LiveChat: A Large-Scale Personalized Dialogue Dataset Automatically Constructed from Live Streaming

Jingsheng Gao[1,2]*, Yixin Lian[2], Ziyi Zhou[2], Yuzhuo Fu[1]†, Baoyuan Wang[2]†
[1] School of SEIEE, Shanghai Jiao Tong University, China
[2] Xiaobing.AI
{gaojingsheng, yzfu}@sjtu.edu.cn
{lianyixin, zhouziyi, wangbaoyuan}@xiaobing.ai

Abstract

Open-domain dialogue systems have made promising progress in recent years. While the state-of-the-art dialogue agents are built upon large-scale text-based social media data and large pre-trained models, there is no guarantee these agents could also perform well in fast-growing scenarios, such as live streaming, due to the bounded transferability of pre-trained models and biased distributions of public datasets from Reddit and Weibo, etc. To improve the essential capability of responding and establish a benchmark in the live open-domain scenario, we introduce the LiveChat dataset, composed of 1.33 million real-life Chinese dialogues with almost 3800 average sessions across 351 personas and fine-grained profiles for each persona. LiveChat is automatically constructed by processing numerous live videos on the Internet and naturally falls within the scope of multi-party conversations, where the issues of Who says What to Whom should be considered. Therefore, we target two critical tasks of response modeling and addressee recognition and propose retrieval-based baselines grounded on advanced techniques. Experimental results have validated the positive effects of leveraging persona profiles and larger average sessions per persona. In addition, we also benchmark the transferability of advanced generation-based models on LiveChat and pose some future directions for current challenges. [1]

1 Introduction

Building dialogue systems to converse naturally with humans has been one of the longest-running goals in artificial intelligence (Zhou et al.; Roller et al., 2021). To usher that chatbot response properly in diverse scenarios, it is desirable to train a conversational agent based on massive large-scale datasets with multiple domains. Current dialogue datasets mainly leverage online forum posts to build reply-to relationships between users, such as Reddit (Mazaré et al., 2018; Zhong et al., 2020) and Weibo (Zheng et al., 2019; Qian et al., 2021). Despite the scalability and diversity of current dialogue corpora, dialogue models pre-trained on these conversation datasets can not perform effectively when applied to a completely new domain, such as live streaming. The reason lies in the intrinsic domain gap between online-post constructed data and those required in downstream conversational tasks. Even recent state-of-the-art (SOTA) dialogue models built upon large pre-trained language models (PLMs) like LaMDA (Thoppilan et al., 2022) and ChatGPT[2] heavily rely on publicly available text-only data. These large pre-trained models' distributions remain different across domains (Zeng et al., 2022) and are distinct from those of models learning the information contained in other modalities, video as an example.

Video is also an important dialogue data source in the wild with great diversity. As a form of popular video-based conversations, streaming is a broadcasting scenario that transcribes and broadcasts at

Figure 1: A session example of LiveChat. A streamer will respond to one audience's comment from the comments area.

*Work done during an internship at Xiaobing.AI
†Corresponding Author
[1]The code and dataset will be publicly available at https://github.com/gaojingsheng/LiveChat.
[2]https://openai.com/blog/chatgpt

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15387–15405
July 9-14, 2023

Dataset	Data Source	Dialogues	Persona	Addressee	Avg. Sessions	Language
PersonaChat (Zhang et al., 2018b)	Crowdsourced	10,907	✔	✗	8.69	English
PCR (Mazaré et al., 2018)	Online Posts	700,000,000	✔	✗	53.0	English
PersonalDialog (Zheng et al., 2019)	Online Posts	20,830,000	✔	✗	6.64	Chinese
PEC (Zhong et al., 2020)	Online Posts	355,000	✔	✗	26.0	English
PchatBot (Qian et al., 2021)	Online Posts	198,875,796	✔	✗	7.58	Chinese
MSC (Xu et al., 2022b)	Crowdsourced	5,001	✔	✗	42.9	English
DuLemon (Xu et al., 2022c)	Crowdsourced	27,501	✔	✗	16.3	Chinese
Linux-IRC (Elsner and Charniak, 2008)	Online Chatroom	2,500	✗	✔	-	English
Ubuntu-IRC (Kummerfeld et al., 2019)	Online Chatroom	77,563	✗	✔	-	English
INTERVIEW (Majumder et al., 2020)	Interview Transcripts	105,000	✗	✗	-	English
RealMedDial* (Xu et al., 2022a)	Short Videos	2,637	✔	✗	44.7	Chinese
LiveChat (ours)	Live Videos	1,332,073	✔	✔	3795	Chinese

Table 1: Comparison between our dataset and other existing open-domain dialogue datasets (mainly for tasks of personalized dialogue generation and addressee recognition). * for the medical domain. Persona represents whether there are personal profiles in the dataset. Addressee means if the dataset contains reply-to labels for addressee recognition problem in MPCs. Avg. Sessions denotes the average session number per persona and - means it is not mentioned in the dataset. Note that LiveChat can automatically and continuously construct dialogue sessions from videos while other video-sourced works like RealMedDial depend on crowdworkers.

the same time, which involves entertainment, life-sharing, education and so on (Wongkitrungrueng and Assarut, 2020). Such video-based conversations are one of the main ways human beings spread and exchange information efficiently in their daily lives and are naturally in line with the way people communicate. They are also the desired sources of dialogue datasets that are vitally significant in training large-scale dialogue models for homologous downstream virtual human scenarios, such as Virtual YouTubers, Virtual Employees, and Virtual Celebrities. Nevertheless, works that extract data from online videos do not receive enough attention although video-sourced dialogues are more life-oriented and naturally abundant.

Current video-sourced spoken corpora can be separated into two main categories (Mahajan and Shaikh, 2021): scripted and unscripted. The former refers to planned dialogues such as movie and TV scripts (Danescu and Lee, 2011; Li et al., 2016). The latter means spontaneous conversations in real situations, for instance, the interview dataset of Majumder et al. (2020). However, these previous video-sourced dialogues can not meet the scale of training a satisfied chatbot, owing to the trouble of continuously obtaining and processing various kinds of videos, and troubles of extracting valid dialogue sessions from them. For example, it is challenging to build valid dialogue sessions automatically from movies without human annotators. Thus, a large-scale video-sourced dialogue dataset in live streaming is essential for facilitating research in this area. The live broadcast is a typical one-to-many chat scene, which generally involves one streamer and multiple audiences. The challenge of building such a dataset lies in retrieving the reply-to relationships between the streamers and audiences. Unlike post-based social media with clear links between posts and replies, the streamer's responses in the live scene have no explicit reply-to relationships with audiences' comments.

To tackle the aforementioned problems, in this paper, we propose a novel and automatic video-sourced dialogue-constructing method and build a large-scale personalized dialogue dataset from the live streaming domain, named **LiveChat**. It is a non-trivial work since this dataset originates from a video-based source, distinct from most previous text-sourced data. Meanwhile, as far as we know, this is almost the only work that can effectively and endlessly extract dialogue sessions from videos.

As illustrated in Huang et al. (2020), one of the main challenges of existing open-domain chatbots is lacking a consistent personality as these agents are trained over different dialogues each with no or limited speaker information, while LiveChat naturally contains distinctive persona features (especially for streamers). To promote research in this field, we collect publicly available information for each streamer and add manual annotations to create the persona profiles, with individual information anonymized for privacy concerns. Compared to the previous personalized dialogue datasets (Zhang et al., 2018b; Mazaré et al., 2018; Zheng et al., 2019; Zhong et al., 2020; Qian et al., 2021; Xu et al., 2022c), our dataset provides more

fine-grained persona profiles, and more importantly, the average session number of each speaker exceeds previous ones extraordinarily, as shown in Table 1. This proves to be beneficial for personalized dialogue modeling.

Moreover, live streaming is also a multi-party conversation (MPC) scene involving more than two interlocutors. An example of LiveChat is illustrated in Figure 1. During the streaming process, a streamer naturally has to recognize which audience to reply to. We collect public live videos and process the streamer's responses and all audiences' comments to form multiple sessions of dialogues where each session contains a streamer's response and multiple candidates of addressee comments. A reply-to-whom matching method is brought forward to accurately find the correct candidate for a streamer's response. In this way, we can leverage the reply-to-whom relationship to build datasets for two classical tasks: response modeling and addressee recognition. Our proposed two classical dialogue tasks in LiveChat can help solve the MPC problem in a unified dataset, essential for building a practical dialogue agent in live streaming.

To sum up, our main contributions are as follows:

- We propose a large-scale personalized dialogue dataset LiveChat with a unique automatic dialogue-constructing method for countless live streams in the wild. To the best of our knowledge, our LiveChat is not only the largest video-sourced dialogue dataset, which contains detailed persona profiles and the largest average sessions per persona, but also the largest MPC dataset for addressee recognition released to the community.

- Sufficient experiments on two benchmark tasks: Response Modeling and Addressee Recognition, prove that our persona selection method is beneficial and larger average sessions per persona do help the modeling of the dialogue. We design retrieval baselines with considerable performance on both tasks to facilitate further research and build more genuine live-domain dialogue systems.

- We further investigate transfer learning of generation models and illustrate that pre-trained dialogue models perform poorly under the video-sourced data after fine-tuning, while large PLMs exhibit richer informativeness but worse relevance under few-shot settings. This arouses the interest in exploring domain adaptation with large PLMs in such video-sourced datasets.

2 Related Work

Dialogue Datasets A qualified open-domain dialogue model is usually trained on sufficient supervised datasets. Due to the accessibility and characteristics of social media, the current large-scale open-domain dialogue datasets are mainly constructed from text-based social media, such as Reddit (Mazaré et al., 2018; Zhong et al., 2020), Douban (Wu et al., 2017), and Weibo (Qian et al., 2021). Besides, a large-scale dataset with persona annotations is essential in building a personalized dialogue system. The persona profiles utilized in current persona datasets can be generally classified into two categories: basic profiles and text profiles. The basic profiles in Zheng et al. (2019) and Qian et al. (2021) are composed of personality traits like age, gender, and location. The text profiles are mainly composed of crowdsourced (Zhang et al., 2018b; Xu et al., 2022c) or automatically collected (Mazaré et al., 2018; Zhong et al., 2020) descriptive persona sentences. In LiveChat, we collect more fine-grained basic profiles and text profiles, with extraordinarily larger average sessions per persona than in previous works.

Furthermore, multi-party dialogue datasets are crucial when occurring conversations consisting of more than two speakers. However, most existing MPC datasets (Danescu and Lee, 2011; Lowe et al., 2015; Firdaus et al., 2020) have no explicit reply-to-whom annotations, and thus can not be leveraged in addressee recognition. Elsner and Charniak (2008) manually group sentences of disentangled conversations into separated sessions in `Linux` IRC. Kummerfeld et al. (2019) propose a larger MPC dataset manually annotated with reply-to structure from the `Ubuntu` IRC channel, which extremely prompts the research in MPC problems. Our LiveChat naturally originates from a multi-party scenario, whose size also remarkably exceeds previous ones, credit to the automatically reply-to-whom matching method.

As for those spoken dialogue corpora (Xu et al., 2022a; Majumder et al., 2020; Li et al., 2016; Danescu and Lee, 2011), most are pre-scripted or manually transcribed, intrinsically difficult to scale up because of the restricted video- or audio-based

sources where people can effortlessly extract valid dialogue sessions.

Personalized Response Modeling Early works use explicit persona profiles from predefined information or implicit persona vectors from dialogue history to generate personality-coherent responses. Explicit models use persona descriptions, attributes, or extracted profiles to learn personalized response modeling. Kim et al. (2014) leverages a persona knowledge base to extract predefined triples and entities in a retrieval-based dialogue system. Qian et al. (2018) propose an explicit persona model to generate personalized responses based on a pre-specified user profile. Song et al. (2019) propose a memory-augmented architecture to exploit persona information from context to generate diverse and sustainable conversations. On the other hand, implicit methods like Zhang et al. (2019) generate consistent responses by maintaining certain features related to topics and personas, while Li et al. (2021) encodes all the dialogue history of a speaker into the implicit persona. Zhong et al. (2022) design a personality selecting module to obtain abundant and accurate persona information from the user dialogue history. In LiveChat, we leverage explicit persona information to maintain persona consistency.

Addressee Recognition Addressee recognition which is also named explicit addressee modeling aims at understanding who speaks to whom in a multi-party conversation. Previous works mainly focus on predicting the targeting addressee of the last utterance in one conversation (Ouchi and Tsuboi, 2016; Zhang et al., 2018a). Later on, a who-to-whom model for predicting all the missing addressees to understand the whole conversation was introduced by Le et al. (2019a). Gu et al. (2021) further leverages a pre-trained language model for learning this problem in a unified manner. We follow this learning paradigm, and furthermore, are able to investigate personalized addressee recognition in LiveChat attributed to the available persona profiles.

3 Dataset Construction

3.1 Dataset Overview

The raw data constructed in LiveChat are collected from Douyin[3] (Chinese Tiktok), one of the largest Chinese live streaming and short video platform with over 10 million streamers and around 800 million users. We selected 351 representative streamers that interact and chat with the audiences frequently. By capturing the publicly available streamers' live videos and the audiences' comments in the broadcast room for a long time, we retrieved massive video clips with a huge amount of comments.

The whole dialogue construction process is shown in Figure 2, consisting of three steps. The first two steps are to construct dialogue sessions by processing videos and matching audience comments with streamer responses, and the last step is to enrich the dataset with fine-grained persona profiles, including basic profiles and text profiles.

Algorithm 1 Dialogue construction through reply-to-whom matching method.

Input: The streamer responses $\mathcal{R}$ and audience comments $\mathcal{C}$; each sentence is accompanied with timestamp T; max response time interval Δt; length ratio threshold τ; matching function $\mathcal{F}$.
Output: Matched dialogues $\mathcal{D}$.

Step 1: $c_i \leftarrow \mathcal{C}$ ▷ *Traverse all comments*
$r_j \leftarrow \mathcal{R}$ where $0 \leq T_{rj} - T_{ci} \leq \Delta t$ ▷ *Traverse the responses during time interval*
$c_i \rightarrow \mathcal{M}_j$ if $\mathcal{F}(c_i, r_j) = 1$ ▷ *Record all matched comments of response j in a set* $\mathcal{M}_j$
Step 2: $r_m \leftarrow R$ ▷ *Traverse all responses*
if $\mathcal{M}_m \neq \oslash$, $c_n \leftarrow \mathcal{M}_m$ **then** ▷ *Traverse matched comments in reverse order.*
  if $r_m[-1] =$. or ? **then** ▷ *Detect if the response with an ending punctuation*
    if $\frac{len(r_m)}{len(c_n)} > \tau$ **then**
      $(c_n, r_m) \rightarrow \mathcal{D}$, break ▷ *Add matched dialogue pairs.*
    else $r_m \rightarrow r_{m+1}$ ▷ *Merge current response sentence into next one*
  else $r_m \rightarrow r_{m+1}$ ▷ *Merge current response sentence into next one*
return $\mathcal{D}$

3.2 Dialogue Construction

Firstly we have to collect the raw spoken texts of the streamers. Since the original data are in the form of video clips, we need to transcribe them into text utterances. A video format converter is utilized to extract the voice content. Then we leverage an automatic speech recognition (ASR) model[4] to transcribe these voice clips into texts with times-

[3] https://www.douyin.com

[4] https://www.volcengine.com

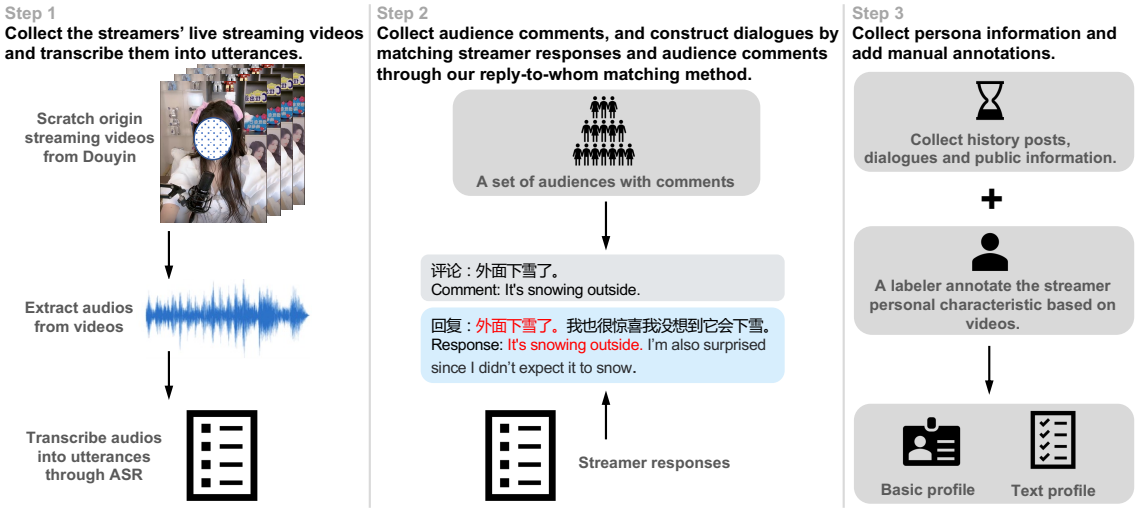

Figure 2: The whole construction process of LiveChat.

tamps, and this model is fine-tuned on a large-scale pan-entertainment dataset. Consequently, the raw data is transcribed into the streamer's spoken texts. Details of ASR are illustrated in Appendix A.

Secondly, we collect the raw audience comments and propose a reply-to-whom matching method to retrieve the reply-to relationships between streamers and audiences. Our proposed matching method is mainly based on the observations particularly apt to the streaming scenario: the streamer will reply to one audience in the comments area after that audience sent the message for a while. And usually, the streamer will repeat or summarize the audience's comment before responding to it, which helps the rest of the audiences understand what the streamer is talking about. We simply focus on extracting valid dialogue sessions based on the above observations and filter out others that are not satisfied. On this basis, the pseudocode of the whole matching process is illustrated in Algorithm 1. For each audience comment, we go through all the transcribed spoken utterances by the streamer within one minute. If there exists a repetition or summarization of this comment in the transcribed streamer's utterance, they will be recorded as a matched pair. Note that we apply a combination of BOW (bag of words) and pre-trained Chinese BERT (Cui et al., 2021) as the matching function. After retrieving the matched pairs, we iteratively concatenate the transcribed streamer's utterances to meet the ending punctuation and satisfy the required threshold τ for sufficient length, because the transcribed response from the ASR tool can sometimes be a broken sentence from what the streamer originally expresses. In addition, if a response matches several comments, we choose the closest one in time.

For each constructed dialogue pair, the response will repeat the comment. To prevent models from overfitting in this kind of manner, we remove the repetition prefix of each response. Besides, considering the specificity of this scenario, we filter out noisy pairs such as "谢谢**(Thanks to **)" or "欢迎**(Welcome **)" which miss valuable dialogue information. Finally, we can construct the dataset based on such matched pairs.

3.3 Persona Extraction

The last step is to construct detailed persona profiles in LiveChat, which are composed of basic profiles and text profiles. Following the work of PersonalDialog (Zheng et al., 2019) and Pchatbot (Qian et al., 2021), the basic profiles contain age, gender, and location. Except these, the basic profile in LiveChat also includes streamer characters and live room information such as live time, fans number, live streaming style, and so on. Part of this information can be retrieved from the live room or the streamers' homepages, besides, we crowdsource a set of questions and each annotator is required to label those missing contents by watching these streamers' streaming videos. Details about data privacy and annotators are elaborated in Ethical Consideration and Appendix A.

The text profile is composed of several sentences which describe the streamer's personal habits or characteristics. Sentences in the text profile are extracted in two ways: rules-based and classifier-

based. Similar to Mazaré et al. (2018) and Zhong et al. (2020), we collect persona sentences from all history spoken utterances and posts the streamer spoke or wrote on Douyin by rules. The final selected sentences must satisfy the following requirements: 1) between 4 and 20 words; 2) the contents include "我(I)"; 3) at least one verb; 4) at least one noun or adjective. Besides this, we train an additional persona classifier to further refine the text profiles. In detail, the classifier-based method means to discriminate if a single sentence contains persona facts by a learned classifier, which in our case is trained from DuLemon (Xu et al., 2022c).

3.4 LiveChat

We combine each pair of audience comments and streamer responses along with each streamer's corresponding persona to create LiveChat, the first large-scale personalized dialogue dataset from the live streaming domain. It is worth noting that each session in LiveChat contains not only the pairs of comments and responses but also several comments candidates within the same period, details illustrated in the appendix A. Although the LiveChat we discussed in this paper consists of single-turn-only dialogues, the multi-turn dialogues can be easily built by continuously tracing the interaction between the streamer and the same audience in a range of time. Data privacy in LiveChat including persona profiles is assured by carrying out the transformation, deletion, and anonymization of personal information as illustrated in Ethical Consideration.

With LiveChat, we propose that two benchmark tasks should be considered: (1) Response Modeling; (2) Addressee Recognition. The matched dialogue pairs can be directly leveraged in response modeling, while the other candidates of comments can be grouped together for training the addressee recognition task.

4 Models

4.1 Task Definition

Response Modeling Suppose we have a dialogue dataset $\mathcal{D} = \{(C_i, R_i, P_i)\}_{i=1}^{n}$, where $\forall i \in 1, ..., n$, C_i is the input dialogue context, R_i is the response, and P_i is the corresponding persona profile for the respondent of C_i. The goal is to learn a dialogue model g from $\mathcal{D}$, where for any new input context C_j, g can generate a response R_j based on its given persona P_j.

Previous works chiefly include retrieval-based and generation-based methods. To study the quantitative influence of our proposed persona profiles, we apply the retrieval-based architecture for the main experiments. As for the study of the transferable performance of advanced models in LiveChat, most generation-based ones are investigated.

Addressee Recognition Given a streamer S_i with persona profile P_i, a response R_i, and a set of comments $C_{i1}, C_{i2}, ..., C_{im}$, where $\forall j \in 1, ..., m$, each comment C_{ij} is associated with an audience A_j. The goal is to recognize which C_{ij} (or A_j) the R_i targets. Note that the purpose of this task is to identify the appropriate addressee comment instead of the appropriate streamer reply in response modeling. Dataset details about the settings of candidate comments can be seen in Appendix A.

4.2 Architecture

To investigate how existing dialogue baseline models can be leveraged in LiveChat, we build three retrieval-based models for response modeling and addressee recognition. Besides, five generation-based pre-trained language models (PLMs) are taken into account to study transfer learning on LiveChat. Details of our utilized models in this paper are described below.

4.2.1 Retrieval-based models

CoBERT The overall architecture of our retrieval-based persona model is depicted in Figure 3, which is inspired by Zhong et al. (2020).

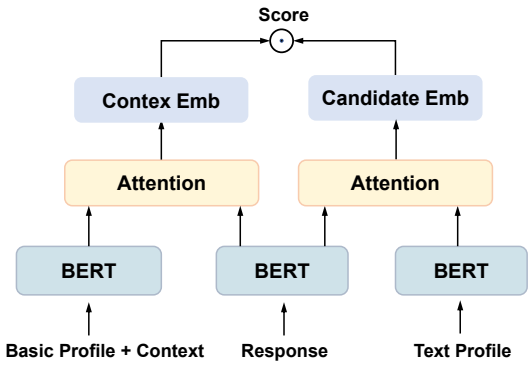

Figure 3: Our retrieval-based architecture.

We encode context, response, and text profile by separated BERT (Devlin et al., 2019). Given an input user context, we leverage the basic profile as the streamer's initialized embedding, and a [SEP] token is added between the basic profile and context. During our experiments, we only use the streamer

ID information instead of all annotations. As for the multiple text profile sentences, we concatenate them with [SEP] to meet the length of maximum input tokens. After retrieving three individual representations, two co-attention modules (Zhong et al., 2020) are implemented for better feature fusion. Finally, we obtain context embedding and candidate response embedding, then apply dot product to compute the matching score and calculate cross-entropy loss to optimize the full network.

TwinBERT Current advanced retrieval-based models can be generally classified into context-response matching double-stream frameworks (Humeau et al., 2019; Lu et al., 2020) and PLMs-based single-stream frameworks (Gu et al., 2020). To keep the bi-encoder model consistent with CoBERT, we also adopt the attention module into TwinBERT (Lu et al., 2020), but without extra inputs of persona profiles to compare the effects of personal information.

BERT BERT (Devlin et al., 2019) is a typical single-stream network. The interaction and aggregation operations can be performed in a unified way by feeding the concatenation of the context and the response candidate into the model. During the inference stage, we can sort the output scores between the context and all response candidates to finally obtain the matched response. Note that in experiments of CoBERT, TwinBERT, and BERT, we use the pre-trained BERT checkpoint of the Chinese version.

4.2.2 Generation-based models

BART (Shao et al., 2021) is a denoising autoencoder for pre-training sequence-to-sequence model and pre-trained by reconstructing the original text from the arbitrary corrupting text, which has been a universal transformer-based baseline PLM.

CDialGPT Wang et al. (2020) proposed a Chinese GPT pre-trained from a large version of the open-domain dialogue dataset. The dataset sources originate from Chinese online forums, such as Weibo and Douban.

EVA2.0 is an encoder-decoder PLM for open-domain dialogue modeling (Gu et al., 2022), whose architecture is similar to BART. This model is pre-trained on a 60GB high-quality dialogue dataset, which is composed of WDC-Dialogue (Zhou et al., 2021) and some extra copra, like movie scripts or crowdsourcing datasets. WDC-Dialogue is sourced from Chinese social media and is the main training dataset of EVA2.0.

GLM (Du et al., 2022) is a large-scale model based on autoregressive blank infilling to unify all language tasks. The original Chinese GLM owns 10 billion parameters pre-trained on a Chinese corpus.

GPT3 (Brown et al., 2020) is an autoregressive language model with 175 billion parameters, which has shown engaging performance on many NLP tasks and exhibits powerful abilities in multilingual zero-shot, one-shot, and few-shot settings.

5 Experiments

We train retrieval baselines for two tasks as described in Section 4.1: response modeling and addressee recognition. We also investigate transfer learning of current popular generation-based models on LiveChat. Experimental settings including training details and evaluation metrics can be found in Section B.

5.1 Results of Response Modeling

In this session, we fully investigate the influence of our persona profiles, the extraction methods for text profiles, and the impact of larger average sessions per persona. The main architecture follows the work of CoBERT (Zhong et al., 2020). Note that CoBERT without extra persona profile input is equal to TwinBERT (Lu et al., 2020).

Impact of Personas The test performance of retrieval-based response modeling is shown in Table 2. Obviously, CoBERT with text profile and basic profile achieves the best performance in our experimental settings, indicating both text profile and basic profile will facilitate the modeling of response. We attribute this to the fact that the basic profile is significant in denoting the corresponding speaker, and the text profiles include detailed personal descriptions which may have correlations with the candidate responses. An exclusive text profile achieves a higher score than a single basic profile, that is, detailed persona features of text profiles retrieve a more essential influence on model performance.

Impact of Average Sessions To study the influence of the length of average sessions per persona on the model performance, we conduct experiments on different settings of data scales and the number of persona IDs based on CoBERT along with complete persona profiles. Since the data scale is equal

Model	Recall@1	Recall@2	MRR
CoBERT	68.72	75.58	76.25
+ *text profile*	70.04	77.43	77.66
+ *basic profile*	69.43	76.58	77.06
+ *text & basic profile*	**72.18**	**79.58**	**79.63**

Table 2: Comparison of automatic evaluation metric results (%) among different retrieval-based settings.

Data Scale	ID Num	Recall@1	Recall@2	MRR
400k	150	**69.39**	**77.87**	**77.67**
100k	150	67.86	74.99	75.63
100k	50	67.65	75.95	75.95
100k	15	68.78	77.25	77.09
40k	150	64.01	71.57	72.50

Table 3: Test performance (in %) under different data scales and number of persona IDs.

to the persona ID number times the average session number by person, and the same number of persona IDs with larger data scales and the same data scales with fewer IDs both indicate that there are more average sessions per persona. To reduce the influence of different scales of training data and make a fair comparison, we also keep the same data scale (100k) while decreasing the number of IDs from 150 to 15 as shown in Table 3. We make sure the persona IDs of the test set are all seen before. Consequently, all of our testing persona IDs are incorporated into the training settings.

Experimental results demonstrate: (1) Obviously, more average sessions with the same number of IDs will enhance the model to capture the speaker's personalized response. (2) The average number of sessions is more significant than the number of IDs for response modeling. The priority of the number of sessions per persona also proves the superiority of our proposed dataset to other existing ones since LiveChat exceeds others extraordinarily in this indicator.

Influence of Text Profiles For the extraction of our text profiles, we empirically analyze the effect of different extraction methods as illustrated in Table 4. The *random from user* means we randomly select sentences by the streamer as his or her text profiles, and *random from dataset* refers to randomly selected in the whole dataset. The *Length* represents the maximum truncation length for all concatenated text profiles. We can see that the rules and classifier both improve the model performance, indicating rules can filter the noisy sentences to some extent and persona definition in `DuLemon` is effective for training a classifier to further refine text profiles. Besides, the increase in persona sentence length will also enrich persona profiles and improve the results.

Persona Selection	Length	Recall@1	MRR
-	0	69.43	77.06
rules + classifier	256	71.09	78.49
random from user	512	69.49	77.27
random from dataset	512	69.46	76.92
rules	512	71.07	78.55
classifier	512	71.19	78.61
rules + classifier	512	**72.18**	**79.63**

Table 4: Test performance (in %) among different persona selection methods.

Model	Recall@1	Recall@2	MRR
BERT	62.29	75.38	74.59
TwinBERT	58.76	72.52	71.92
CoBERT	59.27	73.04	72.43

Table 5: Test performance (in %) among different addressee recognition models.

5.2 Results of Addressee Recognition

Previous works (Gu et al., 2021; Le et al., 2019b) adopt BERT to classify the relationship between the streamer response and multiple user comments, and we adopt a similar approach with a step further to explore the benefits of persona profiles. TwinBERT, compared with BERT, is utilized to study the difference between single-stream and double-stream architecture, and CoBERT is for investigating the influence of our collected persona profiles.

Table 5 presents the results of addressee recognition. It shows that single-stream BERT outperforms double-stream TwinBERT. The reason is that by feeding the concatenation of the context and the response into a unified BERT, the interaction and aggregation operations can be performed through the attention mechanism sufficiently. Besides, CoBERT retrieves a better performance than TwinBERT, demonstrating our persona profiles are also beneficial to addressee recognition.

6 Transfer Learning

To further investigate the performance of the pre-trained dialogue model on our LiveChat, we fine-tune BART, Chinese CDialGPT, and EVA2.0 to study whether pre-trained dialogue corpora can contribute to the learning of our case. The latter two are trained on dialogue data from text-based

	Pre-trained model	Parameters	ROUGE1	ROUGE-L	BLEU1	BLEU4	+2	+1	+0	Score
Fine-tuning	BART	220M	**31.64**	**29.95**	**35.02**	**12.46**	3.2%	**81.4%**	**15.4%**	0.878
	EVA2.0	300M	25.18	23.29	31.60	8.25	1.5%	67.6%	30.9%	0.706
	CDialGPT	104M	18.98	17.42	28.54	7.42	2.9%	38.5%	58.6%	0.443
1-Shot	GLM	10B	18.44	16.99	29.48	7.26	12.6%	61.7%	25.7%	0.868
	GPT3	175B	13.87	12.10	23.98	5.84	11.4%	56.3%	32.3%	0.791
8-Shot	GLM	10B	20.72	19.22	28.78	7.70	**14.9%**	65.0%	20.1%	**0.949**
	GPT3	175B	18.87	16.80	29.05	7.69	10.8%	66.3%	22.8%	0.880

Table 6: Automatic and human evaluations from different pre-trained generative models. The 2/1/0 score schema is elaborated in Appendix B.2. Score is the average score.

social media. Furthermore, we conduct in-context learning on GLM and GPT3 to explore the few-shot transferability of large language models (LLMs) on this video-sourced dataset. The data utilized in Table 6 and Figure 4 are dissimilar, and the details of the training data as well as our in-context templates are expounded upon in Appendix B.1.

Table 6 shows the results. First, the performance of BART is better than EVA2.0 and Chinese DialGPT. It confirms that the domain of our LiveChat is far away from the domains of those dialogue datasets utilized in existing pre-trained dialogue models. Therefore, it is challenging work to directly transfer from models trained on other dialogue domains. LLMs, nevertheless, offer a solution to this problem due to their great ability to generalization. Although the automatic evaluation results of fine-tuned models are better than LLMs by the reason that fine-tuning enables the models to learn the intrinsic distribution of LiveChat. We discover that the percentage of score 2 in human evaluation results of LLMs is dramatically larger than fine-tuned ones, which means better performance in terms of rich informativeness. We attribute this to the massive knowledge contained in LLMs and the few-shot demonstrations to elicit such knowledge. Yet despite this, we see a performance gap in score 1 with BART, which indicates a large room to increase contextual coherence through ways like parameters-efficient domain adaptation of LLMs to LiveChat, simultaneously maintaining their original powerful capabilities.

As a supplement, we also have performed a series of experiments of in-context learning on different shots to study the influence of demonstrations. The ROUGE1 and BLEU1 results are depicted in Figure 4. The performances keep growing as the shots gradually increase. However, when the number of demonstrations exceeds 8 shots, the performances of the LLMs slightly decrease due to the random manual selection of demonstrations.

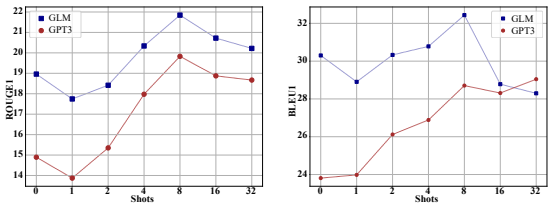

Figure 4: In-context learning results of GLM and GPT3 on different shots.

7 Conclusion

In this paper, we propose LiveChat, a Chinese video-sourced and personalized dialogue dataset from the live streaming domain with detailed persona profiles. It maintains the largest average sessions per persona and is also the largest MPC dataset for addressee recognition since live streaming is a natural MPC scenario. This is achieved owing to the reply-to-whom matching method that enables automatically extracting dialogue sessions from live videos, while most video extraction methods can not. Experimental results on two benchmark tasks show that the selected persona profiles and the larger number of average sessions per persona are advantageous in learning the speaker's personalized response and addressee decision. In addition, the comparisons between BART with other pre-trained dialogue models and LLMs have unveiled the distinctiveness of this video-sourced dialogue domain and we expect further research on parameters-efficient transfer learning of LLMs for LiveChat.

Limitations

There exist some limitations in our work. LiveChat is a Chinese-originated dataset involving unique cultures and abundant replying styles. However, this intensifies the difficulty of fully understand-

ing the content of this dataset. Fortunately, the same data construction pipeline can be applied to streaming platforms of other languages, like TikTok. And currently, our LiveChat is only sourced from 351 streamers on Douyin, not sufficient to train a general chatbot. We believe that LiveChat helps get one's foot in the door to the wonderful and diversified live scenarios and a dialogue model pre-trained on the considerable amount of video-sourced dialogue data among cross-platforms is promising. Besides, LiveChat contains some noisy spoken language segments that are not easy to read after transcribing from the ASR tool. The upper bound data quality is limited by such third-party tools. The future work to concatenate such text segments to restore the content of the original expression by streamers is highly anticipated. As for the dialogue-matching method, we simply implement a combination of BOW and BERT for semantic matching, which needs further optimization.

Other limitations from the training perspective can also be highlighted. For example, contextual background information is not considered in our modeling. That includes history dialogues in multi-turn settings and information from other modalities, like the streamer eating in front of the camera. In addition, we have not explored enough of our annotated basic profiles. In our primary experiments, we found that directly adding basic information such as age, gender, location, and other room information has limited influence on the model performance. We account for the fact that these basic profiles have limited connections with reply styles and contents in LiveChat. Also, note that we remove the repetition part of a streamer's response before training, while it is useful to maintain this pattern in practical application.

Ethical Consideration

This work presents LiveChat, a free and open Chinese dataset for the research community to study personalized open-domain dialogue generation and addressee recognition. Our dataset contains well-processed dialogues, and annotations (basic profiles and text profiles).

Data Privacy The original live-streaming clips and streamers' profiles of LiveChat are collected from Douyin, one of the largest Chinese live-broadcasting platforms. Similar to previous dialogue data from Reddit (Mazaré et al., 2018) and Weibo (Qian et al., 2021), LiveChat is an open-domain dialogue dataset that crossover multiple topics and users. Since all streamers must comply with platform rules during their online live streaming under the strict supervision of the Chinese government, their topics do not contain any pornographic, violent, reactionary, or discriminatory statements. Besides, due to the property of streaming, historically broadcast videos are no longer available when finished. Therefore it is not traceable from LiveChat to the identity of real streamers. Moreover, we clean the raw data with transformation, anonymization, and deletion to ensure there is no disclosure of private information and the identity of the streamers or audiences can not be inferred from it. Thus, all the collected data (including persona profiles) is publicly available and does not contain any private information of streamers and audiences, such as emails, phone numbers, and real user names. Although we collect the Age and Location information, in our basic profile, the Age is expressed as an interval range that doesn't represent the real age of the streamers, and the Location only contains the province's information. Besides, all the attributes of our basic profiles are re-indexed as numbers in the final released dataset. Thus, both our raw data and persona profiles do not create additional ethical risks. Moreover, we are sure that all the collected data is consistent with the platform usage rules and protocols. LiveChat will only be allowed to be used for academic research. At last, our construction of LiveChat was approved by an internal review board (IRB).

Annotators In terms of basic profile annotation and manual evaluation, all the annotators are Chinese undergraduates specifically responsible for annotation work in our institution. They are informed of the ongoing research and well known the way the curated data will be used. All the annotated information and evaluation results do not contain any private information.

References

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Cristian Danescu and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

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

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2014. Acquisition and use of long-term memory for personalized dialog systems. In *International workshop on multimodal analyses enabling artificial agents in human-machine interaction*, pages 78–87. Springer.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019a. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. 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 1909–1919, Hong Kong, China. Association for Computational Linguistics.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019b. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. 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 1909–1919, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–25.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2645–2652.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: A large-scale open-source corpus of media dialog. *arXiv preprint arXiv:2004.03090*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143, Austin, Texas. Association for Computational Linguistics.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2470–2477.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Apiradee Wongkitrungrueng and Nuttapol Assarut. 2020. The role of live streaming in building consumer trust and engagement with social commerce sellers. *Journal of Business Research*, 117:543–556.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Bo Xu, Hongtong Zhang, Jian Wang, Xiaokun Zhang, Dezhi Hao, Linlin Zong, Hongfei Lin, and Fenglong Ma. 2022a. RealMedDial: A real telemedical dialogue dataset collected from online Chinese short-video clips. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3342–3352, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022b. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022c. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Andy Zeng, Adrian S. Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Peter R. Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv*, abs/2204.00598.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018a. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93.

A Dataset Construction Details

Our constructed dataset are composed of 1332073 dialogues, and each dialogue consists of one streamer response and several audience comments. The overall statistics of the LiveChat and raw data are illustrated in Table 7.

Details of Automatic Speech Recognition Our HuoShan ASR tool is from Chinese company ByteDance. The ASR is pretrained on a large entertainment dataset that includes domains such as fashion, food, games, and singing. After testing on a 64k Chinese video-based recognition dataset from various domains, the ASR achieved a Character Error Rate (CER) of 3.17%.

Dialogue samples in response modeling In response modeling, we select all the matched dialogue pairs from our raw conversation dataset. Several constructed dialogue cases are shown in Figure 5. Each audience comment is associated with a streamer response. During our retrieval-based response modeling experiments, given an audience comment, all the responses in one batch are negative responses.

Persona Annotations Our persona annotations include the basic profile and text profile, and a persona profile sample of one streamer is shown in Figure 6. Text profiles are collected from the history posts and dialogues based on the rules and a persona classifier, and basic profiles are collected and annotated by crowdworkers who are native Chinese speakers and familiar with live streaming. Apart from the basic information on the streamer's

Category	Size
Raw Audiences Comments	13,745,577
Raw Total Video Num.	182,943
Raw Total Videos Hours	30,248
Raw Streamer Sentences	35,475,979
Dialogues	1,332,073
Utterances	9,416,573
Streamer Num.	351
Audience Num.	1,058,595
Avg. Sessions per Streamer	3,795
Avg. Length of Utterances	10.22
Avg. Sentences of Text Profiles	69

Table 7: Statistic of LiveChat.

homepage, the crowdworkers are required to label some extra information that may have an influence on the streamer's speaking style. We present our annotation interface in Figure 7. For each streamer, the annotator is required to answer these questions based on the provided live streaming videos.

Selection of candidate audiences A streamer in LiveChat will respond to one audience selectively, and the segmentation of all audience comments is shown in Figure 8. We noted the timestamp of the matched comments and responses among all the comments. The comments between matched $(i-1)$-th comment and i-th comment are the candidate comments of the streamer's i-th response. In addressee recognition, the streamer aims to retrieve which comment among these candidates to respond to.

Figure 5: A conversation between one streamer and several audiences in LiveChat.

Basic Profile

Age: 18-24
Gender: Female
Location: Guangdong
Character: Active, Warm
Skill: Sing
Live Streaming Time: Forenoon
Audiences number: Less than 1000

Text Profile

1. 我长得像高中生。
I look like a high school student.
2.我觉得紫色好看。
I think purple is beautiful.
3.我是一个晚婚的人。
I am a late married person.
4.我喜欢吃手抓饼。
I like to eat finger biscuits.
5.我是广东人在广州。
I am Cantonese in Guangzhou.
6.我以前领养过一只小猫是朋友家猫妈妈生的。
I used to adopt a kitten from a friend's cat.
...

Figure 6: The annotated basic profile and collected text profile of one streamer. Note that in the final released dataset, all basic profiles are re-indexed as numbers for privacy concerns.

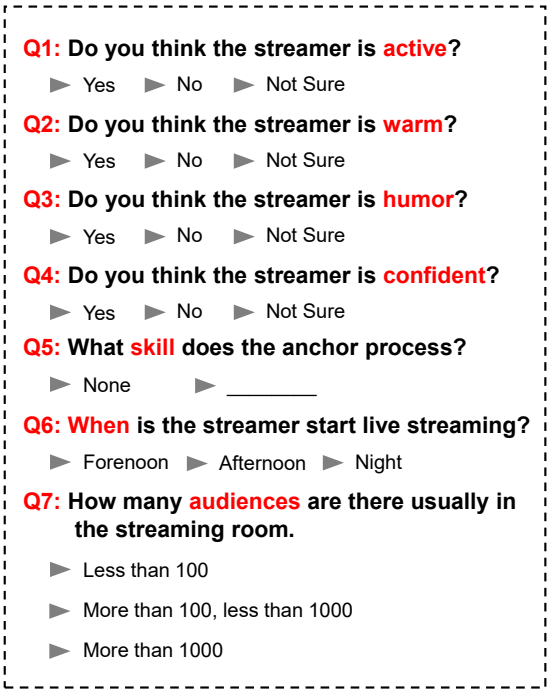

Figure 7: Annotation User Interface.

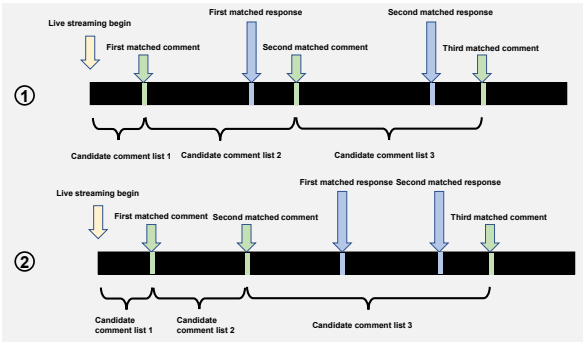

Figure 8: Segmentation for candidates of comments.

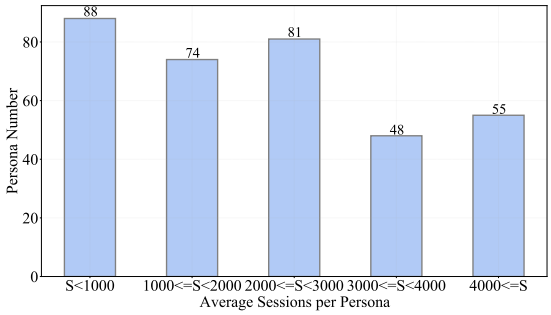

Figure 9: Session length distribution in LiveChat.

B Training and Evaluation Details

B.1 Training Details

Retrieval-based models Figure 9 provides the distribution of session length for each persona. There exist some persona IDs without enough sessions, thus we filter those IDs with more than 2400 sessions to study the influence of the average session number and persona profiles in a more clear setting. In this way, we retrieve 150 persona IDs in total. During our training process, we use 400k dialogues for training and 10k dialogues for testing in all retrieval-based dialogue experiments if there is no declaration before. The batch size is set to 24, which also means the size of the dynamic searching library of response modeling is 24.

In addressee recognition, the number of candidate comments ranges from one to hundreds. Thus, we process each session into one response and 10 candidate comments. If comments are too many, we select the last 10 comments, where the final sentence is the corresponding comment. And if the number of comments in one session is less than 10, we add comments in the front sessions to keep the total comment number to 10 in each session. The batch size we set here is also 24.

During training, we set the max input length and output length as 64, the max text profiles length as 512, and the epoch number and learning rate are set to 30 and 1e-5. All the experiments in the above two dialogue tasks are conducted on Nvidia Tesla V100s.

Generation-based models During the process of fine-tuning the pre-trained language models, we keep the most original experimental settings from their initial training parameters, and the utilized GPT3 version is text-davinci-002. In Table 6, the training dataset for fine-tuning is 400k, and the test dataset is 10k. Due to the cost of the GPT3 API, we only evaluate 1k samples for each experiment of GPT3 in Figure 4. In order to keep in line with GPT3, all data utilized in GLM is the same as GPT3. Thus, the results in Table 6 are inconsistent with those in Figure 4.

As for the in-context learning of GLM and GPT3, the template of n-shots is formulated as "我是一名线上直播间的主播，爱好是唱歌、与粉丝聊天等。以下是我在直播间和粉丝的互动。粉丝说：[CONTEXT-1]。我说：[RESPONSE-1]。...粉丝说：[CONTEXT-N]。我说：[RESPONSE-N]。以下是另一段我在直播间和粉丝的互动。粉丝说：[CONTEXT-TEST]。我说：[RESPONSE-TEST]" ("I am a streamer of an online live room, hobbies are singing, chatting with fans and so on. Followings are my interactions with fans in the live room. One fan says: [CONTEXT-1] I say: [RESPONSE-1] ... One fan says: [CONTEXT-N] I say: [RESPONSE-N]. Here is another interaction I have with my fans in the live room. One fan says: [CONTEXT-TEST] I say: [RESPONSE-TEST]).

The [CONTEXT-K] and [RESPONSE-K] ($0 < k <= n$) is the n-shot cases provided for LLMs. The [CONTEXT-TEST] and [RESPONSE-TEST] are the two utterances of one test dialogue pair, where the LLMs are required to return the [RESPONSE-TEST].

B.2 Metrics

Retrival-based **Recall@k** is a commonly used metric for evaluating whether the correct response exists among the top k candidates out of all the candidate responses. **MRR** (Mean Reciprocal Rank) is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries and is formulated as the average of the reciprocal ranks of results.

Generation-based **BLEU-n** measures the ratios of the co-occurrences of n-grams between the gen-

erated and real text. **ROUGE-n** measures the text quality by counting the overlapping n-grams between the generated and real text, and **ROUGE-L** means leveraging the longest common subsequence.

Human Evaluation We employ crowd workers to evaluate the responses generated by different models, and 1000 samples for each model. Our evaluation schema scores each sentence according to the following rules, inspired by Wang et al. (2020):

1. **Relevance** If a fluent response is logically consistent and relevant to the content of the comment, it will get 1. Otherwise, it will get 0.

2. **Informativeness** If a response has achieved the requirements of 1 score and is additionally rich in content, it will get 2.

C Case Study

To concretely demonstrate response ability in generative-based models, we further provide several sample cases (between BART, C-DialGPT, EVA2.0, GLM, and GPT3) in Table 8 and Table 9. BART, C-DialGPT, and EVA2.0 are finetuned on our LiveChat. LLMs present the results of 1-shot and 8-shot in-context learning.

Relevance Seen in these cases, all the responses generated by BART have strong connections with input contexts. Compared to EVA and C-DialGPT, BART maintains a higher relevance. For example, in Case 2-"走错直播间了？ (In the wrong live streaming room?)" from Table 9, we can find the response of C-DialGPT is not logically relevant to the comment, and the response of EVA2.0 is also not reasonable.

Informativeness Pre-trained models generally contain knowledge inside themselves. We can see that LLMs reply with more informative content in some cases, which means the richness and abundant knowledge of LLMs will be leveraged in dialogue generation.

Context	搞笑女来了。 The comedian girl has arrived.	我一个月的工资都刷给你了。 I have spent all of my salaries for the month on you.
BART	我不是搞笑女我是颜值主播。 I am not a comedian, I am a good-looking anchor.	你一个月工资多少啊? How much do you make a month?
EVA2.0	搞笑女孩我来了。 The funny girl is here.	你这个工资是不是很高啊。 Your salary is not very high.
C-DialGPT	可是我很少活的很有感情。 But I rarely live with much feeling.	你为什么可以刷到9,000? Why can you achieve 9,000?
GLM (1-shot)	对呀,我是搞笑女。 Yeah, I am a comedian girl.	你刷得有点多哦。 You are doing a little too much.
GPT3 (1-shot)	是谁来了? Who is there?	你辛苦了，我会努力争取更多粉丝。 Thank you. I'll try to get more fans.
GLM (8-shot)	搞笑女来了,快进来,快进来。 The comedian girl is here, come in quickly, come in quickly.	刷?你不许刷。 Spend? You are not allowed to do that.
GPT3 (8-shot)	你要搞笑，我可不搞笑. You want to be funny, I do not want to be funny.	你一个月工资都刷给我了？你不用上班了? You spent your entire month's salary on me? You don't have to work anymore?
Expected Response	不不皮了。 No, do not joke.	是听到我们的号召了吗? Is that for hearing the call of us?

Table 8: Generated responses by different pre-trained language models.

Context	跳个舞蹈吧。 Have a dance.	走错直播间了? In the wrong live streaming room?
BART	你们想看什么舞蹈? What dance do you want to see?	没有走错直播间。 Haven't gone to the wrong live streaming room.
EVA2.0	你给我跳个跳舞吧。 I want you to dance for me.	没有啊,他们都说我在直播。 No, they all say that I am live streaming.
C-DialGPT	感谢落落谢谢无心球。 Thank Luoluo, thank Wuxin Ball.	你好，仙人掌花开放26度。 Hello, the cactus is blooming at 26 degrees.
GLM (1-shot)	好呀好呀,不过呢,现在得先唱歌。 Great, however, now we have to sing first.	没有。 No.
GPT3 (1-shot)	不行，我不会跳舞。 No, I don't know how to dance.	你没有走错，这里是我的直播间， 你回去收看我的直播吧。 You did not go wrong, this is my live streaming room, please go back and watch my live streaming.
GLM (8-shot)	我现在心情不是很好。 I am not feeling well now.	走错直播间了吗? Have you gone into the wrong live streaming room?
GPT3 (8-shot)	跳个舞蹈要不然? Why not have a dance?	怎么你问这个，走错我的直播间了吗? Why are you asking this, did you accidentally go to the wrong live streaming room of mine?
Expected Response	我不跳要不你给我跳一个看看。 I don't dance and you can dance for me.	你没有走错，这是在下。 You haven't taken a wrong turn, this is me.

Table 9: Generated responses by different pre-trained language models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Section: Limitations

☑ A2. Did you discuss any potential risks of your work?
In Section: Ethical Consideration

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Abstract and Section 1: Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

In Sections 5 & 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Sections 5 & 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Sections 5 & 6 & Appendix

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

In Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
In Appendix

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
In Appendix

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
In Ethical Consideration

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In Appendix