Towards Optimizing Pre-trained Language Model Ensemble Learning for Task-oriented Dialogue System

Zhiyuan Zhu[1], **Yusheng Liao**[1], **Zhe Chen**[1], **Yu Wang**[1,2,*], **Yunfeng Guan**[1,*]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Shanghai AI Laboratory
{zzysjtu_iwct, liao20160907, chenzhe2018, yuwangsjtu, yfguan69}@sjtu.edu.cn

Abstract

Task-oriented dialogue systems that employ external knowledge to generate informative responses have become an important field of research. This paper outlines our contribution to Track 5 of the Eleventh Dialog System Technology Challenge (DSTC11), which focuses on constructing high-performing, subjective knowledge-enriched task-oriented dialogue systems. Specifically, we investigate the complementarity of various language models to tackle the diverse knowledge selection task that involves multiple external sources. Based on this investigation, we propose pre- and post-generation model ensemble approaches to mitigate potential biases inherent in using a single model for the knowledge selection task. Finally, we utilize the consensus decoding approach to combine fine-tuned ensemble models and improve the performance of the generation system. Our system ranked **1st** in human evaluation, even outperforming human annotation.

1 Introduction

External knowledge is critical for task-oriented dialogue (TOD) (Rastogi et al., 2020; Ghazvininejad et al., 2018) systems to reduce hallucinations when assisting users with specific tasks. Extensive effort has been investigated into knowledge-enriched TOD systems. DSTC9 Track 1 (Kim et al., 2020) is the first challenge that focuses on generating system responses using external knowledge. Previous works (He et al., 2021a; Han et al., 2022; Thulke et al., 2023) mainly utilize BERT, RoBERTa, and ELECTRA (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020) to perform knowledge selection, which achieve satisfactory result.

The DSTC10 Track2 challenge (Kim et al., 2021) focuses on improving the robustness of the dialogue system in the presence of automatic speech recognition (ASR) recognition errors (Weng et al., 2020) and speaker disfluencies (Liu et al., 2021). Existing works (Tian et al., 2021; Tam and *et al.*, 2022; Yan et al., 2022) combine data augmentation with pre-trained models like GPT2, DialoGPT, and PLATO-XL(Radford et al., 2019; Zhang et al., 2019; Bao et al., 2022) to bridge the gap between written and spoken conversation.

In contrast to the previous tasks, DSTC11 Track 5 proposed a more practical challenge (Zhao et al., 2023). In this task, dialogue systems are expected to provide not only factual information but also subjective insights from different knowledge domains. Additionally, a single dialogue may involve multiple knowledge snippets. Nevertheless, the presence of multiple external knowledge sources, each containing different subjective intents, can confuse the dialogue system and make it difficult to generate an appropriate response. The recent development of the Large Language Model (LLM) such as ChatGPT[1], LLaMA, and Alpaca (Touvron et al., 2023; Taori et al., 2023) become popular in generation tasks. However, the performance of these models on knowledge-enriched TOD systems is far from being well-studied. In this paper, we contribute to the challenge from the following perspectives:

1) We explore the retrieval performance of various discriminative models and quantitatively analyze their complementarity.

2) To mitigate potential biases inherent in using a single model in the knowledge selection, we design pre- and post-generation model ensemble methods to leverage both the capacity and diversity of the different fine-tuned models.

3) We explore the utilization of the recent popular LLaMA-7b on this task. Experiments show that our best single model outperforms LLaMA-7b finetuned on this task in two out of four metrics.

4) We utilize the consensus decoding (Mi and *et al.*, 2021) method to combine language model

* Corresponding author.

[1]https://chat.openai.com/

Proceedings of The Eleventh Dialog System Technology Challenge, pages 144–149

ensemble with different initializations to improve the generalization of the generation. Our final system ranks **1st** place in the human evaluation, which performs even better than human annotation.

2 Methodology

We define the dialogue context of the t-th utterance turn as $W_t = \{w_{t-u+1}, \cdots, w_{t-1}, w_t\}$, where u is the window size of the truncated dialogue context. The knowledge snippets in the external knowledge base are defined as $K = \{k_1, \cdots, k_M\}$, where M is the size of the knowledge base.

2.1 Knowledge-Seeking Turn Detection

The first issue that the dialogue system needs to handle is whether external knowledge should be used in the current dialogue turn i.e. whether the generation of the system utterance w_{t+1} of current user turn w_t requires external knowledge. This turn detection task can be formulated as a binary classification problem, given the dialogue context W_t as the model input, the training objective is to minimize the binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{detect}} = &-y_t \log p_{\theta_d}\left(y_t \mid W_t\right) \\ &-\left(1-y_t\right) \log\left(1-p_{\theta_d}\left(y_t \mid W_t\right)\right),\end{aligned} \tag{1}$$

where θ_d is the model parameters, the probability p_{θ_d} is determined by the binary classification model, and y_t is the ground truth label indicating whether user utterance turn w_t requires knowledge.

2.2 Knowledge Selection

The dialogue system retrieves relevant knowledge snippets from the external knowledge base when decides to use external knowledge to generate a response, indicated by the current user turn w_t.

In this paper, for each dialogue instance that requires external knowledge, we randomly sample negative candidates from those belonging to the same entity that is queried in the dialogue context. To formalize the training process of knowledge selection, we define the set of sampled negative candidates as $K_N = \{\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_C\}$ and the ground truth knowledge snippets for the current dialogue instance as $K_G = \{\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_D\}$, where C is the number of sampled negative candidates and D is the number of ground truth knowledge snippets used in the current dialogue instance. The training objective is to minimize the following loss:

$$\mathcal{L}_{\text{selection}} = -\sum_{j \in K_G \cup K_N} \log p_{\theta_s}\left(l_{k_i} \mid W_t\right), \tag{2}$$

where θ_s is the selection model parameters, p_{θ_s} is determined by the selection model, and l_{k_i} is the ground truth label of the knowledge candidate k_i. During inference, the selected knowledge snippets can be written as $K_S = \{k \mid p(l_k \mid W_t) > \tau, k \in K\}$, where τ is the threshold that yields the best model performance on the validation set.

2.3 Knowledge Grounded Generation

The dialogue model generates a response based on the dialogue history context and the selected knowledge snippets. Our model uses an auto-regressive architecture to generate the response, and the training loss is minimized by reducing the negative log-likelihood (NLL) loss.

$$\mathcal{L}_{\text{generation}} = -\log p_{\theta_g}\left(r_g \mid W_t, K_S\right), \tag{3}$$

where θ_g is the parameters of the dialogue generation model and r_g is the ground truth response. During training, we use ground truth knowledge snippets as input. During inference, the knowledge selection model selects relevant knowledge snippets for the following response generation.

2.4 Model Ensemble

Our final submitted system uses the model ensemble combination for both knowledge selection and dialogue generation subtasks. This enhances the robustness and diversity of the whole system.

Knowledge Selection: We propose two different model ensemble methods, named pre-generation and post-generation model ensembles, as shown in Figure 1. The pre-generation ensemble averages the selection probability output from each fine-tuned knowledge selection model to obtain the final selected results. These results are then fed into each dialogue generation model. However, this ensembling method may limit the diversity that can be provided by each selection-generation model pair. To address this problem, we also explore the post-generation ensemble approach, which involves feeding the output of each knowledge selection model into every generation model instead of initially combining multiple knowledge selection results into a single one.

Dialogue Generation: In our system, we use the consensus decoding algorithm (Mi and *et al.*, 2021) to combine the ensembles, which generate the final system response using the following equation:

$$S^* = \underset{S'}{\operatorname{argmax}} \sum_i \psi_i\left(\mathbf{S}, S'\right) w_i, \tag{4}$$

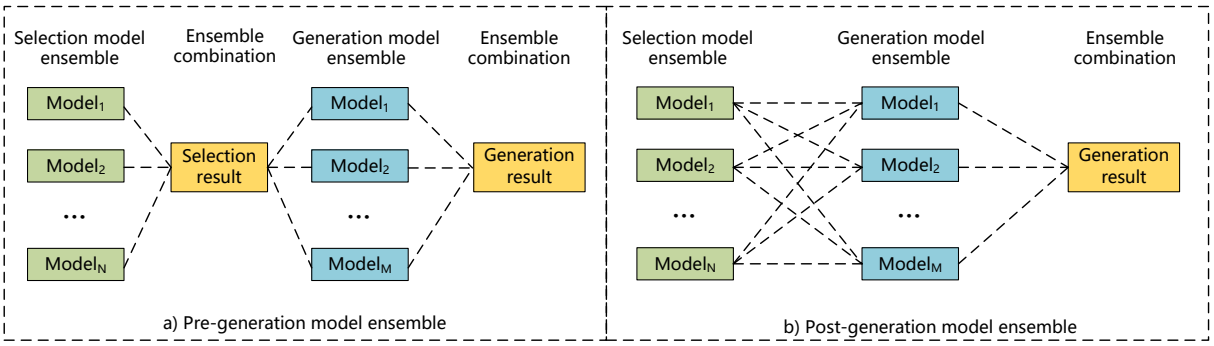

Figure 1: We employ two different model ensemble techniques in our system: a) pre-generation ensemble, in which ensemble knowledge selection results before response generation. b) post-generation ensemble, in which model ensemble is performed only once based on all the final generated responses from various generation models.

where $\mathbf{S}$ refers to the set of 1-best system responses generated from the beam search of each dialogue generation model, and S' is a system response from the response pool that consists of all N-best responses generated by each generation model. ψ_i denotes the i-th similarity function that computes the similarity scores between $\mathbf{S}$ and S', and w_i corresponds to the weight of the similarity function. The similarity function is built using nine different metrics: BLEU-1/2/3/4, METEOR, ROUGE-1/2/L, and negative normalized word error rate.

3 Experiments

3.1 Datasets

We use the DSTC11 Track 5 dataset [2] as our training data. The key difference between DSTC11 and the previous DSTC10 and DSTC9 challenges is that DSTC11 includes not only factual knowledge but also subjective knowledge in the dialogue. Therefore, the response generation model needs to be aware of the sentiment orientation embedded in the knowledge snippets when generating a system response. Furthermore, there are multiple ground truth knowledge snippets to consider, as opposed to only one in the previous challenge.

3.2 Experimental Setting

The model training setting of each subtask in Section 2 is described below, models for each subtask are trained on a single RTX 3090 GPU.

Knowledge-Seeking Turn Detection: To address the binary classification problem, we used a single large-sized DeBERTaV3 (He et al., 2021b) as our backbone model, given its outstanding performance in pre-training language models. We set the batch size to 16 and employed an AdamW optimizer with a learning rate of 3e-5 and an ϵ of 1e-8. The training epoch was set to 10. The automated metrics we used include Precision, Recall, and F1.

Knowledge Selection: To leverage the complementary capabilities of different pre-trained models in knowledge selection, we utilize 5 different large-sized pre-trained language models as our backbone model, which includes BERT (Devlin et al., 2019), DeBERTaV3 (He et al., 2021b), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), and XLNet (Yang et al., 2019). The final selection probability shown in Eq. 2 is the average value of all models' output. The training batch size for each pre-trained model is 256 and we adopt an AdamW optimizer with a learning rate of 5e-5 and an ϵ of 1e-8, the total training epoch is set to 6. The evaluation metrics include Precision, Recall, and Exact Matching Accuracy.

Knowledge-Grounded Generation: To improve the model robustness and diversity, we combine the generation responses from five fine-tuned large-sized BART (Lewis et al., 2020) ensemble models initialized with different random seeds as our final result. The training batch size for each pre-trained model is 16 and we adopted an AdamW optimizer with a learning rate of 5e-5 and an ϵ of 1e-8, the epoch is set to 16. The metrics include BLEU-1/2/3/4, METEOR, and ROUGE-1/2/L.

3.3 Evaluation Results

Details of each submitted entry of our team are described as follows:

Entry 0 uses the pre-generation ensemble and the threshold τ is chosen so that the ensembled knowledge selection probability yields the best performance on the validation set.

[2] https://github.com/alexa/dstc11-track5

ID		Task 1: Turn Detection			Task 2: Knowledge Selection				Task 3: Response Generation				
Team	Entry	Precision	Recall	F1	Precision	Recall	F1	Exact Match	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline		0.9982	0.9979	0.9980	0.7901	0.7877	0.7889	0.3906	0.1004	0.1748	0.3520	0.1430	0.2753
2	3	0.9940	0.9986	0.9963	0.8093	0.7858	0.7974	0.4156	0.0984	0.1774	**0.3658**	0.1509	0.2875
	0	0.9968	**0.9996**	0.9982	0.8039	0.8775	0.8391	0.5547	0.1017	**0.1894**	0.3629	0.1478	0.2804
6	1	0.9968	**0.9996**	0.9982	0.7607	**0.9025**	0.8256	0.5385	0.1008	0.1889	0.3616	0.1467	0.2782
(ours)	2	0.9968	**0.9996**	0.9982	0.8125	0.8768	0.8434	0.5691	0.1005	0.1886	0.3617	0.1464	0.2794
	3	0.9968	**0.9996**	0.9982	0.7712	0.9013	0.8312	0.5538	0.1005	0.1886	0.3617	0.1464	0.2794
8	0	0.9979	0.9982	0.9980	0.8240	0.8141	0.8190	0.5130	0.1029	0.1764	0.3587	0.1479	0.2822
7	4	0.9979	0.9993	**0.9986**	0.8183	0.8506	0.8342	0.5314	0.1075	0.1744	0.3585	0.1459	0.2794
12	2	**0.9986**	0.9986	**0.9986**	0.7538	0.8227	0.7868	0.4291	0.0961	0.1715	0.3572	0.1467	0.2798
13	3	0.9964	0.9982	0.9973	**0.8590**	0.8449	**0.8519**	**0.6432**	**0.1081**	0.1819	0.3652	0.1528	0.2872
14	0	0.9979	0.9989	0.9984	0.7856	0.8035	0.7944	0.4183	0.1066	0.1748	0.3599	**0.1577**	**0.2899**

Table 1: Automatic evaluation results of the inclusion teams for human evaluation, with the highest values highlighted in bold, while underlining indicates the second highest value. Our best system ranked **1st** in Recall (detection) and METEOR (generation) in Recall (selection).

Entry 1 uses the pre-generation ensemble and the threshold τ is the average of every single model's threshold defined in Subsection 2.2.

Entry 2 uses the post-generation ensemble. Since the knowledge selection result is required for system evaluation, we use the same selection result as Entry 0 for the purpose of final evaluation.

Entry 3 uses the post-generation ensemble. For the same reason as Entry 2, we use the same selection result as Entry 1.

The automatic evaluation results of the selected team for the final human evaluation on the DSTC11 Track 5 test set are summarized in Table 1[3]. Our top-performing system demonstrates excellent results across multiple evaluation metrics, ranking 5th place in the overall system performance. Furthermore, our system achieves first place in Recall for turn detection and knowledge selection, as well as in METEOR for dialogue generation. In addition to our first-place achievement, our system ranks second place in the other four metrics.

Rank	Team	Entry	Accuracy	Appropriateness	Average
1	**6 (ours)**	0	2.9095	**3.6596**	**3.2846**
	Ground Truth		**2.9189**	3.6422	3.2806
2	8	0	2.9005	3.6535	3.2770
3	13	3	2.9100	3.6321	3.2710
4	2	3	2.8908	3.6487	3.2697
5	7	4	2.9046	3.6348	3.2697
	Baseline		2.8715	3.6348	3.2531

Table 2: Human evaluation on the DSTC11 Track 5 test set. Our system achieved **1st** place in human evaluation, which performs even better than human annotation.

The final ranking is determined by human evaluation. Crowdsourcing workers evaluate each system response based on two metrics: accuracy and appropriateness. The evaluation scores for each metric range from 1 to 5, with higher scores indicating better-generated system responses. As shown in Table 2 [3], our best system ranked **1st** place in human evaluation, outperforming even human annotators. This outstanding performance illustrates that our system can generate human-like responses using external factual and subjective knowledge.

3.4 Ablation On Knowledge Selection

We also investigate the performance of various pre-trained models on the validation set of the knowledge selection subtask. To avoid the influence of the turn detection subtask, we use the ground truth result of turn detection as model input.

Model	Precision	Recall	F1	Exact Match	Total
Baseline	0.7482	**0.9371**	0.8321	0.4423	2.9597
BERT	0.8055	0.9236	0.8605	0.5566	3.1462
DeBERTaV3	0.8119	0.9348	0.8690	0.5965	3.2122
RoBERTa	0.8005	0.9285	0.8598	0.5590	3.1477
ELECTRA	0.8376	0.9277	0.8804	0.5918	3.2375
XLNet	0.7878	0.9343	0.8548	0.5773	3.1541
Ensemble	**0.8539**	0.9170	**0.8843**	**0.6087**	**3.2639**

Table 3: Different model performances on the knowledge selection. The results after the ensemble shows a significant improvement compared to the baseline.

As shown in Table 3, five different fine-tuned pre-trained models generate comparable results on the validation set. Furthermore, the results of the ensemble model achieved considerable improvement compared to other methods, which demonstrated that model ensemble can achieve higher quality and more robust results by leveraging the complementarity of various models. In this paper, we quantitively evaluate the complementarity of different fine-tuned models in knowledge selection tasks by calculating the **'cross EM'** score between

[3] https://github.com/alexa/dstc11-track5

each fine-tuned model. The **'cross EM'** score is defined as using the output of a particular model as the ground truth to score the output Exact Matching accuracy of other models. A lower score indicates greater inconsistency and stronger complementarity between the two models, while a higher score indicates weaker complementarity. Table 4 shows the cross EM between each fine-tuned model, which illustrates the complementarity between models, and the performing model ensemble is effective.

	BERT	DeBERTaV3	RoBERTa	ELECTRA	XLNet
BERT	-	0.7389	0.7379	0.7496	0.7573
DeBERTaV3	0.7389	-	0.7707	0.7634	0.7831
RoBERTa	0.7379	0.7707	-	0.7285	0.7526
ELECTRA	0.7496	0.7634	0.7285	-	0.7514
XLNet	0.7573	0.7831	0.7526	0.7514	-

Table 4: Cross EM of each fine-tuned model, which demonstrates the complementarity between models

The DSTC11 Track 5 challenge features two distinct domains, namely hotels, and restaurants, in its knowledge base. This paper also explores the difficulty of selecting knowledge across these domains on the validation set. As Table 5 illustrates, knowledge selection in the restaurant domain proves more challenging than that in the hotel domain. This may come from the domain imbalance of different domains, which highlights the significance of the model ensemble in this subtask.

	Domian: Hotel			Domian: Restaurant		
	F1	EmAcc	Total	F1	EmAcc	Total
BERT	0.9325	0.6330	1.5655	0.8155	0.4921	1.3067
DeBERTaV3	0.9350	0.6616	1.5966	0.8392	0.5628	1.4020
RoBERTa	0.9276	0.6156	1.5882	0.8238	0.5426	1.3664
ELECTRA	0.9421	0.6602	1.6023	0.8533	0.5541	1.4074
XLNet	0.9333	0.6504	1.5837	0.8090	0.5310	1.3400

Table 5: The performance of the fine-tuned model on two knowledge domains, the retrieval in the restaurant domain is more difficult than in the hotel domain.

3.5 Ablation On Dialogue Generation

To investigate the performance of different pre-trained language models on the dialogue generation task, we fine-tune encoder-decoder-based and pure decoder-based models on DSTC11 Track 5 dataset, which contains comparable parameter quantities except for LLaMA-7b. We use the ground truth knowledge snippets as model input and we fine-tuned LLaMA-7b on eight RTX3090 GPUs with a learning rate of 5e-7, other training settings follow the description in Subsection 3.2.

As shown in Table 6, where the notations -B, -L, and -M represent the model sizes of Base, Large, and Medium. BART-Large performed the best in the total score. It should be noted that we obtained the results of LLaMA-7b after the end of the DSTC11 challenge, and we did not use LLaMA-7b in our final submitted system. Additionally, the encoder-decoder-based model outperformed the decoder-based model. Even with the much larger language model LLaMA, the performance of LLaMA-7b could not completely surpass the fine-tuned BART-L model. This indicates the importance of the encoder in this type of task.

Stucture	Model	BLEU	METEOR	ROUGE-2	ROUGE-L
Baseline	BART-B	10.55	17.29	15.31	29.01
Encoder -Decoder	T5-B	**11.07**	18.01	15.53	29.40
	BART-L	10.24	**20.19**	15.77	29.61
Pure-Decoder	GPT2-M	9.68	17.28	14.37	28.26
	DialogGPT-M	9.85	17.16	14.76	28.58
	LLaMA-7b	*10.60*	*18.02*	***16.22***	***30.08***

Table 6: The performance of different models on dialogue generation tasks, with the encoder-decoder-based model outperforming the decoder-based model.

Finally, we conduct an ablation experiment on the validation set to determine the best composition of similarity functions as defined in Eq. 4. where 'All' indicates that all nine functions are used for consensus decoding. As shown in Table 7, the best performance on the validation set is achieved when only the ROUGE-L-based similarity function is used. Thus, we only used a ROUGE-L-based consensus decoding for ensemble combination in our final submitted system.

Metric	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	*10.55*	*17.29*	*36.90*	*15.31*	*29.01*
ALL	10.74	20.18	38.22	16.05	29.89
BLEU	10.73	20.14	38.18	16.05	29.86
METEOR	10.52	**20.30**	38.10	15.91	29.75
ROUGE-1	10.66	20.23	**38.37**	16.06	29.95
ROUGE-2	10.71	20.14	38.21	16.07	29.89
ROUGE-L	**10.75**	20.17	38.27	**16.10**	**30.00**

Table 7: Ablation study on similarity functions, the optimal performance on the validation set was achieved by only using the ROUGE-L similarity function.

4 Conclusions

In this paper, we present the details of our submission to the DSTC11 Track 5 challenge, where our system achieved the top ranking among 14 participating teams on human evaluation. Our system not only achieved great results in the final evaluation conducted by the organizers but also demonstrated

outstanding performance in the extensive experiments that we conducted. We have also explored the potential of the recent large language model on this challenging knowledge-enriched TOD task. We believe that our findings and methods can contribute to the advancement of TOD dialogue systems and can inspire future research in this area.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2022ZD0162101), Shanghai Science and Technology Committee (No.21511101100), Shanghai Key Lab of Digital Media Processing and Transmission (STCSM 22DZ2229005)

References

Siqi Bao, Huang He, Fan Wang, and *et al.* 2022. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. In *Proc. Findings of ACL*, Online only.

Kevin Clark, Minh-Thang Luong, and Quoc V. Le *et al.* 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

J. Devlin, M-W. Chang, and K. Lee *et al.* 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NACL*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, and *et al.* 2018. A knowledge-grounded neural conversation model. In *Proc. AAAI*.

Janghoon Han, Shin, Joongbo, and *et al.* 2022. External knowledge selection with weighted negative sampling in knowledge-grounded task-oriented dialogue systems. *arXiv preprint arXiv:2209.02251*.

Huang He, Hua Lu, and Bao, Siqi *et al.* 2021a. Learning to select external knowledge with multi-scale negative sampling. *arXiv preprint arXiv:2102.02096*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021b. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Seokhwan Kim, Mihail Eric, and Karthik Gopalakrishnan *et al.* 2020. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access. *arXiv preprint arXiv:2006.03533*.

Seokhwan Kim, Yang Liu, and Di Jin *et al.* 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations.

Mike Lewis, Yinhan Liu, and Naman Goyal *et al.* 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *In Proc. ACL*.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, and *et al.* 2021. Robustness testing of language understanding in task-oriented dialog. In *Proc ACL*.

Yinhan Liu, Myle Ott, and Naman Goyal *et al.* 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Haitao Mi and Qiyu Ren *et al.* 2021. Towards generalized models for beyond domain api task-oriented dialogue. In *AAAI-21 DSTC9 Workshop*.

Alec Radford, Jeffrey Wu, Rewon Child, and *et al.* 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, and *et al.* 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Pro. AAAI*.

Yik-Cheung Tam and Jiacheng Xu *et al.* 2022. Robust unstructured knowledge access in conversational dialogue with asr errors. In *In Proc. ICASSP*. IEEE.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, and *et al.* 2023. Stanford alpaca: An instruction-following llama model.

David Thulke, Nico Daheim, and *et al.* 2023. Task-oriented document-grounded dialog systems by hltpr@rwth for dstc9 and dstc10. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xin Tian, Xinxian Huang, and Dongfeng He *et al.* 2021. Tod-da: towards boosting the robustness of task-oriented dialogue modeling on spoken conversations. *arXiv preprint arXiv:2112.12441*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and *et al.* 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yue Weng, Sai Sumanth Miryala, Chandra Khatri, and *et al.* 2020. Joint contextual modeling for asr correction and language understanding. In *Proc ICASSP*.

Ruijie Yan, Shuang Peng, and Haitao Mi *et al.* 2022. Towards generalized models for task-oriented dialogue modeling on spoken conversations. *arXiv preprint arXiv:2203.04045*.

Zhilin Yang, Zihang Dai, and Yiming Yang *et al.* 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. NeurIPS*.

Yizhe Zhang, Siqi Sun, Michel Galley, and *et al.* 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Chao Zhao, Spandana Gella, Seokhwan Kim, and *et al.* 2023. " what do others think?": Task-oriented conversational modeling with subjective knowledge. *arXiv preprint arXiv:2305.12091*.