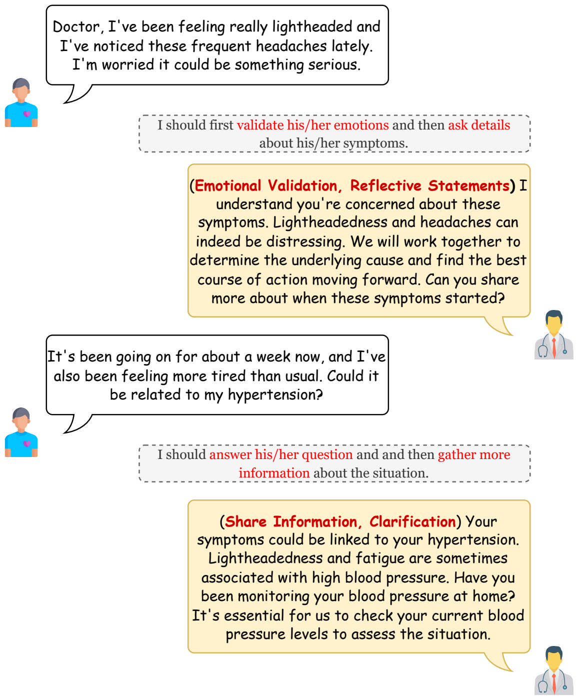
EHDChat: A Knowledge-Grounded, Empathy-Enhanced Language Model for Healthcare Interactions

Shenghan Wu **Wynne Hsu** **Mong Li Lee**
Institute of Data Science
National University of Singapore
shenghan@nus.edu.sg, {whsu, leeml}@comp.nus.edu.sg

Abstract

Large Language Models (LLMs) excel at a range of tasks but often struggle with issues like hallucination and inadequate empathy support. To address hallucinations, we ground our dialogues in medical knowledge sourced from external repositories such as Disease Ontology and DrugBank. To improve empathy support, we develop the Empathetic Healthcare Dialogues dataset, which utilizes multiple dialogue strategies in each response. This dataset is then used to fine-tune an LLM, and we introduce a lightweight, adaptable method called Strategy Combination Guidance to enhance the emotional support capabilities of the fine-tuned model, named EHDChat. Our evaluations show that EHDChat significantly outperforms existing models in providing emotional support and medical accuracy, demonstrating the effectiveness of our approach in enhancing empathetic and informed AI interactions in healthcare.



Doctor, I've been feeling really lightheaded and I've noticed these frequent headaches lately. I'm worried it could be something serious.

I should first validate his/her emotions and then ask details about his/her symptoms.

(**Emotional Validation, Reflective Statements**) I understand you're concerned about these symptoms. Lightheadedness and headaches can indeed be distressing. We will work together to determine the underlying cause and find the best course of action moving forward. Can you share more about when these symptoms started?

It's been going on for about a week now, and I've also been feeling more tired than usual. Could it be related to my hypertension?

I should answer his/her question and and then gather more information about the situation.

(**Share Information, Clarification**) Your symptoms could be linked to your hypertension. Lightheadedness and fatigue are sometimes associated with high blood pressure. Have you been monitoring your blood pressure at home? It's essential for us to check your current blood pressure levels to assess the situation.

Figure 1: Example chat from the proposed EHD dataset depicting a patient (left) receiving emotional support from a doctor (right). Conversation strategies are highlighted in red.

1 Introduction

The integration of technology into healthcare continues to foster innovations aimed at enhancing patient outcomes. The emergence of Large Language Models (LLMs) has significantly advanced their ability to provide responses to user queries concerning medications, dispense general health and wellness advice, and offer support for mental health issues (Sweeney et al., 2021; Lai et al., 2023; Abbasian et al., 2023; Ghosh et al., 2024). However, the effectiveness of these LLMs is contingent upon the quality of the training data they receive.

One limitation is the tendency of LLM to hallucinate and generate erroneous or fabricated medical knowledge which may have serious consequences. Another limitation is their frequent disregard for emotional cues, which can result in interactions that feel insincere and disconnected from the user's emotional state. This issue highlights a gap in the ability of LLMs to engage in genuinely empathetic communication, which is essential for addressing the nuanced needs of patients (Lai et al., 2021).

To address this gap, we need high-quality training data that can provide accurate responses and reflects human emotions. Existing emotional support datasets such as ESConv (Liu et al., 2021a), ExTES (Zheng et al., 2023c), and Reddit mental health forums (Sharma et al., 2020) offer valuable insights into empathetic interactions. However, these resources often lack comprehensive medical knowledge, limiting their applicability in healthcare contexts. Conversely, medical datasets like MedDialog (Zeng et al., 2020), CovidDialog (Ju et al., 2020) and CMtMedQA (Yang et al., 2024) provide extensive healthcare information but frequently lack the depth of genuine empathetic interactions.

Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), pages 141–151
November 16, 2024 ©2024 Association for Computational Linguistics

We develop an Empathetic Healthcare Dialogue (EHD) dataset[1] to train smaller models in generating human-like empathetic responses within the healthcare domain. To construct this dataset, we leverage LLMs, known to have gathered a wide array of information from vast human corpora and specialized databases. However, recognizing the potential for LLM hallucination, we ground our dialogues in external verified medical knowledge repositories, thereby enhancing the reliability and the depth of conversations across a wide range of healthcare topics and concerns.

We also devise a method to incorporate conversational strategies such that each response can utilize multiple strategies. We fine-tune an LLM using the EHD dataset and introduce a lightweight, plug-and-play method, called Strategy Combination Guidance (SCG) to guide the model in selecting a good combination of strategies to improve response quality. We further integrate Retrieval-Augmented Generation (RAG) technology (Lewis et al., 2020), which provides relevant context for more accurate and contextually appropriate responses, resulting in our EHDChat.

Quantitative experiment results show that EHDChat is far superior in providing emotional support and giving accurate medical information. Human evaluation reveals a preference for the responses generated by EHDChat. Case studies also illustrate that employing a variety of conversational strategies leads to more natural dialog.

2 Related Work

Research on medical dialogue systems has primarily focused on developing question-answering systems (Wu et al., 2024; Yang et al., 2022; Han et al., 2023). ChatDoctor (Li et al., 2023) utilizes ChatGPT to generate synthetic medical conversation data, thereby allowing the fine-tuning of a LLaMA model that is able to simulate dialogues between doctors and patients. DoctorGLM (Xiong et al., 2023) adapted this approach to develop a system that simulates doctor-patient conversations in Chinese. These dialogue systems generally do not provide adequate emotional support due to the scarcity of datasets that include effective emotional support conversations (Rashkin et al., 2019).

Early efforts to curate emotional support datasets primarily focused on transcribing therapist sessions (Liu et al., 2023; Shen et al., 2020) or compiling emotional question-answer pairs from online platforms (Garg et al., 2022; Sharma et al., 2020; Sun et al., 2021; Medeiros and Bosse, 2018). However, these datasets often lack proper design and consist mainly of single-turn responses.

ESConv (Liu et al., 2021b) incorporates conversation strategies and utilize crowdsourcing to create high-quality and multi-turn emotional support datasets. Building upon ESConv, ExTES (Zheng et al., 2023c) employs LLMs to synthesize additional scenarios and dialogues that incorporate more emotional support strategies, further expanding the available data for training these specialized dialogue systems.

Current medical dialogue datasets lack emotional content, while empathetic datasets lack medical knowledge. Our work aims to create a new dataset that combines empathetic dialogues with accurate medical information. This fills a gap in existing resources for doctor-patient conversations, providing a more complete dataset for training AI in healthcare communication.

3 EHD Dataset

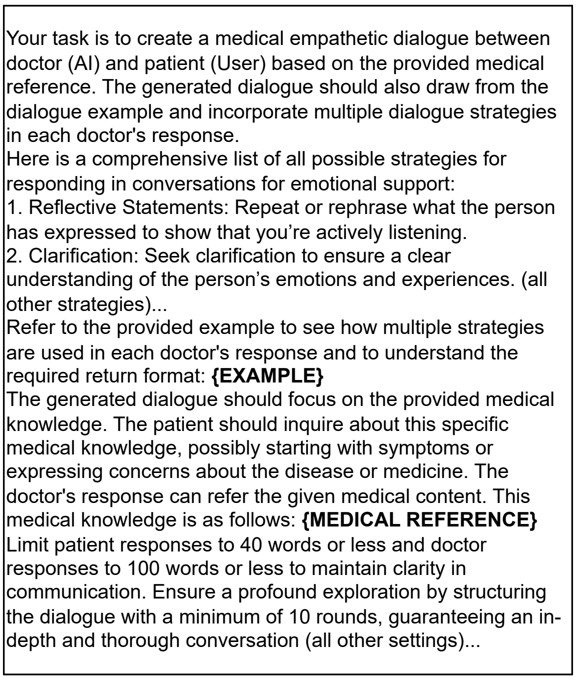
Our goal is to generate a dataset comprising of a wide range of synthetic, multi-turn dialogues between doctors and patients that are not only emotionally supportive, but also clinically informative. We produce dialogues via the self-chat technique (Xu et al., 2023) and reference-based dialogue generation methods (Dai et al., 2022; Yang et al., 2023; Kim et al., 2023).

Previous research (Liu et al., 2021b; Zheng et al., 2023c) has shown that emotional support strategies can boost empathy levels (Hill, 2020; Organization et al., 2020). To enhance the empathetic nature of the dialogues, we incorporate a mix of known emotional support strategies, drawing on research that demonstrates the increased effectiveness of combining two strategies within a single response (Tu et al., 2022). Figure 2 shows the prompt template used for dialogue generation which includes basic instructions, a medical knowledge reference, dialogue strategies and an example dialogue.

For medical knowledge on diseases, we utilize the Disease Ontology (Schriml, 2018) as our reference source. As the information available in the Disease Ontology is limited to basic descriptions of diseases, we leverage the external links stored in Disease Ontology to directly extract data from the corresponding web pages. For knowledge on

---

[1] https://huggingface.co/datasets/ericw955/EHD.



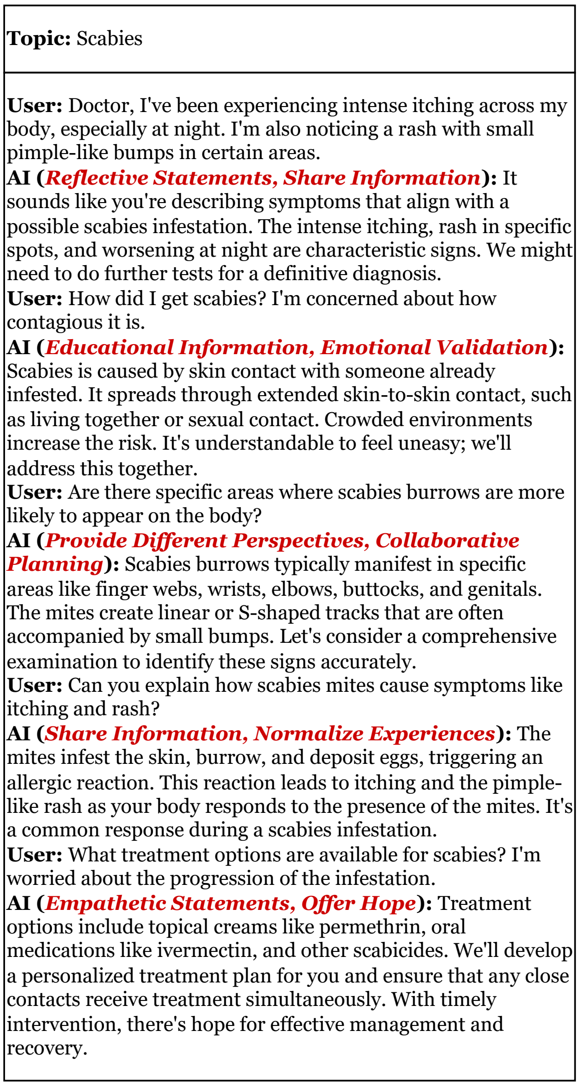
Your task is to create a medical empathetic dialogue between doctor (AI) and patient (User) based on the provided medical reference. The generated dialogue should also draw from the dialogue example and incorporate multiple dialogue strategies in each doctor's response.
Here is a comprehensive list of all possible strategies for responding in conversations for emotional support:
1. Reflective Statements: Repeat or rephrase what the person has expressed to show that you're actively listening.
2. Clarification: Seek clarification to ensure a clear understanding of the person's emotions and experiences. (all other strategies)...
Refer to the provided example to see how multiple strategies are used in each doctor's response and to understand the required return format: **{EXAMPLE}**
The generated dialogue should focus on the provided medical knowledge. The patient should inquire about this specific medical knowledge, possibly starting with symptoms or expressing concerns about the disease or medicine. The doctor's response can refer the given medical content. This medical knowledge is as follows: **{MEDICAL REFERENCE}**
Limit patient responses to 40 words or less and doctor responses to 100 words or less to maintain clarity in communication. Ensure a profound exploration by structuring the dialogue with a minimum of 10 rounds, guaranteeing an in-depth and thorough conversation (all other settings)...

Figure 2: Prompt for generating dialogues.

drugs, we utilize the public dataset called DrugBank (Wishart DS, 2017) to extract the description, effect and dosage of drugs. In total, we compile 4,223 entries related to disease knowledge and 3,229 entries on drug knowledge.

We randomly select 100 dialogues from ExTES (Zheng et al., 2023c), and retain the first sentence of each dialogue as the initial prompt. Then we instruct GPT-4 using the prompt in Figure 2 to generate responses that incorporate two emotional support strategies per response. The medical reference mentioned in the prompt is selected from a predetermined set of commonly encountered diseases or drugs information. Figure 3 shows an example of the dialogue generated. The conversation is focused on a specific disease Scabies and each response employs at least 2 strategy combinations, highlighted in red.

After generation, we evaluate the accuracy of the medical content and the correct integration of multiple emotional support strategies in each response. Dialogues identified as medically inaccurate or lacking in emotional support strategies are removed. The remaining generated dialogues form the example dialogues in our prompt to generate additional dialogues, with each dialogue centered around a single medical concept.



Topic: Scabies

User: Doctor, I've been experiencing intense itching across my body, especially at night. I'm also noticing a rash with small pimple-like bumps in certain areas.
AI (*Reflective Statements, Share Information*): It sounds like you're describing symptoms that align with a possible scabies infestation. The intense itching, rash in specific spots, and worsening at night are characteristic signs. We might need to do further tests for a definitive diagnosis.
User: How did I get scabies? I'm concerned about how contagious it is.
AI (*Educational Information, Emotional Validation*): Scabies is caused by skin contact with someone already infested. It spreads through extended skin-to-skin contact, such as living together or sexual contact. Crowded environments increase the risk. It's understandable to feel uneasy; we'll address this together.
User: Are there specific areas where scabies burrows are more likely to appear on the body?
AI (*Provide Different Perspectives, Collaborative Planning*): Scabies burrows typically manifest in specific areas like finger webs, wrists, elbows, buttocks, and genitals. The mites create linear or S-shaped tracks that are often accompanied by small bumps. Let's consider a comprehensive examination to identify these signs accurately.
User: Can you explain how scabies mites cause symptoms like itching and rash?
AI (*Share Information, Normalize Experiences*): The mites infest the skin, burrow, and deposit eggs, triggering an allergic reaction. This reaction leads to itching and the pimple-like rash as your body responds to the presence of the mites. It's a common response during a scabies infestation.
User: What treatment options are available for scabies? I'm worried about the progression of the infestation.
AI (*Empathetic Statements, Offer Hope*): Treatment options include topical creams like permethrin, oral medications like ivermectin, and other scabicides. We'll develop a personalized treatment plan for you and ensure that any close contacts receive treatment simultaneously. With timely intervention, there's hope for effective management and recovery.

Figure 3: Sample dialogue generated.

Table 1 gives the characteristics of the dataset. For comparison, we also include the statistics of existing ESConv and ExTES datasets. In total, we generate 33,303 dialogues, with an average of 12.0 utterances per dialogue and each utterance averaging 32.6 words. Table 2 shows the distribution of 16 conversation strategies used in the dataset. We see that 'Share Information' and 'Offer Hope' are the top-2 most prevalent.

Table 3 displays the top-10 most frequently occurring strategy combinations out of a total of 240 possibilities. The most frequently used combination is 'Reflective Statements+Share Information' with a count of 11,042 instances. This suggests that reflective listening, coupled with providing information, is a prevalent method for offering emotional support in our dataset. Our dataset, featuring doctor-patient dialogues, emphasizes 'Sharing Information' as a crucial aspect of emotional support. In the interactions, the synergy of informative com-

Table 1: Characteristics of EHD dataset, compared to existing ESConv, ExTES datasets.

Attribute	ESConv	ExTES	EHD
# Dialogues	1,053	11,177	**33,303**
# Utterances	31,410	200,393	**393,678**
Avg. length of dialog.	29.8	18.2	**11.8**
Avg. length of utter.	17.8	26.0	**34.5**
# conversation strategies	8	16	**16**
Healthcare-related	No	No	**Yes**
Knowledge-verified	No	No	**Yes**

Table 2: Statistics of conversation strategy in EHD.

Conversation Strategy	Count
Share Information	72,780
Offer Hope	51,309
Emotional Validation	34,473
Reflective Statements	31,985
Provide Perspectives	28,943
Affirmation	28,765
Collaborative Planning	29,193
Suggest Options	20,079
Empathetic Statements	19,241
Promote Self-Care Practices	17,712
Clarification	17,305
Normalize Experiences	14,875
Stress Management	7,482
Others	6,099
Reframe Negative Thoughts	5,710
Avoid Judgment and Criticism	4,056

Table 3: Top-10 combinations of conversation strategies within the same response.

Strategy Combination	Count
Offer Hope + Share Info.	11711
Reflection + Share Info.	11573
Emotional Validation + Reflection	10435
Provide Perspective + Share Info.	8867
Emotional Validation + Share Info.	8382
Clarification + Share Info.	5897
Normalize Experiences + Share Info.	5095
Collaborative Planning + Offer Hope	4830
Share Info. + Suggest Options	4679
Clarification + Reflection	4157

Table 4: Top-5 conversation strategy transitions in EHD.

Strategy Transition	%
Emotional Validation + Reflection → Offer Hope + Share Info.	2.06
Reflection + Share Info. → Provide Perspective + Share Info.	1.10
Reflection + Share Info. → Emotional Validation + Share Info.	1.00
Emotional Validation + Reflection → Clarification + Share Info.	0.95
Clarification + Reflection → Emotional Validation + Share Info.	0.90

munication with other support strategies is a natural and essential pairing, ensuring responses are both empathetic and informative.

Table 4 gives the top-5 strategy transitions in the generated dialogues. We see that the transitions strategies used in the EHD dataset are diverse and often affirm emotions first before providing advice. This approach offers comfort and support to alleviate emotional distress.

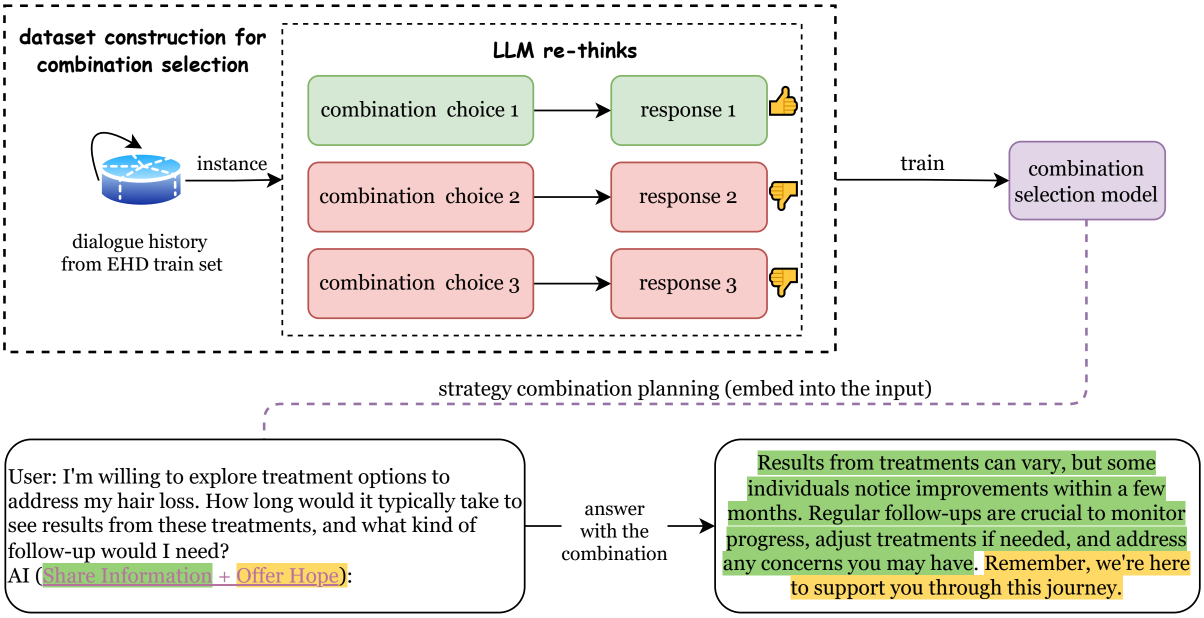
3.1 Dialogue Quality Evaluation

The efficacy of fine-tuning smaller models is heavily dependent on the quality of the training dataset. To demonstrate the superiority of our EHD dataset, we conduct a comparative analysis against two existing datasets: ESConv and ExTES.

Our evaluation framework, based on established research methodologies (Liu et al., 2021b; Zheng et al., 2023a), assessed five key aspects of dialogue quality: **Informativeness** evaluates the degree to which the supporter provides relevant and correct information to the seeker. **Understanding** measures the supporter to correctly interpret and respond to the seeker's intent, emotions, and experiences. **Coherence** gauges whether the dialogue maintains a logical flow. **Helpfulness** checks the extent to which the supporter assists the seeker in resolving their issue and meeting their needs. **Consistency** evaluates the extent to which participants maintain their designated roles throughout the dialogue and demonstrate behaviors that are logically coherent and free from contradictions. All metrics were measured using a Likert Scale, with scores ranging from 1 (lowest) to 5 (highest), ensuring a standardized and comprehensive evaluation. We apply the GPTScore methodology (Fu et al., 2023) to evaluate the quality of dialogues at the dialogue level. The prompt contains the evaluation settings, definitions of each metric and the dialogue.

Table 5 shows the GPTScore for ESConv, ExTES and EHD. Our results indicate that both ExTES and EHD demonstrate a high capacity for comprehending seekers' emotional states and intentions, which surpass ESConv. These datasets provide substantial assistance while maintaining consistent performance throughout the dialogue.

Language toxicity has been a critical concern in NLP research and dialogues between doctors and patients should not contain aggressive or offensive language (Pavlopoulos et al., 2020; Gehman et al., 2020). In assessing the toxicity of synthetic dialogues, we employ the wide-used Perspective API


dataset construction for combination selection
dialogue history from EHD train set
instance
LLM re-thinks
combination choice 1
combination choice 2
combination choice 3
response 1
response 2
response 3
train
combination selection model
strategy combination planning (embed into the input)
answer with the combination

User: I'm willing to explore treatment options to address my hair loss. How long would it typically take to see results from these treatments, and what kind of follow-up would I need?
AI (Share Information + Offer Hope):

Results from treatments can vary, but some individuals notice improvements within a few months. Regular follow-ups are crucial to monitor progress, adjust treatments if needed, and address any concerns you may have. Remember, we're here to support you through this journey.

Figure 4: EHDChat with Strategy Combination Guidance.

Table 5: Comparison of dialogue quality.

Metric	ESConv	ExTES	EHD
Informativeness	2.99	4.11	4.76
Understanding	3.57	4.98	4.98
Coherence	3.50	4.98	4.99
Helpfulness	3.54	4.94	4.98
Consistency	3.46	4.98	4.98

Table 6: Comparison of toxicity.

Category	ESConv	ExTES	EHD
Toxicity	0.190	0.112	**0.053**
Severe Toxicity	0.016	0.006	**0.003**
Identity Attack	0.036	0.017	**0.008**
Insult	0.065	0.032	**0.017**
Profanity	0.141	0.076	**0.029**
Threat	0.033	0.017	**0.010**

to measure the textual toxicity across six indicators for the ESConv, ExTES and EHD datasets. The results are shown in Table 6. Each indicator is scored on a scale from 0 to 1, where lower scores indicate reduced toxicity. Our findings show that our dataset has the lowest scores in all the indicators.

4 EHDChat

Building on the EHD dataset, we develop an empathetic healthcare dialogue system called EHDChat that is able to provide accurate responses to medical queries while demonstrating empathy and understanding of the user's emotional states. This system is specifically designed to handle the dual challenges of medical accuracy and emotional empathy, essential in healthcare settings.

We randomly partition the EHD dataset into a training set comprising of 7,175 medical terms across 31,918 dialogues, and a test set containing the remaining 300 medical terms in 1,385 dialogues. For each dialogue in the training set, we generate training instances where the target output is the next response, and the preceding conversation provides the context. These instances are used to fine-tune the LLaMA-1 (7B) model (Touvron et al., 2023), leveraging its capabilities to handle complex dialogue scenarios.

In order to generate more precise and effective responses, we introduce a Strategy Combination Guidance (SCG) module to optimize the selection of emotional support strategy combinations for each response. We achieve this by taking 5,000 dialogue rounds in the EHD dataset and employ LLM to determine potential strategy combinations for each round. For each strategy combination, we prompt the LLM to generate a corresponding response. The most effective strategy combination is then selected based on which generated response achieves the highest GPTScore.

Figure 4 shows the strategy combination selection process. We use these instances to further refine the fine-tuned LLaMA-1 model, culminating in the development of the proposed EHDChat (Touvron et al., 2023). Note that SCG is a lightweight, plug-and-play module that can be applied to any existing emphathetic healthcare dialogue systems to guide their response generation process.

5 Experiments

We carry out a series of experiments to evaluate the performance of our proposed EHDChat in terms of providing emotional support as well as the accuracy of the shared medical knowledge during the conversation. We implement the following baselines for comparison:

- ESConvChat. This is a fine-tuned LLaMA-1 model using ESConv dataset.
- ExTESChat. This is a fine-tuned LLaMA-1 model using ExTES dataset.
- ChatDoctor. This is the state-of-the-art medical chatbot that uses datasets from medical forums, disease databases, and synthetic datasets of single-turn dialogues between doctors and patients to fine-tune LLaMA-1.

We fine-tuned the models on two A100 80G GPUs with a batch size of 16, using a 3% warmup phase and cosine learning scheduler.

5.1 Emotional Support Capability Evaluation

We first compare the emotional support capabilities of the various models. Zheng et al. (2023b) has demonstrated that LLMs, when used as an evaluator, can achieve high consistency with human evaluation. We employ GPT-4 as the judge, and design evaluation prompts based on the following metrics (Liu et al., 2021a), to assess the emotional support capabilities of the generated responses:

- **Suggestion** gauges the extent of pertinent information provided by the model.
- **Identification** measures the model's comprehension of the context, including its grasp on users' concerns and emotions.
- **Fluency** assesses the logical flow and contextual appropriateness of the responses.
- **Comforting** examines the model's capacity to provide emotional support.
- **Overall** provides a comprehensive evaluation of the responses' general effectiveness.

Table 7 shows the results for the different test sets. We see that **EHDChat achieves the highest scores for all the metrics across all the test sets.** This indicates that incorporating multiple dialogue strategies within a single response can significantly enhance the performance and emotional support ability of the model's replies. Although ESConvChat is fine-tuned on ESConv dataset which has incorporated diverse emotional support tactics, its scores are the lowest across all the metrics. Closer examination reveals that ESConv consists of human dialogues that tend to be brief and concise. This may hamper the ability to generate high-caliber empathetic responses.

Table 7: GPTScore for emotional support capabilities. S(uggestion), I(dentification), F(luency), C(omforting)

Model	S	I	F	C	Overall
ESConvChat	2.20	2.76	3.50	2.71	2.68
ExTESChat	2.80	3.61	4.24	3.92	3.59
ChatDoctor	2.51	2.71	2.94	2.61	2.58
EHDChat	**3.43**	**4.04**	**4.56**	**4.55**	**4.05**

(a) ESConv Test Set

Model	S	I	F	C	Overall
ESConvChat	2.51	3.25	3.99	3.31	3.16
ExTESChat	3.76	4.37	4.83	4.52	4.36
ChatDoctor	3.26	3.51	3.93	3.50	3.45
EHDChat	**3.96**	**4.42**	**4.85**	**4.63**	**4.42**

(b) ExTES Test Set

Model	S	I	F	C	Overall
ESConvChat	2.92	3.06	3.82	2.73	3.06
ExTESChat	3.77	4.08	4.74	3.97	4.08
ChatDoctor	3.50	3.15	3.30	2.63	3.10
EHDChat	**4.12**	**4.45**	**4.92**	**4.38**	**4.42**

(c) EHD Test Set

5.2 Knowledge Correctness Evaluation

Next, we examine the correctness of the medical knowledge in the model's generated responses. Here we use the MedQuAD (Ben Abacha and Demner-Fushman, 2019), a comprehensive medical dataset including various question-answer pairs created from 12 NIH websites, as the test set.

In addition to using the standard metrics such as BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE-L (Lin, 2004) to measure the degree of overlap between the model's responses and the ground truth, we also employ GPT-4 to evaluate the correctness of the medical knowledge in the responses as GPT-4 shows great medical abilities in recent researches (OpenAI, 2023; Nori et al., 2023). This is because in the medical field, minor discrepancies in the responses (high BLEU, METEOR or ROUGE scores) may not necessarily imply that the response is correct.

Table 8 shows the results. We see that **EHDChat exhibits superior correctness compared to the other models.** EHDChat significantly enhances the

Table 8: Knowledge correctness in MedQuAD test set.

Model	%Correct	BLEU	MET	ROUGE
ESConvChat	17.34	0.45	4.03	7.26
ExTESChat	36.14	1.14	9.78	13.98
ChatDoctor	49.46	**1.67**	**16.50**	**18.04**
EHDChat	66.09	0.69	10.38	14.18

Table 9: Human evaluation of EHDChat vs. ChatDoctor.

	EHDChat win	Tie	EHDChat lose
Suggestion	34%	30%	**36%**
Identification	**43%**	40%	17%
Fluency	**76%**	21%	3%
Comforting	**65%**	32%	3%
Overall	**62%**	24%	14%

model's precision in responding to medical queries by leveraging external knowledge for improved response correctness in the domain of medical dialogue. Although ChatDoctor's responses has the highest similarity to the ground truth in MedQuAD, its correctness is lower than EHDCHAT as ChatDoctor's responses tend to be more noisy with unverified content which resulted in a decline in the model's precision in delivering correct responses.

5.3 Human Evaluation

We also conduct a human evaluation following the methodology in (Li et al., 2022; Sabour et al., 2022). We compare EHDChat with ChatDoctor as its objectives is most closely aligned with ours. We randomly sample 100 dialogues from the EHD test set and use these as dialogue history for both EHDChat and ChatDoctor to generate responses.

We recruit five undergraduate students from diverse backgrounds to conduct a human evaluation of the generated responses. The evaluation involves six indicators: suggestion, identification, fluency, comforting, empathy, and overall quality. For each indicator, the students rate the responses as a win, tie, or loss for our model compared to ChatDoctor.

Table 9 shows the results. Both ChatDoctor and EHDChat show comparable proficiency in offering suggestions (Suggestion). However, ChatDoctor is less effective in providing emotional support (Comforting) and maintaining dialogue quality (Identification, Fluency and Overall). This demonstrates the effectiveness of EHDChat in providing emotional support in the healthcare domain.

6 Case Studies

Finally, we present case studies to show the emotional support capabilities and medical knowledge correctness of EHDChat compared to ExTESChat, ESConvChat, and ChatDoctor.

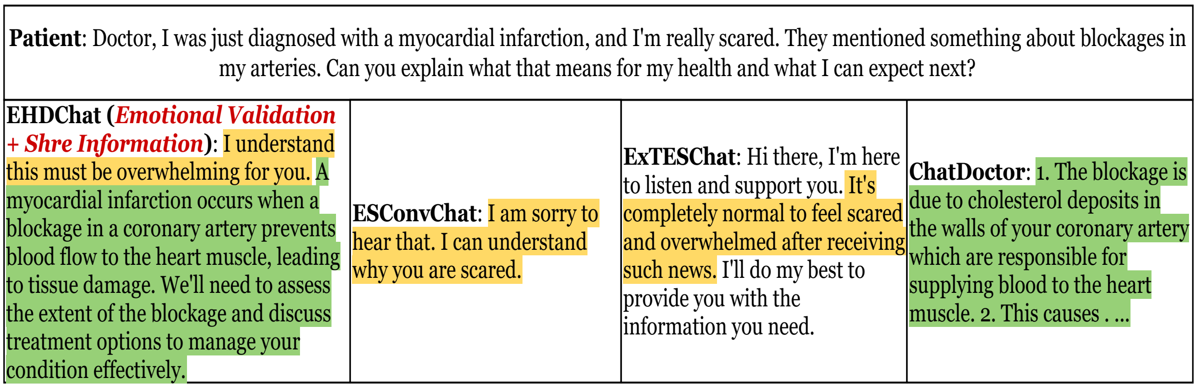
Figure 5 shows the responses generated by the various models given the same dialogue input. We observe that ExTESChat focuses solely on emotional support (highlighted in yellow), failing to address patient query about his diagnosis of myocardial infarction. On the other hand, ChatDoctor's response exclusively focuses on the impact of myocardial infarction on health (highlighted in green), neglecting the patient's emotional state. In contrast, EHDChat employ a combination of dialogue strategies (Emotional Validation+Share Information) to first acknowledge the patient's emotion before addressing his medical query, leading to a more balanced response.

Figure 6 demonstrates the strategy transitions used in EHDChat and how the transitions are able to provide better emotional support and answer medical queries raised by users. We see that EHDChat is able to employ a variety of conversational strategy and the dialogues are more natural offering both emotional support and answering the medical queries raised by users. In contrast, ESConvChat repeatedly responses with "I understand..", while ExTESChat primarily offers emotional support with occasional superficial medical advice, and ChatDoctor's responses only focus on addressing the medical queries.

7 Conclusion

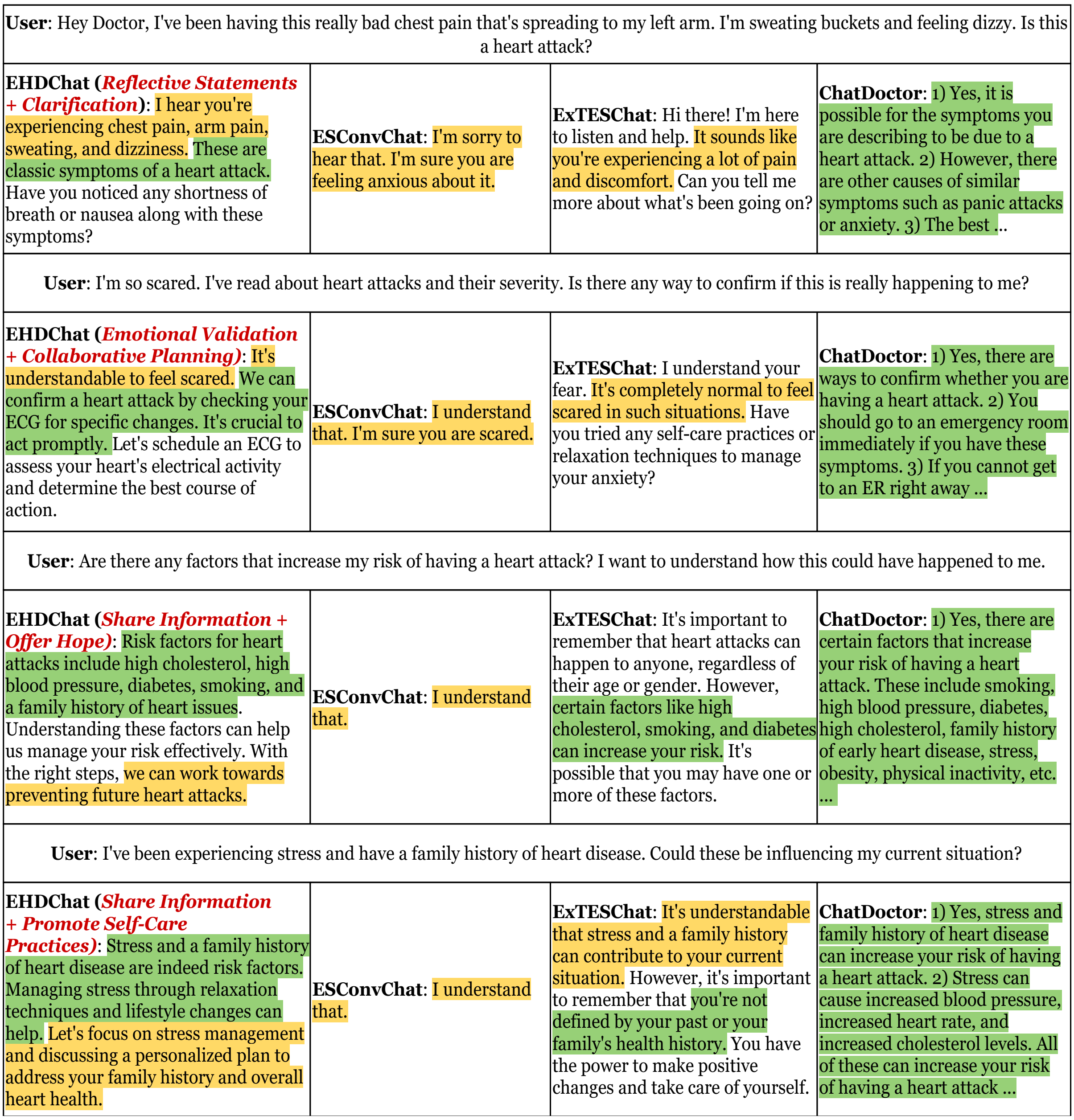
We have described a new approach for generating empathetic multi-turn dialogues for the healthcare domain. By allowing a single response to contain multiple conversation strategies, we are able to enhance the level of emotional support to the user. Our approach incorporates verified medical references and employs a robust verification method to reduce errors and ensure the authenticity of the information in the generated dialogues. Our analysis confirms that the EHD dataset is both empathetic and medically accurate. Our EHDChat, trained on the EHD dataset, has shown improvement both in terms of emotional support and medical proficiency, surpassing state-of-the-art chatbots.

Despite its strengths, EHDChat may face challenges when dealing with complex medical issues, underscoring the urgent need for further research to reduce inaccuracies and enhance the reliability of chatbots in critical applications.



Patient: Doctor, I was just diagnosed with a myocardial infarction, and I'm really scared. They mentioned something about blockages in my arteries. Can you explain what that means for my health and what I can expect next?			
EHDChat (*Emotional Validation + Shre Information*): I understand this must be overwhelming for you. A myocardial infarction occurs when a blockage in a coronary artery prevents blood flow to the heart muscle, leading to tissue damage. We'll need to assess the extent of the blockage and discuss treatment options to manage your condition effectively.	**ESConvChat**: I am sorry to hear that. I can understand why you are scared.	**ExTESChat**: Hi there, I'm here to listen and support you. It's completely normal to feel scared and overwhelmed after receiving such news. I'll do my best to provide you with the information you need.	**ChatDoctor**: 1. The blockage is due to cholesterol deposits in the walls of your coronary artery which are responsible for supplying blood to the heart muscle. 2. This causes . ...

Figure 5: Case study to show the responses generated by the various models given the same dialogue input.



User: Hey Doctor, I've been having this really bad chest pain that's spreading to my left arm. I'm sweating buckets and feeling dizzy. Is this a heart attack?			
EHDChat (*Reflective Statements + Clarification*): I hear you're experiencing chest pain, arm pain, sweating, and dizziness. These are classic symptoms of a heart attack. Have you noticed any shortness of breath or nausea along with these symptoms?	**ESConvChat**: I'm sorry to hear that. I'm sure you are feeling anxious about it.	**ExTESChat**: Hi there! I'm here to listen and help. It sounds like you're experiencing a lot of pain and discomfort. Can you tell me more about what's been going on?	**ChatDoctor**: 1) Yes, it is possible for the symptoms you are describing to be due to a heart attack. 2) However, there are other causes of similar symptoms such as panic attacks or anxiety. 3) The best ...
User: I'm so scared. I've read about heart attacks and their severity. Is there any way to confirm if this is really happening to me?			
EHDChat (*Emotional Validation + Collaborative Planning)*: It's understandable to feel scared. We can confirm a heart attack by checking your ECG for specific changes. It's crucial to act promptly. Let's schedule an ECG to assess your heart's electrical activity and determine the best course of action.	**ESConvChat**: I understand that. I'm sure you are scared.	**ExTESChat**: I understand your fear. It's completely normal to feel scared in such situations. Have you tried any self-care practices or relaxation techniques to manage your anxiety?	**ChatDoctor**: 1) Yes, there are ways to confirm whether you are having a heart attack. 2) You should go to an emergency room immediately if you have these symptoms. 3) If you cannot get to an ER right away ...
User: Are there any factors that increase my risk of having a heart attack? I want to understand how this could have happened to me.			
EHDChat (*Share Information + Offer Hope)*: Risk factors for heart attacks include high cholesterol, high blood pressure, diabetes, smoking, and a family history of heart issues. Understanding these factors can help us manage your risk effectively. With the right steps, we can work towards preventing future heart attacks.	**ESConvChat**: I understand that.	**ExTESChat**: It's important to remember that heart attacks can happen to anyone, regardless of their age or gender. However, certain factors like high cholesterol, smoking, and diabetes can increase your risk. It's possible that you may have one or more of these factors.	**ChatDoctor**: 1) Yes, there are certain factors that increase your risk of having a heart attack. These include smoking, high blood pressure, diabetes, high cholesterol, family history of early heart disease, stress, obesity, physical inactivity, etc. ...
User: I've been experiencing stress and have a family history of heart disease. Could these be influencing my current situation?			
EHDChat (*Share Information + Promote Self-Care Practices)*: Stress and a family history of heart disease are indeed risk factors. Managing stress through relaxation techniques and lifestyle changes can help. Let's focus on stress management and discussing a personalized plan to address your family history and overall heart health.	**ESConvChat**: I understand that.	**ExTESChat**: It's understandable that stress and a family history can contribute to your current situation. However, it's important to remember that you're not defined by your past or your family's health history. You have the power to make positive changes and take care of yourself.	**ChatDoctor**: 1) Yes, stress and family history of heart disease can increase your risk of having a heart attack. 2) Stress can cause increased blood pressure, increased heart rate, and increased cholesterol levels. All of these can increase your risk of having a heart attack ...

Figure 6: Case study to show the responses generated by various models over dialogues that span mutiple rounds.

Limitations

While EHDChat has demonstrated good ability to answer medical queries, it is still limited in handling complex medical issues. The augmentation of external knowledge via RAG does not fully mitigate the risk of providing incorrect answers. This may have serious implications in the medical domain which should be addressed in future iterations to ensure safety and trust in medical settings.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001-2B).

References

Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2023. Conversational health agents: A personalized llm-powered agent framework. *arXiv preprint arXiv:2310.02374*.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning*, pages 4558–4586. PMLR.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J Dorr. 2022. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. *arXiv preprint arXiv:2207.04674*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, and Pengtao Xie. 2020. Covididialog: Medical dialogue datasets about covid-19. *https://github.com/UCSD-AI4H/COVID-Dialogue*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Yuanyuan Lai, Eleni Lioliou, and Panos Panagiotopoulos. 2021. Understanding users' switching intention to ai-powered healthcare chatbots. In *ECIS*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021b. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS 2018, Toledo, Spain, June 20–22, 2018, Proceedings 16*, pages 196–209. Springer.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

World Health Organization et al. 2020. Mental health and psychosocial support aspects of the covid-19 response. Technical report, WHO Regional Office for the Western Pacific.

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

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

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

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Mitraka E. Munro J. Tauber B. Schor M. Nickle L. Schriml, L. M. 2018. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research.*

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

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.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *arXiv preprint arXiv:2106.01702*.

Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D Mulvenna, Siobhan O'neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, et al. 2021. Can chatbots help support a person's mental health? perceptions and views from mental healthcare professionals and experts. *ACM Transactions on Computing for Healthcare*, 2(3):1–15.

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

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. *Preprint*, arXiv:2203.13560.

Guo AC Lo EJ Marcu A Grant JR Sajed T Johnson D Li C Sayeeda Z Assempour N Iynkkaran I Liu Y Maciejewski A Gale N Wilson A Chin L Cummings R Le D Pon A Knox C Wilson M Wishart DS, Feunang YD. 2017. Drugbank5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.*

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

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

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan.

2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

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

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023c. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.