Context Aggregation with Topic-focused Summarization for Personalized Medical Dialogue Generation

Zhengyuan Liu, Siti Umairah Md Salleh, Pavitra Krishnaswamy, Nancy F. Chen
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

In the realm of dialogue systems, generated responses often lack personalization. This is particularly true in the medical domain, where research is limited by scarce available domain-specific data and the complexities of modeling medical context and persona information. In this work, we investigate the potential of harnessing large language models for personalized medical dialogue generation. In particular, to better aggregate the long conversational context, we adopt topic-focused summarization to distill core information from the dialogue history, and use such information to guide the conversation flow and generated content. Drawing inspiration from real-world telehealth conversations, we outline a comprehensive pipeline encompassing data processing, profile construction, and domain adaptation. This work not only highlights our technical approach but also shares distilled insights from the data preparation and model construction phases.

1 Introduction

Medical dialogue systems hold significant potential for improving the efficiency of clinical workflows (Xu et al., 2021). As a specialized form of task-oriented dialogue, medical dialogue typically involves the completion of multiple tasks, including diagnosis, question answering, and consultation (Althoff et al., 2016; Tian et al., 2019; Xia et al., 2020; Gupta et al., 2020). There has been significant progress in this research field of the dialogue system in past years with the development of contextualized representation learning and neural language generation (Xu et al., 2019; Palanica et al., 2019). However, the general-purpose conversational interactive systems are proven to be inadequate, as they cannot adapt their responses to the unique medical histories and the diverse user preferences and personalities (Li et al., 2016; Mazaré et al., 2018). Personalized dialogue systems, tailored to the specific needs and characteristics of different users, can potentially bridge this gap (Ghosh et al., 2018; Schloss and Konam, 2020). By leveraging patient profiles, such as medical records, demographic information, and previous interactions, the personalized systems can facilitate more nuanced, empathetic, and context-aware conversations. This level of personalization not only enhances patient engagement and satisfaction, but also has the potential to improve healthcare outcomes by fostering adherence to treatment plans and providing tailored health education.

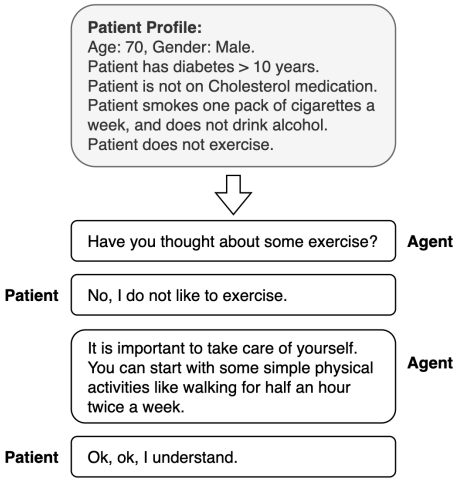

Figure 1: One dialogue example for "physical activity customized coaching" based on the personalized medical dialogue generation.

In this work, we conduct a case study on a clinical conversation scenario. Because of the chronic nature of diabetes and its associated complications, it requires constant attention and regular follow-up operation (Piette et al., 2000; Lawson et al., 2005). In practice, nurses schedule calls with patients to track their compliance status and health condition, provide general education, and customized coaching and lifestyle advice (Piette et al., 2001; Kivelä et al., 2014). To facilitate the communication process and deliver more efficient health management, the follow-up calls are organized according to a

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 310–321

medical protocol and telecarers adjust the conversation topics based on the patient's lifestyle management status and medication records (Kirkman et al., 1994; Taylor et al., 2003). This renders the follow-up call a representative use case for personalized dialogue generation. For instance, customized coaching is an effective patient education method (Kivelä et al., 2014), and its sub-topics are strongly correlated to the patient profile (as the example shown in Figure 1). The challenges of developing a personalized medical dialogue system come from three fundamental aspects: the lack of domain-specific data (Zhou et al., 2022); the complexity of modeling medical context and persona information (Liu et al., 2022a); and how to extensively evaluate the system (Abbasian et al., 2023). Moreover, due to the verbal nature of human spoken dialogues, the follow-up calls are often lengthy by covering various topics, which results in a low information density. The noisy long context also poses challenges for modeling and generation. We thus propose and adopt topic-focused summarization to distill and aggregate core information of the dialogue context, and use such information to guide the subsequent conversation flow and content generation.

In practice, to bootstrap the data-driven approaches, we construct a sample set derived from human spoken conversations, and we leverage the advancements in Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023b) for developing the dialogue system, which have demonstrated their exceptional language understanding and generation capabilities in the medical domain (Singhal et al., 2023). We add user profile information to produce personalized conversation, and improve the generation coherence based on topic-level context aggregation. Experiments show that our proposed method can substantially improve the generation quality, especially in the long context setting. This work not only highlights the technical approach but also shares distilled insights from the data preparation and model construction phases.

2 Related Work

Medical Dialogue Generation Medical dialogue systems aim to provide medical services for patients (Xu et al., 2021). As one specialized form of a task-oriented dialogue system, many previous studies focus on making diagnostic predictions after gathering patients' information of symptoms (Wei et al., 2018; Xu et al., 2019; Zhou et al., 2021), and healthcare counseling (Cao et al., 2019; Shen et al., 2020). Data-driven approaches and methods are proposed and applied for medical dialogue generation upon the development of large-scale medical dialogue datasets such as MedDialog (Zeng et al., 2020) and MedDG (Liu et al., 2022a), and the scarcity of domain-specific data still poses this task as a low-resource challenge (Lin et al., 2021).

Personalized Dialogue Systems One-size-fits-all approaches to human-machine communication have shown limitations in accommodating the diverse needs, preferences, and contexts of individual users. By contrast, personalized dialogue systems (Li et al., 2016; Mazaré et al., 2018) offer the potential to transcend these limitations by tailoring interactions to unique characteristics and requirements, thus raising much research interest. In particular, improving the modeling of persona or user information is one of the key points, and there are different approaches proposed in previous studies, such as explicitly utilizing pre-defined persona attributes to generate conditional responses (Qian et al., 2018; Olabiyi et al., 2019), constructing user embeddings to enhance personalized dialogue generation (Li et al., 2016; Chan et al., 2019), and building implicit user information from dialogue history (Al-Rfou et al., 2016; Ma et al., 2021).

Language Models as Conversational Agent Leveraging pre-trained language backbones for building conversation agents has seen remarkable progress recently (Liao et al., 2023), and the recent large language models have demonstrated impressive capabilities in both open-domain and task-oriented scenarios (Zhang et al., 2020; Thoppilan et al., 2022). Instruction tuning is one efficient and effective way to enable the conversational capabilities of large language models, such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023). It has been proved that using reinforcement learning with human feedback can further optimize language models for human-machine interaction, and the LLMs not only take conversation in a human-like manner, but also can do task solving and complex reasoning (Ouyang et al., 2022). Furthermore, LLMs demonstrate strong language understanding and generation capabilities in various downstream tasks that require certain domain knowledge (Wang et al., 2022; Hendrycks et al., 2020) (even in the zero-shot setting), which benefits from their large-scale pre-training (Touvron et al., 2023a).

Intent	Topic Type	Example
Information Gathering	Identification, Medical Experience, Appointments, Programme, Vitals, Insulin, Hyper/Hypo Incident, Base Compliance	[**Topic**: Vitals] **Nurse**: Can you tell me your blood sugar level four hours after dinner? **Patient**: If I remember correctly, it was around 13.4. **Nurse**: And what about your post-dinner reading? **Patient**: Ah, yes. After dinner, it was around 23 to 24, if I'm not wrong.
General/Customized Coaching	Self-Monitoring, Diet Management, Insulin, Physical Activity, General Education	[**Topic**: Diet Management] **Nurse**: From a dietary perspective, do you have any issues? **Patient**: no no **Nurse**: Are you okay with your diet? **Patient**: Yes, I'm fine. **Nurse**: Okay, good. A bit difficult, but you have to control it. **Patient**: I know, I have to be disciplined for my own health.
Other	Introduction, Social Chatting, Financial and Social Aid	[**Topic**: Social Chatting] **Nurse**: Never mind, this computer is taking a while to respond. **Patient**: Okay, Okay. **Nurse**: We'll have to wait for a bit. **Patient**: Ok, no problem.

Table 1: List of the dialogue topics and their intent categorization.

3 Personalized Dialogue Generation: Data Preparation & Refinement

In this work, we conduct a case study on personalized follow-up calls for diabetes patients. Diabetes is a chronic metabolic disorder characterized by abnormal glucose regulation, and effective management of diabetes is essential to mitigate its associated complications and improve patients' overall quality of life (Lawson et al., 2005). In practical use cases, the general-purpose messages may not adequately address the unique needs of individual patients. For example, customized coaching of physical activity should take into account factors such as the patient's age, comorbidities, lifestyle, and psychosocial aspects. By recognizing the heterogeneity of diabetes patients and offering tailored coaching interventions, it is useful for improving health management.

3.1 Raw Data Collection and Statistics

The raw data are extracted from call recordings of diabetes health management conversations (Liu et al., 2023) and fully anonymized.[1] Speech transcribers are employed for manual speech-to-text conversion to ensure quality. Speaker roles (e.g., nurse, patient, caregiver) are added to each utterance, and the informal and spontaneous styles of spoken dialogues such as back-channeling, hesitation, and repetition are preserved. The dialogue segmentation and topic categorization are manually performed.[2] Our linguistic annotators are familiar with clinical conversations, and have finished a training session on diabetes health management. Topic categories are built on the medical protocol refined by the healthcare provider. Moreover, there have been interactions for the corpus construction, where we collect feedback from nurses, refine the annotation scheme, and update the whole corpus.

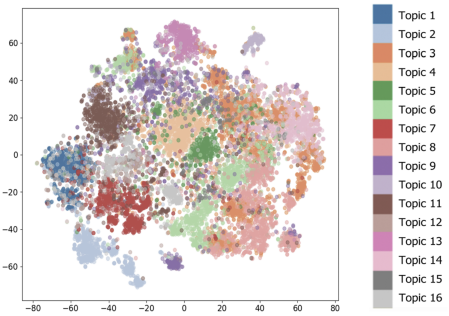

Figure 2: Feature visualization of segment embeddings via t-SNE. The colored points denote topically coherent segments labeled with different topics.

The transcribed dataset contains 856 transcripts. Depending on the patient's medical history and phases of the healthcare programme, nurses schedule their follow-up calls differently, and this results in length and topic variation. We obtain the segment representations from an unsupervised sentence embedding model (Gao et al., 2021), and use t-SNE (Van der Maaten and Hinton, 2008) to illustrate their distribution in a 2-dimensional space. As shown in Figure 2, dialogue utterances in different topics are semantically diverse and distinct. Moreover, there are two major types of dialogue

[1] This research study was approved by the SingHealth and A*STAR Institutional Review Boards. Participants enrolled in the healthcare programme consented to use of anonymized versions of their data for research.

[2] All dialogue examples in this manuscript are dummy data for demonstration purposes.

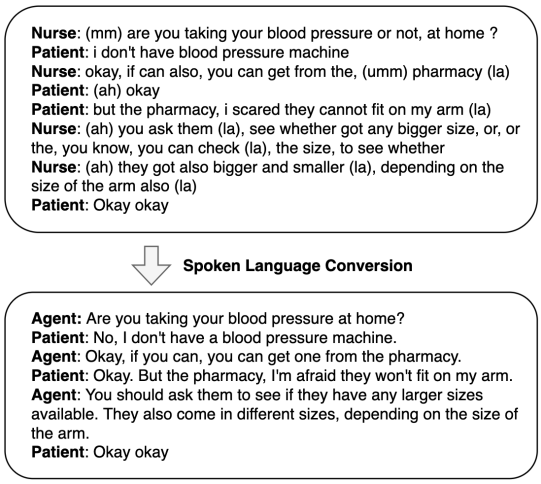

Figure 3: One dummy example of the spoken language conversion. Sentences are normalized and adjacent utterances with the same speaker are combined.

intent: information gathering and general and customized coaching. As shown in Table 1, there are four topics that are strongly related to customized coaching: physical activity, diet management, insulin, and self-monitoring, which usually shows a strong dependency on the dialogue context, as nurses will adjust the dialogue content based on the patient's response and feedback.

3.2 Spoken Language Conversion

While both human-human and human-machine medical conversations are task-oriented and topically organized, they demonstrate distinct linguistic characteristics, especially from the lexical and syntactic perspectives (Bernsen et al., 1996). More specifically, compared with real-world spoken dialogues, there is much less informal and colloquial wording in the human-machine interaction (Hill et al., 2015). Directly training on the raw transcripts will result in issues such as verbose sentences, unnecessary repetition, and incomplete utterances. Therefore, to improve the formality and readability of machine-generated responses, we conduct a spoken language conversion on the transcribed samples. As the example shown in Figure 3, there are three basic pre-processing steps: (1) We adopt an off-the-shelf text normalization model to process the utterances (Liu et al., 2022b). The colloquial sentences are paraphrased and the grammar errors are corrected. (2) We further normalize the utterances by reducing other common spoken language features such as repetition, pauses, and fillers. (3) To construct the turn-by-turn interaction for human-machine conversation, adjacent utterances with the same speaker are combined.[3] In our corpus preparation, we observe that the normalization step brings substantial changes in most utterances, and the processed sample set is significantly distinct from the raw dialogue data.

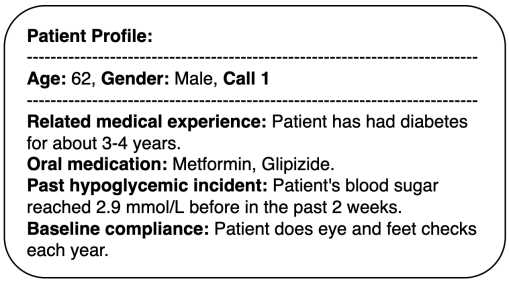

Figure 4: One dummy example of the patient profile. The basic information and summary from information gathering topics are collected.

3.3 Patient Profile Construction

Considering each patient's health condition and personal preferences, telecarers adjust their health management advice and provide general and customized coaching (Piette et al., 2000; Lawson et al., 2005). For instance, when discussing the type and frequency of physical activity, nurses should ask patients who have hypoglycemia symptoms to pay more attention to their sugar levels during exercise. Therefore, a feature-rich profile should include both basic demographic information, and up-to-date health condition of patients. To this end, aside from the basic information (e.g., age, gender, scheduled call phase) extracted from a structured database,[4] we also collect the key discussed points from the information gathering topics, as shown in Figure 4. In our clinical data, the gathered information from each follow-up call is recorded in a human-written summary. When such manually collected information is not available, automated approaches such as entity and event extraction can also be used for information extraction.

4 Context Aggregation via Topic-focused Summarization

Due to the complexity and verbose nature of human spoken dialogues (Sacks et al., 1978), and the necessity to cover multiple topics in clinical follow-up calls, nurse-to-patient conversations are

[3]Since our raw data contain topic-level annotation, we conduct the normalization process on each topic segment.

[4]Both language and structured data are fully anonymized, without any identifiable personal information.

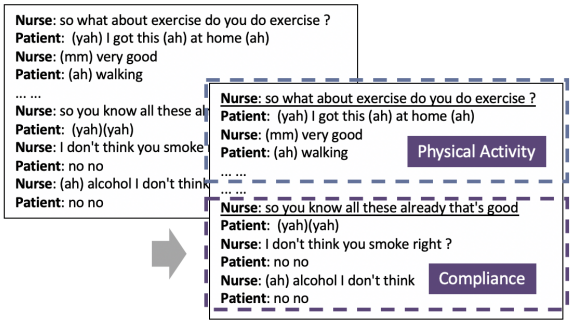

Figure 5: One dummy dialogue example in two topics. Frames indicate topically-coherent segments, and their corresponding label is highlighted.

often lengthy and thus characterized by lower information density than other document formats. For instance, in our transcribed calls, the maximum, median, and minimum utterance numbers are 1996, 221, and 21, respectively; the maximum, median, and minimum number of words are 16701, 1684, and 70, respectively. Nearly 5% samples (at the 95% quantile) are comprised of more than 800 utterances (6000 words). This requires models to precisely capture the core information from the long dialogue context and poses challenges for dialogue systems in both modeling and generation. In this work, we propose and adopt topic-focused summarization, to distill and aggregate the salient pieces from a noisy dialogue context. The refined context is then leveraged to guide the subsequent generation, and improve relevance and coherence. More specifically, we leverage the large language models to generate dialogue summaries for each dialogue snippet about a certain topic, and concatenate them as the history context. We conduct the following steps to build samples for training the data-driven approach:

4.1 Topic Segmentation and Categorization

First, each dialogue is processed with topic segmentation and topic categorization, as shown in Figure 5. This step is to parse the conversation into coherent segments, and helps identify the underlying structure of the dialogue. Here we use the manual annotated information in both the training and testing process: each training sample is to generate one coherent dialogue segment with a topic label and previous dialogue context, and it ends with a '*<topic-end>*' token for boundary modeling and a topic label of next segment prediction, which is a supervised approach for the dialogue topic modeling.

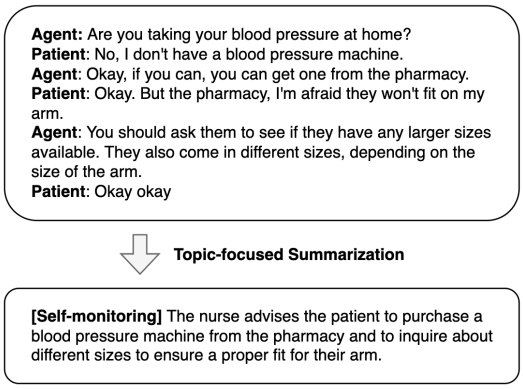

Figure 6: One dummy example of topic-focused summarization. The corresponding topic label is in brackets.

4.2 Topic-focused Summarization

For each identified segment, we then distill the core information by using a dialogue summarization model. In our preliminary study, we found that prompting large language models can produce reasonable dialogue summaries in the clinical scenario. We thus employ a state-of-the-art open model (i.e., Mistral-7B-Instruct-v0.2) for this step.[5] As shown in Figure 6, the summarizer is able to capture salient spans in the dialogue, and generate a concise version. Moreover, to better incorporate the dialogue topic information (Liu et al., 2019), we add their corresponding topic label before each summary.

4.3 Dialogue Generation Integration

The generated summaries serve as the historical context for the dialogue system. Since there is more than one topic segment in the conversation, we concatenate all summaries as one context and feed it into the system for subsequent generations. The response generation process is informed by a concentrated version of the dialogue history, emphasizing relevance and topic coherence. This enables the system to generate responses that are not only contextually appropriate but also enriched with the distilled essence of the prior conversation.

5 Personalized Dialogue Generation: Training & Evaluation

5.1 Task Definition

In a multi-turn human-machine conversation, we define C_i as the profile of the user i, and at a turn t, U_t is the user input and S_t is the system's response.

[5]The user prompt for the summarization step is *"Given the following nurse-patient dialogue about <topic-label>, please write a concise summary: <dialogue-content>."*

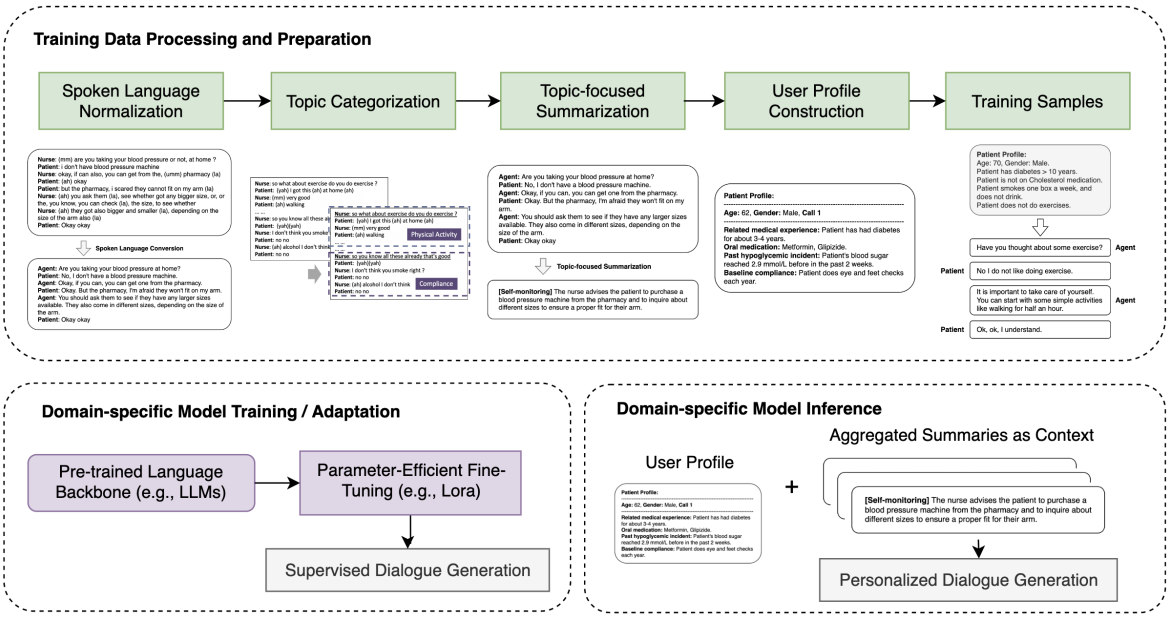

Figure 7: Overview of the pipeline for training and inference with personalized medical dialogue generation.

Basically, for modeling the dialogue history, all previous turns are concatenated and fed to the system as input: $H = [U_0, S_0, U_1, S_1, ..., U_{t-1}, S_{t-1}]$. In our framework of personalized dialogue generation with context aggregation, the user profile C and topic-focused summaries $H_{summary}$ are also part of context information. Therefore, at a turn t, the system's response S_t is conditioned on profile information C_i, summarized context $H_{summary}$, in-topic context H_{topic} and user's current utterance U_t, which are concatenated as a single sequence. To allow for handling descriptive profiles, we retain the profile C_i in the form of natural language text, in contrast to previous studies that encode the profile features via one-hot encoding and limit the model's accessibility to various features.

5.2 Adapting LLMs as Conversation Agents

Large language models have been shown to achieve remarkable performance across a variety of natural language tasks. Aside from their versatile capabilities of language understanding and generation where expert knowledge is not required, LLMs also show impressive results in medical document processing and decision support, and obtained comparable scores in medical examinations to human (Singhal et al., 2023). By learning from large volumes of text data to predict the subsequent tokens, LLMs with the auto-regressive framework can generate coherent, fluent, and reasonable responses to diverse prompts, and they are adopted as the conversation agents via in-context learning and instruction tuning (Chiang et al., 2023). To leverage the large-scale language backbone and adapt it to our domain-specific use case, we conduct experiments on some representative large language models, such as LLaMA (Touvron et al., 2023b) and Mistral (Jiang et al., 2023) on the profile-aware dialogue samples[6], and improve the efficiency of the training process from data and model perspective.

5.2.1 Parameter-Efficient Training

One major challenge of utilizing LLMs is the high demand for computational resources for adaptive training. To fine-tune LLMs in a low-resource setting, here we employ parameter-efficient approaches: Low-rank adaption (LoRA) (Hu et al., 2021) and QLoRA (Dettmers et al., 2024). Previous studies show that the over-parameterized models in fact reside on a low intrinsic dimension. Compared with full-parameter training, LoRA and QLoRA update to the weight matrices with a low-rank matrix factorization, and significantly reduces the number of trainable parameters, and speeds up training with little impact on the final performance.

5.2.2 Dialogue-level Efficient Training

Given one multi-turn dialogue sample, at the fine-tuning stage, generally, only the system responses are used for loss calculation and weight updating. In practice, if we split a n-turn dialogue into n

[6]All open models used in this work are only for research use. We follow their corresponding license in our experiments.

Model Type	BLEU-2	BLEU-3	ROUGE-1	ROUGE-2	ROUGE-L	SimCSE
LLaMA-2 7B	4.314	2.517	12.75	2.500	13.09	28.90
+ Utterance Normalization	5.752	3.521	17.04	4.052	18.24	41.89
+ Context Aggregation	7.087	4.533	18.49	4.183	20.36	44.40
LLaMA-2-Chat 7B	4.530	2.788	13.00	2.553	13.18	28.46
+ Utterance Normalization	7.849	5.205	18.84	6.001	21.05	42.95
+ Context Aggregation	9.313	7.344	20.27	6.492	21.96	44.38
LLaMA-2-Chat 13B	4.526	2.625	12.78	2.711	13.85	29.77
+ Utterance Normalization	8.160	5.596	20.99	5.205	23.88	45.13
+ Context Aggregation	10.53	7.227	22.77	6.544	26.16	49.03
Mistral-7B	4.434	2.406	13.06	2.501	14.28	30.06
+ Utterance Normalization	8.782	6.441	19.74	6.353	21.75	42.97
+ Context Aggregation	11.29	8.248	21.68	10.11	25.34	48.98
Mistral-7B-Instruct-v0.2	4.878	2.957	14.34	2.901	13.28	28.86
+ Utterance Normalization	7.942	5.341	18.96	6.541	21.88	46.24
+ Context Aggregation	11.76	8.358	22.36	9.783	26.40	53.19

Table 2: Experimental results with automated evaluation metrics on topically-coherent dialogue generation.

turn-level samples, the learning step increases by a factor of n. To improve training efficiency, here we leverage the properties of causal language models since each token only depends on its precedent tokens. Therefore, we feed the entire dialogue sequence to the decoder-only model, and mask out the user utterances, and compute the loss of all system responses in parallel.

5.2.3 Balanced Data Sampling

Since the sample number of customized coaching is limited, we mixed dialogue segments from other topics for training data augmentation. The frequency distribution of different topics is imbalanced. For instance, compared with the topic "oral medication", the "general education" is more frequently discussed and presents a larger utterance number. When fine-tuning the language backbone, a diverse and balanced sample set can bring higher performance, we thus construct the training set by sampling a balanced ratio at the topic level.

6 Experiments and Results

6.1 Experimental Setting

The processed conversational data (5.0K topic-level dialogue samples) are used for training, and we randomly select 10% for validation and testing (0.5K samples) respectively. The maximum length of the dialogue sequence is set at 2048. *AdamW* optimizer is used with a learning rate of 1e-5, the batch size with gradient accumulation is set at 64, and the epoch number is 5. Best checkpoints are selected based on validation results using cross-entropy loss. Models are implemented with PyTorch[7] and HuggingFace Transformers[8]. Parameter-efficient fine-tuning is applied with PEFT (Mangrulkar et al., 2022), and the rank k in LoRA adaptation is set at 16. Following previous work, we add the projection layers of the Transformer network to the LoRA training process, and the trainable parameter sizes of LLaMA-2-7B/Mistral-7B and LLaMA-2-13B are 2.32M and 3.63M, respectively. All experiments are run on a single Nvidia A100 GPU with 40G memory.

6.2 Evaluation Metrics

Following previous work (Shen et al., 2020), we use two lexical automated evaluation metrics: BLEU (BLEU-2 and BLEU-4) and ROUGE (ROUGE-1, ROUGE-2 and ROUGE-L) (Lin, 2004), as well as the embedding-based metrics SimCSE (Gao et al., 2021). All reported scores are rescaled to percentage values. For each topically coherent dialogue segment ended with '<*topic-end*>', we calculate the averaged evaluation scores of each nurse's utterance. Speaker role tokens (e.g., *Nurse*, *Patient*) and model-generated special tokens (e.g., *</s>*, *[INST]*) are not included.

6.3 Evaluation Results & Analysis

We use a hold-out test set to evaluate the generated nurse responses. In our experiments, we indicate gold topic labels for model comparison. Since personalized dialogue generation is mainly for delivering customized education or consultation, we thus focus on evaluating the four customized coaching

[7] https://pytorch.org

[8] https://github.com/huggingface/transformers

Model Type	BLEU-2	BLEU-3	ROUGE-1	ROUGE-2	ROUGE-L	SimCSE
LLaMA-2-Chat 7B	7.012	4.336	17.79	3.848	19.98	41.80
- Context Aggregation	5.997	3.409	16.67	3.101	17.42	35.75
- Patient Profile	4.473	3.166	11.24	2.630	11.02	29.85
LLaMA-2-Chat 7B	8.334	5.554	18.40	5.760	21.24	43.20
- Context Aggregation	6.019	3.757	17.65	3.221	19.65	38.28
- Patient Profile	4.509	3.190	11.53	2.981	11.63	30.42
Mistral-7B	9.636	7.022	21.53	7.319	23.14	48.65
- Context Aggregation	6.914	5.065	15.68	4.751	18.47	35.72
- Patient Profile	5.303	3.682	12.52	2.673	13.79	34.78
Mistral-7B-Instruct-v0.2	10.19	7.282	21.31	7.520	24.19	48.30
- Context Aggregation	6.062	4.252	15.28	3.808	17.30	36.12
- Patient Profile	5.136	3.508	12.06	2.351	13.51	35.11

Table 3: Ablation study on the context aggregation via topic-focused summarization at the inference stage.

topics: self-monitoring, diet management, insulin, and physical activity.

6.3.1 Dialogue Generation Evaluation

Table 2 shows the results of dialogue generation by training the representative open LLMs (e.g., LLaMA, Mistral). Here we report evaluation results of modeling training with our proposed enhancements: the human spoken dialogue data refinement (i.e., utterance normalization) and context modeling and aggregation (i.e., utilizing topic-focused summarization). As shown in Table 2, the generation quality benefits a lot from adopting utterance normalization on all tested models and at all metrics. This is because human conversations contain many spoken linguistic features such as fillers, thus training on the original noisy spoken data affects the generation quality significantly, models tend to produce less meaningful and fluent sentences. Therefore, to build reasonable human-machine conversational interaction, it is necessary to include the normalization step in the spoken dialogue samples. On the other hand, compared with other language generation tasks such as machine translation, the overall evaluation scores of dialogue response generation are at a low level, this is mainly due to the utterance diversity in the nurse-patient conversations.

Moreover, adding context aggregation with topic-focused summarization also significantly improves the scores, demonstrating its effectiveness of coherent personalizing response generation. Considering the scoring alignment between lexicon-based and embedding-based metrics, the overall evaluation ranks are consistent across the three tested metrics: BLEU, ROUGE, and SimCSE. Upon the summarization process, the historical context length can be reduced to 20% of the original length, with dense information in a formal wording. This is also beneficial for the model to capture important features to organize the subsequent generations. Surprisingly, in our experimental setting, we observe that instruction-tuned models (e.g., LLaMA-2-Chat, Mistral-7B-Instruct) did not show substantial gain over the pre-training foundation models, and scores become even lower in some metrics (e.g., BLEU-2, ROUGE-2) when training with utterance normalization. As LLMs contain massive prior knowledge from large-scale pre-training, both model types could achieve the same dialogue modeling and generation capabilities after domain-specific adaptation on one downstream task.

6.3.2 Leveraging LLMs as Evaluator

Recent work shows that the LLMs can be used as evaluators for various NLP tasks, and present a high correlation with human preference (Li et al., 2023). Here we use GPT3.5-turbo for the automatic evaluation. We feed generated utterances from our trained Mistral-7B-Instruct-v0.2, and compare the vanilla model with our model upon normalization and context aggregation, by predicting which response is better. We sampled 30 utterances for evaluation, and the winning rate of the enhanced model is 0.80, demonstrating the effectiveness of our proposed methods.

6.3.3 Ablation Study on Context Aggregation

We conduct an ablation study on the context aggregation of topic-focused summarization. In our preliminary experiment, we observe that the first three utterances from the nurse of each topically-coherent dialogue segment show more dependency on the historical context, due to the explicit topic shift (e.g., from symptom checking to customized

coaching of insulin). Therefore, at the inference stage, we collect the first-3 generated utterances of each topic, and compared models with and without adding the aggregated summaries. As shown in Table 3, all evaluation scores drop significantly when the historical summaries are removed, for all tested models (e.g., LLaMA, Mistral). This demonstrates that nurse dynamically change their topic during the conversation, the topic-specified questions in certain topics depend on the information they collect from the patient.

6.3.4 Ablation Study on Patient Profile

We conduct an additional ablation study on the patient profile. Following the previous step, at the inference stage, we still collect the first-3 generated utterances of each topic, and compared models with and without adding the patient profile information. As shown in Table 3, the generation performance for all tested models (e.g., LLaMA, Mistral) drops significantly when no profile is provided. For instance, in the topic *'Diet'* we observe that models tend to generate common questions (e.g., *"how is your diet?"*) when there is no profile and dialogue context. In comparison, models can ask more targeted questions (e.g., *"What about your sugar intake? Do you consume sweetened beverages?"*), which are more informative, especially at the beginning of each topic segment.

7 Conclusion

In this work, we investigated the feasibility and effectiveness of leveraging language language models for personalized medical dialogue generation. We conducted a case study on healthcare follow-up calls for diabetes management. Inspired by real-world conversations, we built a data preparation and refinement pipeline for spoken conversation processing, user profile construction, and proposed topic-focused summarization to distill and aggregate the historical context. To exploit the potential of LLMs, we applied efficient model training methods for domain adaptation. Our experimental results showed that context aggregation via topic-focused summarization is beneficial for long-context modeling and coherent generation.

Limitations

The data and model used in this work are in English, thus to apply the approach to other languages, it will require training data on the specified language or using multilingual language backbones. While our proposed methods are general, when adopt them to other conversational data, in-domain annotation is required to obtain reliable results. Moreover, the hallucination made by large language models is an open problem, and the system generations in clinical scenarios still need human verification and intervention if necessary.

Ethics and Impact Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. The in-domain samples used in this work are fully anonymized. The original data are collected under consent for academic research purposes. Our proposed framework and methodology in general do not create a direct medical implication, and are intended to be used to improve the model accuracy and robustness for downstream applications.

Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore under its Industry Alignment Pre-Positioning Fund (Grant No. H19/01/a0/023 - Diabetes Clinic of the Future). We thank Ai Ti Aw and Rosa Qi Yue So at the Institute for Infocomm Research (I^2R) for their support and assistance, and thank Siti Maryam Binte Ahmad Subaidi, and Nabilah Binte Md Johan for linguistic resource construction, and Sakinah Binte Yusof and Helen Erdt for data-related discussion. We gratefully acknowledge valuable inputs from Joan Khoo, Anne Teng Ching Ching, Winnie Soo Yi Ling, Lee Yian Chin, Angela Ng Hwee Koon, Sharon Ong Yu Bing, Bryan Choo Peide, and Oh Hong Choon at the Changi General Hospital, Singapore. We thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, et al. 2023. Foundation metrics: Quantifying effectiveness of healthcare conversations powered by generative ai. *arXiv preprint arXiv:2309.12444*.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Con-

versational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Niels Ole Bernsen, Hans Dybkjær, and Laila Dybkjær. 1996. Cooperativity in human-machine and human-human spoken dialogue. *Discourse processes*, 21(2):213–236.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders. 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 1931–1940.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. *Stud Health Technol Inform*, 252:51–56.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

M Sue Kirkman, Morris Weinberger, Pamela B Landsman, Gregory P Samsa, E Anne Shortliffe, David L Simel, and John R Feussner. 1994. A telephone-delivered intervention for patients with niddm: effect on coronary risk factors. *Diabetes care*, 17(8):840–846.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Margaret L Lawson, Nini Cohen, Christine Richardson, Elaine Orrbine, and Ba' Pham. 2005. A randomized trial of regular standardized telephone contact by a diabetes nurse educator in adolescents with poor diabetes control. *Pediatric Diabetes*, 6(1):32–40.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B 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.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022a. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F Chen. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821. IEEE.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022b. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy, and Nancy Chen. 2023. Joint dialogue topic segmentation and categorization: A case study on clinical spoken conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 185–193.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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.

Oluwatobi Olabiyi, Anish Khazane, Alan Salimov, and Erik Mueller. 2019. An adversarial learning framework for a persona-based multi-turn dialogue model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of medical Internet research*, 21(4):e12887.

John D Piette, Morris Weinberger, Frederic B Kraemer, and Stephen J McPhee. 2001. Impact of automated calls with nurse follow-up on diabetes treatment outcomes in a department of veterans affairs health care system: a randomized controlled trial. *Diabetes care*, 24(2):202–208.

John D Piette, Morris Weinberger, Stephen J McPhee, Connie A Mah, Fredric B Kraemer, and Lawrence M Crapo. 2000. Do automated calls with nurse follow-up improve self-care and glycemic control among vulnerable patients with diabetes? *The American journal of medicine*, 108(1):20–27.

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 27th International Joint Conference on Artificial Intelligence*, pages 4279–4285.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Benjamin Schloss and Sandeep Konam. 2020. Towards an automated soap note: classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

C Barr Taylor, Nancy Houston Miller, Kelly R Reilly, George Greenwald, Darby Cunning, Allison Deeter, and Liana Abascal. 2003. Evaluation of a nurse-care management system to improve outcomes in patients with complicated diabetes. *Diabetes care*, 26(4):1058–1063.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1062–1069.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Lu Xu, Leslie Sanders, Kay Li, James CL Chow, et al. 2021. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR cancer*, 7(4):e27850.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

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

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. 2021. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 886–896, Online. Association for Computational Linguistics.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2022. Towards enhancing health coaching dialogue in low-resource settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 694–706, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.