# Balancing Knowledge Delivery and Emotional Comfort in Healthcare Conversational Systems

**Shang-Chi Tsai Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
d08922014@ntu.edu.tw y.v.chen@ieee.org

## Abstract

With the advancement of large language models, many dialogue systems are now capable of providing reasonable and informative responses to patients' medical conditions. However, when patients consult their doctor, they may experience negative emotions due to the severity and urgency of their situation. If the model can provide appropriate comfort and empathy based on the patient's negative emotions while answering medical questions, it will likely offer a more reassuring experience during the medical consultation process. To address this issue, our paper explores the balance between knowledge sharing and emotional support in the healthcare dialogue process. We utilize a large language model to rewrite a real-world interactive medical dialogue dataset, generating patient queries with negative emotions and corresponding medical responses aimed at soothing the patient's emotions while addressing their concerns. The modified data serves to refine the latest large language models with various fine-tuning methods, enabling them to accurately provide sentences with both emotional reassurance and constructive suggestions in response to patients' questions. Compared to the original LLM model, our experimental results demonstrate that our methodology significantly enhances the model's ability to generate emotional responses while maintaining its original capability to provide accurate knowledge-based answers.[1]

## 1 Introduction

A healthcare conversational system is a dialogue-based framework specifically developed for the medical domain. Its primary purpose is to interact with patients, systematically collect supplementary symptom information, facilitate preliminary diagnostic processes, and provide automated recommendations for treatment plans (Tang, 2016; Wei et al., 2018; Liao et al., 2022; Zhong et al., 2023). Healthcare conversational systems demonstrate significant potential to enhance the efficiency of diagnostic procedures while reducing the costs associated with patient information collection (Chen et al., 2023a; Wang et al., 2023b).

In recent years, large language models (LLMs), e.g. ChatGPT (OpenAI, 2023), LLaMA (Touvron et al., 2023a,b), Qwen (Yang et al., 2024a) have been widely applied in medical fields and catalyzed substantial technological transformations in dialogue systems. Through high-quality instruction fine-tuning and reinforcement learning based on human feedback (RLHF) (Ouyang et al., 2022), large language models (LLMs) possess the ability to generate human-like responses with exceptional accuracy, thereby redefining the foundational framework of healthcare dialogue systems.

While large language models fine-tuned for medical dialogues have demonstrated the ability to produce knowledge-intensive and contextually appropriate responses (Wu et al., 2023; Han et al., 2023; Chen et al., 2023b; Li et al., 2023b; Singhal et al., 2022, 2023; Toma et al., 2023), a discrepancy persists between the generated responses and the appropriate real-world answers in certain medical consultation scenarios, particularly in emotion-related contexts. For example, in Figure 1, when patients exhibit intense negative emotions during consultations due to the severity of their condition, employing a comforting tone while delivering solutions can significantly mitigate their psychological distress. However, existing LLM-based medical dialogue systems primarily focus on the rationality and accuracy of the responses' knowledge. In contrast, emotion-focused large language models (Liu et al., 2024) emphasize emotion recognition without prioritizing the acquisition of medical knowledge.

In this paper, we aim at developing an approach that effectively trains a model to deliver knowl-

[1]The source code is available at https://github.com/MiuLab/EmoDoctor.



*Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 115–128
May 27–30, 2025. ©2025 Association for Computational Linguistics

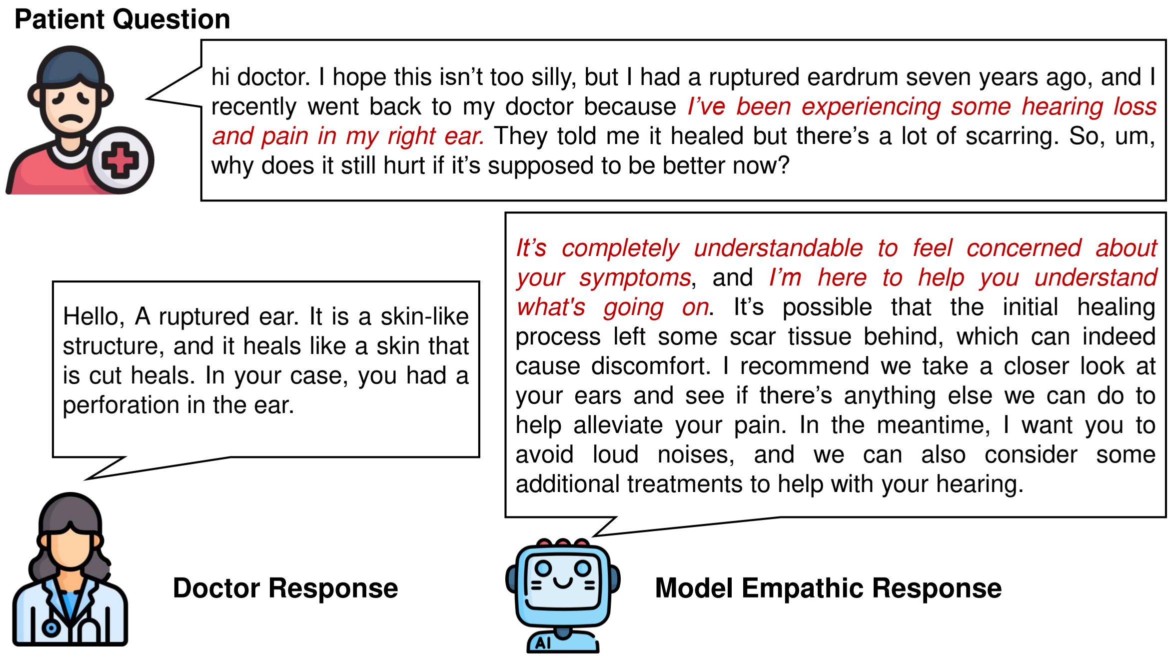


Figure 1: Illustration of an example about the emotional healthcare dialogue system.

edgeable responses while maintaining a balance of emotional comfort, thereby enabling more realistic and human-centric interactions. Inspired by the exceptional creativity of large language models (Tsai et al.; Angel et al.), we first utilized them to modify the emotional tone of real-world doctor-patient dialogues. This approach generated patient queries infused with specific negative emotions, alongside medical responses designed to soothe the patients' negative emotional states. We then applied three distinct approaches to fine-tune the base model using the aforementioned modified dialogues. The three fine-tuning methods are: 1) SFT (Supervised fine-tuning) (Wei et al., 2022), 2) DPO (Direct Preference Optimization) (Rafailov et al., 2024), 3) KTO (Kahneman-Tversky Optimization) (Ethayarajh et al., 2024). These approaches have been validated as effective strategies for aligning large language models to specific tasks. By integrating these techniques, the fine-tuned model can generate responses that balance knowledge delivery with emotional soothing. The effectiveness of our proposed methodology is verified through experiments on another doctor-patient dialogue with emotion-specific scenarios. We further analyze several factors that affect the performance of LLM, including fine-tuning methods, modified datasets, emotional categories, and evaluation models. To the best of our knowledge, this is the first LLM-based medical dialogue system to explore how to balance knowledge expression and empathy in real-world medical conversations. Additionally, our work enables medical dialogue systems to foster more meaningful interactions by addressing both the informational and emotional needs of patients, creating a more supportive consultation experience.

The contributions of this paper are as follows:

- We utilized a large language model to rewrite and generate patient consultations with negative emotions and medical responses aimed at soothing those emotions.
- We experimented with three fine-tuning approaches to enable the model to learn how to balance knowledge delivery and emotional soothing.
- We tested and analyzed the model's performance to determine whether it could effectively balance knowledge and emotional expression on real-world medical dialogue dataset.

## 2 Related Work

### 2.1 Healthcare Conversations System

Healthcare conversational system is an important yet challenging task in the medical domain. In recent advancements, large language models have exhibited remarkable capabilities in downstream tasks, reshaping the foundation of medical dialogue systems. According to the existing literature (Shi et al., 2024), the medical dialogue system can be broadly categorized into two groups based on their association with the emergence of large language models. The methods before the emergence of LLM are divided into three categories: retrieval-based methods, generation-based methods, and hybrid methods (Wang et al., 2023c). Retrieval-based medical dialogue systems are designed to select appropriate responses from the pre-built index (Tao et al., 2021; Zhu et al., 2022). Generation-based methods can be categorized into two approaches: pipeline and end-to-end. Pipeline methods generate system responses by utilizing multiple subcomponents (Zhang et al., 2020; Naseem et al., 2022), whereas end-to-end methods produce system responses directly from dialogue history and the associated knowledge base (Zhou et al., 2021; Zhao et al., 2022). Hybrid methods combine both approaches, using retrieval for efficiency and generative methods for flexibility (Yang et al., 2021; Li et al., 2018). Medical dialogue methods based on LLMs can be divided into two categories: prompting and fine-tuning general LLMs. Prompting methods give instructions to prompt LLMs to perform a task efficiently (Wang et al., 2023d; Gao et al., 2023; Tang et al., 2024; Singhal et al., 2022, 2023). The method of fine-tuning foundation models on medical data could align the LLMs with medical scenarios. (Ye et al., 2024; Toma et al., 2023; Wu et al., 2023; Li et al., 2023b; Han et al., 2023; Huang et al., 2022; Chen et al., 2023b; Liu et al., 2023; Wang et al., 2023b; Xiong et al., 2023; Wang et al., 2023a)

### 2.2 Emotion Language Model

Even though large language models demonstrate remarkable language understanding and generation capabilities, there is a considerable gap between the Emotional Intelligence (EI) capabilities of existing LLMs and humans. (Wang et al., 2023e; Sabour et al., 2024; Paech, 2024) propose comprehensive frameworks for Emotional Intelligence, including assessments of emotional understanding and application. (Li et al., 2023a; Liu et al., 2024; Xu et al., 2024) enhanced the LLMs with prompt or fine-tuning to improve the performance of Emotional Intelligence.

## 3 Methodology

To develop a model to deliver knowledge-rich responses while simultaneously addressing emotional comfort for emotion-sensitive healthcare conversations, we first construct a dataset tailored to this specific scenario. Then, we fine-tuned a base model to the constructed dataset with three renowned fine-tuning methods to enhance its ability. The details of the components are described in the following sections.

### 3.1 Data Modification

We constructed an emotional healthcare dialogue dataset, which consists of Empathetic Response(ER) and Emotional Question(RQ) + Soothing Response(SR). The objective of the Empathetic Response (ER) is to enable the model to generate responses that exhibit empathy, even in the context of standard medical inquiries. On the other hand, the Emotional Question (EQ) + Soothing Response (SR) seeks to equip the model with the ability to handle patient consultations involving negative emotions by delivering informative responses alongside emotional reassurance. Both types of emotional dialogues are structured as single-turn utterances.

We first divided an existing real-world single-turn medical dialogue dataset, which is collected from internet platforms, into two parts. Then, we designed distinct, tailored prompts to utilize a large language model for modifying the doctor's responses in each dialogue of both parts because doctors often respond very briefly through internet platforms, lacking emotional tone. For the Empathetic Response(ER) part, the large language model was prompted to generate responses that exhibit empathy and compassion while retaining medical knowledge based on the given dialogue. For the Emotional Question (EQ) + Soothing Response (SR) part, the large language model was prompted to rewrite the given dialogue into patient queries with negative emotions and responses that are reassuring yet maintain medical knowledge.

Below is the prompt template we used for EQ+SR data.

```
You will be given a dialogue between
a patient and a dotor.
Please rewrite the patient's question
ensuring that it retains the original
information while expressing a sense
of {emotion}. At the same time,
rewrite the doctor's response
to retain the original information
while soothing the patient's
{emotion}.
```

### 3.2 Supervised Fine-Tuning

Supervised Fine-Tuning, which can also be referred to as instruction tuning (Zhang et al., 2024), is a crucial technique to enhance the capabilities and controllability of large language models. It involves further training LLMs using (INSTRUCTION, OUTPUT) pairs, where instructions serve to constrain the model's outputs to align with the desired response characteristics or domain knowledge. We chose the LLaMA3 model (Grattafiori et al., 2024) as the base LLM architecture for further fine-tuning, since it is open source and has excellent language understanding and generation with relatively fewer parameters. We conducted SFT on the base model using the dataset we constructed in Section 3.1 to improve its abilities in emotion comprehension and soothing.

Considering each prompt $X_i = [x_{i,1}, x_{i,2}, \ldots]$ as well as its corresponding response $Y_i = [y_{i,1}, y_{i,2}, \ldots]$ from the healthcare dialogue dataset, the loss function of SFT stage can be defined as follows:

$$L_{SFT}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log\left[P(y_{i,t+1} \mid X_i, y_{i,1\ldots t}, \theta)\right], \tag{1}$$

where $N$ denotes the total number of training instances and $\theta$ denotes model parameters.

### 3.3 Direct Preference Optimization

Based on the previously validated training methods for LLMs (Ouyang et al., 2022), fine-tuning large language models using human preferences significantly improves their behavior on a wide range of tasks and shows promising generalization. One prominent approach is Reinforcement Learning with Human Feedback (RLHF), which employs reward models from response rankings to optimize the training of LLMs. However, RLHF is complex and prone to instability, requiring extensive hyperparameter optimization. To enhance stability, we utilized Direct Preference Optimization (DPO) to align the outputs of the SFT model with human preferences. Compared to RLHF, DPO offers a simpler and more efficient approach, as it eliminates the need for explicit reward modeling or reinforcement learning.

To convert the dataset we constructed in Section 3.1 into the format required for DPO, we treated the modified soothing responses as the preferred responses and the original doctor responses as the rejected responses. Each training sample is a triplet consisting of a prompt, a preferred response, and a rejected response. For the $i$-th prompt $X_i$, our objective was to compute the log probabilities of the preferred response $Y_{i,1}$ and the rejected response $Y_{i,2}$ generated by the current model. Subsequently, we fine-tuned the model parameters to increase the likelihood of the preferred responses $Y_{i,1}$ while reducing the likelihood of the rejected responses $Y_{i,2}$. This optimization process was guided by a loss function below:

$$\begin{aligned} L_{DPO}(\theta) = -\sum_{i} \log\sigma\Big[\beta \log \frac{P(Y_{i,1} \mid X_i, \theta)}{P(Y_{i,1} \mid X_i, \theta^0)} \\ -\beta \log \frac{P(Y_{i,2} \mid X_i, \theta)}{P(Y_{i,2} \mid X_i, \theta^0)}\Big], \end{aligned} \tag{2}$$

where $\sigma$ denotes the sigmoid function, $\theta^0$ means the initial parameters, $\beta$ serves as a hyperparameter that regulates the relative weighting of the two terms.

### 3.4 Kahneman-Tversky Optimization

Another preference optimization called Kahneman-Tversky Optimization (KTO) is a cost-effective method to align large language models with human feedback, enhancing performance without relying on preference pairs. To convert the dataset we constructed in Section 3.1 into the format required for KTO, we treated the modified soothing responses as the preferred responses and the original doctor responses as the rejected responses. In contrast to DPO, KTO does not need training data containing both preferred and rejected responses simultaneously. Each training instance consists of a prompt, a preferred or rejected response, and a binary label indicating whether the response is preferred or rejected. This optimization process was guided by a loss function below:

$$r_\theta(x, y) = \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}. \tag{3}$$

$$z_0 = \mathrm{KL}\big(\pi_\theta(y' \mid x) \,\big\|\, \pi_{\text{ref}}(y' \mid x)\big), \tag{4}$$

$$v(x,y) = \begin{cases} \lambda_D \, \sigma\big(\beta\,(r_\theta(x,y) - z_0)\big), \\ \quad \text{if } \mathrm{Regex}\big(y, y_x^\star\big) = 1 \\ \lambda_U \, \sigma\big(\beta\,(z_0 - r_\theta(x,y))\big), \\ \quad \text{if } \mathrm{Regex}\big(y, y_x^\star\big) = 0 \end{cases} \tag{5}$$

$$L_{\mathrm{KTO}}(\pi_\theta, \pi_{\mathrm{ref}}) = \mathbb{E}_{x,y\sim D}\big[\lambda_y - v(x,y)\big]. \tag{6}$$

## 4 Experiments

To evaluate the effectiveness of our proposed pipeline, we conducted experiments using the dataset introduced in prior work (Li et al., 2023b), which consists of real-world conversations between patients and doctors. This dataset includes a 100k training set sourced from `HealthCareMagic.com` and a 7k testing set from `icliniq.com`. We employed `llama3` models (Grattafiori et al., 2024) with various fine-tuning methods to assess the efficacy of our approach.

### 4.1 Setup

The training set was divided into two subsets, each rewritten with an emotion-specific focus using an LLM:

- **Empathetic Response (ER)**: Approximately 60k entries from the training set were rewritten to transform original doctor responses into empathetic and compassionate replies. This modification was facilitated using the LLaMA3.1 model.
- **Emotional Question (EQ) + Soothing Response (SR)**: The remaining 50k entries were adapted by rephrasing patient questions to convey specific negative emotions. The corresponding doctor responses were rewritten to address the questions while mitigating these emotions. To create realistic scenarios, prompts representing five distinct negative emotions—fear, anxiety, embarrassment, frustration, and distrust—were used to guide the rewrites, leveraging the gpt-4o mini model (OpenAI et al., 2024).

For our experiments, we selected `llama-3.2` as the base model, a multilingual LLM optimized for dialogue in multilingual contexts. Specifically, we used its instruction-tuned generative variant with 1B parameters for fine-tuning. The base models (Zheng et al., 2024) were fine-tuned for one epoch on our emotion-enhanced dataset, with hyperparameters largely aligned with those used for the original `llama-3.2` model. The training input consisted of task instructions and the patient's medical inquiry, with the objective of maximizing the likelihood of generating the correct medical response. This process was carried out on a V100 GPU with 32GB of memory.

To evaluate the fine-tuned models, we measured accuracy on a test set adapted using the same methodology as the **EQ+SR** subset of the training set. This ensured consistency in assessing the model's ability to address queries expressing negative emotions and provide corresponding alleviating responses.

### 4.2 Evaluation

To assess whether the fine-tuned model could balance knowledge delivery and emotional support, we employed task-specific instructions and two large language models as evaluators: `Qwen2.5-7B-instruct` (Yang et al., 2024b), which excels across diverse NLP benchmarks, and `Emollama-chat-7b` (Liu et al., 2024), which specializes in emotion recognition tasks. Additionally, we used ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores to measure the n-gram similarity between generated responses and original doctor responses.

### 4.3 Results

We present the results of our evaluations below. Baseline comparisons included the original `llama-3.2` model and a prompt-based approach for generating emotional responses. The model's performance in mitigating negative emotions and its ability to deliver medical knowledge are discussed in Sections 4.3.1 and 4.3.2, respectively.

#### 4.3.1 Emotion Score

Table 1 presents the results of an evaluation where EmoLLaMA assigned numerical scores to the emotional intensity of responses. Higher values indicate stronger emotional content. Metrics were calculated for three key emotions—empathetic, comforting, and reassuring—as well as their average and maximum values. Our fine-tuned models consistently outperformed the original model and the prompt-based approach across all metrics.

Among the methods tested, fine-tuning with DPO demonstrated the most significant improvements. DPO not only increased the likelihood of generating emotionally rich responses but also minimized the probability of producing emotionally

| Method | Empathetic | Comforting | Reassuring | Mean | Max |
|---|---|---|---|---|---|
| llama3.2-1B | 0.55 | 0.52 | 0.55 | 0.54 | 0.58 |
| + prompt | 0.66 | <u>0.61</u> | 0.65 | 0.64 | 0.67 |
| + ER (sft) | <u>0.68</u> | 0.60 | <u>0.67</u> | <u>0.65</u> | <u>0.69</u> |
| + EQ + SR (sft) | 0.67 | 0.60 | 0.64 | 0.64 | 0.68 |
| + EQ + SR (dpo) | **0.70** | **0.63** | **0.68** | **0.67** | **0.70** |
| + EQ + SR (kto) | 0.67 | 0.59 | 0.64 | 0.63 | 0.68 |
| + ER(sft) + EQ + SR (sft) | 0.67 | 0.60 | 0.64 | 0.63 | 0.68 |
| + ER(sft) + EQ + SR (dpo) | <u>0.68</u> | <u>0.61</u> | 0.66 | <u>0.65</u> | <u>0.69</u> |
| + ER(sft) + EQ + SR (kto) | 0.67 | 0.60 | 0.64 | 0.63 | 0.68 |

Table 1: Emotional intensity on the test set with Emollama as the evaluator. **Bold**: the highest score; <u>underlined</u>: second highest.

| Method | BLEU | BLEU-1 | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|---|---|
| ***Doctor's response** as label* | | | | | |
| llama3.2-1B | 0.91 | 12.8 | 0.17 | **0.02** | 0.16 |
| + prompt | 0.92 | 12.3 | 0.17 | **0.02** | 0.16 |
| + ER (sft) | 1.05 | 13.8 | 0.18 | **0.02** | 0.17 |
| + EQ + SR (sft) | 1.84 | 27.3 | **0.21** | **0.02** | **0.19** |
| + EQ + SR (dpo) | 1.41 | 23.8 | 0.18 | 0.01 | 0.16 |
| + EQ + SR (kto) | 1.89 | **27.5** | **0.21** | **0.02** | **0.19** |
| + ER(sft) + EQ + SR (sft) | 1.86 | 27.3 | **0.21** | **0.02** | **0.19** |
| + ER(sft) + EQ + SR (dpo) | 1.22 | 17.6 | 0.19 | **0.02** | 0.17 |
| + ER(sft) + EQ + SR (kto) | **1.90** | **27.5** | **0.21** | **0.02** | **0.19** |
| ***Modified response** as label* | | | | | |
| llama3.2-1B | 1.86 | 17.6 | 0.23 | 0.04 | 0.21 |
| + prompt | 2.68 | 17.5 | 0.25 | 0.05 | 0.23 |
| + ER (sft) | 3.45 | 19.8 | 0.27 | 0.06 | 0.25 |
| + EQ + SR (sft) | 9.85 | 44.5 | **0.34** | **0.11** | 0.31 |
| + EQ + SR (dpo) | 5.91 | 37.7 | 0.29 | 0.07 | 0.26 |
| + EQ + SR (kto) | 9.79 | 44.5 | **0.34** | **0.11** | **0.32** |
| + ER(sft) + EQ + SR (sft) | **9.93** | **44.6** | **0.34** | **0.11** | **0.32** |
| + ER(sft) + EQ + SR (dpo) | 4.81 | 26.7 | 0.29 | 0.07 | 0.27 |
| + ER(sft) + EQ + SR (kto) | 9.14 | 44.1 | **0.34** | **0.11** | 0.31 |

Table 2: BLEU and Rouge scores on the test set. **Bold**: the highest score.

deficient ones. Direct fine-tuning using the EQ+SR context proved particularly effective, achieving superior results with a smaller dataset. Specifically, fine-tuning with EQ+SR data using DPO improved the average and maximum metrics by 0.03 and 0.13, respectively, compared to the prompt-based approach and the base model. These results confirm that our revised dataset and fine-tuning process significantly enhance the emotional soothing capabilities of the dialogue system.

#### 4.3.2 Knowledge Score

To ensure the model retained essential medical knowledge, we compared its generated responses against the original doctor responses and the emotionally modified responses using ROUGE and BLEU scores (Table 2). The fine-tuned model consistently outperformed both the original base model and the prompt-based approach across all evaluation metrics.

Notably, KTO and SFT approaches achieved better performance than DPO. This may be attributed to the fact that paired responses in DPO's training data already contain substantial knowledge, limit-

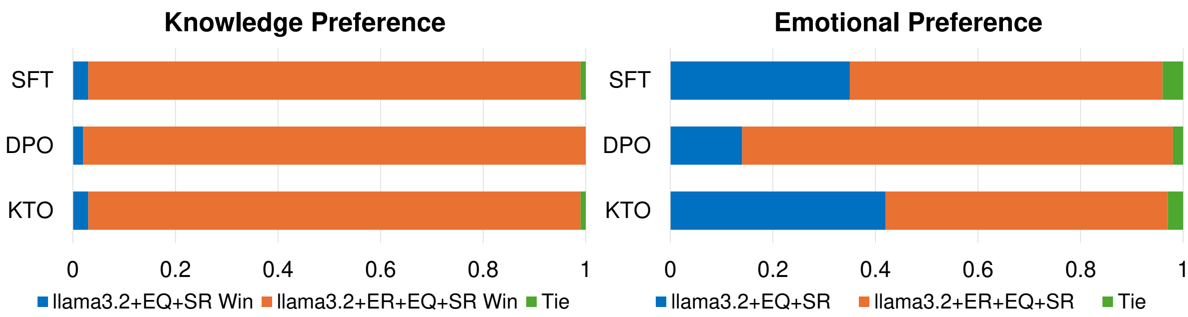


Figure 2: Preference selection based on the knowledgeable and emotional dimensions of Qwen's responses.

ing its ability to enhance further. In contrast, SFT's focus on a single correct response allows it to better capture and internalize the required knowledge. Fine-tuning with ER+EQ+SR data using SFT and KTO yielded a 27-point improvement in BLEU-1 scores compared to both the prompt-based approach and the base model when evaluated against the modified responses. Similar trends were observed for comparisons against original doctor responses, with a 15-point improvement.

These results demonstrate that our approach effectively integrates emotional support with the accurate medical knowledge necessary to address patient inquiries.

### 4.4 Ablation Study

To compare the quality of responses from different methods, we presented the various responses to the Qwen2.5 model simultaneously, allowing it to select the most knowledgeable or empathetic response. In the left part of Figure 2, we plotted the preference selections of the Qwen model across different methods based on the richness of knowledge in the responses. In the right part of Figure 2, we visualized the preference selections of the Qwen model across different methods based on the level of reassurance provided in the responses.

In these two charts, we compared the impact of using two different datasets, specifically examining the effect of incorporating ER data for pre-fine-tuning. Our finding indicates that regardless of the training method used—SFT, DPO, or KTO—models pre-fine-tuned with the ER dataset consistently demonstrated greater preference in both knowledge and emotional selection criteria. This is particularly evident in the emotional selection, as the ER dataset is specifically designed to enable the model to provide empathetic responses even when addressing standard informational content.

### 4.5 Qualitative Analysis

In Table 3, there are some examples of emotional questions and soothing responses generated by our fine-tuned models. Based on the analysis of the models' responses in case (a), it is evident that all three approaches initially focus on alleviating patient anxiety and demonstrating empathy, followed subsequently by the provision of medical knowledge and recommendations. As discussed in the previous section, the DPO approach is particularly effective in fostering the ability to provide emotional reassurance and, therefore, tends to emphasize empathetic expression in its responses. However, this heightened focus on emotional support can occasionally lead to a diminished emphasis on knowledge transmission, as exemplified by the response to case (b). Conversely, the SFT and KTO approaches facilitate more robust knowledge acquisition, resulting in improved informational clarity, while still maintaining an appropriate balance of empathetic language.

## 5 Conclusion

In this paper, we develop a method that enables the model to provide timely emotional comfort in response to the patient's negative emotions during healthcare conversations, while simultaneously offering knowledge-based solutions to address their concerns. We design two kinds of prompts for generating emotionally-aware medical conversations by rewriting existing real-world medical dialogues using a large language model. The first involves modifying doctors' responses into sentences infused with empathy and compassion. The second adds negative emotional tones to patients' statements and generates corresponding comforting responses. Next, we fine-tuned a base model on our

**(a) Emotional Question**

Doctor, *I'm really scared*. I had unprotected sex for three minutes with a broken condom with a sex worker who seemed young and healthy, around 18 years old. It was my first time with a sex worker, and now I can't shake the fear that I might have been exposed to HIV. From the moment the condom broke, I've been filled with dread. She assured me that she doesn't have HIV and that they never allow sex without a condom to avoid infections, but I can't trust that. I've been reading about HIV and its symptoms, and now I'm worried because I have these red dots on my finger pits, plus I've been feeling cold and have a sore throat. What could these red dots be? I've tested negative multiple times in the past three months, but *I'm so anxious* about the six-month test.

**Soothing Response with SFT**:

*I understand that you're feeling anxious, and it's completely normal to have concerns in this situation*. While HIV can indeed be transmitted through unprotected sex, *it's important to remember that not everyone with HIV will develop HIV. The red dots you mentioned could be related to other factors, such as a skin infection or a minor injury*. I recommend that you get tested for HIV to put your mind at ease. In the meantime, try to stay calm and focus on your health. If you have any more questions or need further support, please don't hesitate to reach out.

**Soothing Response with DPO**:

*I understand how unsettling these circumstances can be, and I'm here to help you feel more at ease. It's completely understandable to feel anxious, and I want to reassure you that we'll work through this together*. I recommend taking some time to relax and focus on your health. You've taken the first step by seeking help, which is truly commendable. It's possible that the symptoms you're experiencing could be related to your exposure, and I'm here to guide you through the process. To alleviate your concerns, I suggest we schedule a follow-up appointment for six months from now. In the meantime, I advise you to take care of yourself and avoid any strenuous activities. *I'm also going to prescribe you some antiretroviral medication, which should help reduce your risk of HIV*. Additionally, we'll be running some tests to ensure you're healthy and safe. Please don't worry; we'll get through this together. You're not alone in this, and I'm here to support you every step of the way. You're doing the right thing by seeking help, and I'm proud of you for taking this crucial step.

**Soothing Response with KTO**:

*I understand that you're feeling anxious, and it's completely normal to have concerns after such an experience*. *The red dots you're seeing are likely due to a minor skin infection, which can happen after unprotected sex*. It's important to keep the area clean and avoid any further irritation. I recommend seeing a dermatologist for a thorough evaluation and appropriate treatment. Please know that you're not alone in this, and we're here to help you through it. If you have any more questions or need further reassurance, don't hesitate to reach out.

---

**(b) Emotional Question**

Doctor, *I'm really scared. I've been feeling so low for two years with these digestive problems*, and despite being diagnosed with candida and dysbiosis, my regular doctor seems lost about what's really wrong with me. I've been under a lot of stress and have struggled with anxiety since I was 16. I recently started taking Cymbalta, but after just a few days, I felt this burning in my feet, and I stopped taking it about a month ago. My feet seemed okay for a while, but now, out of nowhere, I have this terrible pain, burning, tickling, and a pins-and-needles sensation. I'm not diabetic or an alcoholic, and I have no idea what's causing this or if it can be fixed. I'm so anxious because I have to be on my feet for my job every morning for four and a half hours. Is there a way to test if my nerve cells are dead or just damaged, and can this ever be reversed? I feel like this is going to ruin my life. Please help me.

**Soothing Response with SFT**:

*I understand that you're feeling overwhelmed and scared right now, and it's completely normal to feel this way given the circumstances*. *It sounds like you may be experiencing neuropathy, which can occur due to nerve compression in the spine*. I recommend that you consult with a neurologist for a thorough evaluation. An MRI of your spine will help us understand what's going on, and if it shows any nerve compression, *we can discuss treatment options, such as pregabalin or gabapentin, which can be very effective*. Please know that you're not alone in this, and I'm here to support you. If you have any more questions or need further assistance, don't hesitate to reach out.

**Soothing Response with DPO**:

*I understand how unsettling this situation is for you, and I'm here to help. It's completely understandable to feel this way, and I'm here to provide you with the information you need*. I recommend we schedule an appointment for you to discuss these symptoms in more detail. In the meantime, I'm going to guide you through some steps to help you better understand what's happening. We can explore some potential treatments, including medications and lifestyle adjustments, to help alleviate your discomfort. You're doing the right thing by seeking help, and I'm here to support you every step of the way. Please, don't worry; we'll work together to find a solution.

**Soothing Response with KTO**:

*I understand that you're feeling overwhelmed, and it's completely normal to be concerned about these symptoms.* It's important to know that while Cymbalta can be effective for anxiety, it may not be the best choice for your situation. *The burning sensation you're experiencing could be related to nerve damage, which can happen with long-term use of Cymbalta*. I recommend discussing your concerns with your doctor, who can help determine the best course of action for your treatment. Remember, you're not alone in this, and there are options available to help you feel better. If you have any more questions or need further assistance, please don't hesitate to reach out.

Table 3: The case study from the testing dataset using LLMs with different fine-tuning methods. Sentences in red are related to emotional comfort, while sentences in blue are related to knowledge transmission.

curated dataset using methods such as SFT, DPO, and KTO. After fine-tuning, we tested the model on real-world emotional conversations to evaluate its performance. Our experimental results show that the fine-tuned model demonstrates significant improvements in both emotional expression and knowledge delivery. Additionally, our work can help medical dialogue systems interact with patients in a more humanized manner, providing not only professional consultation but also emotional comfort to support their well-being.

## Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5, and Google's PaliGemma Academic Program for the GCP Credit Award. We thank the National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

## References

Shang-Chi Tsai Seiya Kawano Angel, Garcia Contreras Koichiro Yoshino, and Yun-Nung Chen. Asmr: Augmenting life scenario using large generative models for robotic action reflection.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023a. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. Preprint, arXiv:2310.15896.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. Preprint, arXiv:2311.16079.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. Preprint, arXiv:2402.01306.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction. Preprint, arXiv:2308.14321.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

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. Preprint, arXiv:2304.08247.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. Large language models understand and can be enhanced by emotional stimuli. Preprint, arXiv:2307.11760.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. Preprint, arXiv:1805.08298.

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

Kangenbei Liao, CHENG ZHONG, Wei Chen, Qianlong Liu, zhongyu wei, Baolin Peng, and Xuanjing Huang. 2022. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning.

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

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. 2023. Radiology-llama2: Best-in-class large language model for radiology. Preprint, arXiv:2309.06419.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. arXiv preprint arXiv:2401.08508.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022. Incorporating medical knowledge to transformer-based language models for medical dialogue generation. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 110–115, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan

Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. Preprint, arXiv:2410.21276.

OpenAI. 2023. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Preprint, arXiv:2203.02155.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models. Preprint, arXiv:2312.06281.

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 on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. Preprint, arXiv:2305.18290.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In Findings of the Association for Computational Linguistics: ACL 2024, pages 2840–2861, Bangkok, Thailand. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. Preprint, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. Preprint, arXiv:2305.09617.

Kai-Fu Tang. 2016. Inquire and diagnose : Neural symptom checking ensemble using deep reinforcement learning.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and

Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. Preprint, arXiv:2311.10537.

Chongyang Tao, Jiazhan Feng, Chang Liu, Juntao Li, Xiubo Geng, and Daxin Jiang. 2021. Building an efficient and effective retrieval-based dialogue system via mutual learning. Preprint, arXiv:2110.00159.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. Preprint, arXiv:2305.12031.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. Preprint, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

Shang-Chi Tsai, Seiya Kawano, Angel Fernando Garcia Contreras, Koichiro Yoshino, and Yun-Nung Chen. Asmr: Augmenting life scenario using large generative models for robotic action reflection.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. Preprint, arXiv:2306.09968.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Huatuo: Tuning llama model with chinese medical knowledge. Preprint, arXiv:2304.06975.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023c. A survey of the evolution of language model-based dialogue systems. Preprint, arXiv:2311.16789.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023d. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. Preprint, arXiv:2302.07257.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023e. Emotional intelligence of large language models. Preprint, arXiv:2307.09042.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. Preprint, arXiv:2109.01652.

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

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. Preprint, arXiv:2304.14454.

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. Preprint, arXiv:2304.01097.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(1):1–32.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. arXiv preprint arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024b. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5000–5009, Online. Association for Computational Linguistics.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, and Zhenhua Guo. 2024. Qilin-med: Multi-stage knowledge injection advanced medical large language model. Preprint, arXiv:2310.09089.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. Preprint, arXiv:2308.10792.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent advances and challenges in task-oriented dialog system. Preprint, arXiv:2003.07490.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022. Medical dialogue response generation with pivotal information recalling. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, page 4763–4771, New York, NY, USA. Association for Computing Machinery.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. Preprint, arXiv:2403.13372.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. Hierarchical reinforcement learning for automatic disease diagnosis. Preprint, arXiv:2004.14254.

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

Ying Zhu, Shi Feng, Daling Wang, Yifei Zhang, and Donghong Han. 2022. Knowledge-enhanced interactive matching network for multi-turn response selection in medical dialogue systems. In Database Systems for Advanced Applications, pages 255–262, Cham. Springer International Publishing.