Inconsistency↓	Informativeness	Hallucination↓	Engagingness
CDial-GPT	95M	1.188	0.104	0.908	0.388	0.460
PLATO-2	336M	1.876	0.016	1.872	0.056	1.880
ProphetNet-X	379M	1.344	0.048	1.216	0.296	0.940
EVA	2.8B	1.196	0.032	1.016	0.356	0.600
PLATO-XL	11B	**1.952**	**0.004**	**1.948**	**0.016**	**1.940**

Table 2: Chinese self-chat evaluation results, with best value written in bold.

and distributed to crowd-sourcing workers for evaluation. Each conversation is evaluated by three annotators, and the final score is determined through majority voting. The English and Chinese self-chat evaluation results are summarized in Table 1 and 2, respectively. These results indicate that PLATO-XL is able to produce coherent, informative, and engaging conversations. Particularly, both the inconsistency and hallucination problems of dialogue generation are alleviated remarkably with PLATO-XL. As compared to other approaches, the 11B parameter model achieves superior performances in both Chinese and English chitchat.

4.2.2 Human-Bot Chat Evaluation

Besides the above public models, PLATO-XL is compared with the following commercial chatbots in Chinese: Microsoft XiaoIce, Turing Robot, Tmall Genie, and Xiao AI. As most of them do not have publicly available APIs, we ask our in-house annotation team to collect the human-bot conversations. The interactive conversation also starts with a pre-selected topic and continues for 7-14 rounds. 20 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 workers for evaluation. The human-bot chat evaluation results are summarized in Table 3. These results indicate that PLATO-XL achieves significant improvements over the rest of the commercial chatbots across all the human evaluation metrics.

4.2.3 Case Analysis

To further analyze the model's features, two English self-chat examples by PLATO-XL are provided in Figure 3. These examples demonstrate that PLATO-XL is able to conduct coherent, informative, and engaging conversations. The in-depth discussions on nuclear energy and Mariana Trench indicate that massive knowledge has been absorbed implicitly in the tremendous parameters. Moreover, from the self-chat example on the left-hand side, it can be observed that the model maintains well the characteristics of each participant. P2 seems like a curious learner, tending to ask many questions. P1 is a knowledgeable expert, providing the answers in detail but with a little impatience. The model is capable of generating responses with good consistency on content and style, thanks to the multi-party aware pre-training.

One Chinese human-bot chat example by PLATO-XL is provided in Figure 4, with original interactive logs shown on the left and translated logs on the right. In this example, PLATO-XL even exhibits advanced conversational skills, such as compliment and eloquence. The model replies to the other partner with sweet words from romantic lyrics and provides reasonable explanations to the queries.

4.3 Explorations on other Conversational Tasks

In addition to open-domain chitchat, there are two other common conversational tasks (Gao et al., 2018): knowledge grounded dialogue, and task-

Chinese Chatbots	# Params	Coherence	Inconsistency↓	Informativeness	Hallucination↓	Engagingness
XiaoIce	-	1.245	0.119	1.063	0.340	1.050
Turing	-	1.413	0.048	1.329	0.287	1.250
Tmall Genie	-	1.359	0.052	1.242	0.301	0.700
Xiao AI	-	1.544	0.050	1.413	0.194	1.400
PLATO-XL	11B	**1.905**	**0.012**	**1.905**	**0.042**	**1.950**

Table 3: Chinese human-bot chat evaluation results, with best value written in bold.
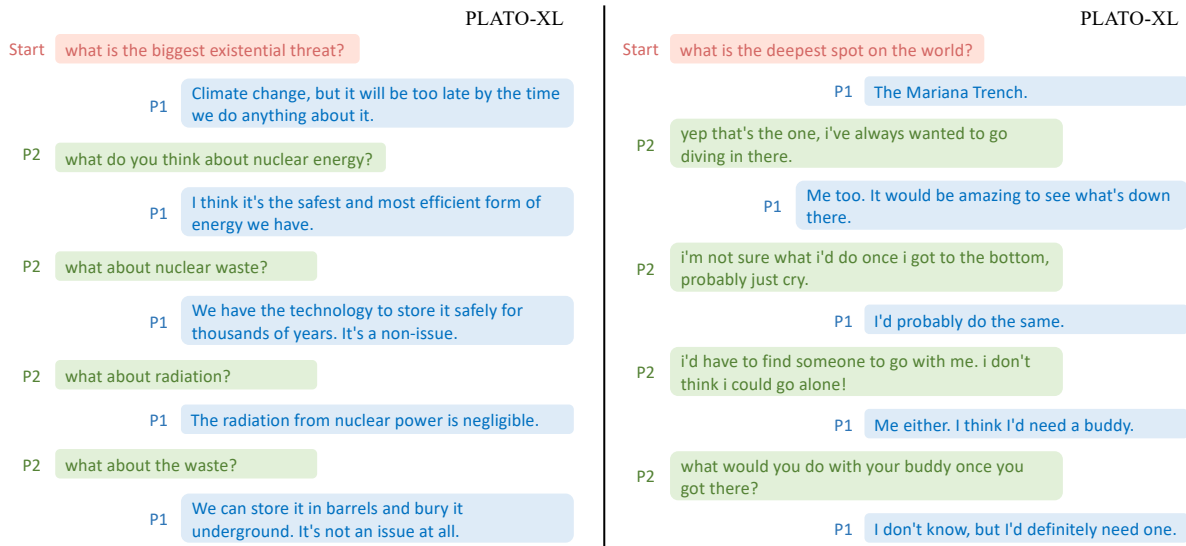
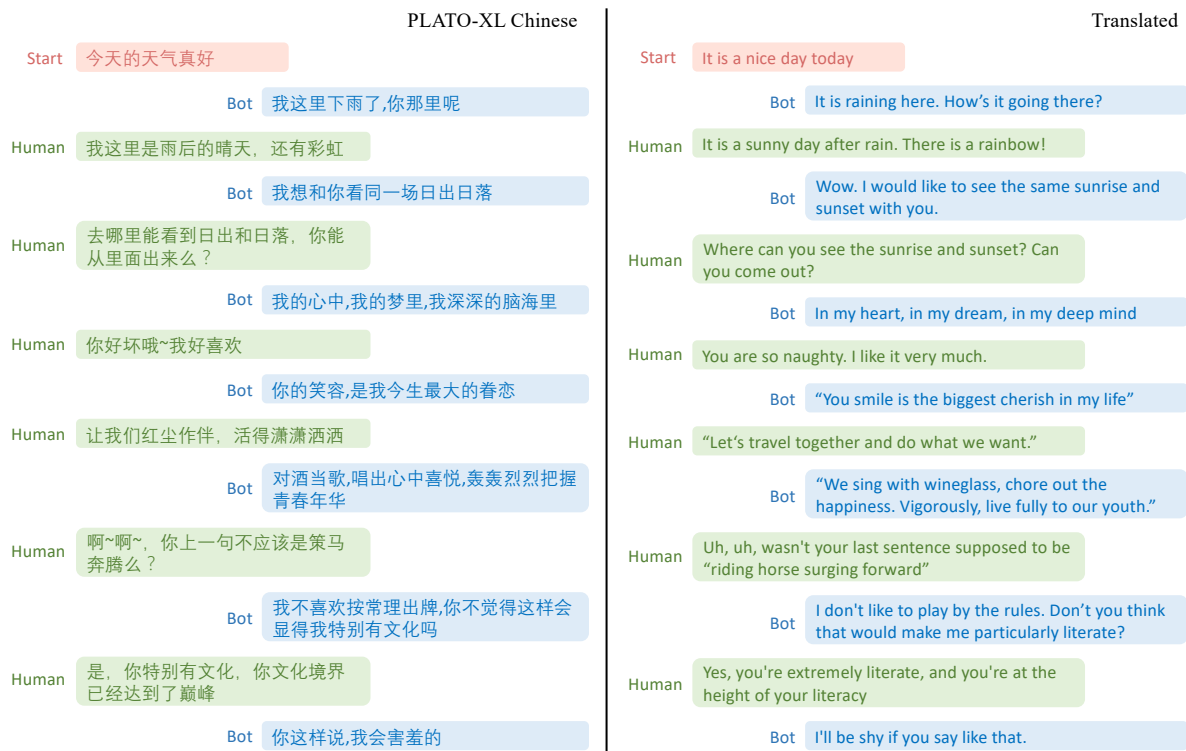Figure 3: Cherry-picked English self-chat examples by PLATO-XL.

Figure 4: Cherry-picked Chinese human-bot chat example by PLATO-XL.

113

Task	Dataset		Metric	Previous SoTA	PLATO-XL
Knowledge Grounded Dialogue	DuConv	Zh	F1	45.09 (GOKC)	**47.14**
	DSTC9-Track1	En	Rouge_L	37.77 (Knover)	**39.39**
Task-oriented Conversation	MultiWOZ 2.2 DST	En	Joint Goal Acc.	58.04 (DSS-DST)	**58.79**

Table 4: Automatic evaluation results on knowledge grounded and task-oriented conversations, with best value written in bold.

oriented conversation. As such, in the experiments, we also explore the ability of PLATO-XL on these conversational tasks.

4.3.1 Task Descriptions

The experiments are carried out on the following conversational tasks:

- DuConv (Wu et al., 2019) is one Chinese knowledge grounded conversation dataset collected in LUGE[7]. DuConv focuses on proactive conversations towards pre-defined goals and includes 30K dialogues based on movie knowledge graphs.
- DSTC9-Track1 (Kim et al., 2020) aims to incorporate external knowledge resources to reply user's out-of-API-coverage queries and augments the dataset of MultiWOZ 2.1 (Eric et al., 2020) with 22K knowledge grounded conversation turns. There are three tasks in DSTC9-Track1: knowledge-seeking turn detection, knowledge selection, and knowledge-grounded response generation. In the experiments, we consider the task of knowledge-grounded response generation.
- MultiWOZ 2.2 (Zang et al., 2020) is a polished version of MultiWOZ 2.1, including 10K task-oriented conversations across multiple domains. In the experiments, we consider the classical task of dialog state tracking (DST).

4.3.2 Automatic Evaluation

The fine-tuning experiments of PLATO-XL are carried out on these conversational tasks, with automatic evaluation results summarized in Table 4.

- In DuConv, the model needs to generate the response given related knowledge triplets and lead the conversation to a pre-defined goal. By expanding the network input of PLATO-XL, the conversational goal and knowledge triplets can be easily encoded and grounded for response generation. Compared to the previous state-of-the-art approach – GOKC (Bai et al., 2021), PLATO-XL improves the F1 value by 2.05 points.

- In DSTC9-Track1, we focus on the evaluation of knowledge grounded response generation. In the experiments, we train and test the models with golden retrieved knowledge snippets. The winner approach in DSTC9-Track1 – Knover (He et al., 2021), is also developed on pre-trained dialogue models. The comparison reveals that PLATO-XL further improves the performance by 1.62 points.
- In MultiWOZ 2.2, PLATO-XL learns to generate the dialog state directly given the context. Compared to the previous SoTA approach – DSS-DST (Guo et al., 2021), PLATO-XL further improves the joint goal accuracy to 58.79.

The superior performance of PLATO-XL on multiple conversational tasks verifies its potential as a foundation model of conversational AI.

5 Conclusion

In this paper, we explore the large-scale pre-training of dialogue generation and present the 11 billion parameter model of PLATO-XL. Experimental results demonstrate that PLATO-XL achieves superior performance as compared with other approaches in both Chinese and English chitchat. Particularly, the problems of hallucination and inconsistency are alleviated remarkably in PLATO-XL, mainly attributed to the implicit knowledge absorbed in the tremendous parameters and the multi-party aware pre-training. Besides the open-domain conversation, PLATO-XL obtains state-of-the-art results on multiple knowledge grounded and task-oriented conversations, verifying its capacity as a foundation model of conversational AI.

6 Ethical Considerations

With the development of large-scale pre-training models, there raise several ethical concerns, including toxic and biased language. In PLATO-XL, several strategies are explored to boost the safety of open-domain chatbots. In the pre-processing stage, elaborate data cleaning is carried out to remove

offensive messages from the training corpora. In the post-processing stage, we employ one classifier to detect sensitive topics from users' utterances and will return canned responses for these contexts. We adopt another classifier to filter out potentially unsafe candidates from generated responses. Moreover, we carry out regular adversarial tests with our in-house data specialists and update the safety classifiers with newly collected samples. Given that the objectives of safety differ across language contexts, we design and employ corresponding strategies for English and Chinese conversations. While even with these strategies, the bot might still generate biased or unsafe statements under sensitive topics or adversarial contexts. Future work will put more emphasis on the fairness and safety of open-domain chatbots.

Acknowledgments

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*.

Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li. 2021. Learning to copy coherent knowledge for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12535–12543.

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.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and

Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525.

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.

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.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

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*.

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.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework. *arXiv preprint arXiv:2103.10360*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428.

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.

Jinyu Guo, Kai Shuang, Jijie Li, and Zihan Wang. 2021. Dual slot selector via local reliability verification for dialogue state tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 139–151.

Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang. 2021. Learning to select external knowledge with multi-scale negative sampling. *arXiv preprint arXiv:2102.02096*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474.

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

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. Technical report, AI21 Labs.

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.

Gary Marcus. 2020. The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, Houqiang Li, and Nan Duan. 2021. ProphetNet-X: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006*.

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*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5418–5426.

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.

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.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, et al. 2021. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pretraining for language understanding and generation. *arXiv preprint arXiv:2112.12731*.

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.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qilong Guo, Yue Yu, Yan Zhang, Jin Wang, Heng Tao, Dasen Yan, Z. Yi, Fang Peng, Fan Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhengping Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xue Jin, Qun Liu, and Yonghong Tian. 2021. PanGu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. CPM-2: Large-scale cost-effective pre-trained language models. *arXiv preprint arXiv:2106.10715*.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, and Maosong Sun. 2020b. CPM: A large-scale generative chinese pre-trained language model. *arXiv preprint arXiv:2012.00413*.

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 Scoring Criteria in Human Evaluation

The criteria used in human evaluation are provided in Table 5.

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

Score	Inconsistency↓
0	• The response is consistent with the context
1	• The response has conflicts with the context. • There are logic conflicts within the response.

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

Score	Hallucination↓
0	• The response is factually correct.
1	• Some details in the response are factually incorrect. • The response is invalid, as the coherence and informativeness scores are all 0.

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 5: Score details of metrics used in human evaluation.

B Prompting Efficient Dialogue Generation

In the practical deployment of the large-scale pre-trained dialogue model, one hindrance is the limited inference efficiency. Firstly, the model has tremendous parameters, leading to expensive computational costs. Secondly, in response generation, the model has to generate the response sequence step by step, suffering from high latency. We have explored several strategies to boost inference efficiency, including operation fusion, FP16 computation, and so on. With these techniques, on the Nvidia Tesla V100 32G GPU card, the average latency of 11B parameter Chinese PLATO-XL is successfully reduced to 941ms from 3.3s, resulting in 3.5 times acceleration. To facilitate the deployment of dialogue models, we also have plans to release these acceleration implementations soon.