EVA2.0: Investigating Open-domain Chinese Dialogue Systems with Large-scale Pre-training

Yuxian Gu[1,2]*   Jiaxin Wen[1,2]*   Hao Sun[1,2]*   Yi Song[1,2]   Pei Ke[1,2]

Chujie Zheng[1,2]   Zheng Zhang[1,2]   Jianzhu Yao[2]   Lei Liu[3]

Xiaoyan Zhu[1,2]   Minlie Huang[1,2]

[1] The Conversational AI Group, Tsinghua University, Beijing 100084, China

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[3] Department of Electrical Engineering and Computer Science, York University, Toronto M3J1P3, Canada

Abstract: Large-scale pre-training has shown remarkable performance in building open-domain dialogue systems. However, previous works mainly focus on showing and evaluating the conversational performance of the released dialogue model, ignoring the discussion of some key factors towards a powerful human-like chatbot, especially in Chinese scenarios. In this paper, we conduct extensive experiments to investigate these under-explored factors, including data quality control, model architecture designs, training approaches, and decoding strategies. We propose EVA2.0, a large-scale pre-trained open-domain Chinese dialogue model with 2.8 billion parameters, and will make our models and codes publicly available. Automatic and human evaluations show that EVA2.0 significantly outperforms other open-source counterparts. We also discuss the limitations of this work by presenting some failure cases and pose some future research directions on large-scale Chinese open-domain dialogue systems.

Keywords: Natural language processing, deep learning (DL), large-scale pre-training, dialogue systems, Chinese open-domain conversational model.

1 Introduction

Recently, large-scale pre-training[1, 2] has become a mainstream approach to building open-domain dialogue systems, both in English[3–5] and Chinese[6–8]. These works mostly construct large dialogue corpora from social media platforms and then pre-train the model with generative objectives. Similar to the pre-trained models for general NLP tasks[9–11], pre-trained dialogue models acquire general conversational skills and versatile knowledge from large dialogue corpora during pre-training. Then, they can be easily fine-tuned to fit into various downstream dialogue scenarios, outperforming those trained from scratch.

However, building a powerful dialogue model is more than simply scaling up the model size and dialogue corpora. There are some other key factors that significantly

impact the final performance. Although Adiwardana et al.[4] and Roller et al.[5] explore some of them, including the pre-training tasks, decoding strategies, and evaluation metrics in English scenarios, some under-explored key factors remain, especially in Chinese. For example, many works report an overview of the pre-training data they use but do not provide the data collection, cleansing procedures, and quality control details. Another example is the decoding strategy. Existing works on Chinese pre-trained dialogue models generally focus on the model training phase but only give a rough analysis of different parameter settings of decoding. We argue that due to the intrinsic differences among languages, lessons about decoding strategies in English may not be directly transferred to Chinese scenarios.

Therefore, in this paper, we investigate how to build an open-domain Chinese dialogue system based on large-scale pre-training. We provide a detailed analysis of the pre-training corpus and conduct extensive experiments on model design, pre-training methods, and decoding strategies. First, we comprehensively analyze the quality of the largest Chinese open-domain dialogue corpus WDC-dialogue[6]. We find that the corpus suffers from severe problems of context-response relevance, language fluency, and domain diversity despite its large scale.

Second, we explore several variants of model architectures, pre-training approaches, and decoding strategies. We empirically find that these factors do have a non-trivial impact on the pre-trained model.

Putting all these together, we first design a pipeline for data cleansing and construct a 60 GB high-quality open-domain dialogue dataset for large-scale pre-training. Then, based on this dataset, we build EVA2.0, an open-domain dialogue model with 2.8 B parameters and two variants with 300 M and 970 M parameters. Fig. 1 shows an example of a conversation between a human and the 2.8 B EVA2.0 model. In both automatic and human evaluations, the 2.8 B EVA2.0 model significantly outperforms other open-source generative dialogue models. We notice that even the 300 M variant performs on par with the 2.8 B EVA1.0 model[6] in the automatic evaluation while requiring only 11% parameters. We also provide case studies to analyze the conversational ability of EVA2.0 from different aspects to shed light on future research directions of the large-scale pre-trained open-do-

main Chinese dialogue systems. Our work provides foundation models for the research on Chinese dialogue modeling, which we believe will significantly benefit the dialogue research community.

2 Related work

2.1 Large-scale pre-trained language models

In the past few years, pre-trained language models such as the GPT family[9, 12, 13], BERT[10], XLNet[14], and BART[15] have greatly promoted the progress of the NLP community. These models are commonly pre-trained on massive textual data with self-supervised learning objectives to capture general language features. Many recent works have shown that scaling up the model sizes and the pre-training corpora leads to dramatic improvement[16]. For example, RoBERTa[17] improves the performance of

Fig. 1   An example of the conversation between a human and the 2.8 B EVA2.0 model

BERT by simply increasing the training corpus size and optimizing the pre-training details. T5[11] scales up the model size to 11 billion parameters for the first time and shows superior performance on both language understanding and generation tasks. GPT3[13], with 175 billion parameters and pre-trained on 570 GB filtered data, has been proven effective under various few-shot and zero-shot scenarios.

Numerous large-scale pre-trained models have also emerged in Chinese. The CPM family[18, 19] pioneers the Chinese pre-trained models. PanGu-α[20] and Yuan 1.0[21] boost the power of Chinese language models by pushing the model sizes to 200 B and 245 B. Mengzi[22] instead builds a lightweight yet still powerful Chinese model and is computation-friendly for deployment. CPT[23] and ERNIE3.0[24] explore unified frameworks for language understanding and generation.

2.2 Pre-trained conversational models

Besides general language understanding and generation, conversational pre-training is receiving increasing attention. For instance, DialoGPT[3] and Meena[4] pre-train the models on massive English Reddit data to acquire open-domain conversational ability. BlenderBot[5] and LaMDA[25] establish better conversational skills and more attractive traits by fine-tuning the pre-trained models on high-quality crowdsourced datasets[26–28]. In addition, they present studies on different factors that affect the model, including the model sizes, pre-training, and decoding details to guide future works.

In Chinese, there are also dialogue models pre-trained on large-scale social media data. For example, CDial-GPT[29] adopts generative pre-training on the data collected from the Weibo Comments. PLATO[30] and PLATO-2[7] leverage the discrete latent variable and curriculum learning to improve generation diversity and quality. EVA1.0[6] and PLATO-XL[8] scale the model sizes up to 2.8 B and 11 B and show impressive conversational ability. However, most of these works do not involve many details of how a powerful dialogue model is built. In this paper, besides the final model evaluation, we also focus on the key recipes toward a large-scale pre-trained Chinese dialogue model.

3 Data

Data quality essentially influences the model performance in large-scale pre-training. In this section, we define several automatic metrics to comprehensively measure the quality of the dialogue corpus obtained from social media. We also design a data refinement pipeline to construct the high-quality pre-training data based on these metrics.

3.1 Data quality evaluation

We define the data quality scores in the following three aspects.

Relevance score. The relevance score between context and response is an essential metric that reflects the coherence of dialogues. We adopt both untrained and trained metrics to compute this score.

For the untrained metric, we treat the word-level overlap between the context and response as an aspect of relevance. We assign higher scores to data samples where the overlapped words are further apart in a session for the preference of the long-dependency property. Formally, the relevance score of the context (C) and response (R) is defined as

$$S_1 = \sum_{w_i \in C, w_j \in R} \text{dist}(w_i, w_j)^\tau \boldsymbol{I}(w_i = w_j) \qquad (1)$$

where $\text{dist}(w_i, w_j)$ means the index distance between the utterances containing w_i and w_j, τ is a hyper-parameter.

For the trained metric, we fine-tune a BERT$_{\text{BASE}}$ binary classifier[10] on the LCCC dataset[29] to recognize whether a response is appropriate for a given context. The log-probability of the "Appropriate" class can be treated as the relevance score:

$$S_2 = \log P(\text{"Appropriate"} \mid C, R). \qquad (2)$$

Compared to the untrained metric, the trained metric better estimates the semantic relevance.

Fluency score. We compute the probability of each utterance in the dialogue corpus using statistical models based on kenlm[1]. The mean fluency score of a dialogue session is defined as

$$S_3 = -\frac{1}{n} \sum_{i=1}^{n} \log P(w_1^i w_2^i \cdots w_{|u_i|}^i) \qquad (3)$$

where n is the utterance number of the session and $u_i = w_1^i w_2^i \cdots w_{|u_i|}^i$ is the i-th utterance.

Entertainment tendency. The Chinese social media platforms contain many undesired information exchanges about entertainment, which are uncommon in daily conversations. Therefore, we compute the ratio of dialogues involving Chinese stars to measure the entertainment tendency as an aspect of the domain diversity of the dataset.

3.2 Data refinement

Dataset-level filtering. We find that some dialogue datasets are not suitable for open-domain conversations. Training on them will result in undesired behaviors like the tone of the E-commerce customer service. Therefore, we remove these datasets like the JDDC[31].

Context-level filtering. Since our datasets are primarily from social media platforms, some contexts cor-

---

[1] https://github.com/kpu/kenlm

respond to a considerable number of responses (e.g., Weibo posts and their comments). These responses are very similar in format and can severely harm the performance of language models[32]. Therefore, we set a max response number for each context during filtering.

Rule-based filtering. We strengthen the rule-based filtering procedure in [6]. For example, we transform traditional Chinese characters into simplified ones and remove unreasonable multiple successive punctuation marks. Details of the rules we use can be found in Appendix A.2.

Classifier-based filtering. For each dialogue in the corpus, we compute the scores defined in Section 3.1 and filter out those samples whose scores are lower than a threshold. The overall score of a session is defined as $S = \alpha S_1 + \beta S_2 + \gamma S_3$. In practice, we empirically choose different thresholds for different data sources to fit their data distributions and make the final dataset balanced.

3.3 Data statistics

We construct our final pre-training data: EVA2.0-dataset from various sources using the above data refinement pipeline. The detailed data source information and the hyper-parameter values can be found in Appendix A. In Tables 1 and 2, we illustrate the basic statistics and quality evaluations of EVA2.0-dataset compared with the largest open-domain dialogue corpus: WDC-dialogue[6], using the metrics defined in Section 3.1. We can see that WDC-dialogue suffers from severe problems of context-response relevance, language fluency, and domain diversity. Although EVA2.0-dataset amounts to only less than a third of WDC-dialogue, its quality is significantly better, which verifies the effectiveness of our data refinement pipeline. Moreover, EVA2.0-dataset also contains dialogue sessions with more utterances, which resembles the daily multi-turn conversations better. In Section 5.3, we will see that despite its small amount, EVA2.0-dataset brings better model performance, owing to its high data quality.

4 Method

4.1 Model

We adopt a Transformer-based architecture combined with a bidirectional encoder and a unidirectional decoder for dialogue modeling. Different from EVA1.0[6] and T5[11], we add the $\sqrt{d}$ scale to the attention scores in the Transformer, which reduces the demand for careful initialization before pre-training. The dialogue histories are fed into the encoder as context, and the decoder generates the response in an auto-regressive manner based on the encoded context. We train models with three different sizes, whose configurations are shown in Table 3.

Table 1   Basic statistics of EVA2.0-dataset and WDC-dialogue[6]. "#Sess.", "#Uttr." and "#Token" indicate the total number of sessions, utterances, and tokens.

Dataset	#Sess.	#Uttr.	#Token	Storage
WDC-dialogue[6]	1.4 B	3.0 B	78.3 B	181 GB
EVA2.0-dataset	0.4 B	1.1 B	22.4 B	60 GB

Table 2   Quality evaluations of EVA2.0-dataset and WDC-dialogue[6]. "Relevance", "Fluency", and "Entertainment" indicate the relevance score, the fluency score and the entertainment tendency.

Dataset	Relevance ↑	Fluency ↑	Entertainment ↓
WDC-dialogue[6]	55.2	−7 147	7.0%
EVA2.0-dataset	93.8	−3 237	6.2%

Although similar model designs are also used in previous works, including both general pre-trained models[11, 18] and dialogue-specific pre-trained models[4−6], they do not include the discussions of some variants of the model that can have a non-trivial impact on the final performance. Therefore, we consider two aspects of the model configuration.

Layer numbers. Both BlenderBot and Meena adopt the encoder-decoder architecture to model dialogues. However, different from the models pre-trained on the long documents, which usually use balanced encoder and decoder layers[11, 15], these dialogue models use decoders much deeper than the encoders. Intuitively, a deeper decoder may be beneficial for generation tasks. However, a deeper encoder can better understand the dialogue histories in dialogue modeling, which improves relevance and consistency between the generated response and the dialogue context. Therefore, we try different layer ratios of the encoder and decoder while keeping the same parameter numbers.

Role information. Recent work[25] has pointed out that current pre-trained dialogue models can confuse their roles in long conversations because the model is pre-trained on approximated dialogues from social media. Therefore, it is intuitive to add the role information into the dialogue model to improve the role consistency. For example, PLATO-XL[8] introduces role embeddings to encode multi-party dialogues. However, the pre-training corpus of PLATO-XL contains the role identifiers by nature, while many dialogue corpus based on social media, such as WDC-dialogue and our EVA2.0-dataset, do not include this information. Although we can assume the data as two-party dialogues and add the pseudo role information to the input, it is unclear whether this approximation works. Therefore, we follow [29] to incorporate role identifier tokens and the role embeddings as role information to the model and test its effectiveness.

Table 3   Model information of EVA2.0 with different sizes. n_{params} is the parameter number. n_{enc}-layers and n_{dec}-layers are the numbers of layers of the encoder and decoder, respectively. d_{model} is the hidden state size. d_{ff} is the size of the feedforward layer. n_{heads} is the number of attention heads and d_{head} is the dimension of the attention head.

Model	n_{params}	n_{layers}	d_{model}	d_{ff}	n_{heads}	d_{head}
EVA2.0$_{\mathrm{Base}}$	300 M	12	768	3 072	12	64
EVA2.0$_{\mathrm{Large}}$	970 M	24	1 024	4 096	16	64
EVA2.0$_{\mathrm{xLarge}}$	2.8 B	24	2 048	5 120	32	64

4.2 Pre-training

We train our models with the sequence-to-sequence language modeling[33]. The maximum lengths of the context and the response are 128, and the models see 1.05 M tokens in a forward pass. We set the learning rate to 0.01, the batch size to 4 096, the warmup steps to 10 K, and use the Noam scheduler[34] to adjust the learning rate. To improve the training efficiency, we use the ZeRO-stage-1[35] from DeepSpeed[36] and follow [6] to pack as many samples as possible into a single sequence.

We study two pre-training approaches: pre-training from scratch on the dialogue corpus or further pre-training from a long-document pre-trained generative model. Intuitively, further pre-training yields better knowledge skills by inheriting versatile knowledge from long documents, which is scarce in social media. However, the distributions of the dialogue utterances and the document sentences differ significantly. It is unclear whether this difference causes catastrophic forgetting[37] or negative transfer[38] during further pre-training.

4.3 Decoding

We study various decoding strategies in this work. Although Roller et al.[5] have conducted experiments on the commonly used decoding approaches on English chatbots, we argue that the choice of decoding strategies is language-specific, and the conclusion can be different in Chinese.

Greedy search. Greedy search is a simple decoding strategy in which tokens are generated iteratively from left to right. A new token y_t is selected to maximize the probability conditioned on the previously generated tokens $y_{<t}$ and the dialogue history x, which is computed by the output logits h_t: $y_t = \arg\max_y P(y \mid y_{<t}, x)$, where $P(y \mid y_{<t}, x) = \mathrm{softmax}(h_t)$.

Sampling. Previous works find that greedy search can result in severe repetition and degradation in the generated texts. Therefore, sampling-based approaches are proposed to improve the generation quality by stochastic sampling from the word distribution: $y_t \sim P(y \mid x; y_{<t}) = \mathrm{softmax}(h_t/T)$, where T controls the sharpness of the distribution. In this work, we study an improved variant: Top-p Sampling[39], which filters out the low-probability tokens from the vocabulary and samples y_t from the re-

normalized probability.

Beam search. Beam search[40] is an extension to greedy search, which finds the most likely sentence from a larger searching space. It can also be coupled with the sampling approaches mentioned above to diversify the generation.

Length control. Vanilla beam search favors short generation over the long ones since a negative score is added at each step, leading generic responses. Therefore, we combine beam search with length control strategies. In Minimal length constraint, the probability of $\langle \mathrm{EOD}\rangle$[2] token is set to 0 until the generated response reaches a minimal length. In Length penalty, the score of each candidate in beam search is divided by l^α where l is the prefix length, and higher α encourages longer responses.

Handling repetitions. Repetition is a commonly observed phenomenon in current generative language models, which severely affects the generation quality. Hence, we consider the No-Repeat-N-Gram strategy, where we forbid the second time generation of any previously appeared n-grams in the generated prefix and the dialogue history.

5 Experiment

5.1 Evaluation setup

Datasets. We conduct response generation and self-chat experiments using automatic and human evaluation. For response generation, we adopt three test sets, Single, Multi, and KnowQA. Single and Multi contain single-turn and multi-turn dialogues collected from social media, which have no overlap with the pre-training data. KnowQA contains knowledge queries manually created from Chinese open-domain commonsense questions. For self-chat, we give the model a starting utterance and let it chat with itself to reach a maximum utterance number. The starting utterances are translated from the English self-chat query set used in [30]. The data statistics are shown in Table 4.

Metrics. We use uni-gram F1 (F1), ROUGE-L (R-L), BLEU-4 (B-4), and distinct 4-grams (D-4) for automatic evaluation. For human evaluation, we hire three volunteers to annotate each sample. On the test sets from social media (Single and Multi), we report sensibleness

---
2 We use the $\langle \mathrm{EOD}\rangle$ token to indicate the end of a sentence.

Table 4   Test dataset statistics. "Auto." / "Human" indicates the dataset is used for automatic / human evaluation. "#Sess." means the session number. "$\overline{\#Uttr}$" means the average utterance number per session. "$\overline{\#Token}$" means the average token number per utterance.

Test set	Evaluation	#Sess.	$\overline{\#Uttr}$	$\overline{\#Token}$
Single	Auto.	10 000	2.00	18.4
	Human	50	1.00	19.9
Multi	Auto.	10 000	4.17	15.3
	Human	50	3.38	15.0
KnowQA	Human	50	1.00	11.1
Self-Chat	Human	50	1.00	12.7

(Sensible.) and specificity (Specific.) used in [4]. We also add a consistency (Consist.) dimension to examine whether the model generates contradictory sentences. On KnowQA, we require the annotator to determine if the model's response matches the factual knowledge. For self-chat evaluation, apart from Sensibleness, Specificity, and Consistency, we also consider engagingness (Engaging.) as suggested in [41].

5.2 Strategies comparison

In this section, we compare different approaches described in Section 4. We use the mark ★ to highlight our final choice in each table.

Balanced VS. unbalanced layers. We compare models with different encoder and decoder layers. Specifically, we use the 300 M version of the model to save the computational cost. We test the balanced layers (12–12) and two unbalanced variants of our model: 18 encoder layers + 6 decoder layers (18–6) and 6 encoder layers + 18 decoder layers (6–18). From the results in Table 5, we conclude that the model with balanced layers performs the best in automatic evaluations because it balances the dialogue context understanding and response generation. Therefore, we adopt the balanced layers in the rest of our experiments.

Whether to add role information. We test the effect of the role information based on the 300 M model, and the results are shown in Table 5. Comparing the lines "12–12" and "+role", we can see that the role information hurts the model's performance. At first glance, this phenomenon seems to contradict that in [8] which claims that the additional role embeddings help the model to maintain the role consistency. However, in [8], the roles in the data are naturally distinguishable, which enables them to regard the utterances from social media as multi-party dialogues. In our data (and most data from public social media platforms), the role identifiers are not available. This forces us to roughly assume that the conversations are carried out between two characters. We argue this assumption introduces additional noise to the data

Table 5   Results of balanced/unbalanced layers and role information. "6–18" means 6 encoder layers and 18 decoder layers; "18–6" means 18 encoder layers and 6 decoder layers. "12–12" means balanced layers, which we finally adopt in EVA2.0$_{Base}$. "+role" means we add role information based on the balanced model.

Test Set	Model	F1	R-L	B-4	D-4
Single	6–18	15.6	13.3	1.48	49.4
	18–6	15.5	13.4	1.52	50.0
	12–12 ★	**16.2**	**13.8**	**1.63**	**53.4**
	+role	13.3	11.3	1.29	45.6
Multi	6–18	16.1	13.7	1.54	46.2
	18–6	16.2	13.9	1.43	45.6
	12–12 ★	**16.6**	**14.3**	**1.74**	**50.2**
	+role	14.4	12.0	1.31	42.3

and makes the optimization more difficult, which explains our results.

Training from scratch or not. We compare the model pre-trained from scratch on the dialogue data to the model further trained from CPM[19], a long-document pre-trained generative model. From the results in Tables 6 and 7, we observe that further pre-training outperforms pre-training from scratch in KnowQA but is worse in almost other evaluation metrics. This suggests that although further pre-training inherits the knowledge in CPM, it sacrifices basic conversational skills. Since we focus on building a chit-chat bot in this work, we choose to pre-train from scratch on the dialogue corpus for our final model.

Decoding approaches. We compare different decoding strategies in Table 8 for automatic evaluations and Table 9 for human evaluations. We incrementally combine other techniques with beam search to validate their influence. We combine no-repeat-n-gram with greedy search by default since simple greedy search often leads to repetition in the generated text. Through the automatic and human evaluations, we conclude that 1) no decoding strategy outperforms others consistently across all evaluation metrics; 2) sampling tends to diversify the responses but fails to maintain the sensibleness; 3) simple greedy decoding with the no-repeat-n-gram strategy yields surprising good performance in the human evaluation; 4) the model tends to generate self-contradict responses and hurt the consistency score in the human evaluation with the minimal length constraint, which is different from the English scenarios[5]; 5) when combined with sampling, repetition control, and length penalty, beam search has relatively balanced performance. As a result, we choose this as our final decoding strategy.

5.3 Final model evaluations

By putting the lessons learned from the previous ex-

Table 6   Automatic evaluation results of the pre-training approaches. "Scratch" represents pre-training from scratch on dialogue data. "Further" represents further pre-training from the CPM model.

Test set	Pre-training	F1	R-L	B-4	D-4
Single	Scratch ★	**17.0**	**14.9**	**2.23**	67.7
	Further	16.1	13.9	1.77	**68.2**
Multi	Scratch ★	**17.8**	**15.4**	**2.89**	**66.4**
	Further	16.6	14.3	1.84	59.7

Table 7   Human evaluation results of the pre-training approaches. "Scratch" and "Further" have the same meanings as in Table 6.

Pre-training	Sensible.	Specific.	KnowQA
Scratch ★	**0.76**	**0.70**	0.16
Further	0.74	0.62	**0.50**

Table 8   Automatic evaluation results of different decoding strategies. The score marked as **bold** means the best performance. The score marked with an underline means the second best performance.

Test Set	Decoding	F1	R-L	B-4	D-4
Single	Greedy	16.4	14.1	2.09	63.1
	Sampling	12.2	10.4	1.20	**91.6**
	Beam search	16.5	14.7	<u>2.80</u>	43.3
	+Sampling	16.3	14.5	2.21	<u>75.4</u>
	+Len_penalty	**17.4**	**15.4**	**3.23**	66.2
	+No-repeat ★	<u>17.0</u>	<u>14.9</u>	2.23	67.7
	+Min_len	16.4	14.2	2.04	62.3
Multi	Greedy	16.5	14.2	2.76	64.2
	Sampling	12.5	10.7	1.99	**91.5**
	Beam search	16.9	15.0	<u>3.50</u>	46.0
	+Sampling	16.4	14.6	2.59	<u>73.2</u>
	+Len_penalty	**17.8**	**15.7**	**3.79**	64.9
	+No-repeat ★	**17.8**	<u>15.4</u>	2.89	66.4
	+Min_len	17.1	14.9	2.47	62.8

Table 9   Human evaluation results of different decoding strategies. The score marked as **bold** means the best performance. The score marked with an underline means the second best performance.

Decoding	Sensible.	Specific.	Consist.
Greedy	**0.80**	<u>0.70</u>	0.96
Sampling	0.60	0.54	**1.00**
Beam search	0.66	0.60	0.94
+Sampling	0.72	0.68	<u>0.98</u>
+Len_penalty	0.70	0.66	0.96
+No-repeat ★	<u>0.76</u>	<u>0.70</u>	**1.00**
+Min_len	0.74	**0.74**	0.92

trained based on a Chinese GPT model on the LCCC dataset, which contains $12\,M$ dialogue sessions. It contains $95.5\,M$ parameters and is further pre-trained for 30 epochs. Compared with CDial-GPT, our model is much larger and is pre-trained on dialogue data from scratch to ensure its chit-chat ability.

2) EVA1.0[6], a $2.8\,B$ model with encoder-decoder architecture. It is pre-trained from scratch on the WDC-dialogue Corpus for $20\,K$ steps. It uses top-p sampling[39] strategy for decoding. Unlike EVA1.0, our model is trained on a dialogue corpus with a much better data-cleansing procedure. Besides, our model also adopts the most suitable decoding strategy found in Section 5.2.

Automatic evaluation. The results of the automatic evaluation are shown in Table 10. We can see that EVA2.0$_{xLarge}$ outperforms the baselines on both the relevance and diversity metrics. Note that although EVA2.0$_{Base}$ is nine times smaller than EVA1.0 and uses three times less data, it still performs comparably with EVA1.0, which highlights the importance of careful data refinement.

Observational human evaluation. The observational human evaluation results in Table 11 suggest that the generated response of EVA2.0$_{xLarge}$ is preferable to humans in terms of sensibleness, specificity, and consistency.

Self-chat human evaluation. Since human-model interactive evaluations are time-consuming and expensive, self-chat has been widely adopted to evaluate dialogue systems. Given a starting utterance, the model chats with itself for nine turns, and the generated sessions are assessed by the annotators. Results in Table 12 show that EVA2.0 consistently achieves the best performance in all the evaluated dimensions, producing the most coherent, informative, and engaging conversations.

5.4   Failure cases and future directions

Although EVA2.0 achieves good performance in both automatic and human evaluations, there is still room for improvement. We examine EVA2.0's limitations and elab-

periments together, we build the final EVA2.0 models whose configurations are shown in Table 3. We train the models on EVA2.0-Dataset from scratch without the role information. We use beam search + top-p sampling for decoding where beam_size = 4, top–p = 0.9, and $T = 0.9$. We set the length penalty to 1.6 and the no-repeat-n-gram to 4. In the following sections, we denote our $2.8\,B$ model as EVA2.0$_{xLarge}$, the $970\,M$ model as EVA2.0$_{Large}$, and the $300\,M$ model as EVA2.0$_{Base}$.

Baselines. We compare our model with all open-sourced Chinese dialogue models with large-scale pre-training:

1) CDial-GPT[29], a decoder-only model further pre-

Table 10   Automatic evaluation of EVA2.0 models and the baselines

Test set	Model	F1	R-L	B-4	D-4
Single	CDial-GPT	9.9	8.6	0.67	61.2
	EVA1.0	13.1	11.3	1.27	50.7
	EVA2.0$_{Base}$	16.2	13.8	1.63	53.4
	EVA2.0$_{Large}$	16.4	14.0	1.67	55.8
	EVA2.0$_{xLarge}$	**17.0**	**14.9**	**2.23**	**67.7**
Multi	CDial-GPT	11.9	10.3	0.88	63.9
	EVA1.0	15.3	13.2	1.94	56.3
	EVA2.0$_{Base}$	16.6	14.3	1.70	50.2
	EVA2.0$_{Large}$	17.2	15.1	2.27	58.9
	EVA2.0$_{xLarge}$	**17.8**	**15.4**	**2.90**	66.9

Table 11   Observational human evaluation results. We use the xLarge (2.8 B) version of EVA2.0.

Model	Sensible.	Specific.	Consist.
CDial-GPT	0.50	0.40	1.00
EVA1.0	0.68	0.60	0.96
EVA2.0	**0.76**	**0.70**	**1.00**

orate on three issues as follows. We present typical cases for each issue in Fig. 2.

Consistency. A common issue of dialogue models is that they tend to forget the information in the context

Table 12   Self-chat human evaluation results. We use the xLarge version (2.8 B) of EVA2.0.

Model	Sensible.	Specific.	Consist.	Engaging.
CDial-GPT	1.18	0.88	1.77	0.79
EVA1.0	1.21	1.11	1.82	1.00
EVA2.0	**1.71**	**1.55**	**1.89**	**1.27**

and occur inconsistency during the conversation. As shown in the first case in Fig. 2, our model occasionally contradicts itself. Although some works try to address the problem[42, 43], it is not yet entirely solved, especially in languages other than English.

Knowledge. Pre-trained dialogue models often generate responses containing factual error (the second case in Fig. 2). This is likely due to the knowledge-sparse training data obtained from social media. To tackle this problem, some works release knowledge-grounded dialogue datasets[27, 44] or augment the model with extra knowledge bases[45]. However, applying these approaches to Chinese scenarios is still not fully explored.

Safety. The real-world deployment of neural dialogue systems brings safety challenges. As discussed in [46], neural models tend to have various types of unsafe behaviors, such as social bias and the ignorance of suicide risk. EVA2.0 also sometimes generates socially prejudiced responses (the third example in Fig. 2). Detecting and limiting these behaviors is crucial for the practical application of neural dialogue models.
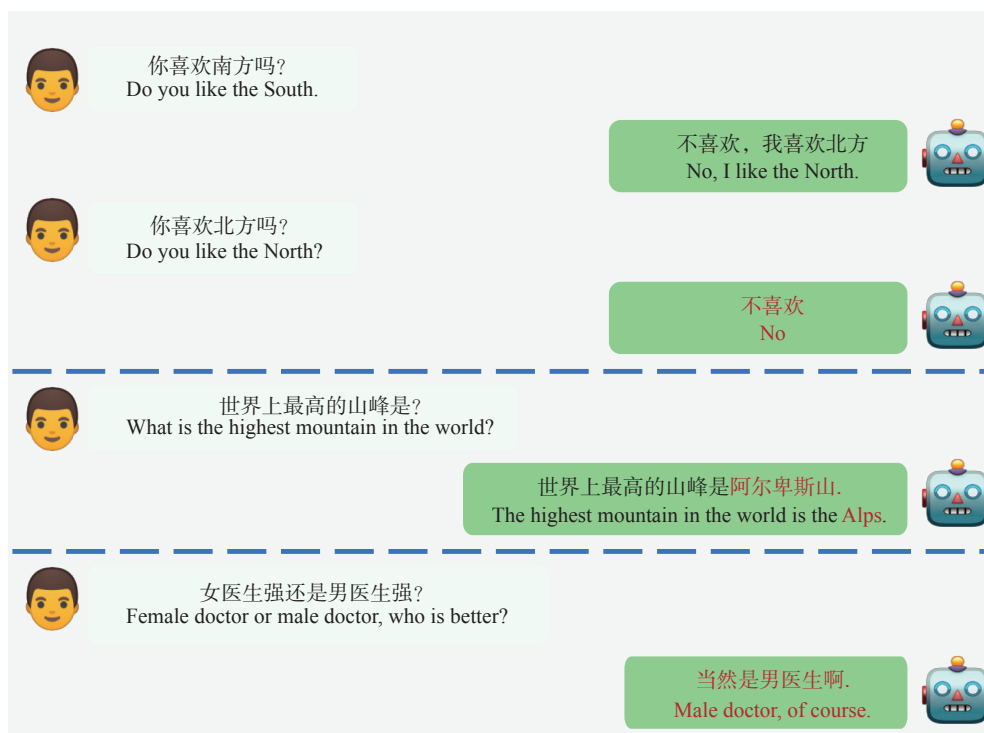

Fig. 2   Failure cases of EVA2.0$_{xLarge}$: Consistency (first case, deny liking the "North" after saying "I like the North"), knowledge (second case, the highest mountain in the world is Mount Everest), safety (third case, prefer male doctors).

6 Conclusions

This work investigates how to build a Chinese open-domain dialogue system via large-scale pre-training. We conduct extensive experiments to explore some critical factors of the model training and inference, including data quality control, model architectures, pre-training approaches, and decoding strategies. We share the insights in the experiments and build the EVA2.0 model, which significantly outperforms existing open-source baselines in both automatic and human evaluations. We also comprehensively analyze the failure cases of EVA2.0 and discuss some important future directions. Our work will facilitate future research and the application of open-domain Chinese dialogue systems.

Acknowledgements

Appendix A More data information

A.1 Data source

The data sources of EVA2.0-dataset consist of two parts: WDC-dialogue[6] and data from extra public sources: 1) Dialogues extracted from subtitles in the movies or TV plays[47]3; 2) Dialogues extracted from novels and stories[48]; 3) Zhidao Q&A pairs4; 4) LCCC Corpus[29]; 5) Existing crowdsourcing corpora including Du-Conv[49], KdConv[44], DuRecDial[50], and NaturalConv[51].

A.2 Data cleansing details

We first use rule-based methods to clean raw data from social media platforms like Weibo5 and Douban6. The rule-based methods are mostly based on the library clean-dialog7. The process includes: 1) removing the utterance containing words in our blacklist, which records illegal and vulgar words; 2) removing emoji like [smile]; 3) removing some special platform characters like #, which are largely applied in social media; 4) transforming tradi-

---

3 https://www.opensubtitles.org/

4 https://zhidao.baidu.com

5 https://weibo.com/

6 https://www.douban.com/

7 https://github.com/thu-coai/CDial-GPT

8 https://github.com/brightmart/roberta_zh

---

tional Chinese characters into simplified ones; 5) removing unreasonable multiple successive punctuation marks such as ",,,"; 6) removing some sensitive and private information like URLs, phone numbers, email addresses, QQ numbers, people name by regular expression; 7) deduplicating the utterances in one conversation and deduplicating the conversations. We also use context-level filtering described in Section 3.2 to control the max response number for each context. The number is set as 1 000.

After adopting rule-based cleansing, we apply classifier-based filtering to further process the conversational data. The cleansing module primarily comprises the relevance scorer described in Section 3.1. The relevance scorer is a binary classifier based on the Chinese RoBERTa$_{BASE}$ fine-tuned on LCCC dataset[29] to classify whether an utterance is appropriate for the context. We use the log-probability of the "Appropriate" class as the relevance score, and we use Chinese base version pre-trained model RoBERTa[17]8. Besides relevance scorer, we also compute perplexity and word overlap to get more coherent and relevant data described in Section 3.1.

A.3 Hyper-parameters in data cleansing

We set τ in (1) to 0.5 and [α, β, γ] in Section 3.2 (Classifier-based filtering) to [0.1, 0.8, 0.1].

Appendix B Training details

The 28 B version of EVA2.0 is trained on 64 NVIDIA V100 GPUs for 30 days. The 960 M and the 300 M variants are trained on 32 NVIDIA V100 GPUs for 30 days. We adopt the Adam optimizer[52] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and wirght_decay = 0.01. For hyper-parameter search, we fix the batch size as 4 096 and search for the learning rate in [1E–3, 5E–3, 1E–2, 5E–2] to choose the one yielding the minimal loss after 20 K pre-training steps.

References

[1] X. Han, Z. Y. Zhang, N. Ding, Y. X. Gu, X. Liu, Y. Q. Huo, J. Z. Qiu, Y. Yao, A. Zhang, L. Zhang, W. T. Han, M. L. Huang, Q. Jin, Y. Y. Lan, Y. Liu, Z. Y. Liu, Z. W. Lu, X. P. Qiu, R. H. Song, J. Tang, J. R. Wen, J. H. Yuan, W. X. Zhao, J. Zhu. Pre-trained models: Past, present and future. *AI Open*, vol. 2, pp. 225–250, 2021. DOI: 10.1016/j.aiopen.2021.08.002.

[2] T. X. Sun, X. Y. Liu, X. P. Qiu, X. J. Huang. Paradigm shift in natural language processing. *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022. DOI: 10.1007/s11633-022-1331-6.

[3] Y. Z. Zhang, S. Q. Sun, M. Galley, Y. C. Chen, C. Brockett, X. Gao, J. F. Gao, J. J. Liu, B. Dolan. 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*, pp. 270–278, 2020. DOI: 10.18653/v1/2020.acl-demos.30.

[4] D. Adiwardana, M. T. Luong, D. R. So, J. Hall, N. Fiedel,

R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. F. Lu, Q. V. Le. Towards a human-like open-domain chatbot. [Online], Available: https://arxiv.org/abs/2001.099 77, 2020.

[5] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. H. Liu, J. Xu, M. Ott, E. M. Smith, Y. L. Boureau, J. Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, 2021. DOI: 10.18653/v1/2021.eacl-main.24.

[6] H. Zhou, P. Ke, Z. Zhang, Y. X. Gu, Y. H. Zheng, C. J. Zheng, Y. D. Wang, C. H. Wu, H. Sun, X. C. Yang, B. S. Wen, X. Y. Zhu, M. Huang, J. Tang. EVA: An open-domain Chinese dialogue system with large-scale generative pre-training. [Online], Available: https://arxiv.org/abs/ 2108.01547, 2021.

[7] S. Q. Bao, H. He, F. Wang, H. Wu, H. F. Wang, W. Q. Wu, Z. Guo, Z. B. Liu, X. C. Xu. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Proceedings of the Findings of the Association for Computational Linguistics*, pp. 2513–2525, 2021. DOI: 10. 18653/v1/2021.findings-acl.222.

[8] S. Q. Bao, H. He, F. Wang, H. Wu, H. F. Wang, W. Q. Wu, Z. H. Wu, Z. Guo, H. Lu, X. X. Huang, X. Tian, X. C. Xu, Y. Z. Lin, Z. Y. Niu. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. [Online], Available: https://arxiv.org/abs/2109.09519, 2021.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. Language Models are Unsupervised Multitask Learners, OpenAI Technical Report, San Francisco, USA, [Online], Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, 2019.

[10] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, USA, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Q. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, vol. 21, no. 1, Article number 140, 2020.

[12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever. Improving Language Understanding By Generative Pre-training, OpenAI Technical Report, San Francisco, USA, [Online], Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article number 159, 2020.

[14] Z. L. Yang, Z. H. Dai, Y. M. Yang, J. Carbonell, R.

Salakhutdinov, Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 517, 2019.

[15] M. Lewis, Y. H. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020. DOI: 10. 18653/v1/2020.acl-main.703.

[16] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei. Scaling laws for neural language models. [Online], Available: https://arxiv.org/abs/2001.08361, 2020.

[17] Y. H. Liu, M. Ott, N. Goyal, J. F. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. [Online], Available: https://arxiv.org/abs/1907. 11692, 2019.

[18] Z. Y. Zhang, Y. X. Gu, X. Han, S. Q. Chen, C. J. Xiao, Z. B. Sun, Y. Yao, F. C. Qi, J. Guan, P. Ke, Y. Z. Cai, G. Y. Zeng, Z. X. Tan, Z. Y. Liu, M. L. Huang, W. T. Han, Y. Liu, X. Y. Zhu, M. S. Sun. CPM-2: Large-scale cost-effective pre-trained language models. *AI Open*, vol. 2, pp. 216–224, 2021. DOI: 10.1016/j.aiopen.2021.12.003.

[19] Z. Y. Zhang, X. Han, H. Zhou, P. Ke, Y. X. Gu, D. M. Ye, Y. J. Qin, Y. S. Su, H. Z. Ji, J. Guan, F. C. Qi, X. Z. Wang, Y. N. Zheng, G. Y. Zeng, H. Q. Cao, S. Q. Chen, D. X. Li, Z. B. Sun, Z. Y. Liu, M. L. Huang, W. T. Han, J. Tang, J. Z. Li, X. Y. Zhu, M. S. Sun. CPM: A large-scale generative Chinese pre-trained language model. *AI Open*, vol. 2, pp. 93–99, 2021. DOI: 10.1016/j.aiopen.2021.07.001.

[20] W. Zeng, X. Z. Ren, T. Su, H. Wang, Y. Liao, Z. W. Wang, X. Jiang, Z. Z. Yang, K. S. Wang, X. D. Zhang, C. Li, Z. Y. Gong, Y. F. Yao, X. J. Huang, J. Wang, J. F. Yu, Q. Guo, Y. Yu, Y. Zhang, J. Wang, H. T. Tao, D. S. Yan, Z. X. Yi, F. Peng, F. Q. Jiang, H. Zhang, L. F. Deng, Y. H. Zhang, Z. Lin, C. Zhang, S. J. Zhang, M. Y. Guo, S. Z. Gu, G. J. Fan, Y. W. Wang, X. F. Jin, Q. Liu, Y. H. Tian. Pangu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. [Online], Available: https://arxiv.org/abs/2104.12369, 2021.

[21] S. H. Wu, X. D. Zhao, T. Yu, R. G. Zhang, C. Shen, H. L. Liu, F. Li, H. Zhu, J. G. Luo, L. Xu, X. W. Zhang. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. [Online], Available: https://arxiv. org/abs/2110.04725, 2021.

[22] Z. S. Zhang, H. Q. Zhang, K. M. Chen, Y. H. Guo, J. Y. Hua, Y. L. Wang, M. Zhou. Mengzi: Towards lightweight yet ingenious pre-trained models for Chinese. [Online], Available: https://arxiv.org/abs/2110.06696, 2021.

[23] Y. F. Shao, Z. C. Geng, Y. T. Liu, J. Q. Dai, H. Yan, F. Yang, L. Zhe, H. J. Bao, X. P. Qiu. CPT: A pre-trained unbalanced transformer for both Chinese language understanding and generation. [Online], Available: https://arxiv.org/abs/2109.05729, 2021.

[24] S. H. Wang, Y. Sun, Y. Xiang, Z. H. Wu, S. Y. Ding, W. B. Gong, S. K. Feng, J. Y. Shang, Y. B. Zhao, C. Pang, J. X. Liu, X. Y. Chen, Y. X. Lu, W. X. Liu, X. Wang, Y. F. Bai, Q. L. Chen, L. Zhao, S. Y. Li, P. Sun, D. H. Yu, Y. J. Ma, H. Tian, H. Wu, T. Wu, W. Zeng, G. Li, W. Gao, H. F.

Wang. ERNIE 3.0 Titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. [Online], Available: https://arxiv.org/abs/2112.12731, 2021.

[25] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. G. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. P. Huang, M. Krikun, D. Lepikhin, J. Qin, D. H. Chen, Y. Z. Xu, Z. F. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Q. Zhou, C.C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, Q. Le. LaMDA: Language models for dialog applications. [Online], Available: https://arxiv.org/abs/2201.08239, 2022.

[26] S. Z. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston. 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 (Volume 1: Long Papers)*, Melbourne, Australia, pp. 2204–2213, 2018. DOI: 10.18653/v1/P18-1205.

[27] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[28] H. Rashkin, E. M. Smith, M. Li, Y. L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 5370–5381, 2019. DOI: 10.18653/v1/P19-1534.

[29] Y. Wang, P. Ke, Y. Zheng, K. Huang, Y. Jiang, X. Zhu, M. Huang. A large-scale Chinese short-text conversation dataset. In *Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing*, Springer, Zhengzhou, China, pp. 91–103, 2020. DOI: 10.1007/978-3-030-60450-9_8.

[30] S. Q. Bao, H. He, F. Wang, H. Wu, H. F. Wang. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 85–96, 2020. DOI: 10.18653/v1/2020.acl-main.9.

[31] M. Chen, R. X. Liu, L. Shen, S. Z. Yuan, J. Y. Zhou, Y. Z. Wu, X. D. He, B. W. Zhou. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 459–466, 2020.

[32] K. Lee, D. Ippolito, A. Nystrom, C. Y. Zhang, D. Eck, C. Callison-Burch, N. Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 8424–8445, 2022. DOI: 10.18653/v1/2022.acl-long.577.

[33] I. Sutskever, O. Vinyals, Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 3104–3112, 2014.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[35] S. Rajbhandari, J. Rasley, O. Ruwase, Y. X. He. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, Atlanta, USA, Article number 20, 2020. DOI: 10.1109/SC41405.2020.00024.

[36] J. Rasley, S. Rajbhandari, O. Ruwase, Y. X. He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020. DOI: 10.1145/3394486.3406703.

[37] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 13, pp. 3521–3526, 2017. DOI: 10.1073/pnas.1611835114.

[38] S. Panigrahi, A. Nanda, T. Swarnkar. A survey on transfer learning. In *Proceedings of International Conference on Innovative Computing and Communication 2019: Intelligent and Cloud Computing*, Springer, Singapore, pp. 781–789, 2021. DOI: 10.1007/978-981-15-5971-6_83.

[39] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

[40] A. Graves. Sequence transduction with recurrent neural networks. [Online], Available: https://arxiv.org/abs/1211.3711, 2012.

[41] M. Li, J. Weston, S. Roller. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. [Online], Available: https://arxiv.org/abs/1909.03087, 2019.

[42] Y. X. Nie, M. Williamson, M. Bansal, D. Kiela, J. Weston. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1699–1713, 2021. DOI: 10.18653/v1/2021.acl-long.134.

[43] M. Li, S. Roller, I. Kulikov, S. Welleck, Y. L. Boureau, K. Cho, J. Weston. Don′t say that! Making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, 2020. DOI: 10.18653/v1/2020.acl-main.428.

[44] H. Zhou, C. J. Zheng, K. L. Huang, M. L. Huang, X. Y. Zhu. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7098–7108, 2020. DOI: 10.18653/v1/2020.acl-main.635.

[45] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih, T. Rocktäschel, S. Riedel, D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Pro-*

cessing Systems, Vancouver, Canada, Article number 793, 2020.

[46] H. Sun, G. X. Xu, J. W. Deng, J. L. Cheng, C. J. Zheng, H. Zhou, N. Y. Peng, X. Y. Zhu, M. L. Huang. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Proceedings of the Findings of the Association for Computational Linguistics*, Dublin, Ireland, pp. 3906–3923, 2022. DOI: 10.18653/v1/2022.findings-acl.308.

[47] P. Lison, J. Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia, pp. 923–929, 2016.

[48] J. Guan, Z. E. Feng, Y. M. Chen, R. L. He, X. X. Mao, C. J. Fan, M. L. Huang. LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 434–451, 2022. DOI: 10.1162/tacl_a_00469.

[49] W. Q. Wu, Z. Guo, X. Y. Zhou, H. Wu, X. Y. Zhang, R. Z. Lian, H. F. Wang. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 3794–3804, 2019. DOI: 10.18653/v1/P19-1369.

[50] Z. M. Liu, H. F. Wang, Z. Y. Niu, H. Wu, W. X. Che, T. Liu. Towards conversational recommendation over multitype dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1036–1049, 2020. DOI: 10.18653/v1/2020.acl-main.98.

[51] X. Y. Wang, C. Li, J. Q. Zhao, D. Yu. NaturalConv: A Chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 14006–14014, 2021. DOI: 10.1609/aaai.v35i16.17649.

[52] D. P. Kingma, J. Ba. Adam: A method for stochastic optimization. [Online], Available: https://arxiv.org/abs/1412.6980, 2015.

Yuxian Gu received the B. Eng. degree in computer science and technology from Tsinghua University, China in 2021. Currently, he is a Ph. D. degree candidate in computer science and technology at Department of Computer Scence and Technology, Tsinghua University, China.

His research interests include natural language processing, pre-trained language models, and dialogue systems.

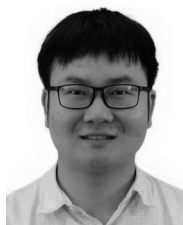E-mail: guyx21@mails.tsinghua.edu.cn

ORCID iD: 0000-0002-4607-7025

Jiaxin Wen received the B. Eng. degree in computer science and technology from Tsinghua University, China in 2022. He is currently a master student in computer science and technology at Department of Computer Science and Technology, Tsinghua University, China.

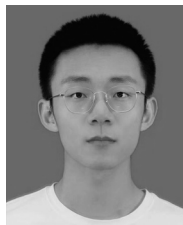His research interests include pre-trained language models and dialogue systems.

E-mail: wenjx22@mails.tsinghua.edu.cn

Hao Sun received the B. Eng. degree in computer science and technology from Shanghai Jiao Tong University, China in 2016. He is a master student in computer science and technology at Department of Computer Science and Technology, Tsinghua University, China.

His research interests include natural language generation and dialogue systems.

E-mail: h-sun20@mails.tsinghua.edu.cn

Yi Song received the B. Eng. degree in computer science and technology from Beijing Institute of Technology, China in 2021. He is currently a master student in computer science and technology at Department of Computer Science and Technology, Tsinghua University, China.

His research interests include natural language generation and dialogue systems.

E-mail: y-song21@mails.tsinghua.edu.cn

Pei Ke received the Ph. D. degree in computer science and technology from Tsinghua University, China in 2022. He is currently a postdoctoral researcher at Department of Computer Science and Technology, Tsinghua University, China.

His research interests include natural language generation, dialogue systems and sentiment analysis.

E-mail: kepei1106@outlook.com

Chujie Zheng received the B. Sc. degree in physics from Tsinghua University, China in 2020. He is a Ph. D. degree candidate in computer science and technology at Department of Computer Science and Technology, Tsinghua University, China.

His research interests include natural language generation and dialogue systems.

E-mail: zcj16@tsinghua.org.cn
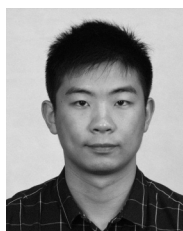
Zheng Zhang received the B. Eng. and Ph. D. degrees in computer science and technology from Department of Computer Science and Technology, Tsinghua University, China in 2015 and 2021, respectively. He is now a postdoctoral researcher at Tsinghua University, China.

His research interests include natural language processing, dialogue systems and text generation.

E-mail: zhangz.goal@gmail.com

Jianzhu Yao is an undergraduate student in computer science and technology at Department of Computer Science and Technology, Tsinghua University, China.

His research interests include natural language generation and dialogue systems.

E-mail: cnyaojz@gmail.com

Lei Liu received the M. Sc. degree in computer science from Central China Normal University, China in 2019. He is currently a Ph. D. candidate in the Graduate Program of Electrical Engineering and Computer Science, York University, Canada.

His research interests include dialogue systems and natural language generation.

E-mail: lliu@eecs.yorku.ca

Xiaoyan Zhu received the B. Sc. degree in computer science and technology from University of Science and Technology Beijing, China in 1982, the M. Sc. degree in computer science and technology from Kobe University, Japan in 1987, and the Ph. D. degree in computer science and technology from Nagoya Institute of Technology, Japan in 1990. She is currently a professor with Department of Computer Science and Technology, Tsinghua University, China. She has authored more than 100 peer-reviewed articles in leading international conferences (SIGKDD, IJCAI, AAAI, ACL) and journals (TOIS, *Bioinformatics, Genome Biology*).

Her research interests include intelligent information processing, machine learning, natural language processing, question and answering system and Bioinformatics.

E-mail: zxy-dcs@tsinghua.edu.cn

Minlie Huang received his Ph.D. degree in engineering of physics from Tsinghua University, China in 2006. He is currently an associate professor with Department of Computer Science and Technology, Tsinghua University, China. He has published more than 60 papers in premier conferences and journals (ACL, EMNLP, AAAI, IJCAI, WWW, SIGIR, etc.). His work on emotional chatting machines was reported by MIT Technology Review, the Guardian, and many other mass media. He serves as standing reviewer for TACL, area chairs for ACL 2020/2016, EMNLP 2019/2014/2011, and Senior PC members for AAAI 2017–2020 and IJCAI 2017–2020, and reviewers for TASLP, TKDE, TOIS, TPAMI, etc. He is a Nominee of ACL 2019 Best Demo Papers, the Recipient of IJCAI 2018 Distinguished Paper Award, CCL 2018 Best Demo Award, NLPCC 2015 Best Paper Award, Hanvon Yougth Innovation Award in 2018, and Wuwenjun AI Award in 2019. He was supported by a NSFC key project, several NSFC regular projects, and many IT companies.

His research interests include natural language processing, particularly in dialog systems, reading comprehension, and sentiment analysis.

E-mail: aihuang@tsinghua.edu.cn (Corresponding author)

ORCID iD: 0000-0001-7111-1849