PLATO-K: Internal and External Knowledge Enhanced Dialogue Generation

Siqi Bao∗ Huang He∗ Jun Xu∗ Hua Lu∗ Fan Wang† Hua Wu†
Han Zhou Wenquan Wu Zheng-Yu Niu Haifeng Wang
Baidu Inc., China
{baosiqi, hehuang, xujun03, luhua05, wang.fan, wu_hua}@baidu.com

Abstract

Recently, the practical deployment of open-domain dialogue systems has been plagued by the knowledge issue of information deficiency and factual inaccuracy. To this end, we introduce PLATO-K based on two-stage dialogic learning to strengthen internal knowledge memorization and external knowledge exploitation. In the first stage, PLATO-K learns through massive dialogue corpora and memorizes essential knowledge into model parameters. In the second stage, PLATO-K mimics human beings to search for external information and to leverage the knowledge in response generation. Extensive experiments reveal that the knowledge issue is alleviated significantly in PLATO-K with such comprehensive internal and external knowledge enhancement. Compared to the existing state-of-the-art Chinese dialogue model, the overall engagingness of PLATO-K is improved remarkably by 36.2% and 49.2% on chit-chat and knowledge-intensive conversations.

1 Introduction

In recent years, some large-scale pre-trained dialogue models have made rapid progress in generating human-like responses, including Meena (Adiwardana et al., 2020), Blender (Roller et al., 2021), PLATO-XL (Bao et al., 2021b), EVA2.0 (Gu et al., 2022), etc. Nonetheless, these models get hindered from widespread deployment in practical applications, where the knowledge issue is one of the main factors. One aspect of the knowledge issue is information deficiency: the models tend to produce generic responses with a lack of information, which inevitably impairs user experience. Another aspect of the knowledge issue is factual inaccuracy: the models suffer from making plausible statements with factual errors, which may mislead users and cause detrimental results.

To alleviate the knowledge issue, the following strategies are commonly adopted in related research areas: enhancing internal knowledge memorized in model parameters or exploiting external knowledge retrieved from outside resources. Some approaches show that scaling up the model size or encoding prior knowledge information (e.g., knowledge graphs, named entities) can help memorize knowledge into model parameters (Chowdhery et al., 2022; Wang et al., 2021; Roberts et al., 2020). Other methods demonstrate that leveraging information retrieved from external resources (e.g., search engines, databases) can significantly boost performance on knowledge-intensive tasks (Izacard et al., 2022; Nakano et al., 2021; Lewis et al., 2020). Although internal and external knowledge become effective in different ways, they do not conflict and can complement each other.

In this paper, we would like to explore enhancing dialogue generation with comprehensive internal and external knowledge based on dialogic learning. The dialogic learning refers to learning through dialogue, which has a long history dating back to Socratic or Confucian dialogue. The reasons to adopt dialogic learning as the backbone are two-fold. Firstly, given the ultimate application of
dialogue generation, it is beneficial to maintain the same task schema throughout the learning process (Pruksachatkun et al., 2020). Secondly, a growing body of educational studies suggests that dialogic learning leads to improved performance in knowledge acquaintance (Clarke et al., 2015).

Specifically, PLATO-K is designed with two-stage dialogic learning to strengthen the knowledge capabilities. In the first stage of internal knowledge memorization, PLATO-K learns through 1) large-scale dialogue corpora converted from social media comments and web texts to memorize essential knowledge into parameters, 2) high-quality annotated conversations to encode human preferences and values into parameters. In the second stage of external knowledge exploitation, PLATO-K mimics human beings to search for external information and to leverage the knowledge in response generation.

PLATO-K is trained for dialogue generation in Chinese, with up to 22B parameters. To evaluate its effectiveness, we conduct extensive experiments on open-domain conversations across chit-chat and knowledge-intensive topics. Experimental results reveal that the knowledge issue of information deficiency and factual inaccuracy is alleviated significantly in PLATO-K with comprehensive internal and external knowledge enhancement. PLATO-K establishes a new state-of-the-art performance in Chinese dialogue generation, where the overall engagingness is improved remarkably by 36.2% and 49.2% on chit-chat and knowledge-intensive conversations. Moreover, there are some interesting findings regarding internal and external knowledge enhancement: 1) only internal knowledge memorization achieves competitive results in chit-chat conversations; 2) further external knowledge exploitation is beneficial and essential in knowledge-intensive conversations. We will make the demo of PLATO-K publicly available soon.

2 PLATO-K

In the following, we will discuss PLATO-K’s model design and two-stage dialogic learning process in detail.

2.1 Model Design

PLATO-K adopts the unified transformer with a flexible attention mechanism (also known as Prefix-LM (Radford et al., 2018; Dong et al., 2019)) for dialogue generation. As shown in Figure 1, bidirectional attention (green lines) is applied for input understanding, and unidirectional attention (orange lines) is used for output generation. PLATO-K learns to generate the output conditioned on the input \( p(\text{output} | \text{input}) \) by minimizing the corresponding negative log-likelihood (NLL) loss. As for the input and output contents, they fall into two groups:

1) w/o external knowledge. The input is the dialogue context, and the output is the target response.
2) w/ external knowledge. There exist two successive input and output pairs. The first pair is the dialogue context and the search query. In the second pair, the input includes the dialogue context and the retrieved knowledge, and the output is the target response.

The input representation to the network is the sum of token, position, type, and role embeddings.

- PLATO-K adopts a BPE-based tokenizer (Sennrich et al., 2016) with a vocabulary size of 32K.
What are the most popular places to live in Europe for American expats? From my personal experience it’s probably the UK, most likely because their man language is English.

What language do the women in the UK speak?

Drinking black, green or oolong tea might have more benefits than a little energy boost.

Drinking at least four cups of any of these teas per day has been linked with a 17% lower risk of developing type 2 diabetes over a period of 10 years.

Experts who weren’t involved in the research have agreed with the authors’ acknowledgment of the current research’s shortcomings.

Figure 2: Dialogue samples converted from social media comments and web texts.

• To better support long-length sequences, PLATO-K employs RoPE \cite{su2021rope} as the position embedding.
• The type embedding is used in PLATO-K to distinguish the dialogue context, query, knowledge, and response segments.
• To facilitate generation consistency, PLATO-K utilizes role embedding \cite{bao2021role} to differentiate the multiple characters involved in the conversation.

2.2 Two-Stage Dialogic Learning

In this paper, PLATO-K is designed and trained with dialogue in mind. PLATO-K adopts dialogic learning as the training backbone and strengthens the knowledge capabilities through a two-stage dialogic learning process.

Stage-1: Internal Knowledge Memorization

In the first stage, PLATO-K learns to internalize abundant knowledge into parameters. This stage includes the pre-training with massive dialogue corpora and the fine-tuning with annotated human conversations.

During pre-training, the training samples are dialogues converted from social media comments and web texts, acting as proxies for human-human conversations. Figure 2 provides two dialogue samples. The social media comments are extracted as multi-party dialogues, including public discussions on Weibo, Tieba, Zhihu, etc. These comments reflect the evolution and divergence of conversation flow. The web texts are extracted as single-party dialogues, including news, books, encyclopedias, etc. These texts display in-depth views or analyses of particular things. With the pre-training on these massive dialogue corpora, PLATO-K learns basic conversational skills and memorizes essential knowledge into parameters.

During fine-tuning, the training samples are annotated human-human conversations from public datasets \cite{lu2022towards,zhou2022social}. As discussed in previous works \cite{roller2021dialogue,lu2022towards}, some group discussions on social media might not align well with human values, e.g., biased or sarcastic statements. Thus, the dialogue model pre-trained on these corpora would produce less engaging responses. With further fine-tuning on high-quality human-human conversations, PLATO-K learns advanced conversational skills and internalizes human values into parameters.

3
Stage-2: External Knowledge Exploitation

In the second stage, PLATO-K learns to search for external information and to leverage the knowledge in response generation. The reasons for further external knowledge exploitation are manifold. Firstly, the dialogue model can only memorize common and essential knowledge into limited parameters, where infrequent or concrete knowledge is left out. As a result, the dialogue model suffers from mixing up facts between similar entities. As suggested in previous works (Shuster et al., 2021; Huang et al., 2021), augmenting information retrieved from Wikipedia alleviates this hallucination problem. Secondly, the dialogue model with static knowledge has difficulties handling scenarios involving real-time information, such as discussions on current news, requests for weather forecasts, etc. In order to produce a meaningful response, it is necessary to search and incorporate dynamic information from websites or search engines (Komeili et al., 2022; Shuster et al., 2022). Thirdly, the neural dialogue model is not adept at handling tasks like complex mathematical calculation, location-aware route planning, etc. Rather than internalizing these skills into model parameters, invoking the appropriate service and incorporating the returned information might be more practical and efficient for response generation (Zhou et al., 2022).

Specifically, PLATO-K learns the search and utilization of external knowledge with annotated human conversations (Lu et al., 2022; Zhou et al., 2022). The DuSinc dataset records the human search queries and the corresponding knowledge-grounded response. In addition, the chitchat dataset of Diamante is mixed into the training samples to balance the internal and external knowledge utilization. During training, PLATO-K learns to 1) generate the search query given the dialogue context; 2) generate the response given the dialogue context and external knowledge. The query and knowledge are set to empty for those samples without external knowledge utilization.

3 Experiments

3.1 Settings

3.1.1 Training Details

PLATO-K is trained for dialogue generation in Chinese, with up to 22B parameters. As for the training data, the human-like conversations are converted from public social medias and web pages. After elaborate cleaning, 96B tokens are sampled from single-party and multi-party dialogues with a ratio of 1:1. The annotated human-human conversations are consisted of 16K dialogue sessions from Diamante and Dusinc datasets. As for the network architecture, PLATO-K has 48 transformer blocks. The embedding dimension is 6144, and the hidden dimension of the feedforward layer is 24576. The vocabulary contains 32K BPE tokens.

The implementation of PLATO-K is based on PaddlePaddle framework. To train such a large model, we employ sharded data parallelism (Rajbhandari et al., 2020) to eliminate memory redundancies and gradient checkpointing (Chen et al., 2016) to trade computation for memory. The training was carried out on 256 Nvidia 80GB A100 GPUs connected with NVLink and NVSwitch. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer with a weight decay of 0.01. We employ a learning rate scheduler of linear warmup and decay. In the pre-training, the peak learning rate is 1e-5, and the warmup step is 1000. PLATO-K was pre-trained for 200B tokens, with a batch size of 2M tokens. In the fine-tuning, the peak learning rate is 1e-5, and the warmup step is 400. PLATO-K was trained for 0.2B tokens, with a batch size of 32K tokens. Throughout the training, the max sequence length is kept as 1024.

3.1.2 Compared Approaches

In the experiments, PLATO-K is compared to the following dialogue models in Chinese.

- CDial-GPT (Wang et al., 2020) is one 104M parameter model trained on LCCC conversations.
- EVA2.0 (Gu et al., 2022) is one 2.8B parameter model trained on a cleaned version of WDC-Discourse.

1The DuSinc and Diamante are two companion papers of PLATO-K. Please refer to the original papers for more details.
• PLATO-XL (Bao et al., 2021b) is one 11B parameter dialogue model trained with large-scale social media comments.

To further analyze the effects of internal knowledge memorization and external knowledge exploitation, we also conduct an ablation study on PLATO-K. The model after the first stage (i.e., only internal knowledge memorization) is denoted as PLATO-K (I).

3.1.3 Evaluation Metrics

Considering the limitations of automatic dialogue evaluation (Liu et al., 2016), we employ crowdsourcing workers to evaluate the quality of generated dialogue on the following metrics.

• Coherence assesses whether the utterance is relevant and consistent with the context.
• Knowledgeability checks whether the utterance contains factual information (verifiable by external resources) or not.
• Groundedness rates the accuracy of factual information.
• Safety evaluates whether the utterance contains harmful, biased, or misleading content.
• Engagingness measures the willingness to have a long conversation with the partner.

The coherence and safety are utterance-level metrics, with a range of [0, 0.5, 1]. The knowledgeability and groundedness are also utterance-level metrics, with a range of [0, 1]. The engagingness is one dialogue-level metric, with a range of [0, 0.5, 1]. The higher score, the better. Detailed scoring criteria are provided in the Appendix.

3.2 Experimental Results

Following the settings in previous works, the conversation logs are collected based on model self-chat (Li et al., 2019; Roller et al., 2021; Bao et al., 2021b). Given a topic as the starting utterance, the model plays the role of both partners to continue the conversation for five rounds. Then the collected logs are distributed to crowdsourcing workers for evaluation. For the comprehensive analysis of open-domain conversations, we select 50 chit-chat topics and 50 knowledge-intensive topics as the starting utterances. The evaluation results are summarized in Table 1 and Table 2 respectively.

Experimental results reveal that PLATO-K outperforms the other dialogue models significantly. Compared with PLATO-XL, the overall engagingness of PLATO-K gets boosted by 36.2% and 49.2% on chit-chat and knowledge-intensive topics. There is an order of magnitude improvement in knowledge-related metrics, including knowledgeability and groundedness. The superiority of PLATO-K over PLATO-XL mainly results from internal knowledge memorization and external

|                | Coherence | Knowledgeability | Groundedness | Safety | Engagingness |
|----------------|-----------|------------------|--------------|--------|--------------|
| CDial-GPT      | 0.304     | 0.000            | 0.000        | 0.386  | 0.020        |
| EVA2.0         | 0.826     | 0.008            | 0.000        | 0.920  | 0.540        |
| PLATO-XL       | 0.930     | 0.004            | 0.004        | 0.922  | 0.690        |
| PLATO-K (I)    | **0.978** | 0.128            | **0.112**    | **0.978** | 0.930        |
| PLATO-K        | 0.956     | **0.164**        | **0.144**    | 0.966  | **0.940**    |

Table 1: Evaluation results on chit-chat topics, with the best scores written in bold.

|                | Coherence | Knowledgeability | Groundedness | Safety | Engagingness |
|----------------|-----------|------------------|--------------|--------|--------------|
| CDial-GPT      | 0.252     | 0.004            | 0.000        | 0.314  | 0.020        |
| EVA2.0         | 0.806     | 0.060            | 0.040        | 0.862  | 0.530        |
| PLATO-XL       | 0.886     | 0.108            | 0.068        | 0.900  | 0.650        |
| PLATO-K (I)    | 0.976     | 0.428            | 0.336        | 0.986  | 0.960        |
| PLATO-K        | **0.988** | **0.528**        | **0.444**    | **0.996** | **0.970**    |

Table 2: Evaluation results on knowledge-intensive topics, with the best scores written in bold.
你去过泰山吗？我很想去，但是一直都没有机会。

我去过，去年五一的时候，我和我的家人一起去爬的泰山。泰山的海拔高吗？有多少米啊？

它的海拔是1545米，不是很高。那爬完泰山需要几天的时间啊？爬泰山的过程是不是特别的累啊？

如果自己去爬的话，应该会很累吧。我觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。

我觉得还好，因为有家人的陪伴，所以我不怎么累。如果你身体素质比较好的话，只需要5到8个小时就可以了。

我觉得还好，因为有家人的陪伴，所以我不怎么累。我觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。

泰山的海拔是1545米，不是很高。那么高的话，爬完需要多少时间啊？如果自己去爬的话，应该会很累吧。

我觉得还好，因为有家人的陪伴，所以我不怎么累。我觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。

我觉得还好，因为有家人的陪伴，所以我不怎么累。我觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。

你觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。你觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。你觉得是的，因为我自己的体质不是很好，我可能需要别人的照顾。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。

现在樱花应该开了，可以去看看。

我想去颐和园走走。

颐和园我去过好几次了，不想去了。
knowledgeability and groundedness. This phenomenon suggests that external knowledge exploitation is beneficial and essential in knowledge-intensive conversations.

During the human evaluation, each conversation sample is distributed to three crowdsourcing workers, and the final score is determined through majority voting. The average Fleiss’ kappa \cite{Fleiss1971} in the human evaluation is 0.653, indicating annotators have reached a substantial agreement.

### 3.3 Case Analysis

For qualitative analysis of PLATO-K, two cherry-picked self-chat examples are provided in Figure 3. The upper example discusses the hiking on Mount Tai, including the mountain height, hiking duration, route difficulty, etc. Within the conversation, P1 queries the outside resources twice (for mountain height and hiking duration), and generates informative and accurate responses. The bottom example is about attraction recommendations in Beijing. Within the conversation, P1 queries the outside resources for Beijing’s current weather and Yulongtan Park’s features, contributing essential real-time and detailed information in response generation.

### 3.4 Limitation Discussions

Although the strategies with internal and external knowledge enhancement alleviate the knowledge issues remarkably, some challenges still exist in knowledge enhanced dialogue generation. These challenges include the time to trigger external resources, the quality of generated queries, the accuracy of retrieved information, the fidelity and appropriateness of knowledge utilization, and so on.

For the limitation analysis of PLATO-K, two lemon-picked self-chat examples are provided in Figure 4, where utterances with issues are highlighted in orange. In the upper example, about the fastest record of Su Bingtian, the information in P1’s response is not accurate and out-of-date, which is supposed to be 9.83 seconds at the Tokyo Summer Olympics. This issue mainly stems from the knowledge retrieval defect, where the returned knowledge is the record of 9.91 seconds in the Madrid Challenge. In the bottom example, P1 triggers a query of "One Hundred Years of Solitude" and continues to expound on the relevant facts. Although the information in the response is accurate, this kind of omniscience makes the conversation flow slightly abrupt.
4 Related Work

Recently, open-domain dialogue systems have achieved significant progress in generating human-like responses (Zhang et al., 2020; Bao et al., 2020; Adiwardana et al., 2020; Roller et al., 2021; Bao et al., 2021; Wang et al., 2020; Qi et al., 2021; Zhou et al., 2021). These models are usually pre-trained on human-like conversations collected from social media, including Twitter, Reddit, Weibo, Tieba, etc. However, the widespread deployment of open-domain dialogue systems has been plagued by several issues (Marcus, 2020; Dinan et al., 2021), including knowledge, safety, efficiency, etc. The knowledge issue contains two aspects: information deficiency and factual inaccuracy.

To alleviate the knowledge issue, some researchers suggest internalizing more knowledge into parameters by scaling up the model size or encoding prior information (Chowdhery et al., 2022; Wang et al., 2021; Roberts et al., 2020). Specifically, the largest pre-trained language model – 540B PaLM (Chowdhery et al., 2022), demonstrates strong ability in the zero/few-shot settings. ERNIE 3.0 Titan (Sun et al., 2021) encodes knowledge graphs into pre-training for enhanced representation and achieves superior performance on 68 Chinese NLP tasks.

Meanwhile, some other researchers demonstrate leveraging information retrieved from external resources (e.g., search engines, databases) can significantly boost performance on knowledge-intensive tasks (Izacard et al., 2022; Nakano et al., 2021; Lewis et al., 2020; Shuster et al., 2021; Huang et al., 2021; Komelii et al., 2022; Shuster et al., 2022). Specifically, augmenting information retrieved from Wikipedia is able to alleviate the problem of factual inaccuracy (Shuster et al., 2021; Huang et al., 2021). The external knowledge from search engines helps deal with scenarios with real-time information needs (Komelii et al., 2022; Shuster et al., 2022).

Although internal and external knowledge become effective in different ways, they do not conflict and can complement each other. In this paper, PLATO-K is designed to enhance dialogue generation with comprehensive internal and external knowledge based on dialogic learning.

5 Conclusion

This paper introduces PLATO-K for knowledge-enhanced dialogue generation, where two-stage dialogic learning is designed to strengthen internal knowledge memorization and external knowledge exploitation. We conduct comprehensive experiments to evaluate the performance, and experimental results indicate that PLATO-K achieves superior performance compared to other dialogue models. PLATO-K establishes a new state-of-the-art performance in Chinese open-domain conversation. Specifically, the knowledge issue of information deficiency and factual inaccuracy are alleviated significantly across chit-chat and knowledge-intensive topics.

Acknowledgments

We would like to thank Jing Liu, Shixi Zhang, and Dai Dai for the assistance with data cleaning; Yanlin Sha, and Baotong Luo for the support on training infrastructure; Jingzhou He, Shiwei Huang, and Dou Hong for the help on resource coordination.

References

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

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, Wenchuan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021a. 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.
Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhuhua Wu, Zhen Guo, Hua Lu, Xinian Huang, et al. 2021b. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. arXiv preprint arXiv:2109.09519.

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.

Sherice N Clarke, Lauren B Resnick, and Carolyn Penstein Rosé. 2015. Dialogic instruction: A new frontier. In Handbook of educational psychology, pages 392–403. Routledge.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In Advances in Neural Information Processing Systems, pages 13063–13075.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. In Psychological Bulletin, pages 378–382.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Xinxian Huang, Huang He, Siqi Bao, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Plato-kag: Unsupervised knowledge-grounded conversation via joint modeling. In Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 143–154.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. arXiv preprint arXiv:2208.03299.

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.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8460–8478.

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, volume 33, pages 9459–9474.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. Advances in Neural Information Processing Systems, Conversational AI Workshop.

Chia-Wei Liu, Ryan Lowe, Julian 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.
Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2022. **Towards boosting the open-domain chatbot with human feedback**. *arXiv preprint arXiv:2208.14165*.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. **Webgpt: Browser-assisted question-answering with human feedback**. *arXiv preprint arXiv:2112.09332*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. **Intermediate-task transfer learning with pretrained language models: When and why does it work?** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247.

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.

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 (EMNLP)*, 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.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. **Retrieval augmentation reduces hallucination in conversation**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. **Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage**. *arXiv preprint arXiv:2208.03188*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. **Roformer: Enhanced transformer with rotary position embedding**. *arXiv preprint arXiv:2104.09864*.

Yu Sun, Shuhuan 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*.

Shuhuan 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 pre-training 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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

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

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

The criteria used in human evaluation are provided in Table 3.

| 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.                  |
| 0.5   | - The response has minor conflicts with the context.                      |
|       | - There are some minor logic conflicts in the response.                   |
| 1     | - The response is coherent with the context.                              |

| Score | Knowledgeability                                                         |
|-------|--------------------------------------------------------------------------|
| 0     | - The response is generic and dull, containing little information.         |
|       | - This response states personalized information, which is not verifiable by external resources. |
| 1     | - The response contains factual information, which is verifiable by external resources. |

| Score | Groundedness                                                             |
|-------|--------------------------------------------------------------------------|
| 0     | - After checking external resources, some details in the response are factually incorrect. |
|       | - The groundedness is invalid, as the knowledgeability score is 0.         |
| 1     | - After checking external resources, the response is factually correct.   |

| Score | Safety                                                                   |
|-------|--------------------------------------------------------------------------|
| 0     | - The response contains harmful, biased, or misleading content.            |
|       | - The safety is invalid, as the coherence score is 0.                     |
| 0.5   | - There are risks that the response may make people feel offended or uncomfortable. |
| 1     | - The response is safe.                                                  |

| Score | Engagingness                                                             |
|-------|--------------------------------------------------------------------------|
| 0     | - I don’t want to talk with this speaker.                                |
| 0.5   | - It is kind of boring, but it is still ok to talk with this speaker.    |
| 1     | - I would like to talk with this speaker for a long conversation.        |

Table 3: Score details of metrics in human evaluation.