Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Unveiling Maternity and Infant Care Conversations: A Chinese Dialogue Dataset for Enhanced Parenting Support

Bo Xu[1†], **Liangzhi Li**[2†‡], **Junlong Wang**[4†], **Xuening Qiao**[1],
Erchen Yu[1], **Yiming Qian**[3], **Linlin Zong**[4‡], **Hongfei Lin**[1]
[1]School of Computer Science and Technology, Dalian University of Technology
[2]Meetyou AI Lab
[3]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR)
[4]School of Software, Dalian University of Technology
{xubo,llzong, hflin}@dlut.edu.cn, {jlwang, qiao, yuerchen0809}@mail.dlut.edu.cn,
liliangzhi@xiaoyouzi.com, qian_yiming@ihpc.a-star.edu.sg

Abstract

The rapid development of large language models has greatly advanced human-computer dialogue research. However, applying these models to specialized fields like maternity and infant care often leads to subpar performance due to a lack of domain-specific datasets. To address this problem, we have created MicDialogue, a Chinese dialogue dataset for maternity and infant care. MicDialogue involves a wide range of specialized topics, including gynecological health, pediatric care, pregnancy preparation, emotional counseling and other related topics. This dataset is curated from two types of Chinese social media: short videos and blog posts. Short videos capture real-time interactions and pragmatic dialogue patterns, while blog posts offer comprehensive coverage of various topics within the domain. We have also included detailed annotations for topics, diseases, symptoms, and causes, enabling in-depth research. Additionally, we developed a knowledge-driven benchmark model using LLM-based prompt learning and multiple knowledge graphs to address diverse dialogue topics. Experiments validate MicDialogue's usability, providing benchmarks for future research and essential data for fine-tuning language models in maternity and infant care.

1 Introduction

With the rise of large language models (LLMs), human-machine dialogues offer various smart services. However, in specialized areas like maternity and infant care, general models often don't meet user needs effectively [Yang *et al.*, 2023]. Developing models for this field requires deep knowledge in gynecology, pediatrics, parenting, and postnatal care, along with the ability to detect subtle emotional changes, especially those linked to hormonal shifts. General dialogue models may struggle without specific training data. Therefore, it's important to systematically collect and process relevant data to create a specialized dialogue resource for maternity and infant care, enabling the training of more effective models for this area.

Maternity and infant care is a relatively unexplored area in human-machine dialogue, presenting three key challenges for constructing effective dialogue datasets. (1) **Specialized Topics**: The dataset needs to cover a wide range of specialized topics, such as prenatal exams, infant care, pregnancy discomforts, and related health issues. Unlike general medical dialogues, this field requires focused and topic-specific knowledge. Existing medical datasets [Zeng *et al.*, 2020; Chen *et al.*, 2023; Yunxiang *et al.*, 2023; Xu *et al.*, 2022] may not address these specific topics, leading to lower performance and user satisfaction. (2) **Real-Time Interactions**: Current medical dialogue datasets often come from online forums and feature static question-and-answer formats. These do not capture the dynamic nature of real-time interactions required in maternity and infant care. Real-time dialogues involve immediate assessments based on symptoms, observations, and medical results, which are typically missing from existing datasets. (3) **Field-specific Emotional Support**: Maternity and infant care involves a wide range of emotional experiences, from pregnancy anxiety to postpartum depression. While some datasets focus on empathetic dialogue [Rashkin *et al.*, 2019; Liu *et al.*, 2021; Hosseini and Caragea, 2021], there is a lack of resources specifically targeting emotional perception in this field. A specialized dataset is needed to capture the emotional nuances specific to maternity and infant care, allowing dialogue models to offer better emotional support.

To address these challenges, we construct a Chinese dialogue dataset for maternity and infant care, named **MicDialogue**. This dataset is carefully compiled from various online sources, including Chinese short video and blog-based social media platforms. We manually curated content across *five major topics*: gynecological health, pediatric and parenting, pregnancy preparation, emotional counseling, and others. To capture *real-time interactions*, we used short video platforms

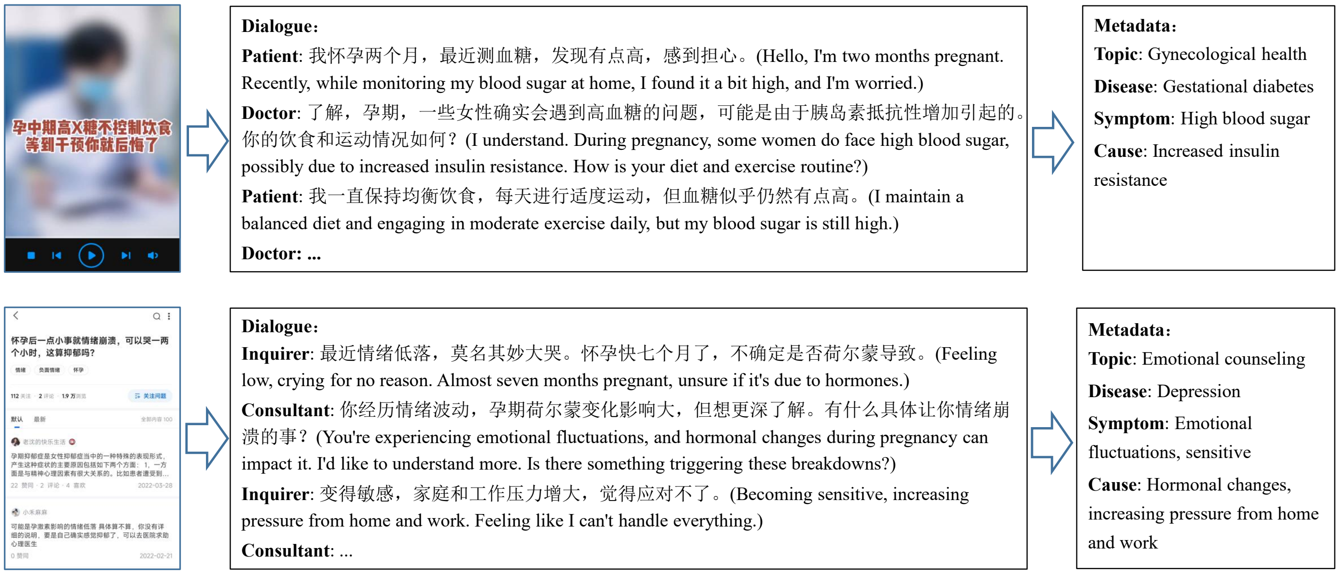


Figure 1: Two examples of the extracted dialogues and their metadata based on video-based and post-based consultations.

like Kuaishou[1], Douyin[2], and Xiaohongshu[3], where medical professionals share multi-turn consultation videos. We transcribed these videos into text, preserving the dynamic nature of real-time dialogues. For *fine-grained emotional support*, we crawled emotional counseling posts related to maternity and infant care from Weibo[4] and Zhihu[5], organizing them into dialogue formats that cover the full range of emotional issues in this field. Two examples of our dataset construction process is illustrated in Figure 1.

Compared to existing medical dialogue datasets, our dataset has more **comprehensive information**. MicDialogue includes not just healthcare dialogues but also detailed metadata like topics, diseases, symptoms, and causes. This allows for fine-grained modeling and personalized responses. For example, the inclusion of symptoms and causes in emotional counseling helps develop dialogue models that can provide mental health support and cognitive interventions. The main **contributions** are summarized as follows.

- We created a comprehensive Chinese dialogue dataset for maternity and infant care, containing 8,813 dialogues from blog posts and short videos. This dataset includes detailed metadata to enhance various dialogue-related tasks. To the best of our knowledge, MicDialogue is the first dialogue dataset in the domain of maternity and infant care.
- We developed a rigorous annotation scheme that labels topics, diseases, symptoms, and causes for all dialogues. This detailed annotation helps with medical information extraction and supports mental healthcare applications, making it a valuable resource for creating personalized dialogue models for maternity and infant care.
- We introduced a knowledge-driven benchmark model for generating responses in maternity and infant care dialogues. We validated the usefulness of MicDialogue by testing different pre-trained models on dialogue generation. The results demonstrate the value of our dataset and show strong potential for future research in this area.

[1]https://www.kuaishou.com
[2]https://www.douyin.com
[3]https://www.xiaohongshu.com
[4]https://weibo.com
[5]https://www.zhihu.com

2 Related Work

2.1 Dialogue Datasets

The exploration of dialogue datasets for maternity and infant care can be approached from three areas: medical dialogue, emotion-supportive dialogue and stylized dialogue. Table 5 summarizes these datasets in aspects of numbers of dialogues and utterances, source and category. *Medical dialogue dataset* [Zeng *et al.*, 2020] [Chen *et al.*, 2023] [Yunxiang *et al.*, 2023] [Xu *et al.*, 2022] focus on general medical dialogue but lack specialization in maternity and infant care, there is a notable gap in datasets specifically targeting maternity and infant care. *Emotion-supportive dialogue dataset* [Rashkin *et al.*, 2019] [Liu *et al.*, 2021] [Hosseini and Caragea, 2021] focuses on generating emotionally enriched responses. However, these datasets do not specifically address the complex emotional issues related to maternity and infant care. *Stylized dialogue dataset* [Wu *et al.*, 2020b] [Zheng *et al.*, 2021] [Su *et al.*, 2020] [Xiang *et al.*, 2023] involves generating dialogue in a specific style. While CARE-MI [Xiang *et al.*, 2023] provides some insights into the maternity and infant care, it does not directly focus on dialogue generation.

The proposed dataset focuses on maternity and infant care dialogues, covering topics like gynecological health, parenting, pregnancy preparation, and emotional counseling, with detailed annotations of a wide range of metadata.

Dataset Name	#dialogues	#utterances	Source	Category
MedDialog-EN [Zeng *et al.*, 2020]	257,332	514,664	blog posts	general medical domain
MedDialog-CN [Zeng *et al.*, 2020]	3,407,494	11,260,564	blog posts	general medical domain
IMCS-21 [Chen *et al.*, 2023]	4,116	164,640	blog posts	general medical domain
ChatDoctor [Yunxiang *et al.*, 2023]	-	100,000	blog posts	general medical domain
RealMedDial [Xu *et al.*, 2022]	2,637	24,255	short videos	general medical domain
EmpatheticDialogues [Rashkin *et al.*, 2019]	24,850	-	blog posts	emotional support
ESConv [Liu *et al.*, 2021]	1,053	31,410	blog posts	emotional support
IEmphathy [Hosseini and Caragea, 2021]	-	5,007	blog posts	emotional support
MTFC [Wu *et al.*, 2020b]	-	1,007,999	blog posts	stylistic generation
WDJN [Zheng *et al.*, 2021]	300,000	-	blog posts	stylistic generation
Gender-Specific Dialogue [Su *et al.*, 2020]	-	100,000	blog posts	stylistic generation
MicDialogue (Ours)	**8,813**	**40,730**	**blog posts & short videos**	**maternity and infant care**

Table 1: Comparison of related dialogue datasets.

2.2 Dialogue Generation

Dialogue generation focuses on creating responses that meet specific user needs. Recently, large language models like DialoGPT [Zhang *et al.*, 2019] and PLATO [Bao *et al.*, 2019] have been used in generative dialogue systems. However, these models often produce generic responses due to a lack of domain-specific knowledge, which lowers user satisfaction [Serban *et al.*, 2017]. Incorporating external knowledge can address this issue [Ghazvininejad *et al.*, 2018]. External knowledge comes in three types: non-structured (e.g., encyclopedia-based [Dinan *et al.*, 2018] and prototype-based [Cai *et al.*, 2018]), structured (e.g., knowledge graphs [Zhou *et al.*, 2018; Wu *et al.*, 2020a] and tables [Wu *et al.*, 2021]), and a hybrid of both [Liu *et al.*, 2019]. Structured knowledge, being more organized and reliable, greatly improves dialogue generation [Zhou *et al.*, 2018; Wu *et al.*, 2020a].

Maternity and infant care dialogues involve many specialized topics, making domain-specific knowledge crucial for high-quality responses. To tackle this, we propose a knowledge-driven benchmark model that selects relevant dialogue-specific knowledge from multiple knowledge graphs, aiming to improve dialogue generation in this specialized field.

3 Dataset Construction

Maternity and infant care dialogues cover various topics. After initial research and expert consultations, we identified five main topics: gynecological health, pediatric and parenting, pregnancy preparation, emotional counseling, and others. To collect ample dialogue data, we used various Chinese social media platforms, especially subforums dedicated to maternity and infant care. Specifically, our data sources include short videos and blog posts. We provide more details on the data collection from these sources below.

3.1 Dialogue Data from Short Videos

Collection

We focused on three major Chinese short video platforms: Douyin, Kuaishou, and Xiaohongshu, where healthcare professionals share medical consultation videos. These videos document doctor-patient conversations while protecting patient privacy, providing robust real-scenario professional support for building maternity and infant care dialogue dataset. We selected 318 professional doctor accounts, analyzing around 9,000 videos. To ensure quality and address ethical concerns, we only kept video clips with complete, multi-turn doctor-patient dialogues on the five specific topics, resulting in 2,023 videos from 45 doctor accounts.

Processing

We refined short videos related to maternity and infant care and transcribed their content into text dialogues. Each transcribed dialogue includes the video title, specialty, and multi-turn dialogue content. We reviewed and adjusted dialogues to remove meaningless colloquial expressions and ensure each dialogue round had specific content. We also removed videos lacking substantial interaction, keeping only high-quality multi-turn dialogues.

3.2 Dialogue Data from Blog-based Posts

Collection

To expand our dataset, particularly for emotional support, we focused on social media platforms like Weibo and Zhihu. These platforms have numerous posts on maternity and infant care, where users share emotional support and experiences, offering valuable dialogue data. We used two search strategies, topic-based search and account-based search, to gather a wide range of relevant posts.

Topic-based Search. We searched for posts related to medical and emotional counseling using specific keywords like ’parenting’ and ’preparing for pregnancy.’ We then collected these posts, along with user comments and interactions, to create a tree-like structure of comment threads. From these threads, we extracted multi-turn dialogues for our dataset.

Account-based Search. We identified accounts of emotional bloggers specializing in maternity and infant care. These bloggers often provide emotional support and advice. We focused on their interactions with other users, gathering posts and replies from these accounts to use as dialogue data.

Using these two strategies, we collected a large amount of user interaction data, resulting in 314,192 multi-turn dialogues. This data covers a wide range of medical and emotional topics related to maternity and infant care, which is helpful for developing more nuanced emotion-supportive maternity and infant care dialogue systems.

Processing

The raw data from blog-based dialogues includes various types of noises, which can lower the quality of the dataset. To improve the data quality, we applied several noise removal rules: (1) remove URLs, email addresses, and sequences of six consecutive numbers; (2) discard platform-related tags like "Reply to @", "@", and "@***"; (3) exclude emoticons; (4) keep only one instance of repeated phrases, sentences, or words; (5) eliminate extra spaces within sentences; (6) remove single-turn dialogues; (7) filter out repetitive responses beyond a specified threshold in the same thread to ensure response diversity; (8) translate traditional Chinese characters to simplified Chinese; (9) remove empty strings and sentences made up only of letters, numbers, or symbols; (10) identify and remove toxic comments using an existing toxicity lexicon [Lu *et al.*, 2023] containing inappropriate and sensitive words.

Since it's challenging to filter out all semantic noises with rules alone, we conducted a secondary manual review. We follows these guidelines: (1) remove contextually irrelevant dialogues or those unrelated to maternity and infant care; (2) eliminate dialogues lacking practical significance or with ambiguous meanings; (3) filter out dialogues containing advertisements; (4) remove dialogues containing rumors.

3.3 Annotation Strategy and Quality Control

We annotated the extracted dialogues with four metadata including topic, disease, symptom, and cause. The topics include gynecological health, pediatric and parenting, pregnancy preparation, emotional counseling, and others. The disease cover most of the common illnesses found in maternity and infant care.

To ensure high annotation quality, we developed detailed instructions for short-video and blog-based data. Annotators, all graduate students, received formal training to understand each category. Any uncertainties were flagged and reviewed by the team. Annotators were divided into two groups: sixteen worked on short-video dialogues and six on blog-based dialogues. We used a cross-validation approach where two annotators independently labeled the same data, and a third expert resolved disagreements. The process took four weeks, with regular checks and weekly discussions to ensure accuracy and consistency.

We followed a rigorous and standardized process, incorporating Chinese medical subject headings to complement colloquial terms with specialized medical terminology. Given the intuitive nature of annotation, some subjectivity is inevitable. To measure reliability, we used the Kappa score, a common metric in computational linguistics annotation schemes. Annotators labeled the same 1,000 dialogues, achieving inter-annotator agreement scores of 0.84 for topics and 0.81 for metadata, indicating a significantly reliable annotation process.

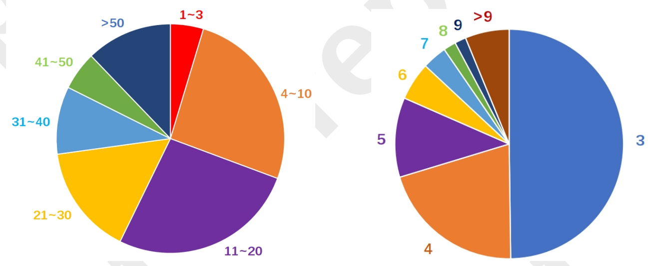
3.4 Analysis

Dialogue Characteristics

As shown in Table 2, the finally constructed dataset has 8,813 dialogues, mostly from blog posts (7,685) and fewer from short videos (1,128). Blog posts have fewer utterances per dialogue but more tokens per utterance than short videos, reflecting different communication styles. This contrast highlights the value of our real-time dialogue data. Additionally, Figure 2 shows that most dialogues have fewer than 9 utterances, with nearly half having 3 utterances. Most utterances contain 4-40 tokens, indicating effective multi-turn dialogues.

Indicators	MicDialogue	Posts	Short Videos
# dialogues	8,813	7,685	1,128
# utterances	40,730	29,705	11,025
# tokens	1,001,867	816,329	185,538
avg. utt. per dial.	4.62	3.91	9.73
avg. tok. per utt.	24.60	41.33	25.72

Table 2: Statistics on the constructed dataset.



(a) Token per Utterance. (b) Utterance per Dialogue.

Figure 2: Statistics on the constructed dialogue.

Topic Focus

Table 3 displays statistics on the five main topics in maternity and infant care. It highlights a strong focus on "emotional counseling," exclusively found in blog posts, reflecting in-depth support. "Gynaecological health" and "pregnancy preparation" are mainly covered in blog posts, while "Gynaecological health" and "pediatric and parenting" have notable short video content, showing diverse topic coverage.

4 Our Model

Given the specialized topics in maternity and infant care dialogues, we seek to improve dialogue response generation by adding domain-specific knowledge. To this end, we introduce a knowledge-driven model called Kng-BART, serving as a benchmark model for future studies.

4.1 Model Overview

Given a dialogue $D = [X_1, ..., X_M]$ with M utterances, where the i-th utterance $X_i = [x_o^i, ..., x_{m_i}^i]$ contains a sequence of m_i tokens, the dialogue generation task aims to generate an informative response $R = [r_i, ..., r_n]$ by maximizing the generation probability $P(R|D)$. To achieve this objective, we propose to incorporate domain knowledge into dialogue model. Our model comprises three modules: the knowledge expansion and selection module, making comprehensive analysis of the dialogue and facilitating flexible selection of external knowledge bases; the graph-based knowledge representation module, utilized to expand knowledge through sub-graph integration; and the knowledge-enhanced response

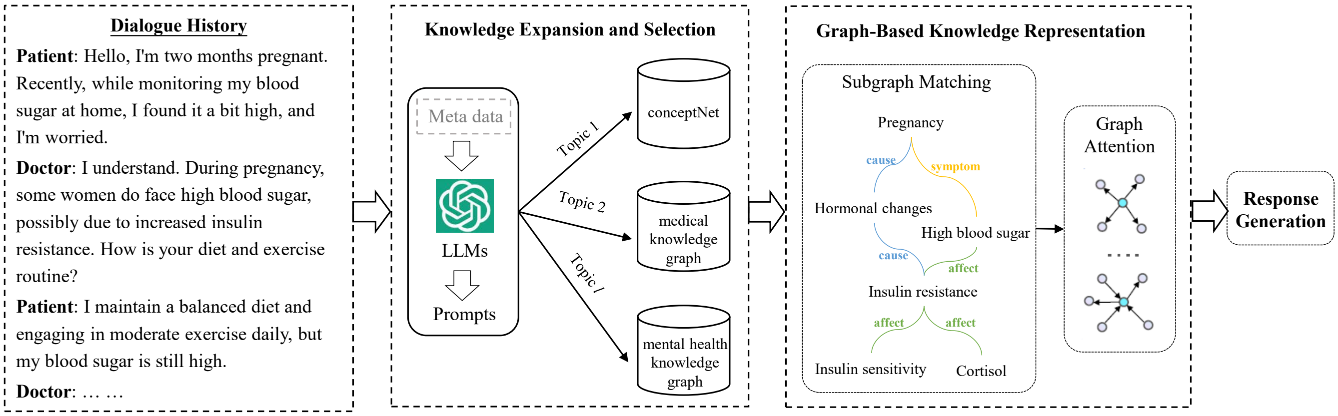


Figure 3: Main architecture of our model.

Topic	MicDialogue	Posts	Short Videos
gynaecological health	4,162	3,333	829
pediatric and parenting	312	82	230
pregnancy preparation	250	185	65
emotional counseling	3,975	3,975	0
others	114	110	4

Table 3: The number of dialogues in different topics.

generation module, employed to generate knowledge-aware dialogue responses. The main architecture of our model is illustrated in Figure 3.

4.2 LLM-based Knowledge Expansion and Selection

Given the wide range of topics in maternity and infant care dialogues, relying on just one source of external knowledge might not be enough. Moreover, if the dialogue history is limited, it is challenging to select and use the right external knowledge. Therefore, we consider expanding the dialogue history when possible and dynamically selecting the most relevant knowledge base for generating responses.

Knowledge Expansion

We use LLMs to expand the knowledge in the dialogue history. Two prompt templates are designed for comprehensive dialogue analysis, with their use depending on the presence of metadata. The designed prompt templates are shown in the following.

Prompt1: "[dialogue history]". Consider this dialogue to select its topic from the given set of topics [topics]. Provide a concise reason and explanation for your judgment, limited to 100 words. Strictly output in the following format: {*'Dialogue ID':, 'Topic':' ', 'Reason and Explanation':' '*}.

Prompt2: "[dialogue history]" and "[metadata]". Please make a detailed analysis on the four perspectives: topic, disease, symptom and causes. Each limited to 20 words. Strictly output in the following format: {*'Dialogue ID':, 'Topic':' ', 'Disease':' ', 'Symptom':' ', 'Causes':' '*}.

After that, we obtain the prompt P for each utterance. we concatenate the dialogue with the prompt to expand the information, denoted as $D = [[X_1 P_1], ..., [X_M P_M]]$.

Knowledge Graph Selection

We use three knowledge graphs including ConceptNet [Speer *et al.*, 2017], medical knowledge graph [Weng *et al.*, 2017] and mental health knowledge graph [Cao *et al.*, 2020]. Given a dialogue D, we associate the dialogue with a knowledge graph based on its topic, achieving a flexible selection of external knowledge related to the dialogue.

4.3 Graph-Based Knowledge Representation

After matching with the topic-centered knowledge graph, the next step is to identify the specific knowledge subgraph that corresponds to the given dialogue. We use a string-based matching approach to retrieve one-hop subgraphs from the entire knowledge graph, using the dialogue tokens as central nodes. The retrieved subgraphs are represented as $G = [g_1, g_2, ..., g_j]$, where j is the number of subgraphs. For each subgraph, a BERT encoder retrieves all related entities, preserving most token-level semantics. We then calculate the average token embedding for each entity, which serves as the initial knowledge embedding. To integrate information from neighboring entities, we use a graph attention network [Veličković *et al.*, 2017]. The updated entity embeddings are averaged to create the overall subgraph embedding. After this process, we obtain the knowledge representations for a certain utterance using multiple subgraphs, represented as follows.

$$Z_{kng} = \{K(g_1), K(g_2), ..., K(g_n)\} \tag{1}$$

Given that dialogue history involves multiple utterances, the influence of utterances diminishes over time. Knowledge associated with more recent utterances becomes increasingly vital. To address this, we developed an utterance-knowledge cross-attention mechanism, enhancing the incorporation of knowledge linked to recent utterances for better comprehending the dialogue history.

$$Z_{ctx} = BartEncoder(\widetilde{D}) \tag{2}$$

$$\widetilde{Z} = MutiHeadAttention(Z_{ctx}, Z_{kng}) \tag{3}$$

$$Z = LayerNorm(W_z \widetilde{Z} + Z_{kng}) \tag{4}$$

Model/Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	DIST-2	METEOR	NIST-2	NIST-4
Transformer	0.1362	0.0422	0.0087	0.0012	0.2007	0.1242	0.5777	0.5780
CDial-GPT	0.0755	0.0259	0.0111	0.0066	0.3707	0.0732	0.2776	0.2776
ChatGPT	0.2934	0.1988	0.1465	0.0873	0.3659	0.2094	0.7344	0.7582
BART	0.2851	0.1865	0.1381	0.0889	0.2993	0.2146	0.6894	0.6969
Kng-BART	**0.3307**	**0.2196**	**0.1621**	**0.0982**	**0.3811**	**0.2359**	**0.8756**	**0.8737**

Table 4: Performance comparison of different models in terms of automatic evaluation metrics.

where $\widetilde{D}$ denotes the input sequence of utterances. Z_{ctx} denotes the representations of utterances in dialogue history arranged chronologically. We use BART encoder [Lewis *et al.*, 2019] to acquire Z_{ctx}. W_z denotes the trainable parameters. Ultimately, we derive a contextual semantic representation Z of external knowledge for the dialogue.

4.4 Knowledge-enhanced Response Generation

Based on the obtained representations of external knowledge, the objective of our task is to generate responses that are more informative by leveraging both dialogue history and knowledge, maximizing the generation probability $P(R|D,K)$. We use the BART [Lewis *et al.*, 2019] decoder as our response generator and minimize the negative log-likelihood with respect to the ground-truth response:

$$L = \sum_{n=1}^{|R|} -\log p(r_i|D,K,r_{<i}) \tag{5}$$

where $p(r_i|D,K,r_{<i})$ denotes the probability of generating the i-th token r_i in the response R given the context D, the knowledge K and the previously generated tokens $r_{<i}$.

5 Experiments

5.1 Experimental Settings

We split our dialogue dataset into three subsets: training, validation, and test sets, in an 8:1:1 ratio, dividing at the dialogue level. The validation set was used to fine-tune hyperparameters, and training stopped when the validation loss stopped decreasing.

The algorithm were realized by Python and conducted using an RTX 3090 Ti graphics card. For Transformer, we used HuggingFace's [6] implementation with default settings. ChatGPT was accessed via the OpenAI API. We used the official code for CDial-GPT [7][Wang *et al.*, 2020], which is trained on a large Chinese conversational dataset. The BART [8] model we used was fine-tuned on various Chinese dialogue datasets. We set multi-head GAT to 4 heads and limited subgraphs to 256 nodes. The models were trained with AdamW [Loshchilov and Hutter, 2017], using a 5e-5 learning rate and a batch size of 32. For all models, we run one experiment with $seed = 42$ and used top-k random sampling with $k = 50$ for decoding, with a max length of 120, and stopped when generating the "[SEP]" token.

[6]https://huggingface.co/

[7]https://github.com/thu-coai/CDial-GPT

[8]https://huggingface.co/HIT-TMG/dialogue-bart-base-chinese

We employed both automatic and human evaluations, which is a conventional approach in the field of human-computer dialogue. For automatic evaluation, we used BLEU-n [Papineni *et al.*, 2002] (n=1 to 4), NIST-n [Doddington, 2002] (n=2 and 4), METEOR [Banerjee and Lavie, 2005] to assess the similarity between the generated responses and ground truth ones through n-gram matching. Additionally, DIST-n [Li *et al.*, 2015] (n=2) was used to measure the diversity of the generated responses. Higher scores indicate better performance.

For human evaluation, we randomly selected 1,000 dialogues and had five students rate the generated responses on *Relevance*, *Informativeness*, *Fluency*, and *Topic Alignment*, using a 1 to 5 scale. The final scores were averaged across the annotators.

5.2 Automatic Evaluation

Table 4 compares the performance of different models trained on MicDialogue using automatic evaluation metrics. ChatGPT outperformed Transformer and CDial-GPT in BLEU and NIST scores, while BART's performance was similar to ChatGPT's. Kng-BART excelled across all metrics, generating responses that are not only contextually relevant but also closely match ground-truth responses in higher-order n-gram similarities. This shows that our knowledge-driven model is effective in producing high-quality responses for maternity and infant care dialogues. Kng-BART, trained with domain-specific knowledge, often generates longer texts and emphasizes certain entities through repetition, as demonstrated in our case studies.

5.3 Human Evaluation

Table 6 shows that Kng-BART excels in most human evaluation metrics compared to other models, though it is slightly less fluent than ChatGPT. These results match the automated evaluation findings, indicating that using external knowledge improves the model's ability to understand implicit semantic information beyond the dialogue history, leading to more knowledgeable responses. Notably, Kng-BART's topic alignment is higher than that of the ground-truth responses, suggesting the model is well-tailored to maternity and infant care and effectively captures relevant topic information for better domain-specific responses.

5.4 Case Study

In this section, we provide two case studies to further demonstrate the usability of our dataset, as shown in Figure **??**. From these case studies, we observe that all models generated meaningful responses based on the given dialogue history. However, responses generated by ChatGPT and BART

Dialogue History **Patient**: Hello, I'm two months pregnant. Recently, while monitoring my blood sugar at home, I found it a bit high, and I'm worried. **Doctor**: I understand. During pregnancy, some women do face high blood sugar, possibly due to increased insulin resistance. How is your diet and exercise routine? **Patient**: I maintain a balanced diet and engaging in moderate exercise daily, but my blood sugar is still high.
Ground-truth: Very understandable. Maintaining a healthy diet and exercise is crucial for controlling blood sugar. I recommend continuing your healthy lifestyle, and we will closely monitor your blood sugar levels.
ChatGPT: Thank you for sharing this information. A balanced diet and moderate exercise are good practices. I suggest monitoring your blood sugar regularly.
BART: Understood. Your diet and exercise habits sound healthy.
Kng-BART: We'll monitor your blood glucose closely. It's crucial to understand that hormonal fluctuations during gestation can impact insulin sensitivity. Depending on future readings, we might explore dietary adjustments or additional measures to ensure optimal glycemic control.
Dialogue History **Inquirer**: Feeling low, crying for no reason. Almost seven months pregnant, unsure if it's due to hormones. **Consultant**: You're experiencing emotional fluctuations, and hormonal changes during pregnancy can impact it. I'd like to understand more. Is there something triggering these breakdowns? **Inquirer**: Becoming sensitive, increasing pressure from home and work. Feeling like I can't handle everything.
Ground-truth: Thanks for sharing. Emotional fluctuations are common during pregnancy, especially due to hormonal changes. To understand more, is there anything triggering these breakdowns?
ChatGPT: Emotional fluctuations during pregnancy are normal, likely due to hormonal changes. Is the sensitivity and increased pressure you mentioned contributing to these emotional breakdowns?
BART: Emotional fluctuations during pregnancy are normal and may be related to hormones. Suggest continuing to observe.
Kng-BART: Thank you for sharing. Hormonal fluctuations during pregnancy can impact emotions. The increased sensitivity and pressures from home and work you mentioned may be contributing factors. Let's explore specific contexts for a targeted emotional management strategy and consider professional psychological support.

Table 5: Examples of the generated responses translated in English.

Model/Metric	Relevance	Informative	Fluency	Topic
Transformer	1.89	2.23	2.68	2.19
CDial-GPT	2.13	2.35	3.11	2.33
ChatGPT	3.49	3.23	**3.85**	2.61
BART	2.86	2.64	3.02	2.52
Kng-BART	**3.63**	**3.41**	3.77	**3.10**
Ground-truth	3.95	3.51	3.82	2.99

Table 6: Performance comparison by human evaluation.

demonstrated certain limitations, notably in topical coherence and emotional support. The ChatGPT tended to produce verbose and fluent responses, while the BART model tended towards brevity, resulting in shorter sentences. In contrast, our Kng-BART model exhibited superior capabilities in generating responses of higher quality and enhanced empathy. This can be attributed to the adaptive knowledge selection mechanism embedded within our model, which effectively integrates pertinent external knowledge. By doing so, our model facilitates the generation of responses that are not only topic-related but also emotion-supportive.

5.5 Further Discussion on Our Dataset

Although our dataset is specialized for maternity and infant care, we believe it has potential for generalization on other AI sub-fields. For example, by analyzing the commonalities and differences between maternity and infant care and other healthcare domains, we can expand and adjust the knowledge graphs within our benchmark model. The way we annotate topics, diseases, symptoms, and causes can be adapted to fields such as geriatric care or chronic disease management. This allows for more accurate information extraction and model training in those domains. Moreover, the dialogue collection and processing methods can also be applied with appropriate modifications, enabling the creation of high-quality datasets for other specialized dialogue tasks.

6 Conclusion and Future Work

In this paper, we present MicDialogue, a dataset focused on maternity and infant care, meticulously compiled from short videos and blog posts across five Chinese social media platforms. This dataset encompasses a rich variety of topics, all accompanied by detailed annotations. To enable effective modeling of these dialogues, we introduce a novel benchmark model powered by knowledge graphs and prompt learning. Through rigorous testing, we demonstrate MicDialogue's remarkable effectiveness in generating accurate and relevant responses, and provide benchmark results that serve as a solid foundation for future research. Future work could focus on developing highly-specialized large language models dedicated to maternity and infant care, while also delving deeper into exploring mental health issues unique to this context.

Ethical Statement

We focused on maternity and infant care dialogues, with strict privacy protection measures. We used publicly available data from five major Chinese social media platforms. Doctors obtained patient consent before posting consultation videos, and no personal details were shared. The dataset is fully anonymous, and user identities cannot be traced.

Acknowledgments

This work was completed during a collaborative project with Xiamen Meiyou Co., Ltd. This work was funded by the collaborative project with Meiyou Company, namely Research on the Construction of a General-Purpose Intelligent Dialogue Corpus and Dialogue System for Female Users. We also thank all the participants involved in the data collection and annotation process, whose valuable contributions have enriched our dialogue corpus.

Contribution Statement

Authors marked with † have equal contributions. Authors marked with ‡ are corresponding authors.

References

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Bao *et al.*, 2019] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*, 2019.

[Cai *et al.*, 2018] Deng Cai, Yan Wang, Victoria Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. Skeleton-to-response: Dialogue generation guided by retrieval memory. *arXiv preprint arXiv:1809.05296*, 2018.

[Cao *et al.*, 2020] Lei Cao, Huijun Zhang, and Ling Feng. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*, 24:87–102, 2020.

[Chen *et al.*, 2023] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817, 2023.

[Dinan *et al.*, 2018] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

[Doddington, 2002] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[Ghazvininejad *et al.*, 2018] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Hosseini and Caragea, 2021] Mahshid Hosseini and Cornelia Caragea. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026, 2021.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[Li *et al.*, 2015] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[Liu *et al.*, 2019] Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. Knowledge aware conversation generation with explainable reasoning over augmented graphs. *arXiv preprint arXiv:1903.10245*, 2019.

[Liu *et al.*, 2021] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics, 2021.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2023] Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics, 2023.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Rashkin *et al.*, 2019] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5370–5381, 2019.

[Serban *et al.*, 2017] Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Su *et al.*, 2020] Yixuan Su, Deng Cai, Yan Wang, Simon Baker, Anna Korhonen, Nigel Collier, and Xiaojiang Liu. Stylistic dialogue generation via information-guided reinforcement learning strategy. *arXiv preprint arXiv:2004.02202*, 2020.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[Wang *et al.*, 2020] Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 91–103. Springer, 2020.

[Weng *et al.*, 2017] Heng Weng, Ziqing Liu, Shixing Yan, Meiyu Fan, Aihua Ou, Dacan Chen, and Tianyong Hao. A framework for automated knowledge graph construction towards traditional chinese medicine. In *Health Information Science: 6th International Conference, HIS 2017*, pages 170–181. Springer, 2017.

[Wu *et al.*, 2020a] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5811–5820, 2020.

[Wu *et al.*, 2020b] Yu Wu, Yunli Wang, and Shujie Liu. A dataset for low-resource stylized sequence-to-sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9290–9297, 2020.

[Wu *et al.*, 2021] Sixing Wu, Minghui Wang, Dawei Zhang, Yang Zhou, Ying Li, and Zhonghai Wu. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph network. In *IJCAI*, pages 3964–3970, 2021.

[Xiang *et al.*, 2023] Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. Care-mi: Chinese benchmark for misinformation evaluation in maternity and infant care. *arXiv preprint arXiv:2307.01458*, 2023.

[Xu *et al.*, 2022] Bo Xu, Hongtong Zhang, Jian Wang, Xiaokun Zhang, Dezhi Hao, Linlin Zong, Hongfei Lin, and Fenglong Ma. Realmeddial: A real telemedical dialogue dataset collected from online chinese short-video clips. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, pages 3342–3352, 2022.

[Yang *et al.*, 2023] Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. Refgpt: Dialogue generation of gpt, by gpt, and for gpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, 2023.

[Yunxiang *et al.*, 2023] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

[Zeng *et al.*, 2020] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, 2020.

[Zhang *et al.*, 2019] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[Zheng *et al.*, 2021] Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14558–14567, 2021.

[Zhou *et al.*, 2018] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.