SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2025, Vol. 68, Iss. 2, 122105:1–122105:14
https://doi.org/10.1007/s11432-023-4069-x

Towards few-shot mixed-type dialogue generation

Zeming LIU[1], Haifeng WANG[2], Zeyang LEI[2], Zheng-Yu NIU[2], Hua WU[2] & Wanxiang CHE[1*]

[1]*Research Center for Social Computing and Information Retrieval, Harbin 150001, China*
[2]*Baidu Inc., Beijing 100193, China*

Received 31 July 2023/Revised 24 October 2023/Accepted 5 May 2024/Published online 17 January 2025

Abstract Building an agent capable of conducting both open-domain and task-oriented dialogues, known as mixed-type dialogues, has been an enduring challenge for the AI community. Previous approaches have focused on constructing large-scale human-annotated datasets for training models. However, annotating these datasets is expensive and hinders the practical application of these models. This paper identifies a novel challenge, few-shot mixed-type dialogue generation. To address this challenge, we first present a pre-trained dialogue generation framework with modular-based architecture and prompt-tuning component. Additionally, we collect a mixed-type dialogue dataset that combines persona-chat with conversational recommendation or task-oriented dialogues within a single dialogue session. Specifically, the modular-based architecture allows us to easily incorporate more supervised signals and human-annotated information, thereby facilitating the learning of session-level dialogue logic. We pre-train this dialogue generation framework using multiple external datasets and then fine-tune it on the mixed-type dialogue dataset we collected. Experimental results demonstrate that the three key designs — modular-based architecture, prompt-tuning component, and model pre-training — significantly enhance the performance of this framework compared to state-of-the-art baselines.

Keywords mixed-type dialogue, dialogue generation, PLATO-prompt, Mixed-FS, few-shot

Citation Liu Z M, Wang H F, Lei Z Y, et al. Towards few-shot mixed-type dialogue generation. Sci China Inf Sci, 2025, 68(2): 122105, https://doi.org/10.1007/s11432-023-4069-x

1 Introduction

One of the goals of artificial intelligence is to build a conversational agent that can perceive the situation when conversing with users, and then conduct natural dialogues to meet their potential needs [1, 2].

Conversational agents should ideally possess three main functionalities. First, they should be able to engage users through persona-chat, a more engaging form of social chit-chat, which is enriched by endowing the agent with a configurable yet persistent persona. This persona is encoded using multiple sentences of textual description, often termed a profile. Second, agents should discuss various topics via in-depth knowledge-grounded dialogues. Finally, agents should be able to offer services through two types of dialogues: conversational recommendation and task-oriented dialogue. The primary distinction between the two lies in their objectives. Conversational recommendation mainly centers on suggesting resources, such as movies or food, emphasizing the recommendation process. In contrast, task-oriented dialogue aims to provide services, like booking movie tickets or restaurants, with a focus on completing tasks. Therefore, conversational agents need to have the ability to automatically plan appropriate dialogue types and conduct engaging dialogues according to the human-computer interaction situation.

In many real-world scenarios, users often have diverse needs that require chatbots to handle multiple types of conversations within a single dialogue in a seamless manner. For instance, users may engage in social chit-chat while also seeking assistance in a task-oriented dialogue or may ask questions and engage in knowledge-grounded conversations during conversational recommendation processes. As a result, there has been a growing interest in research on mixed-type dialogues in recent years. These studies primarily focus on effectively integrating multiple conversation types to create a more natural and coherent dialogue experience. Researchers such as Andrea et al. [3], Roller et al. [4], and Lin et al. [5] have explored the use of various dialogue datasets to train all-in-one conversation models in an end-to-end

* Corresponding author (email: car@ir.hit.edu.cn)

fashion. By leveraging multiple datasets, these models aim to capture the nuances and complexities of different conversation types, enabling them to handle mixed-type dialogues more effectively. On the other hand, Liu et al. [6] proposed a modularized framework to address the challenges posed by mixed-type dialogues. Instead of training a single model on multiple datasets, they develop a modular approach where different components of the dialogue system specialize in handling specific conversation types. This modularized framework allows for more flexibility and adaptability when dealing with mixed-type dialogues. Furthermore, there have been efforts to collect datasets that encompass multiple dialogue skills and conversation types. These datasets serve as valuable resources for training and evaluating models that can handle mixed-type dialogues effectively. Overall, the research in this area aims to develop dialogue systems that can seamlessly handle different conversation types within a single dialogue, enabling more natural and versatile interactions between users and chatbots.

Almost all these studies attempt to construct a large-scale human-annotated dataset to train models for mixed-type dialogues. However, it is expensive to annotate these datasets and then it hinders the use of these models in real-world application scenarios. To this end, we identify the challenge of few-shot mixed-type dialogue generation.

Specifically, we focus on task decomposition and model design to address this challenge. Task decomposition allows us to conveniently introduce supervised signals (dialogue acts) of decomposed subtasks and leverage external corpora related to those subtasks. This enhances the training of each sub-task, which is particularly important in few-shot settings. For instance, we can improve the model's understanding of persona-chat, knowledge-grounded dialogue, conversational recommendation, and task-oriented dialogues by utilizing datasets such as DuLeMon [7], KdConv [8], DuRecDial [6], and RiSAWOZ [9], respectively. To facilitate training and modeling, we first decompose mixed-type dialogues into three subtasks: natural language understanding (NLU), dialogue act planning (DAP), and natural language generation (NLG). This decomposition is based on the perspective of dialogue roles rather than dialogue types [10]. NLU focuses on understanding the context and identifying the dialogue act of the current conversation. DAP plans the next dialogue act based on the current context and the act identified by NLU. NLG generates responses based on the context and the act planned by DAP. In addition, we design a unified dialogue act format (consisting of dialogue type, dialogue topic, and topic attribute) that can be applied to all types of dialogues. This facilitates NLU, DAP, and the convenient use of knowledge bases. Furthermore, we propose a prompt-based mixed-type dialogue generation framework called PLATO-prompt, which aims to address these three subtasks in a unified manner. This framework enables the effective generation of mixed-type dialogues by incorporating prompt engineering techniques. By combining task decomposition, a unified dialogue act format, and the PLATO-prompt framework, we can tackle the challenges posed by mixed-type dialogues and achieve more comprehensive and coherent conversational capabilities.

Moreover, we collect a benchmark for few-shot mixed-type dialogues, which includes a large-scale knowledge base (KG-FS) spanning 10 domains and a novel human-to-human mixed-type dialogue corpus (Mixed-FS), to facilitate the train and test of PLATO-prompt. In Mixed-FS, as shown in Figure 1, we integrate persona-chat into conversational recommendation and task-oriented dialogue to make the dialogue more natural and closer to practical applications. Specifically, we first collect users' preferences by persona-chat, for example, like European and American movies, and then conduct recommendations or help users complete tasks more naturally and efficiently based on the preferences collected so far. Moreover, to be close to the setting in practical applications, the preferences are automatically updated at the end of each turn rather than each dialogue as done in [10], which will affect subsequent recommendations or tasks.

In summary, this work makes the following contributions:

• To our best knowledge, this is the first attempt to identify the challenge of mixed-type dialogue in the few-shot setting.

• To tackle this challenge, we decompose mixed-type dialogue into three subtasks based on dialogue roles and propose a prompt-based mixed-type dialogue generation framework called PLATO-prompt. This framework allows us to handle the three subtasks effectively in a unified manner. Furthermore, we empirically demonstrate the effectiveness of our approach through experimental evaluations.

• Additionally, we make a novel contribution by integrating persona-chat into conversational recommendation and task-oriented dialogue. This integration bridges the gap between different conversation types and enables us to collect a benchmark dataset for few-shot mixed-type dialogues. This benchmark dataset will serve as a valuable resource for future research in this area. Both automatic evaluation metrics and human evaluations confirm the benefits and utility of this benchmark dataset.

Dialogue type	Mixed-type dialogue
Chit chat	Usr: 你好 (Hello)
Persona chat	Bot: 你好,你们什么时候放假呢？(Hi, when do you have a holiday?) Usr:五一,我会和姐姐一起回来(May Day, I will come back with my sister)
Recommend	Bot: 挺好. 假期适合看电影.我在看大话西游1:月光宝盒. 经典的国产老电影.推荐你看看. (Well. It's suitable to watch movies. I'm watching A Chinese Odyssey Part I Pandora's Box, a classic domestic movie. I recommend it to you.)
Persona chat	Usr:我不太喜欢看国产电影(I don't like watching domestic movies) Bot:那你喜欢看美国的电影吗？(Do you like watching American movies?) Usr: 欧美的电影，我都挺喜欢. (I like European and American movies.)
Recommend	Bot:那你可以看看爱情与灵药，轻松幽默的美国电影. (Then you can see Love and Other Drugs, a relaxed and humorous American movie.)
Knowledge	Usr:这是什么类型的电影啊？(What's the genre of this movie?) Bot:是剧情片，男主角非常帅. (A feature movie. The hero is handsome.)
Persona chat	Usr:我不喜欢剧情片,喜欢动作片(I don't like feature films. I like action films.)
Recommend	Bot:那你要不要看看超能陆战队？这个电影讲述了一个呆萌暖男大白拯救世界的故事(Would you like to watch Big Hero 6? The film tells the story of a cute and warm man who saves the world.)
Knowledge	Usr:我看过了. (I've watched it.) Bot:阿凡达1呢,你有没有看过？(What about avatar 1? Have you watched it?) Usr: 没看过，导演是谁呢？(No, who is the director?) Bot:是詹姆斯·卡梅隆，他拿了第67届金球奖电影类最佳导演(It is James Cameron, who win the 67th Golden Globe Award for best director.) Usr:这个导演很有实力. 但是画面怎么呢？(The director is very powerful. But what about the scene?) Bot:画面很华丽,场景也很漂亮，建议你看看. (The scene is gorgeous and beautiful. I suggest you have a look.)
Chit chat	Usr:我肯定要看，感谢推荐啊. (I will. Thanks for your recommendation.) Bot:不客气. (You're welcome.)

Figure 1 (Color online) Example of Mixed-FS.

Overall, our work not only identifies the challenge of mixed-type dialogue in the few-shot setting but also presents a novel approach and a benchmark dataset to address this challenge, paving the way for further advancements in the field.

2 Related work

Multi-skill dialogues. In recent years, there has been a significant increase in the research on mixed-type dialogues, which focuses on integrating multiple dialogue types seamlessly. Several studies, such as Andrea et al. [3], Roller et al. [4], and Lin et al. [5] have developed an all-in-one conversation model to combine multiple dialogue skills. However, these models are trained using large-scale data and are not suitable for low-cost real-world applications.

In contrast to these approaches, our work decomposes mixed-type dialogues into three subtasks and proposes a modular-based framework with prompt-tuning for few-shot mixed-type dialogue generation. This approach allows us to address the challenges of mixed-type dialogues more efficiently and cost-effectively. Furthermore, previous studies by Smith et al. [1], Shuster et al. [2], Sun et al. [11], Chiu et al. [12], Liu et al. [10], and Liu et al. [6] have collected datasets that mix multiple dialogue skills, as shown in Table 1 [1, 2, 6, 10–19]. However, these datasets are designed to fulfill specific needs and are limited in their ability to solve other requirements at a low cost. In contrast to these existing datasets, we have collected a KG-FS and a Mixed-FS. This collection of resources aims to facilitate the study of few-shot mixed-type dialogue generation and provides a more comprehensive solution to address various needs efficiently.

Prompt-based learning for dialogues. Prompt-based learning has gained significant attention in

Table 1 Comparison of Mixed-FS with other datasets. "know.", "rec.", and "Persona" stand for knowledge-grounded dialogue, conversational recommendation, and persona-chat, respectively.

Dataset	Mixed-type?	Persona?	Dialogue types
decalDialogue [2]	✓	✗	Know., chit-chat, QA, empathetic dialogue, image chat
BlendedSkillTalk [1]	✓	✗	Know., empathetic dialogue, chit-chat
ACCENTOR [11]	✓	✗	Chit-chat,task-oriented dialogue
SalesBot [12]	✓	✗	Chit-chat,task-oriented dialogue
DuClarifyDial [10]	✓	✗	Rec., know. chit-chat, QA, task-oriented dialogue
Facebook_Rec [13]	✗	✗	Rec.
GoRecDial [14]	✗	✗	Rec.
OpenDialKG [15]	✗	✗	Rec.
DuRecDial [6]	✓	✗	Rec., chit-chat, QA, task-oriented dialogue
TG-ReDial [16]	✗	✗	Rec.
INSPIRED [17]	✗	✗	Rec.
Persona [18]	✗	✓	Persona
DuLeMon [19]	✗	✓	Persona
Mixed-FS (ours)	✓	✓	Persona, rec., know., chit-chat, QA, task-oriented dialogue

various task-oriented dialogue applications, including NLU [20–22], dialogue state tracking [22, 23], and NLG [21, 22]. Moreover, Zheng et al. [24] have introduced a prompt-based model for both task-oriented and open-domain dialogues. Compared with them, we propose a prompt-based framework for few-shot mixed-type dialogue generation.

3 Task decomposition

Intuitively, unifying the dialogue act for all dialogue types can facilitate them to share the knowledge base and make few-shot mixed-type dialogue generation more effective. In this study, we propose a novel method to unify the dialogue act across all dialogue types, representing it as (dialogue type, dialogue topic, topic attribute). Furthermore, we decompose mixed-type dialogues into three subtasks: NLU, DAP, and NLG. This approach allows us to better understand and generate responses in mixed-type dialogues by leveraging the unified dialogue act representation.

Let $D^k = \{u_1^k, u_2^k, \ldots, u_n^k\}$ denote a dialogue between the user U^k $(0 \leqslant k < N)$ and the bot, where N is the number of users, and u_j^k $(0 < j \leqslant n)$ is dialogue utterance. Recall that we attach each dialogue utterance (say u_j^k) with a dynamic user profile (denoted as $\mathcal{P}_j^{u_k}$), a dialogue act $A_j^k = (g_j^{\text{ty}}, g_j^{\text{tp}}, g_j^{\text{ta}})$, where g_j^{ty} is a candidate dialogue type, g_j^{tp} is a candidate dialogue topic (an entity in KG-FS), and g_j^{ta} is several attributes of the dialogue topic in KG-FS.

NLU. Given a context X with utterances $\{u_j^k\}_{j=1}^{m-1}$ from the dialogue D^k, and a previous dialogue act A_{m-2}^k (if $m-2<0$, $A_{m-2}^k = (-,-,-)$) for the bot, where u_{m-1}^k is an utterance of the user U^k, NLU aims to learn a proper dialogue act $A_{m-1}^k = (g_{m-1}^{\text{ty}}, g_{m-1}^{\text{tp}}, g_{m-1}^{\text{ta}})$ for u_{m-1}^k.

DAP. Given a context X, and a dialogue act A_{m-1}^k for the user U^k, DAP aims to predict a proper dialogue act $A_m^k = (g_m^{\text{ty}}, g_m^{\text{tp}}, g_m^{\text{ta}})$ for the bot response u_m^k.

NLG. Given a context X with utterances $\{u_j^k\}_{j=0}^{m-1}$ from the dialogue D^k, and a proper dialogue act A_m^k for the bot response u_m^k, NLG aims to produce a proper dialogue response $Y = u_m^k$ for completing the dialogue act A_m^k.

4 Framework

To achieve better integration of multiple decomposed subtasks, we present a novel pre-trained mixed-type dialogue model called PLATO-prompt, which builds upon the PLATO-2 [25] and incorporates a prompt tuning mechanism [26]. Unlike the approach of separately fine-tuning individual subtasks, PLATO-prompt eliminates the need for redundant copies of the model's parameters and enables more effective fusion learning across multiple subtasks. This approach greatly benefits the task of few-shot mixed-type dialogue generation, allowing the model to generate coherent and contextually appropriate responses in diverse dialogue scenarios.

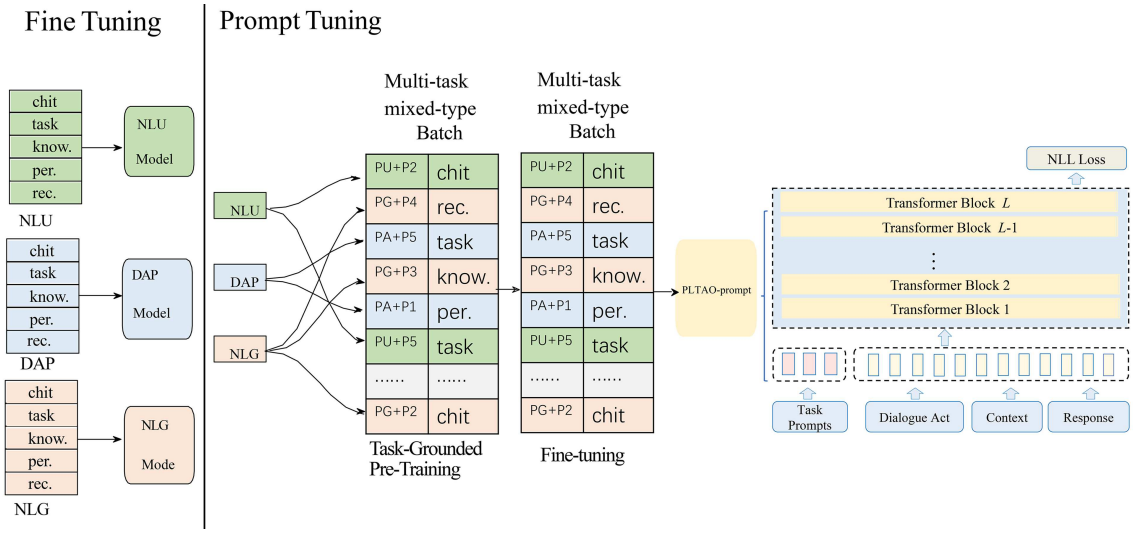

Figure 2 (Color online) Architecture of our framework (denoted as PLATO-prompt). Here, to better understand our prompt embedding settings, we use "P1", "P2", "P3", "P4", "P5", and "P6" to respectively represent different dialogue types such as persona-chat, chit-chat, knowledge-grounded dialogue, conversational recommendation, task-oriented dialogue, and QA. "PU", "PA", and "PG" represent different dialogue processes such as NLU, DAP, and NLG. The combination of the two using '+' signifies the corresponding process of a certain dialogue type. For example, PU+P1 represents the NLU process of the persona-chat dialogue's prompt embedding.

4.1 PLATO-prompt

The overview of our framework PLATO-prompt is shown in Figure 2. Following [25], we use the unified transformer [27] for dialogue act encoding, context encoding, and response generation. Moreover, we add additional task prompt embeddings in the input representation to distinguish different dialogue types and dialogue subtasks. For example, the embeddings of "P1", "P2", "P3", "P4", "P5", and "P6" for persona-chat, chit-chat, knowledge-grounded dialogue, conversational recommendation, task-oriented dialogue, and QA, respectively.

Specifically, PLATO-prompt is first pre-trained by using publicly available dialogue corpora annotated with dialogue acts as Subsection 4.2 to enhance the model performance. Each dialogue sample S in our training data is represented as

$$S = (P, X, A, Y), \tag{1}$$

where P is the task prompts, X is the dialogue context up to the current dialogue turn, A is the dialogue act for the response, and Y is the dialogue response. Due to the knowledge required for dialogue generation, in addition to the dialogue corpora, we also incorporate training data by transforming the knowledge base KG-FS into question-and-answer pairs. Note that, if X is knowledge, A is $(-, -, -)$.

Given training data and knowledge base, our goal is to build a mixed-type dialogue model for few-shot mixed-type dialogue generation. The task can be factorized as

$$p(S) = p(P, X, A, Y) \tag{2}$$

$$= \underbrace{p(Y|P, X, A)}_{\text{NLG}} \underbrace{p(A|P, X)}_{\text{DAP}} p(P, X). \tag{3}$$

Following the (3), we decompose the mixed-type dialogue generation in the few-shot setting into three subtasks: NLU, DAP, and NLG.

Therefore, in addition to distinguishing different dialogue types, task prompts are also used to distinguish different subtasks. For example, the embeddings of "PU", "PA", and "PG" for NLU, DAP, and NLG, respectively.

Then, by combining prompts, we can model different dialogue types for different subtasks. For example, the embeddings of "PU+P1", "PU+P4", and "PU+P5" for the NLU of persona-chat, conversational recommendation, and task-oriented dialogue, respectively.

In the training process, we first post-pretrain PLATO-2 [25] using KG-FS and publicly available dialogue corpora in Subsection 4.2, such as RiSAWOZ [9], KdConv [8], DuRecDial [6], and DuLeMon [7],

with labels of dialogue acts. Then, we use the pre-trained model for further fine-tuning on Mixed-FS in Subsection 4.3[1)].

4.2 Task-grounded pre-training

In this subsection, we describe how to produce data for task-grounded pre-training. The data for task-grounded pre-training includes multiple dialogue corpora and knowledge base KG-FS.

Knowledge-based QA. To facilitate the model's ability to use knowledge for the three subtasks, we convert the knowledge base, KG-FS, into the same format as the dialogue sample. Specifically, we first use the subject and relation in the triplets to manually generate some questions about the object. For example, ["Jay Chou", "constellation", "Capricornus"] can be generated What is Jay Chou's constellation? and What constellation is Jay Chou? Then, we use the query rewriting technology[2)] to generate more questions about the object to train the generalization ability of PLATO-prompt. For example, Do you know Jay Chou's constellation? and I want to ask about Jay Chou's constellation. Finally, the input for model training can be (KG Prompt, Do you know Jay Chou's constellation?, $(-,-,-)$, Capricornus).

Multiple dialogue tasks. Ideally, conversational agents need to have the ability to plan appropriate dialogue types (skills) automatically and conduct engaging dialogues. Specifically, the agent should be able to get to know users by persona-chat or social chit-chat, discuss some topics through in-depth knowledge-grounded dialogues, and provide services to users through conversational recommendation or task-oriented dialogue. Therefore, we enhance the model's ability of social chit-chat, persona-chat, knowledge-grounded dialogue, conversational recommendation, and task-oriented dialogues, by using DuLeMon [7], KdConv [8], DuRecDial [6], and RiSAWOZ [9], respectively.

It should be noted that the input format and task prompts for different subtasks or different dialogue types are the same as before.

4.3 Fine-tuning

When deploying PLATO-prompt to a special task, we can flexibly collect task-specific datasets in the same format as that used for pre-training as (1).

For few-shot mixed-type dialogue generation, we collect a new dataset, Mixed-FS. Note that, each session in Mixed-FS integrates multiple dialogue types, such as social chit-chat, persona-chat, knowledge-grounded dialogues, conversational recommendation, and task-oriented dialogues, which is different from DuLeMon, KdConv, and RiSAWOZ. Therefore, we design different task prompts for different dialogue types at the utterance level. The task prompts for specific dialogue types, such as persona-chat, are the same as Subsection 4.2. Finally, we generate training data for fine-tuning PLATO-prompt in the same format as that used for pre-training as (1).

5 Data collection and analysis

We collect high-quality multi-domain mixed-type dialogue data (Mixed-FS) to facilitate the study of few-shot mixed-type dialogue generation. In Mixed-FS, one person serves as the bot and the other as the user. To make it better suited to real-world applications, we ask the bot to proactively lead the dialogue, get to know the user by persona-chat or social chit-chat, discuss some topics through in-depth knowledge-grounded dialogues, and provide services to the user through conversational recommendation or task-oriented dialogues. Moreover, we collect a KG-FS spanning ten domains to support the achievement of mixed-type dialogues.

In this section, we describe the three steps for knowledge base and dataset construction: (1) constructing the KG-FS for supporting few-shot mixed-type dialogue generation; (2) collecting mixed-type dialogues by crowdsourcing; (3) annotating united dialogue acts for all type dialogues by crowdsourcing.

1) For the zero-shot setting, fine-tuning is not included.
2) https://ai.baidu.com/.

5.1 Knowledge base construction

To facilitate the study of few-shot mixed-type dialogue generation, we build a KG-FS, that includes ten domains: star, movie, music, TV play, TV show, cartoon, attraction, hotel, restaurant, and food.

Specifically, we build KG-FS by crawling publicly available entities and related knowledge information from several related websites for the star[3)4)5)], movie[3)4)5)], music[3)6)7)], TV play[8)9)10)11)], TV show[8)9)10)], cartoon[8)9)10)11)], attraction[12)13)14)], hotel[12)13)14)], restaurant[15)], and food[3)15)] domains.

Then, we clean the knowledge base with some rules, such as filtering Japanese and English, filtering knowledge inconsistent in multiple data sources. Finally, we obtain 154k nodes and 49 attributes (types of edges), resulting in about 1155k triplets (numbers of edges), the accuracy of which is 96%[16)]. Table 2 shows the statistics of KG-FS.

5.2 Dialogue collection

To collect high-quality mixed-type dialogues, we use a strict quality control procedure to guide crowdsourced workers to annotate dialogues based on the above knowledge base (KG-FS). In particular, the procedure includes three stages: (1) selecting crowdsourcing workers, (2) dialogue collection, and (3) quality verification.

In the crowdsourcing workers selection stage, we select four crowdsourcing workers with more than three years of dialogue research experience and give them training about mixed-type dialogues for real-world scenarios.

In the dialogue collection stage, we randomly pair up two workers and set each of them as one role of the user or the bot. To be more suitable for real-world applications, we expect that a dialogue between the bot and the user starts with the real human-bot interaction dialogue, e.g., persona-chat or social chit-chat, and the bot should proactively and naturally guide the dialogue to conversational recommendation (e.g., recommended movies) or task-oriented dialogues (e.g., order a restaurant).

Specifically, the dialogue starts with a real persona-chat or social chit-chat to collect the user coarse coarse-grained preference, e.g., the movie genres the user likes. Then, the bot tries to conduct recommendations based on the preferences collected at the current turn and the knowledge base, and the user's profile is automatically updated at the end of each turn by collecting what the user accepts or rejects to obtain more fine-grained preferences, e.g., the specific movies the user likes. And the fine-grained preferences will affect what to recommend in subsequent turns. In other words, If the user does not accept the recommendation, we will dynamically modify the user preferences and continue to recommend based on the fine-grained preferences until the user accepts it. Moreover, to be closer to real-world application scenarios, if the user struggles to figure out clear and specific goals during conversational recommendation or task-oriented dialogues, e.g., the user cannot decide whether to accept the recommended attractions since he does not know them, the bot will proactively conduct an in-depth knowledge-grounded dialogue to help the user figure out clear and specific goals.

In the quality verification stage, two data specialists will check whether the collected dialogues are natural recommendations (considering the user's preferences and feedback) and whether it is suitable for real-world applications. If a dialogue is considered unqualified, we will ask the two crowdsourcing workers to revise the dialogue until it is qualified.

3) https://baike.baidu.com/.
4) http://www.mtime.com.
5) https://www.douban.com/.
6) https://music.163.com/.
7) https://y.qq.com.
8) https://www.9sha.com/.
9) http://kan.znds.com/.
10) https://kan.2345.com/.
11) https://www.6hdy.com/.
12) https://www.qunar.com/.
13) https://www.ctrip.com/.
14) http://www.mafengwo.cn/.
15) https://www.meituan.com.
16) We randomly sampled 300 triplets and manually evaluated them.

Table 2 Statistics of KG-FS and Mixed-FS.

KG-FS				Mixed-FS		
#Domains	#Entities	#Attributes	#Triples	#Dialogues	#Utterances	#Utterances per dialogue
10	154k	49	1155k	100	3016	30.16

Table 3 Statistics of knowledge base (KG-FS). "Cart.", "Att.", and "Rest." stand for cartoon, attraction, and restaurant, respectively.

	Star	Movie	Music	TV play	TV show	Cart.	Att.	Hotel	Rest.	Food
#Entities	1k	50k	26k	18k	1k	6k	33k	1k	8k	10k
#Attributes	13	11	9	13	10	13	10	12	7	4
#Triples	12k	392k	213k	162k	10k	50k	236k	11k	49k	20k

5.3 Dialogue act annotation

To seamlessly blend mixed-type dialogues for few-shot mixed-type dialogue generation, we design a unified dialogue act (i.e., dialogue type, dialogue topic, and topic attribute) for all types of dialogues. In addition to collecting dialogues, crowdsourcing workers were also required to annotate the dialogue acts.

5.4 Dataset analysis

Tables 2 and 3 provide statistics of Mixed-FS and KG-FS. Following the evaluation method in previous work [6], we conduct human evaluations for data quality. Specifically, a dialogue will be rated "1" if it is wholly suitable for real-world applications and the recommendations are natural, otherwise "0". Then we ask two data specialists to judge the quality of all collected dialogues[17]. Finally, we obtained an average score of 0.95 on this evaluation set.

6 Experiments and results

6.1 Experiment setting

6.1.1 *Datasets*

PLATO-prompt is used for dealing with dialogues that mix multiple dialogue types in each session in the few-shot or zero-shot setting. Therefore, we conducted experiments on two mixed-type dialogue datasets, DuRecDial [6] and Mixed-FS, with three experimental settings. Specifically, We partitioned the Mixed-FS into training, development, and test sets by allocating 30%, 10%, and 60%, respectively. To be more specific, the training set comprises 30 samples, the validation set has 10 samples, and the test set contains 60 samples. We assess the performance of few-shot mixed-type dialogue generation in both the few-shot and zero-shot settings. Then, we randomly select 100 dialogues from the training set of DuRecDial [6] to fine-tune the PLATO-prompt and baselines and use the DuRecDial test set to evaluate the PLATO-prompt and baselines. Besides, to confirm if Mixed-FS has benefits, we use Mixed-FS (100 dialogues in all) to fine-tune the PLATO-prompt and baselines and use the DuRecDial test set to evaluate the PLATO-prompt and baselines. Note, we do not use the train set of Mixed-FS or DuRecDial [6] in the zero-shot setting, as shown in Table 4.

6.1.2 *Baselines*

We carefully select a few strong baselines proposed for mixed-type dialogues [1, 10], open-domain dialogues [25], or open-source large-scale pre-training language models [28, 29]. The details of training data used for all models are shown in Table 4.

BST [1] is a mixed-type dialogue model that can display many skills, and blend them seamlessly and engagingly.

PLATO-2 [25] is a robust pre-trained model as a strong benchmark for open-domain dialogue. We use the released parameters[18].

17) We calculate the averaged weighted kappa value for all dialogues and get a high score of 0.81, demonstrating good agreements between data specialists.

18) https://github.com/PaddlePaddle/Knover/tree/luge-dialog/luge-dialog.

Table 4 Details of training data used for all models. "w/o task", "E2E" and "Zero-shot" stand for PLATO-prompt without task-grounded pre-training, PLATO-prompt without task decomposition and all result in the zero-shot setting, respectively.

Method	Task pre-training (Subsection 4.2)	Fine-tuning (Subsection 4.3)
BST	✓	✓
PLATO-2	✓	✓
PMT	✓	✓
Baichuan-7B	✓	✓
ChatGLM-6B	✓	✓
QWen-7B	✓	✓
PLATO-prompt (ours)	✓	✓
w/o task	✗	✓
E2E	✓	✓
Zero-shot	✓	✗

PMT is PLATO-2 with hard prompts as [10]. The parameters of PMT are the same as those of PLATO-2.

Baichuan-7B is a large-scale pre-training language model developed by BaiChuan-Inc. We use the released parameters[19)].

ChatGLM-6B [30] is an open general bilingual dialogue language model with autoregressive blank infilling, and can serve as a strong baseline.

QWen-7B [31] is a pre-trained large language model proposed by Alibaba Cloud, which can be used as a strong baseline.

6.1.3 *Implementation details*

Our PLATO-Prompt is a model with 12 transformer blocks and 12 attention heads, with an embedding dimension of 768, and the embedding size of the task prompt is also 768. Then, we randomly initialize the representation vectors of task prompts. Besides, we adopt a top$k$-sampling decoding strategy with $k = 5$ during decoding.

To maintain consistency with baseline models such as PMT, the PLATO-2 model we used has 93 million parameters with 12 transformer blocks and 12 attention heads. It has an embedding dimension of 768, and the task prompt also has an embedding size of 768.

6.2 Experiment results for NLU

6.2.1 *Evaluation metrics*

We measure the accuracy of dialogue type (TA), the accuracy of dialogue topic (PA), and the F1, BLEU1(B1), BLEU2(B2) [32], and ROUGE(R) [28] of the topic attribute. Specifically, all the metrics are measured by comparing each label (dialogue type, dialogue topic, topic attribute) to its ground truth label.

6.2.2 *Result*

As shown in Tables 5 and 6, our model outperforms baselines in terms of almost all metrics, indicating better dialogue act understanding ability. Besides, the performance gap between our model and our model without task-grounded pre-training demonstrates that the use of external dialogue data can enhance the ability of the model to understand user requirements. Moreover, by comparing "PLATO-prompt" and "PLATO-prompt$*$", we find that Mixed-FS can facilitate mixed-type dialogues understanding in the few-shot or zero-shot setting.

6.3 Experiment results for DAP

6.3.1 *Evaluation metrics*

Similarly, we also measure the TA, the PA, and the F1, BLEU1, BLEU2 [32], and ROUGE(R) [28] of topic attribute for DAP subtask.

19) https://github.com/baichuan-inc/Baichuan-7B.

Table 5 NLU results on Mixed-FS in the few-shot setting and zero-shot setting. "w/o task" stands for PLATO-prompt without task-grounded pre-training. Best results are highlighted in bold.

Method	Few-shot					Zero-shot				
	TA	PA	F1	B1/B2	R	TA	PA	F1	B1/B2	R
BST	0.61	0.33	0.29	0.10/0.04	0.03	0.32	0.11	0.12	0.06/0.02	0.01
PLATO	0.83	0.52	0.41	0.17/0.10	0.11	0.58	0.33	0.28	0.09/0.05	0.07
PMT	0.88	**0.62**	0.48	0.21/0.14	0.14	0.61	0.39	0.32	0.14/0.09	0.09
Baichuan-7B	0.89	0.59	0.44	0.21/0.12	0.15	0.66	0.40	0.33	0.11/0.07	0.09
ChatGLM-6B	0.81	0.57	0.42	0.19/0.09	0.13	0.60	0.39	0.34	0.10/0.07	0.08
QWen-7B	0.88	0.60	0.45	0.19/0.08	0.16	0.63	0.42	0.36	0.13/0.08	0.10
PLATO-prompt	**0.91**	0.61	**0.52**	**0.23/0.16**	**0.18**	**0.69**	**0.44**	**0.41**	**0.17/0.11**	**0.12**
w/o task	0.82	0.54	0.44	0.20/0.14	0.13	0.11	0.08	0.05	0.01/0.01	0.09

Table 6 NLU results on DuRecDial in the few-shot setting and zero-shot setting. "∗" and "w/o task" stand for PLATO-prompt fine-tuning with Mixed-FS and without task-grounded pre-training, respectively. Best results are highlighted in bold.

Method	Few-shot					Zero-shot				
	TA	PA	F1	B1/B2	R	TA	PA	F1	B1/B2	R
BST	0.56	0.25	0.21	0.08/0.03	0.04	0.27	0.09	0.07	0.03/0.01	0.02
PLATO	0.72	0.44	0.34	0.11/0.09	0.10	0.50	0.29	0.21	0.06/0.03	0.06
PMT	0.79	0.55	0.40	0.16/0.12	0.11	0.52	0.31	0.27	0.09/**0.07**	0.09
Baichuan-7B	0.82	0.55	0.42	0.15/0.08	0.12	0.55	0.30	0.28	0.09/**0.07**	0.08
ChatGLM-6B	0.80	0.51	0.39	0.15/0.07	0.09	0.49	0.27	0.25	0.07/0.05	0.06
QWen-7B	0.83	0.57	0.44	0.16/0.09	0.13	0.52	0.31	0.31	0.10/**0.07**	0.09
PLATO-prompt	0.82	0.57	0.43	**0.21**/0.14	0.17	**0.56**	**0.33**	**0.35**	**0.12/0.07**	**0.11**
w/o task	0.74	0.46	0.41	0.19/0.11	0.12	0.09	0.07	0.05	0.02/0.01	0.07
PLATO-prompt∗	**0.86**	**0.59**	**0.46**	0.20/**0.15**	**0.19**	**0.56**	**0.33**	**0.35**	**0.12/0.07**	**0.11**
w/o task	0.77	0.50	0.42	0.18/0.12	0.09	0.09	0.07	0.05	0.02/0.01	0.06

Table 7 DAP results on Mixed-FS in the few-shot setting and zero-shot setting. "w/o task" stands for PLATO-prompt without task-grounded pre-training. Best results are highlighted in bold.

Method	Few-shot					Zero-shot				
	TA	PA	F1	B1/B2	R	TA	PA	F1	B1/B2	R
BST	0.58	0.31	0.27	0.11/0.05	0.04	0.31	0.12	0.11	0.07/0.03	0.02
PLATO	0.82	0.51	0.40	0.17/0.09	0.12	0.56	0.32	0.28	0.09/0.04	0.09
PMT	0.87	**0.62**	0.46	0.20/0.14	0.15	0.60	0.37	0.31	0.13/0.09	0.09
Baichuan-7B	0.86	0.60	0.43	0.19/0.12	0.13	0.63	0.35	0.29	0.15/0.08	0.08
ChatGLM-6B	0.83	0.55	0.41	0.15/0.09	0.11	0.64	0.33	0.28	0.11/0.07	0.07
QWen-7B	0.88	0.59	0.49	0.18/0.11	0.16	0.61	0.28	0.26	0.10/0.07	**0.10**
PLATO-prompt	**0.89**	0.60	**0.51**	**0.22/0.15**	**0.18**	**0.67**	**0.42**	**0.41**	**0.16/0.11**	**0.10**
w/o task	0.80	0.53	0.44	0.21/0.14	0.12	0.08	0.07	0.03	0.01/0.01	0.06

6.3.2 *Result*

Tables 7 and 8 report the results of PLATO-prompt and all baselines. PLATO-prompt outperforms baselines in terms of almost all metrics, indicating better DAP ability. The performance gap between PLATO-prompt and PLATO-prompt without task-grounded pre-training demonstrates that using external dialogue data can help the model plan dialogue act more effectively. Moreover, by comparing "PLATO-prompt" and "PLATO-prompt∗", we find that Mixed-FS can facilitate the DAP for mixed-type dialogues in the few-shot or zero-shot setting.

6.4 Experiment results for NLG

6.4.1 *Evaluation metrics*

We conduct both automatic evaluation and human evaluation for our model PLATO-prompt and all baselines.

Automatic evaluation metrics. For automatic evaluation, we follow the setting in previous work [6] to use several common metrics such as F1, BLEU (B-1 and B-2) [32], ROUGE(R) [28], and DISTINCT

Table 8 DAP results on DuRecDial in the few-shot setting and zero-shot setting. "∗" and "w/o task" stand for PLATO-prompt fine-tuning with Mixed-FS and without task-grounded pre-training, respectively. Best results are highlighted in bold.

Method	Few-shot					Zero-shot				
	TA	PA	F1	B1/B2	R	TA	PA	F1	B1/B2	R
BST	0.52	0.32	0.24	0.10/0.03	0.05	0.27	0.09	0.08	0.04/0.01	0.02
PLATO	0.71	0.44	0.38	0.13/0.09	0.11	0.47	0.30	0.21	0.06/0.04	0.07
PMT	0.74	0.49	0.39	0.16/0.11	0.12	0.51	0.33	0.28	0.10/0.03	0.07
Baichuan-7B	0.77	0.51	0.36	0.12/0.09	0.11	0.51	0.34	0.27	0.09/0.04	0.06
ChatGLM-6B	0.75	0.44	0.33	0.11/0.07	0.10	0.49	0.31	0.25	0.08/0.03	0.06
QWen-7B	0.78	0.50	0.40	0.15/0.10	0.12	0.52	0.32	0.31	0.11/0.05	0.07
PLATO-prompt	0.80	0.51	0.43	0.18/0.15	0.15	**0.53**	**0.36**	**0.33**	**0.13/0.08**	0.10
w/o task	0.71	0.45	0.39	0.15/0.08	0.09	0.06	0.03	0.03	0.02/0.01	0.06
PLATO-prompt∗	**0.83**	**0.55**	**0.45**	**0.20/0.16**	**0.16**	**0.53**	**0.36**	**0.33**	**0.13/0.08**	**0.11**
w/o task	0.72	0.47	0.38	0.15/0.09	0.10	0.06	0.03	0.03	0.02/0.01	0.08

Table 9 NLG results on Mixed-FS in the few-shot setting. "w/o task" and "E2E" stand for PLATO-prompt without task-grounded pre-training and PLATO-prompt without task decomposition, respectively. Best results are highlighted in bold.

Method	Automatic metrics				Human metrics				
	F1	B-1/2	R	D-1/2	Flu.	Appr.	Info.	Proa.	Cohe.
BST	0.12	0.09/0.04	0.02	0.02/0.01	1.78	0.39	0.19	0.52	0.28
PLATO-2	0.26	0.17/0.13	0.09	0.08/0.06	**1.94**	0.71	0.39	0.98	0.64
PMT	0.29	0.18/0.13	0.11	**0.13/0.09**	1.93	0.77	0.41	1.03	0.66
Baichuan-7B	0.30	0.10/0.07	0.12	0.10/0.07	1.91	0.74	0.33	0.88	0.61
ChatGLM-6B	0.25	0.06/0.02	0.11	0.07/0.02	1.89	0.69	0.35	0.83	0.59
QWen-7B	0.32	0.09/0.06	0.11	0.09/0.05	1.93	0.78	0.38	0.79	0.65
PLATO-prompt (ours)	**0.35**	**0.24/0.19**	**0.15**	0.11/**0.09**	1.91	**0.82**	**0.46**	**1.08**	**0.69**
PLATO-prompt w/o task	0.21	0.09/0.06	0.10	0.06/0.04	1.92	0.57	0.31	0.91	0.51
PLATO-prompt-E2E	0.24	0.14/0.08	0.12	0.05/0.02	1.90	0.66	0.33	0.94	0.53

(D-1 and D-2) [29] to measure the relevance, fluency, and diversity of generated responses.

Human evaluation metrics. The human evaluation is conducted at the level of both turns and dialogues. For turn-level human evaluation, we ask each model to produce a response conditioned on a given context and a dialogue act. The generated responses are evaluated by three persons in terms of fluency, appropriateness, informativeness, and proactivity. For dialogue-level human evaluation, we let each model converse with humans when given dialogue acts and reference knowledge. For each model, we collect 20 dialogues. These dialogues are then evaluated by three persons in terms of coherence, which examines fluency, relevancy, and logical consistency of each response when given the current goal and context. The evaluators rate the dialogues on a scale of 0 (poor) to 2 (good) in terms of each human metric, please see Appendix A for more details. We calculate the averaged weighted kappa value for the human evaluation and get a high score of 0.75, demonstrating good consistency between the three annotators.

6.4.2 *Result*

As shown in Tables 9–12, our model outperforms baselines in terms of almost all metrics, indicating better dialogue generation ability. First, the performance gap between PLATO-prompt and PLATO-prompt-E2E demonstrates the effectiveness of task decomposition and modular-based architecture. Then, the performance gap between PLATO-prompt and PLATO demonstrates the effectiveness of prompt-tuning. Moreover, the performance gap between PLATO-prompt and PLATO-prompt without task-grounded pre-training demonstrates that using external dialogue data can help the model generate responses more naturally and coherently. Finally, by comparing "PLATO-prompt" and "PLATO-prompt∗", we find that Mixed-FS can facilitate the mixed-type dialogues generation in the few-shot setting.

6.5 Training sample analysis

We further study the impact of the number of training samples on our model. Specifically, we vary the number of training samples in the fine-tuning phase, as shown in Figure 3. We observe that with the

Table 10 NLG results on DuRecDial in the few-shot setting. "∗", "w/o task" and "E2E" stand for PLATO-prompt fine-tuning with Mixed-FS, without task-grounded pre-training, and without task decomposition, respectively. Best results are highlighted in bold.

Method	Automatic metrics				Human metrics				
	F1	B-1/2	R	D-1/2	Flu.	Appr.	Info.	Proa.	Cohe.
BST	0.09	0.08/0.05	0.03	0.03/0.01	1.71	0.37	0.21	0.55	0.22
PLATO-2	0.23	0.16/0.10	0.10	0.09/0.05	1.91	0.72	0.35	1.01	0.62
PMT	0.26	0.19/0.12	0.12	0.11/**0.09**	**1.96**	0.71	0.43	1.06	0.63
Baichuan-7B	0.28	0.10/0.05	0.10	0.08/0.03	1.92	0.70	0.39	0.87	0.60
ChatGLM-6B	0.26	0.05/0.02	0.12	0.08/0.01	1.90	0.66	0.36	0.79	0.55
QWen-7B	0.29	0.10/0.04	0.11	0.09/0.04	1.93	0.79	0.40	0.85	0.62
PLATO-prompt (ours)	0.30	0.22/0.15	0.15	0.11/0.08	1.93	0.77	0.45	1.10	0.65
PLATO-prompt w/o task	0.20	0.09/0.06	0.09	0.05/0.04	1.83	0.60	0.33	0.85	0.51
PLATO-prompt-E2E	0.22	0.10/0.08	0.10	0.05/0.03	1.85	0.57	0.31	0.90	0.49
PLATO-prompt∗ (ours)	**0.32**	**0.25/0.16**	**0.17**	**0.12**/0.08	1.93	**0.81**	**0.49**	**1.11**	**0.66**
PLATO-prompt∗ w/o task	0.19	0.08/0.06	0.10	0.06/0.04	1.86	0.61	0.35	0.88	0.53
PLATO-prompt∗-E2E	0.25	0.11/0.09	0.11	0.05/0.02	1.89	0.59	0.32	0.90	0.51

Table 11 NLG results on Mixed-FS in the zero-shot setting. "E2E" stands for PLATO-prompt without task decomposition. Best results are highlighted in bold.

Method	Automatic metrics				Human metrics				
	F1	B-1/2	R	D-1/2	Flu.	Appr.	Info.	Proa.	Cohe.
BST	0.03	0.03/0.01	0.01	0.01/0.01	1.55	0.11	0.01	0.13	0.08
PLATO-2	0.19	0.11/0.09	0.06	**0.08**/0.05	1.85	0.39	0.17	0.63	0.23
PMT	0.21	0.13/0.10	0.08	0.07/0.05	**1.88**	0.42	0.16	0.65	0.27
Baichuan-7B	0.22	0.08/0.06	**0.09**	0.06/0.02	**1.88**	0.41	0.13	0.59	0.28
ChatGLM-6B	0.18	0.05/0.04	0.06	0.05/0.03	1.83	0.33	0.16	0.61	0.24
QWen-7B	**0.23**	0.10/0.07	0.07	0.08/0.03	1.82	0.39	0.18	**0.69**	0.23
PLATO-prompt (ours)	**0.23**	**0.16/0.12**	**0.09**	0.07/**0.06**	1.87	**0.51**	**0.21**	**0.69**	**0.31**
PLATO-prompt-E2E	0.16	0.09/0.06	0.06	0.03/0.02	1.79	0.37	0.14	0.58	0.22

Table 12 NLG results on DuRecDial in the zero-shot setting. "E2E" stands for PLATO-prompt without task decomposition. Best results are highlighted in bold.

Method	Automatic metrics				Human metrics				
	F1	B-1/2	R	D-1/2	Flu.	Appr.	Info.	Proa.	Cohe.
BST	0.02	0.04/0.01	0.02	0.01/0.01	1.51	0.16	0.03	0.09	0.07
PLATO-2	0.13	0.09/0.08	0.06	0.07/0.05	1.78	0.33	0.16	0.59	0.25
PMT	0.18	0.14/0.08	0.06	0.07/0.04	1.81	0.44	0.17	**0.62**	0.24
Baichuan-7B	0.20	0.09/0.07	0.08	0.06/0.03	1.79	0.49	0.11	0.51	0.23
ChatGLM-6B	0.16	0.06/0.04	0.06	0.06/0.03	1.71	0.41	0.15	0.55	0.20
QWen-7B	0.22	0.10/0.08	0.07	0.07/0.04	1.73	0.45	0.12	0.60	0.27
PLATO-prompt (ours)	**0.24**	**0.17/0.09**	**0.09**	**0.09/0.06**	**1.83**	**0.53**	**0.24**	0.58	**0.29**
PLATO-prompt-E2E	0.15	0.08/0.04	0.06	0.06/0.04	1.82	0.33	0.13	0.55	0.19

increase of training samples, almost all metrics are improving, but the range of improvement is getting smaller and smaller. It might be explained that more dialogue samples can enhance the performance of PLATO-prompt.

7 Conclusion

In this paper, We identify the challenge of few-shot mixed-type dialogue generation. We introduce a pre-trained generation framework called PLATO-prompt, which utilizes a modular-based architecture and a prompt-tuning component. We demonstrate the effectiveness of our proposed framework by comparing it with baseline models. Additionally, we present a new benchmark dataset for mixed-type dialogues. This dataset combines persona-chat with conversational recommendation or a task-oriented dialogue within a single dialogue session, providing a natural and comprehensive setting for future research in this domain.

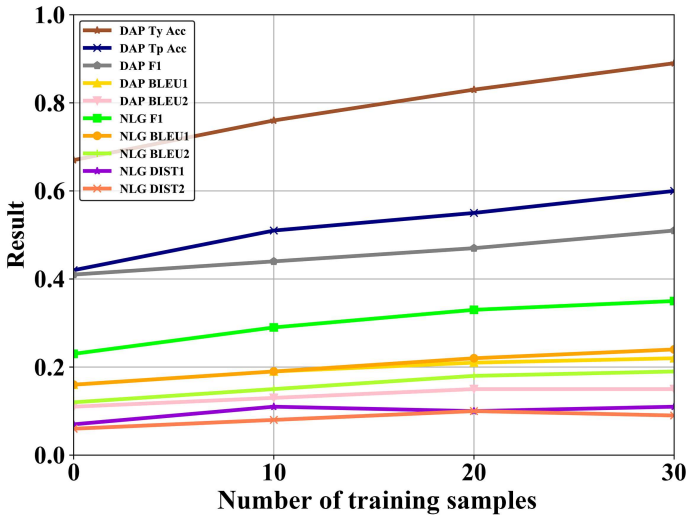

Figure 3 (Color online) Sensitive analysis of the number of training samples.

In our future work, we plan to explore an even more challenging scenario: mixed-type dialogues in the zero-shot setting. This setting involves generating dialogues without any prior exposure to specific types of dialogue tasks or prompts. By tackling this bigger challenge, we aim to further advance the capabilities of dialogue generation models.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2023YFF0725600) and National Natural Science Foundation of China (Grant Nos. 62236004, 62206078, 62441603). Thanks for the insightful comments and feedback from the anonymous reviewers.

References

1 Smith E M, Williamson M, Shuster K, et al. Can you put it all together: evaluating conversational agents' ability to blend skills. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 2021–2030
2 Shuster K, Ju D, Roller S, et al. The dialogue dodecathlon: open-domain knowledge and image grounded conversational agents. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 2453–2470
3 Andrea M, Lin Z J, Wu C S, et al. Attention over parameters for dialogue systems. 2020. ArXiv:2001.01871
4 Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. 300–325
5 Lin Z, Madotto A, Bang Y, et al. The adapter-bot: all-in-one controllable conversational model. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 16081–16083
6 Liu Z M, Wang H F, Niu Z Y, et al. Towards conversational recommendation over multi-type dialogues. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 1036–1049
7 Xu X C, Gou Z B, Wu W Q, et al. Long time no see! Open-domain conversation with long-term persona memory. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2022, 2022. 2639–2650
8 Zhou H, Zheng C J, Huang K L, et al. KdConv: a Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 7098–7108
9 Quan J, Zhang S A, Cao Q, et al. RiSAWOZ: a large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 930–940
10 Liu Z M, Xu J, Lei Z Y, et al. Where to go for the holidays: towards mixed-type dialogues for clarification of user goals. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 1024–1034
11 Sun K, Moon S, Crook P, et al. Adding chit-chat to enhance task-oriented dialogues. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 1570–1583
12 Chiu S, Li M L, Lin Y T, et al. SalesBot: transitioning from chit-chat to task-oriented dialogues. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 6143–6158
13 Dodge J, Gane A, Zhang X, et al. Evaluating prerequisite qualities for learning end-to-end dialog systems. 2016. ArXiv:1511.06931
14 Kang D, Balakrishnan A, Shah P, et al. Recommendation as a communication game: self supervised bot-play for goal-oriented dialogue. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. 1951–1961
15 Moon S, Shah P, Kumar A, et al. OpenDialKG: explainable conversational reasoning with attention-based walks over knowledge graphs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 845–854
16 Zhou K, Zhou Y H, Zhao W X, et al. Towards topic-guided conversational recommender system. In: Proceedings of the 28th International Conference on Computational Linguistics, 2020. 4128–4139
17 Hayati S A, Kang D, Zhu Q, et al. Inspired: toward sociable recommendation dialog systems. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 8142–8152
18 Zhang S Z, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 2204–2213
19 Xu X C, Gou Z B, Wu W Q, et al. Long time no see! Open-domain conversation with long-term persona memory. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 2639–2650
20 Mahdi N, Papangelis A, Tür G, et al. Language model is all you need: natural language understanding as question answering. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. 7803–7807

21 Mi F, Wang Y, Li Y. CINS: comprehensive instruction for few-shot learning in task-oriented dialog systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 11076–11084
22 Peng B L, Zhu C G, Li C Y, et al. Few-shot natural language generation for task-oriented dialog. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020, 2020. 172–182
23 Lin Z J, Liu B, Moon S, et al. Leveraging slot descriptions for zero-shot cross-domain dialogue state tracking. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 5640–5648
24 Zheng C, Huang M. Exploring prompt-based few-shot learning for grounded dialog generation. 2021. ArXiv:2109.06513
25 Bao S Q, He H, Wang F, et al. PLATO-2: towards building an open-domain chatbot via curriculum learning. In: Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021. 2513–2525
26 Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021. 3045–3059
27 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5998–6008
28 Lin C Y. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81
29 Li J W, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. 110–119
30 Du Z X, Qian Y J, Liu X, et al. GLM: general language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. 320–335
31 Bai J Z, Bai X, Chu Y F, et al. Qwen technical report. 2023. ArXiv:2309.16609
32 Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002. 311–318

Appendix A Human evaluation guideline

Fluency measures the fluency of each response:

- score 0 (bad): unfluent and difficult to understand.
- score 1 (fair): there are some errors in the response text but still can be understood.
- score 2 (good): fluent and easy to understand.

Appropriateness examines the relevancy of each response when given the current goal and local context:

- score 0 (bad): not relevant to the current goal and context.
- score 1 (fair): relevant to the current goal and context, but using some irrelevant knowledge.
- score 2 (good): otherwise.

Informativeness examines how much knowledge (goal topics and topic attributes) is provided in responses:

- score 0 (bad): no knowledge is mentioned at all.
- score 1 (fair): only one knowledge triple is mentioned in the response.
- score 2 (good): more than one knowledge triple is mentioned in the response.

Proactivity measures how well the model can introduce new topics with good fluency and relevance:

- score 0 (bad): some new topics are introduced but irrelevant to the context.
- score 1 (fair): no new topics/knowledge is used.
- score 2 (good): some new topics relevant to the context are introduced.

Coherence measures fluency, relevancy, and logical consistency of each response when given the current goal and global context:

- score 0 (bad): more than two-thirds of responses are irrelevant or logically contradictory to the given current goal and context.
- score 1 (fair): more than one-third of responses are irrelevant or logically contradictory to the given current goal and context.
- score 2 (good): otherwise.