Data-Augmented Task-Oriented Dialogue Response Generation with Domain Adaptation

Yan Pan
Technical University of Munich
Frank_panyan@outlook.com

Davide Cadamuro
BMW Group
davide.cadamuro@bmw.de

Georg Groh
Technical University of Munich
grohg@in.tum.de

Abstract

Knowledge-grounded task-oriented dialogue response generation is crucial in helping people solve particular tasks in narrow domains. Instruction-tuned extra-large language models (IXLLMs), like ChatGPT, have shown powerful capabilities in the few-shot setting. However, the required task-related domain knowledge and appropriate style make it challenging for these models to generate suitable task-oriented dialogue responses. Compared to IXLLMs, fine-tuned large language models (FLLMs), like GPT2, learn task-related knowledge from source and target domains. Nevertheless, the size-limited datasets constrain the performance of FLLMs. To overcome the defects of these models by leveraging their respective strengths, we present a novel data-augmented hybrid system with domain adaptation (DAHDA) for task-oriented dialogue response generation. The hybrid system consists of an in-context learning IXLLM and an FLLM. We utilize the IXLLM to synthesize knowledge-grounded dialogues in the target domain, whereas the FLLM captures the task-related knowledge from source and target domains. The synthetic dialogues and task-related knowledge learned by the FLLM support the IXLLM to generate factually-accurate and suitable responses in the target domain.

1 Introduction

The instruction-tuned extra-large language models (IXLLMs) which fluently communicate with humans, like ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023), have caught the attention of the AI research community, as well as the general public. IXLLMs offer a dialogue interaction that is pleasant to users. Given this acceptance, there is a great opportunity to enhance task-oriented dialogue systems with the IXLLM in specific narrow domains (Hudeček and Dušek, 2023), like booking a meeting or searching for a hotel.

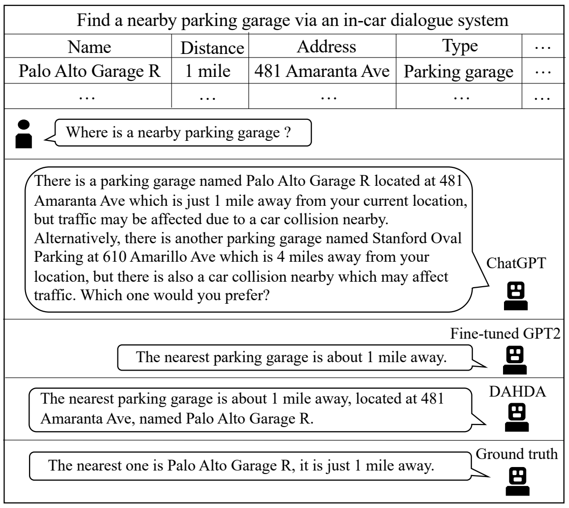

Figure 1: A dialogue example to find a nearby parking garage with an in-car assistant from the navigation domain in KVRET (Eric et al., 2017). Fine-tuned GPT2 and DAHDA are trained based on low-resource task-related weather and schedule source domain datasets.

In previous studies (Budzianowski et al., 2018; Rastogi et al., 2020), task-oriented dialogue systems used relevant domain specific knowledge and task-related schema with ground truth to define the suitable style and facts used in the dialogue response generation task. These ensure the quality of responses in narrow domains. However, recent works show that the IXLLM suffers hallucination problems, for example in the medical and movie domains (Alkaissi and McFarlane, 2023; Bang et al., 2023), and tends to generate inaccurate facts beyond the given domain knowledge (Bang et al., 2023). An example is shown in Figure 1, where ChatGPT ignores the user's request for nearness and gives a less concise response than the ground truth, which represents the suitable style for an in-car assistant. A suitable style indicates the appropriate length, language style, and task-oriented format for an in-car assistant. ChatGPT is not open-sourced, and it works as a black box (Brockman et al., 2020). Learning how to enable the black-box IXLLMs to generate task-oriented dialogue responses with solid task-related domain knowledge and suitable style is the purpose of this study.

In comparison to the IXLLM, smaller language models like GPT2 (Radford et al., 2019) are conveniently fine-tuned to be very proficient in a given dialogue task, as long as enough training data exists. We call these task-specific models fine-tuned large language models (FLLMs). However, the dataset in the target domain is not always publicly available. Moreover, the task-related source domain datasets are sometimes limited in size due to expensive data collection and privacy policies. As exemplified in Figure 1, a fine-tuned GPT2 model with low-resource source domain datasets gives a factually-correct but not suitable response, since the name of the parking garage is missing.

To leverage the strength of both models, especially with low-resource source domains in the zero-shot target domain setting, we propose a data-augmented hybrid system with domain adaptation (DAHDA), which consists of both the in-context learning IXLLM and the FLLM. As exemplified in Figure 1, our proposed DAHDA gives a more appropriate and accurate response in the target domain. To implement DAHDA, we propose a data augmentation technique which makes use of an IXLLM and dialogue paths. A dialogue path is a schematic representation of the utterances that take place in the dialogue, with a focus on informational entities (Moon et al., 2019; Yang et al., 2020; Mehri et al., 2022). With these augmented datasets, we train an FLLM to learn the task-related knowledge and appropriate style. We add the generated response from FLLM as in-context information to support IXLLM to generate suitable and factually-accurate responses in the target domain.

To the best of our knowledge, our paper is the first attempt to explore the combination of an IXLLM and an FLLM for task-oriented dialogue response generation with domain adaptation in a zero-shot setting. In this work, our main contributions are threefold:

- We propose the data-augmented hybrid system with domain adaptation to generate suitable and factually-accurate task-oriented dialogue responses in a zero-shot setting. The hybrid system consists of both FLLM and IXLLM with in-context learning, where the combination captures the benefits and features of each model. The results on two benchmarks, MWOZ (Budzianowski et al., 2018; Eric et al., 2020) and KVRET (Eric et al., 2017), show that our hybrid system displays improvement over the original IXLLM and overall outperforms other domain adaptation strategies with low-resource source domain datasets.
- The designed data augmentation method based on the dialogue path improves the performance of task-oriented dialogue response generation. Specifically, our data-augmented FLLM with rich-resource source domains achieves competitive results.
- The FLLM learned task-related knowledge from the source domain and synthetic target domain datasets. This learned knowledge improves the performance of the IXLLM with in-context learning.

2 Related Work

2.1 Low-resource domain adaptation

Data sparsity is a common problem that makes applying NLP models in the target domain difficult. Although datasets from task-related source domains are available in the training set, the target domain test dataset is different from the source domain training datasets (Gretton et al., 2006; Ramponi and Plank, 2020). This difference leads to worse performance of the model with low-resource source domains in the zero-shot target domain setting. Domain adaptation solves this problem by aiming at generalizing the model into any test samples in the unseen domain (Ramponi and Plank, 2020). One example of domain adaptation is to use new domain descriptions with domain-specific features to generalize a generative dialogue system (Zhao and Eskenazi, 2018). A different approach, gradient-based meta-learning, improves the performance of dialogue response generation in the poor-resource target domains by learning general features from different source domains (Qian and Yu, 2019). However, these methods are not competitive when compared to the given advances brought by recent IXLLMs (OpenAI, 2022, 2023). In comparison to the aforementioned model-centric methods, data-centric methods attract more attention in the research community due to the success of pre-trained models (Ramponi and Plank, 2020; Dai et al., 2023; Mehri et al., 2022). The synthetic

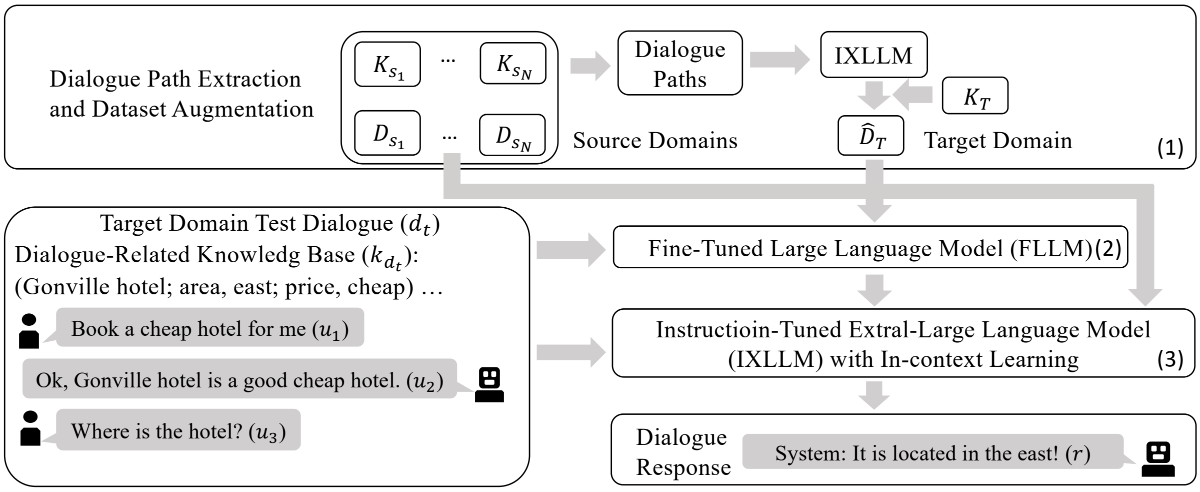

Figure 2: Overall framework.

dataset in the target domain, which is based on a data-centric method, has been proven successful in the opinion mining task (Li et al., 2022). Another study shows that domain-related or task-related datasets benefit the domain adaptation of large language models with adaptive pre-training and fine-tuning (Ramponi and Plank, 2020; Gururangan et al., 2020). However, it is challenging to apply adaptive pre-training and fine-tuning on the recently released instruction-based extra-large language models, which contain hundreds of billions of parameters. In the particular case of ChatGPT and GPT4, the fine tuning is difficult since the models are not open source and they have to be used as black boxes (Hu et al., 2022a; Brockman et al., 2020). How to achieve domain adaptation with these black-box instruction-based extra-large language models is still an open question.

2.2 Large language model

Instruction-tuned extra-large language models (IXLLMs) have shown powerful capabilities in many NLP-related tasks (OpenAI, 2022). In machine translation and question-answering tasks, IXLLMs achieve competitive performance (Bang et al., 2023; Brown et al., 2020). IXLLM performs better with the in-context learning method, in which the instruction and examples in the prompt support IXLLM in generating appropriate texts (Brown et al., 2020; Radford et al., 2019). However, IXLLMs have the problematic tendency to generate toxic or biased text, or make up hallucinations (Bang et al., 2023; Gehman et al., 2020; Tamkin et al., 2021). Using prompts to instruct the LLM and provide direct feedback mitigated the problems (Ouyang et al., 2022), but did not solve them completely (Bang et al., 2023), making it difficult to use the IXLLM directly in task-oriented dialogue systems.

3 Methods

3.1 Problem definition

We formulate the task of knowledge-grounded dialogue response generation with domain adaptation in the zero-shot setting. Given different source domains $\{S_1, ..., S_N\}$, each S_i contains dialogue sessions D_{S_i} and the knowledge base K_{S_i}, whereas the target domain T provides domain-related knowledge base K_T without dialogue sessions. Each dialogue session d from D_{S_i} contains $2L$ utterances $\{(u_1, u_2), ..., (u_{2L-1}, u_{2L})\}$ and the dialogue-related knowledge base k_d. The exchange-level utterances (u_{2l-1}, u_{2l}) represent the communication between a user and the system. Our goal is to use the given dialogue sessions $\{D_{S_1}, ..., D_{S_N}\}$ and knowledge bases $\{K_{S_1}, ..., K_{S_N}, K_T\}$ to build a system that is able to generate a correct and appropriate dialogue response r based on the dialogue history $U_{l_t} = \{u_1, u_2, ..., u_{2l_t-1}\}$ and the dialogue-related knowledge base k_{d_t} from the target domain test dialogue d_t.

3.2 Framework

In the following subsections, we introduce how to generate dialogue responses on the target domain using domain adaptation. As shown in Figure 2, our proposed framework contains (1) the dialogue path extraction and the dataset augmentation with IXLLM on the target domain, (2) the FLLM based on task-related source domain datasets and synthetic target domain dataset, (3) the in-context learning IXLLM with the generated response from the FLLM as well as the retrieved example from the task-related source domain datasets and from the synthetic target domain dataset.

3.3 Dialogue path extraction and dataset augmentation

The dialogue path, which contains the important entities and the entity categories from the domain knowledge base, is a concise representation of a task-oriented dialogue (Moon et al., 2019; Yang et al., 2020; Mehri et al., 2022). As an example in Figure 3, we extract a dialogue path p = [(Moderate, Price0)- (Pizza Express, Name0)-(Pizza Express, Name0)- (West, Area0)] from the source domain dialogue d about finding a restaurant, where (Pizza Express, Name0) means that Pizza Express is the first name entity. In order to generate a similar but diverse dialogue, which is considered as the augmented dataset on the target hotel domain, the dialogue path p is fed to IXLLM with the instruction for data-augmentation and the source dialogue d as an example. After feeding the prompt, the IXLLM generates a target domain dialogue template based on this information. DAHDA randomly chooses the new entities from the same entity categories in the target domain knowledge base K_T. Finally, the placeholders are replaced with the new entities and get the synthetic dialogue $\hat{d}$. Based on the source domain dialogues, this technique generates new synthetic dialogues $\hat{D}_T$ on the target domain.

3.4 Fine-tuned large language model

After generating synthetic dialogues $\hat{D}_T$ on the target domain, the large language model is fine-tuned with source domains and synthetic target domain datasets $D = \{D_{S_1}, ..., D_{S_N}, \hat{D}_T\}$ to generate appropriate dialogue response r_{FLLM} based on the knowledge base k_{d_t} and dialogue history $U_{l_t} = \{u_1, u_2, ..., u_{2l_t-1}\}$. The goal of fine-tuning for domain adaptation is to improve the performance of the large language model on the target domain by learning task-related knowledge from the mixed datasets D.

3.5 In-context learning IXLLM

In our DAHDA, the FLLM trained with the mixed datasets D is able to generate a dialogue response

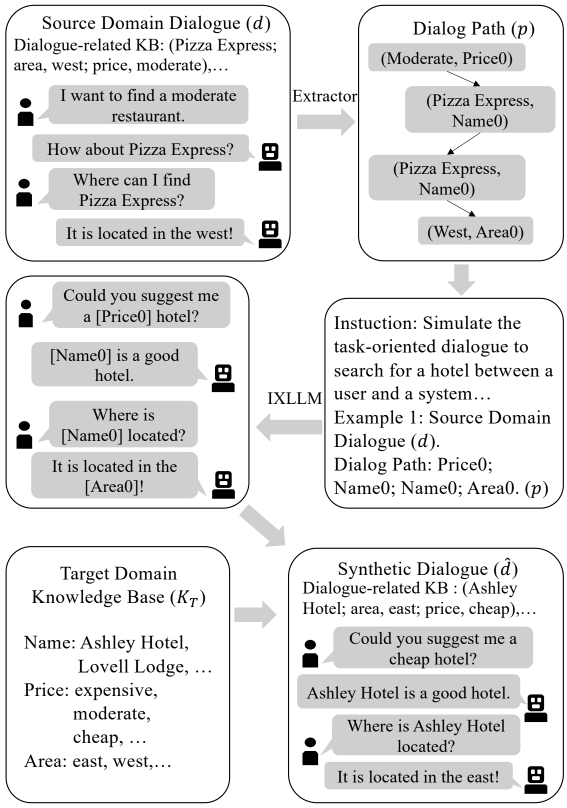

Figure 3: Data augmentation with the dialogue path from source restaurant domain to target hotel domain.

r_{FLLM} for the target domain test dialogue d_t as shown in Figure 4. This dialogue response is added into the prompt for IXLLM because it contains the task-related knowledge and language style that FLLM learns from the mixed source and synthetic target domain datasets. Additionally, a retriever searches the most similar dialogue $d_{t,ins}$ from the mixed datasets D, providing to the in-context learning IXLLM an in-context example for generating a response, as shown in Figure 4. For a test dialogue, the retriever maps its knowledge base and dialogue history into the query input $q_{d_t} = [k_{d_t}, U_{l_t}]$. Then the retriever searches for the $d_{t,ins}$, whose context $c_{t,ins} = [k_{d_{t,ins}}, U_{l_{t,ins}}, u_{2l_{t,ins}}]$ has the highest similarity to q_{d_t}. The final part is task description des, which describes the task requirements and schema in the target domain. In summary, the prompt consists of the task description des, in-context dialogue instance $d_{t,ins}$, the response r_{FLLM}, and the test dialogue d_t. Once the prompt input is created, it is fed into the in-context learning IXLLM and DAHDA gets the final generated dialogue response r.

To find the dialogue with the highest similarity score to the test dialogue, DAHDA uses the re-

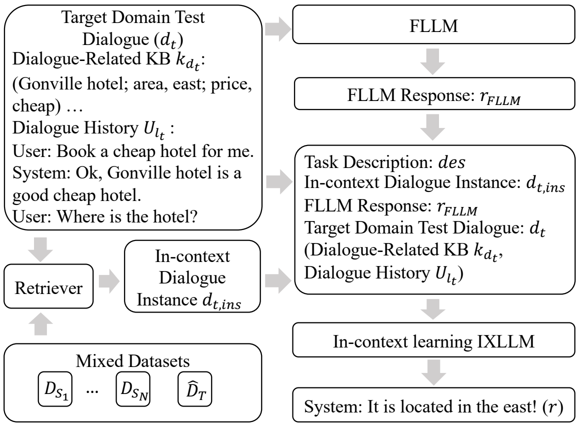

Figure 4: In-context learning IXLLM with the FLLM response and the similar dialogue instance obtained from the mixed datasets with a retriever.

triever model (Reimers and Gurevych, 2019) as a retriever following Hu et al. (2022b). This model maps the query input q into the representation e_q, and the context c of each dialogue from the mixed datasets into e_c. The similarity score is measured by the Euclidean distance between e_q and e_c.

To get better matched representations, the retriever model is fine-tuned with positive and negative samples from the mixed dataset. For each dialogue instance d from mixed dataset with L exchange-level utterances, there is a set of query inputs $\{[k_d, U_1], ..., [k_d, U_L]\}$. For each query input $q = [k_d, U_l]$, the context of positive dialogue instance c^+ is $[k_d, U_l, u_{2l}]$. Moreover, the context of negative dialogue instance c^- is $[k_{d'}, U_{l'}, u_{2l'}]$, where d' is a randomly sampled dialogue instance and l' is a randomly sampled exchange-level utterances number in dialogue instance d'. The retriever model learns the similarity between query input and dialogue context with the contrastive loss $\mathcal{L}_s$ (Reimers and Gurevych, 2019; Khosla et al., 2020). The contrastive loss $\mathcal{L}_s$ aims to minimize the distance between e_q and e_{c+} from positive dialogue instance, and maximize the distance between e_q and e_{c^-} from negative dialogue instance.

4 Experiments

4.1 Datasets

We conduct this framework on two publicly knowledge-grounded task-oriented dialogue datasets, Multi-WOZ 2.1 (MWOZ) (Budzianowski et al., 2018; Eric et al., 2020) and KVRET (Eric et al., 2017). Moreover, we use the sampling KVRET and MWOZ datasets from Raghu et al. (2021), where each dialogue is annotated with a dialogue-related knowledge base. As shown in Table 1, the sampling MWOZ dataset has three domains: restaurant, hotel, and attraction. The sampling KVRET dataset also contains three domains: weather, navigation, and schedule. In comparison to MWOZ, each KVRET domain contains domain-specific categories, like the weather attribute, which only appears in weather domain dialogues. For experiments with the KVRET and MWOZ datasets, we separately adopted each of the three domains as the target domain and the remaining two domains as source domains. In the domain adaptation experimental setting, we sampled 50 dialogues from each source domain regarded as the low-resource source domain datasets $\{D_{S_1}, ..., D_{S_N}\}$. We conducted each experiment three times with different seeds and calculated the average scores as the final results.

Domain	#Dial	Special entity categories
		KVRET
Schedule	1032	agenda, event, room
Weather	995	weather attribute, temperature
Navigate	1000	trafic info, distance, poi
		MWOZ
Restaurant	1311	-
Hotel	636	-
Attraction	150	-

Table 1: Statistic of KVRET and MWOZ. #Dial denotes the number of dialogues in each domain.

4.2 Baselines

As baselines, we used GPT2 for FLLM and ChatGPT (gpt-3.5-turbo) for IXLLM. Moreover, these two models are considered the backbone for our proposed framework to achieve a thorough comparison. In the comparison, we use the following response generation models in the domain adaptation experiments:

- Original ChatGPT: the ChatGPT with only task description and test dialogue information (OpenAI, 2022).
- Fine-tuning: the direct fine-tuning with GPT2 over the task-related source domain datasets for domain adaptation (Radford et al., 2019; Ramponi and Plank, 2020).
- In-context learning: the in-context learning ChatGPT with the example retrieved from

source domain datasets (Hu et al., 2022b).

- BackTransAug: a data-augmented method for domain adaptation, which translates the dialogue utterances into other languages and then back into English (Sennrich et al., 2016).

4.3 Evaluation metrics

According to existing works on knowledge-grounded task-oriented dialogue response generation (Rony et al., 2022), we used BLEU (Papineni et al., 2002), Moverscore (Zhao et al., 2019), and Entity-F1 (Eric et al., 2017) as the evaluation metrics. BLEU is the BLEU-4 score which measures the word overlap between generated and ground truth responses, while Moverscore evaluates their semantic similarity. Entity-F1 calculates the correct and wrong entities in the generated responses based on the gold set of entities. Moreover, Entity-F1 evaluates the capability of the model to generate relevant entities from a knowledge base in the response (Eric et al., 2017).

5 Results and Discussion

5.1 Overall results

Table 2 and 3 present the primary experiment results for the MWOZ and KVRET datasets with the following models in the domain adaptation setting: (1) the original ChatGPT, (2) several models with different domain adaptation strategies, including in-context learning, fine-tuning, BackTransAug, and our DAHDA.

Based on Tables 2 and 3, we observed that our proposed DAHDA obtains overall better performances than the other models, based on the following results: (1) In the experiments with the MWOZ dataset, DAHDA achieves the highest scores of BLEU, Moverscore, and Entity F1 over all three different target domains. In comparison to the original ChatGPT, DAHDA leads to improvements in BLEU of 1.27%–3.47%, Moverscore of 0.74%–1.18%, and Entitiy-F1 of 5.97%–7.21% for the three domains. In the experiments with the KVRET dataset, the Entity-F1 scores from DAHDA are still comparable to those from the original ChatGPT. However, there are slight decreases in Entity-F1 because domain-specific entity categories differ between target and source domains, as shown in Table 1. Moreover, DAHDA presents gains over the original ChatGPT in BLEU and Moverscore of 2.68% and 0.65% in the schedule domain, 6.27% and 3.75% in the weather domain, 2.73% and 0.93% in the navigation domain, respectively. Our DAHDA successfully learned the task-related knowledge and the appropriate style for task-oriented response generation from source domain datasets and target domain knowledge base. This learned information supports our DAHDA to generate more factually-accurate and suitable dialogue responses in the target domain.

(2) Our DAHDA outperforms the single fine-tuned GPT2 and single ChatGPT with in-context learning on the MWOZ dataset because it exploits the best features of both models. On the one hand, our DAHDA obtains improvements of BLEU by 1.12%–5.09%, Moverscore by 0.55%–1.79%, and Entity-F1 by 6.90%–13.97% beyond the fine-tuned GPT2 on the MWOZ dataset. Moreover, we found that fine-tuned GPT2 on the KVRET dataset performs better in the scores of BLEU and Moverscore than the original ChatGPT. The reason is that fine-tuned GPT2 learns the communication schema, length, and language style from source domains, which are similar to those from the target domain. However, the fine-tuned GPT2 is still limited by the low-resource source domain datasets and many unseen entity categories from the target domain in the KVRET dataset. On the other hand, compared to the ChatGPT with in-context learning, DAHDA obtains improvements of BLEU score by 0.50%–3.14%, Moverscore by 0.21%–1.08%, and Entity-F1 by 1.59%–4.79% on the MWOZ dataset. Furthermore, ChatGPT with in-context learning also obtains the average improvement (restaurant, hotel, attraction) in BLEU score by 0.66%, Moverscore by 0.21%, and Entity-F1 by 3.49% beyond the original ChatGPT on MWOZ dataset. Even though the retrieved examples are from source domain datasets, they contain similar task-related styles and entities as the examples from the target domain. However, the size of the retrieved samples in the prompt limits the knowledge provided for ChatGPT with in-context learning. Therefore, in addition to the retrieved sample, the predicted response from the fine-tuned GPT2 is added to the prompt. By combining these information into the prompt, DAHDA is able to capture the strengths of both models.

(3)The DAHDA with data-augmented domain adaptation overall provides better performance in comparison to the original ChatGPT on the MWOZ dataset. This indicates that data-augmentation is

	Restaurant			Hotel			Attraction		
Model	BLEU	Mover	En-F1	BLEU	Mover	En-F1	BLEU	Mover	En-F1
Original ChatGPT	4.41	51.85	22.25	3.20	51.03	15.44	4.65	52.50	28.54
In-context learning	5.30	52.27	26.63	3.97	51.56	19.20	4.98	52.17	30.86
Fine-tuning	5.04	51.97	14.25	3.35	51.22	15.75	3.03	51.46	24.34
BackTransAug	4.78	52.01	18.38	3.51	51.26	19.28	2.71	50.92	22.22
DAHDA	7.72	53.03	28.22	4.47	51.77	22.65	8.12	53.25	35.65

Table 2: Primary results with BLEU, Moverscore, and Entity-F1 metrics on different target domains with low-resource source domain datasets (50 dialogues for each source domain) from MWOZ dataset.

	Schedule			Weather			Navigate		
Model	BLEU	Mover	En-F1	BLEU	Mover	En-F1	BLEU	Mover	En-F1
Original ChatGPT	6.56	52.44	57.96	4.43	50.29	51.48	3.75	50.76	23.65
In-context learning	9.23	53.74	49.20	7.16	51.91	50.80	5.80	51.88	21.98
Fine-tuning	16.76	57.90	17.14	13.77	55.93	12.22	11.49	53.24	4.39
BackTransAug	14.48	57.07	24.40	12.76	55.17	14.34	8.21	52.71	4.45
DAHDA	9.24	53.09	51.48	10.70	54.04	48.77	6.48	51.69	23.90

Table 3: Primary results with BLEU, Moverscore, and Entity-F1 metrics on different target domains with low-resource source domain datasets (50 dialogues for each source domain) from KVRET dataset.

Model	Accuracy	Suitability
Original ChatGPT	4.29	4.43
DAHDA	4.46	4.77

Table 4: Human evaluation results

an efficient method for task-oriented dialogue response with domain adaptation. Further, DAHDA shows an improvement compared to the BackTransAug method on the MWOZ dataset, demonstrating that our data-augmented hybrid system is more effective.

5.2 Human evaluation

We conducted a human evaluation to study the accuracy and style suitability of the generated responses from the original ChatGPT and our DAHDA. 6 judges evaluated 100 randomly sampled dialogue examples from the MWOZ dataset. They scored the examples based on the following questions with the Likert scale from 1 (very bad) to 7 (very good) (Likert, 1932): 1. Is the response factually-accurate? 2. Is the response in a suitable style like the ground truth? As shown in Table 4, our DAHDA obtains an accuracy improvement of 0.17, indicating more factually-accurate responses. Additionally, DAHDA presents a gain in suitability of 0.34. It demonstrates that DAHDA generates responses with more suitable length, language style, and task-oriented format for the target domain.

5.3 Ablation study

We conducted an ablation study to gain a better understanding of the different individual components in our framework. As shown in Table 5, we implemented different simplified versions of our DAHDA by removing a single component of the system on the MWOZ dataset for each version. Overall, our DAHDA outperforms simplified versions within the MWOZ experiment. The improvements of DAHDA from the study mainly come from the following three aspects. First, in the restaurant and attraction domains, removing data augmentation leads to the loss in BLEU by 1.01%–2.30%, Moverscore by 0.08%–0.63%, and Entity F1 by 0.97%–2.45%. In the hotel domain, the model without data augmentation keeps comparable BLEU and Moverscore, but has 1.77% loss in Entity-F1. This drop in performance due to the removal of the data augmentation indicate that the synthetic target dialogue datasets contain the target domain knowledge and language style. Secondly, the predicted responses of FLLM support the in-context learning IXLLM to generate more accurate and appropriate dialogue responses, since removing the FLLM impairs the scores of BLEU by 0.86%–2.52%, Moverscore by 0.53%–0.91%, and Entity-F1 by 3.02%–4.94% on the three domains. It demonstrates that the FLLM captures task-related knowledge from data-augmented do-

	Restaurant			Hotel			Attraction		
Model	BLEU	Mover	En-F1	BLEU	Mover	En-F1	BLEU	Mover	En-F1
DAHDA	7.72	53.03	28.22	4.47	51.77	22.65	8.12	53.25	35.65
-Augmentation	6.71	52.95	25.77	4.56	52.05	20.88	5.82	52.62	34.68
-FLLM	5.20	52.12	25.20	3.61	51.24	17.71	5.84	52.55	31.41
-IXLLM	8.37	53.10	26.27	5.01	52.18	21.87	4.99	52.37	29.41

Table 5: Ablation study with BLEU, Moverscore, and Entity-F1 metrics on different MWOZ target domains.

	Restaurant			Hotel			Attraction		
Model	BLEU	Mover	En-F1	BLEU	Mover	En-F1	BLEU	Mover	En-F1
In-context learning	5.46	52.34	26.28	3.92	51.74	19.48	4.93	52.67	31.70
Fine-tuning	7.38	53.05	24.26	5.69	52.42	21.98	8.15	53.93	40.96
Fine-tuning+aug	7.83	53.13	28.57	7.92	53.39	29.18	8.69	54.31	40.79
DAHDA	6.99	52.82	29.31	5.94	52.45	27.12	8.51	54.13	36.08

Table 6: Results on different MWOZ target domains, when models are trained with all source domain datasets.

main datasets, and integrates the learned knowledge into IXLLM by in-context learning with the generated response from FLLM. Finally, Table 5 shows that removing the IXLLM degrades the average score (restaurant, hotel, attraction) of BLEU by 0.65%, Moverscore by 0.13%, and Entity F1 by 2.99%. This single FLLM with augmented datasets performs worse than our DAHDA in Entity-F1, indicating that the powerful natural language capabilities and general knowledge from IXLLM benefit the response generation with low-resource source domain datasets in the zero-shot target domain setting.

5.4 Impact of dataset size

Based on primary results, we observed that our proposed DAHDA overall performs better than other models with low-resource source domain datasets in the zero-shot target domain setting. To understand the impact of dataset size and explore the performance of the proposed DAHDA, we trained models with all source domain datasets. Specially, we tested the fine-tuned GPT2 with our augmented dataset (fine-tuning+aug) as shown in Table 6. The results present that the fine-tuned GPT2 with our augmented dataset achieves overall better performance, which has improvements than the fine-tuned GPT2 model without our augmented data. It demonstrates that our data augmentation method plays a key role in the performance improvement with rich-resource source domain datasets. Moreover, the fine-tuned GPT2 with our augmented dataset overall performs better than the DAHDA. This indicates that when we have rich-resource source domain datasets containing enough task-related knowledge, the fine-tuned GPT2 model with our data-augmentation is able to achieve competitive results.

5.5 Future work

In this paper, we do not focus on prompt design, but in future work we will explore the different prompts for our system. Additionally, our experiments are limited to the two selected datasets. We are aware that the limited size of the datasets could introduce a bias problem in the model. In the future to avoid the possible problem of bias, the experiments could be conducted with more datasets. Moreover, to the best of our knowledge, our paper is the first work to explore both FLLM and IXLLM as a hybrid system for task-oriented dialogue response generation with domain adaptation and a zero-shot setting. The FLLM and IXLLM in DAHDA can be further replaced by the improved FLLMs and IXLLMs in future work.

6 Conclusion

In conclusion, we build a data-augmented hybrid system with domain adaptation, which can generate factually-accurate and suitable task-oriented dialogue responses. Our proposed DAHDA achieves overall better performances than the other domain adaptation strategies and obtains improvements beyond the original IXLLM, encouraging future work to consider a combination of IXLLMs and FLLMs to address the task-oriented response generation task with domain adaptation.

Acknowledgements

We thank Md Rashad Al Hasan Rony for the support and the reviewers for their valuable feedback.

References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.

Greg Brockman, Mira Murati, Peter Welinder, and OpenAI. 2020. OpenAI API.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022b. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are LLMs All You Need for Task-Oriented Dialogue? *arXiv preprint arXiv:2304.06556*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Junjie Li, Jianfei Yu, and Rui Xia. 2022. Generative Cross-Domain Data Augmentation for Aspect and Opinion Co-Extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4219–4229, Seattle, United States. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. LAD: Language Models as Data for Zero-Shot Dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604, Edinburgh, UK. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain Adaptive Dialog Generation via Meta Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Dinesh Raghu, Atishya Jain, Mausam, and Sachindra Joshi. 2021. Constraint based Knowledge Base Distillation in End-to-End Task Oriented Dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5051–5061, Online. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese Bert-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. DialoKG: Knowledge-Structure Aware Task-Oriented Dialogue Generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2557–2571, Seattle, United States. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online. Association for Computational Linguistics.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-Shot Dialog Generation with Cross-Domain Latent Actions. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Models and prompt examples

For the backbone of DAHDA, we utilize the GPT2 model and the all-mpnet-base-v2 retriever model from Huggingface (https://huggingface.co/models).

Original ChatGPT
Prompt: Response the user utterance as a task-oriented dialogue system called SYSTEM. Your response should be based on the test example's dialogue history and knowledge base. Test example: Knowledge base:...; Dialogue history: USER:... SYSTEM:... USER:...; Response: SYSTEM:
DAHDA
Prompt: Response the user utterance as a task-oriented dialogue system called SYSTEM. There is 1 example provided. Your response should be based on the test example's possible response, dialogue history, and knowledge base. Your response should be same in structure, length, chitchat and tone to the previous SYSTEM Responses in the example 1. Example 1: Knowledge base:...; Dialogue history: USER:...SYSTEM:...USER:...; Response: SYSTEM:...; Test example: Possible Response: ...; Knowledge base:...; Dialogue history: USER:... SYSTEM:... USER:...; Response: SYSTEM:

Table 7: Prompt examples

The ChatGPT (May 3 version) is accessed through the API from OpenAI.

We provide the prompt examples for original ChatGPT and DAHDA in Table 7.