

Building Emotional Support Chatbots in the Era of LLMs

Zhonghua Zheng¹ Lizi Liao² Yang Deng³ Liqiang Nie¹

¹ Harbin Institute of Technology, Shenzhen ² Singapore Management University

³ National University of Singapore

{polang1999, liaolizi.llz}@gmail.com, ydeng@nus.edu.sg, nieliqiang@gmail.com

Abstract

While emotional support in conversational scenarios offers societal benefits, limited data and non-standardized training impede its application. This work endeavors to navigate these challenges by harnessing the capabilities of Large Language Models (LLMs). We introduce an innovative methodology that synthesizes human insights with the computational prowess of LLMs to curate an extensive emotional support dialogue dataset. Our approach is initiated with a meticulously designed set of dialogues spanning diverse scenarios as generative seeds. By utilizing the in-context learning potential of ChatGPT, we recursively generate an **ExTensible Emotional Support** dialogue dataset, named **ExTES**. Following this, we deploy advanced tuning techniques on the LLaMA model, examining the impact of diverse training strategies, ultimately yielding an LLM meticulously optimized for emotional support interactions. An exhaustive assessment of the resultant model showcases its proficiency in offering emotional support, marking a pivotal step in the realm of emotional support bots and paving the way for subsequent research and implementations. The dataset and codes are available here¹.

1 Introduction

In today’s interconnected world, the integration of emotional support into various conversational contexts holds profound significance (Li et al., 2021). Emotional support conversations play a pivotal role in fostering empathy, understanding, and overall well-being among individuals. These conversations create safe spaces where emotions can be openly expressed and validated, allowing individuals to share their joys, sorrows, fears, and challenges. By forging deep connections and nurturing a sense of community, emotional support bots contribute to a more caring and supportive society.

Despite the undeniable importance of Emotional Support (ES) dialogue systems, the real-world applications are largely hindered by the glaring lack of large-scale well-annotated datasets (Sun et al., 2021). Most of existing studies in emotional support conversations prioritize dataset collection from online sources, such as stress-related Twitter interactions (Medeiros and Bosse, 2018), mental health reddit (Sharma et al., 2020a), and online support groups (Hosseini and Caragea, 2021). However, most of these conversations are asynchronous and limited to single-turn interaction scenarios. Contrarily, Liu et al. (2021) introduced the ESConv dataset via questionnaires, highlighting quality collection and multi-turn conversation. Yet, its constraints stem from its modest size and lack of extensive strategy annotations and scenario variety, likely due to the substantial costs associated with its compilation.

Moreover, this domain lacks standardized training methodologies. Earlier works typically settled for small, manually annotated datasets, which were deeply anchored in traditional approaches like face-to-face therapy (Gibson et al., 2016) and motivational interviewing (Pérez-Rosas et al., 2017). The spotlight then was primarily on empathy detection and discerning predictive rationales (Rashkin et al., 2019a; Sharma et al., 2020c). A shift emerged with innovative prototypes like the Emotional chatting machine (Zhou et al., 2018), focusing on emotion-centric dialogue generation. Simultaneously, there’s a surge in studies aimed at controlled text generation. Hu et al. (2017) presented a sentiment and tense-influenced model, while Ghosh et al. (2017) and Cagan et al. (2017) respectively harnessed affective contexts and grammar-focused techniques. Recent evaluations by Liu et al. (2021) on platforms like BlenderBot and DialoGPT and GPT-driven explorations by Chowdhury et al. (2022) mark the field’s evolution. Nonetheless, ranging from small-scale models such as RNNs to

¹<https://anonymous.4open.science/r/ExtESC-2761/>

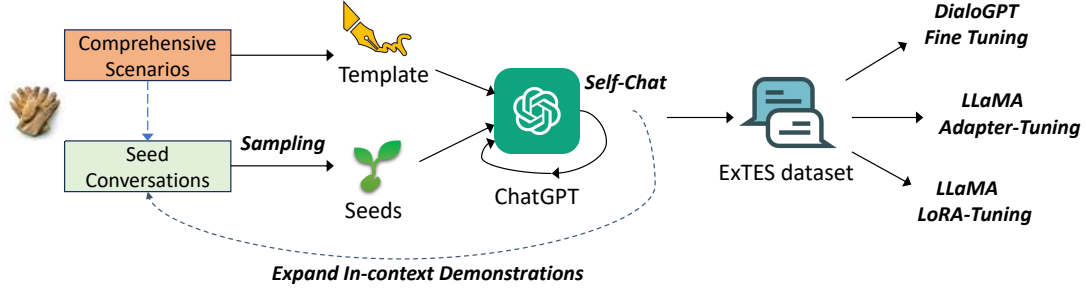


Figure 1: The pipeline for collecting the ExTES conversation dataset via our proposed extendable looping scheme. Based on the dataset, we benchmark and analyze the effect of fine-tuning the chat model with various techniques.

medium-sized ones like GPT-2, a dominant leaning towards sequence-to-sequence models persists.

Recently, Large Language Models (LLMs) have emerged as an epicenter of innovation, exhibiting remarkable generative prowess (Aher et al., 2023). Notably, ChatGPT registers commendable feats across a gamut of NLP challenges and benchmarks (Tan et al., 2023). A salient feature introduced by this model is “In-Context Learning”. This paradigm allows few-shot samples as prompts to facilitate model learning and nuanced generation (Ji et al., 2023). Parallel to this, initiatives like Stanford Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) have sought to emulate ChatGPT’s success, employing data from GPT-3.5 and subsequently refining LLaMA (Touvron et al., 2023). Collectively, these advancements illuminate the tantalizing potential: deploying LLMs for crafting emotional support dialogues and optimizing the LLaMA architecture could conceivably address the twin challenges of data scarcity and training methodology refinement.

Correspondingly, we present a novel study that focuses on developing emotional support dialogue systems with the help of LLMs. As illustrated in Figure 1, the proposed architecture address the pressing concerns of data scarcity by capitalizing on LLMs while sculpting a nuanced emotional support-centric LLM. Initially, we synergize human expertise with LLMs’ capabilities, forging the creation of an **ExTensible Emotional Support** dialogue dataset, termed as **ExTES**. Our approach encompasses the craft of a meticulous and encompassing set of emotion support dialogues, spanning various scenarios and enriched with strategic annotations. These dialogues serve as primordial seeds, prompting ChatGPT to recursively generate expansive dialogues, tapping into its adept in-context learning capabilities. The subsequent phases delve

into the exploration of diverse fine-tuning methodologies aimed at refining our emotional support chatbot model. We undertake exhaustive evaluations via automatic and human evaluations, assessing its proficiency in rendering emotional support. The analyses affirm that (1) The ExTES, when sculpted in concert with LLMs, notably transcends ESConv’s quality across multiple dimensions, laying a fertile groundwork for subsequent inquiries and deployments. (2) The judicious application of parameter-efficient tuning techniques on LLaMA (Touvron et al., 2023) emerges as an optimal blueprint for the evolution of proficient emotional support chatbots.

To summarize, our main contributions in this paper are as follows:

- We innovatively leverage the generative capabilities of LLMs to generate an extensible emotional support dialogues dataset, ExTES, with comprehensive scenarios and strategies, which is released for building robust and generalizable emotional support systems.
- We investigate different fine-tuning strategies to endow LLaMA with effective and flexible emotional support capabilities.
- Extensive evaluations validate the effectiveness and significant values of the ExTES dataset. The successful integration of emotional support dialogue with LLMs can positively impact mental health counseling, social interactions, customer service, and various other domains, contributing to a more compassionate and supportive society.

2 Related Work

Emotional & Empathetic Conversation. Emotion’s role in building engaging dialogue systems

has been extensively studied (Zhou et al., 2018; Li et al., 2017; Zhou and Wang, 2018; Huber et al., 2018; Huang et al., 2020). (Zhou et al., 2018) proposed the Emotional Chatting Machine (ECM) to generate emotional responses based on pre-specified emotions, accurately expressing designated emotions. However, it’s essential to differentiate emotional chatting from emotional support (ES). While emotional chatting expresses emotions like happiness or sadness, ES aims to reduce user emotional distress through proactively guiding the conversation (Liao et al., 2023; Deng et al., 2023a) and leveraging proper support skills. Empathetic responding is another related task in emotional conversation research (Rashkin et al., 2019b; Lin et al., 2019; Majumder et al., 2020; Zandie and Mahoor, 2020; Sharma et al., 2020c; Zhong et al., 2020; Sun et al., 2021). It focuses on understanding users’ feelings (Liao et al., 2021a) and generating appropriate replies. Rashkin et al. (2019b) argued that recognizing interlocutors’ emotions enables more empathetic responses. Empathetic responding is a critical aspect of effective ES but is just one component within the broader framework. Besides empathetic responding, emotional support conversations further require the capability of dialogue strategy learning (Cheng et al., 2022; Tu et al., 2022; Peng et al., 2022; Deng et al., 2023b) to address users’ problems and provide assistance in coping with difficulty. In other words, the key challenge in emotional support conversations is to make strategic decision for handling various emotional issues.

Language Models for Conversations. The field of language models for conversations has seen remarkable progress recently (Liao et al., 2023; Ye et al., 2022a), with various models demonstrating impressive capabilities in open-domain dialogue, such as DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020), LaMDA (Thoppilan et al., 2022), etc. ChatGPT (Ouyang et al., 2022) optimized language models for chat using Reinforcement Learning with Human Feedback, resulting in human-like chat abilities, which was further enhanced into GPT4 with advanced reasoning and multi-modal capabilities in conversations (Liao et al., 2021b; Ye et al., 2022b). Meanwhile, researchers aimed to replicate ChatGPT’s success with open-source foundation models. For example, Stanford Alpaca (Taori et al., 2023) used Self-Instruct (Wang et al., 2023) to collect GPT-3.5 data and fine-tuned the LLaMA model (Touvron et al.,

2023). Vicuna (Chiang et al., 2023) fine-tuned LLaMA on a dialogue corpus from sharegpt.com, enhancing its conversational capabilities. Amidst this evolving landscape, our work contributes by exploring techniques to improve the quality, contextuality, and empathy of emotional support conversations.

Related Datasets for Emotional Support. Early studies on emotional support conversations typically focused on collecting and annotating datasets derived from social media (Medeiros and Bosse, 2018), online forums (Sharma et al., 2020b; Hosseini and Caragea, 2021), or psychotherapy video transcripts (Shen et al., 2020). However, the emotional support strategy is overlooked and the dialogue quality is unsatisfactory in these datasets. To this end, Liu et al. (2021) constructed an Emotional Support Conversation dataset, ESConv, which carefully designs the process of data collection and devises multiple mechanisms to ensure the effectiveness of emotional support strategies in conversations. However, the types of emotional support strategies and the categories of dialogue backgrounds are limited, which is difficult to be generalized to complex situations in reality.

3 Dataset Collection

In this section, we elucidate our approach to constructing a sophisticated multi-turn emotional support chat corpus harnessing ChatGPT (gpt-3.5-turbo). The process bifurcates into two stages. Initially, we delineate comprehensive emotional support scenarios, encompassing response strategies, and meticulously curate exemplar dialogues from extant datasets and online platforms, emphasizing richness and relevance. Subsequently, we leverage ChatGPT to generate an extended set of dialogues using these seed exemplars, followed by a manual correction process. The refined dialogues will then be reintroduced into the model as supplementary exemplars to further enrich the dataset in an iterative fashion. Notably, our strategy offers a marked reduction in human labor, harmonizing efficiency with dataset integrity.

3.1 Comprehensive Scope and Strategies

To ensure the diversity and broad coverage of emotional support conversations, it is important to include comprehensive emotional support conversation scenarios and response strategies. Drawing

Category	Dialogues	Proportion	Category	Dialogues	Proportion
Breakups or Divorce	710	6.3%	Navigating Gender Identity and Transitioning	202	1.8%
Conflicts or Communication Problems	1,109	9.9%	Moving to a New City or Country	202	1.8%
Communication Challenges	1,008	9.0%	Career Transitions	202	1.8%
Coping with the Death of a Loved One	593	5.3%	Parenthood and Parenting Challenges	202	1.8%
Dealing with the Loss of a Pet	601	5.4%	Low Self-Esteem or Lack of Confidence	302	2.7%
Work-related Stress and Burnout	403	3.6%	Body Image Concerns and Eating Disorders	101	0.9%
Financial Worries and Uncertainty	403	3.6%	LGBTQ+ Identity	101	0.9%
Unemployment-related Stress	403	3.6%	Cultural Identity and Belonging	101	0.9%
Academic Stress	403	3.6%	Academic Stress or Pressure	202	1.8%
Spirituality and Faith	202	1.8%	Job Loss or Career Setbacks	202	1.8%
Managing Bipolar Disorder	202	1.8%	Parenting Challenges and Parental Guilt	202	1.8%
Anxiety and Panic	202	1.8%	Sibling Rivalry or Family Conflict	403	3.6%
Depression and Low Mood	403	3.6%	Surviving and Recovering from Physical or Emotional Abuse	101	0.9%
Adjusting to a New Job or Role	302	2.7%	Healing from Sexual Assault or Domestic Violence	101	0.9%
Chronic Illness or Pain Management	302	2.7%	Post-Traumatic Stress Disorder (PTSD)	101	0.9%
Coping with a Diagnosis or Medical Treatment	202	1.8%	Healing from Abuse	202	1.8%
Caregiver Support	202	1.8%	Addiction and Recovery	202	1.8%
Finding Meaning and Purpose in Life	202	1.8%	Support for Loved Ones or Friends	202	1.8%

Table 1: Statistics of all 36 emotional support scenarios covered in our ExTES dataset.

upon a wealth of literature on psychological counseling (Burleson, 2003) and insights from previous emotional support research (Reblin and Uchino, 2008; Meng and Dai, 2021; Shensa et al., 2020; Graham et al., 2019), we create a comprehensive set of emotionally diverse scenarios. Additionally, we referred to emotional support scenes from related works, expanding and refining our collection. In the end, we obtain 36 emotionally impactful scenarios (listed in Table 1), detailed in Appendix D. These scenarios encompass a wide array of everyday life situations, catering to the varied emotional needs of users. Unlike the limited five scenarios in ESConv (Liu et al., 2021), our expanded range now provides a richer array of contexts for emotional support interactions. This expansion has been instrumental in refining the content and scope for providing emotional support.

Similarly, inspired by (Hill, 1999; Organization et al., 2020), we compile 16 emotional support strategies (Table 2). Compared with the eight strategies in ESConv, our emotional support strategies are richer and more conducive to providing users with targeted suggestions.

3.2 In-context Examples and Collection

We initiated our data collection process by manually constructing 87 seed dialogues. These seed dialogues were sourced from real emotion support datasets, including ESConv (Liu et al., 2021), ETMHS (Sharma et al., 2020d), and Reddit (Yeh et al., 2015) datasets. Additionally, we used web crawling to supplement our data collection efforts for certain missing scenarios. By incorporating real

cases from these datasets and web-crawled materials, we ensured a diverse and representative set of emotional support dialogues. Each of the 36 scenarios we identified has at least two seed dialogues, ensuring sufficient coverage. The quality of dialogues can be guaranteed after careful manual correction on the dialogues and response strategy labeling.

3.3 Human Review and Iteration

With meticulously curated exemplar dialogues, we prompt ChatGPT to generate expansive dialogues via self-chat, which has been shown effective in various dataset solicitation works (Xu et al., 2023). It involves utilizing ChatGPT to generate answers for the user and the emotional counseling assistant in the dialog format we set. Specifically, we design a template (shown in Appendix C) to define the format and requirements, allowing the API to continuously generate transcripts of both sides of the conversation, and annotate appropriate emotional support strategies. The dialogue takes the seed dialogue as an example and outputs complete dialogues according to given scenarios. Specifically, We first filled the 87 seed dialogues into the template, collected a total of 1k dialogues with different dialogue scenarios through self-chat, and then continued to collect data by replacing the seed dialogues with these 1k dialogues. The whole process is iterative and expansive, *i.e.* can be expanded easily with new seeds and new scenarios.

Note that the whole collection process heavily rely on ChatGPT’s self-chat, hence the quality of generated dialogues is our major concern.

Although the format and requirements of the output dialogues are clearly specified in the template we defined, we still observe some data format errors, duplication, and failures to meet other requirements. To ensure the quality, we conduct human review on the collected dialogues in every iteration and apply manual correction process. However, we notice that the human intervention requirement is rather limited as compared to existing conversation dataset collection methods such as questionnaire (Liu et al., 2021) or crowd-sourcing (Budzianowski et al., 2018). For instance, we note that fewer than 10% of the generated conversations require human adjustments. Typical discrepancies include formatting issues, absent response strategies, strategy rectifications, and the inclusion of specific details. Conversations needing extensive revisions are directly discarded. After the screening and modification process, we finally obtain around 11k conversations that meet the requirements to form ExTES. This process costs about \$210 to call OpenAI’s API².

Category	Dialogues	Proportion
Reflective Statements (RS)	14,560	14.8%
Clarification (Cla)	2,898	2.9%
Emotional Validation (EV)	19,367	19.8%
Empathetic Statements (ES)	8,482	8.7%
Affirmation (Aff)	16,539	16.9%
Offer Hope (OH)	4,665	4.8%
Avoid Judgment And Criticism (AJC)	1,767	1.8%
Suggest Options (SO)	6,079	6.2%
Collaborative Planning (CP)	3,534	3.6%
Provide Different Perspectives (PDP)	3,322	3.4%
Reframe Negative Thoughts (RNT)	2,050	2.1%
Share Information (SI)	3,181	3.3%
Normalize Experiences (NE)	2,403	2.6%
Promote Self-Care Practices (PSP)	2,686	2.7%
Stress Management (SM)	2,474	2.5%
Others (Oth)	3887	3.9%
Overall	97,893	100%

Table 2: Statistics of response strategies used in ExTES.

4 Dataset Characteristics and Quality

General Statistics Our resultant dataset ExTES comprises 11,177 dialogues in total, with specifics provided in Table 3. On average, a dialogue contains 18.2 utterances. While users often display negative sentiments, the assistants lean towards

²<https://platform.openai.com/docs/api-reference>

Category	ExTES	ESConv
Dialogues	11,177	1,053
Utterances	200,393	31,410
Avg. length of dialogues	18.2	29.8
Avg. length of utterances	26.0	17.8
Num. of support strategise	16	8
Num. of scenarios	36	5

Table 3: The statistics of our ExTES vs. ESConv.

positive expressions, offering emotional support. An illustrative dialogue is displayed in Figure 4 in Appendix A.

The length of these dialogues, averaging 18.2 utterances, suggests that delivering effective emotional support frequently necessitates multiple conversational turns. This is considerably more than what’s observed in prior datasets on emotional chatting (Zhou and Wang, 2018) and empathetic dialogue (Rashkin et al., 2019a). While our dialogues are shorter on average than those in ESConv (29.8 utterances on average), they have a more substantial average utterance length (26.0 versus ESConv’s 17.8). This suggests our dialogues are content-rich. Detailed statistics for other annotations can be found in Table 1 (highlighting ES scenarios) and Table 2 (detailing ES strategies). Notably, the most prevalent emotional challenges arise from communication issues, a trend we identified early in our seed data collection. Following closely are problems stemming from work-related pressures or unemployment, possibly exacerbated by global economic downturns.

	Crowd-sourced	ExTES	κ
Informativeness	2.39	2.53	0.51
Understanding	2.64	2.52	0.46
Helpfulness	2.48	2.61	0.44
Consistency	2.75	2.67	0.39
Coherence	2.38	2.45	0.52

Table 4: Human evaluation of ExTES quality. The scores (from 0 to 3) are averaged over all the samples rated by three annotators. κ denotes Fleiss’ Kappa (Fleiss, 1971), indicating fair to moderate inter-annotator agreement ($0.2 < \kappa < 0.6$).

Dialogue Quality Evaluation We further validate the quality of ExTES via conducting comprehensive human evaluation and comparing the collected ExTES part with our seed dialogues (Crowd-sourced). Referring to (Liu et al., 2021), we adopted the following metrics to evaluate the

quality of our augmented dialogues. **Informative-ness** evaluates the extent to which the individual seeking assistance elaborates on their emotional distress. **Understanding** assesses the level of comprehension the supporter has of the individual’s experiences and feelings. **Helpfulness** determines the degree to which the supporter is successful in alleviating the emotional discomfort of the individual and improving their mood. Moreover, we also evaluate the general quality of the dialogue. **Consistency** checks if the participants’ behavior is in line with their roles, and if there is no contradiction in a participant’s behavior. **Coherence** assesses if the conversation stays focused and in-depth, and if transitions between topics are smooth. All the metrics are rated with the four-level Likert scale (Allen and Seaman, 2007) ranging from 0 to 3 (higher is better). We recruited 50 college students from different majors to serve as annotators for this project. We randomly selected 100 dialogue examples from ExTES and 50 examples from our seed dialogues. At least two different annotators rated each dialogue example.

As shown in Table 4, it demonstrates that our method can generate high-quality emotional support dialogues as in ExTES. Dialogues collected by our method show similar evaluation scores compared to crowdsourced seed dialogues. It is even better than seed dialogues in terms of Informativeness and Helpfulness. According to our observation, this might be because the answers generated by ChatGPT tend to have more substantial and complete content.

Strategy Distribution In our study, we also hope to comprehend the distribution of strategies adopted by ChatGPT in generating responses at various stages of the dialogue. For the purpose of this analysis, we considered a conversation with N utterances in total, where the k -th utterance is from the assistant and adopts the strategy S . The position of this utterance in the conversation is referred to as the conversation phases and is represented as k/N . We evenly divide the conversation progress into four phases.

We examined all the dialogues in our dataset and tallied the proportions of different strategies adopted within these four intervals. The resultant data points provided a snapshot of the strategy distribution at four phases along the conversation progress. As shown in Figure 2, we can observe different trend in the distribution of ES strategies

	Strategy Transition	Proportion
3-Hop	EV → RS → EV	17.19 %
	EV → RS → SO	16.23 %
	EV → RS → ES	14.49 %
	RS → EV → SO	11.03 %
	EV → ES → RS	9.75 %
4-Hop	EV → RS → ES → SO	7.08 %
	EV → RS → SO → Aff	6.61 %
	EV → ES → RS → NE	6.04 %
	RS → Aff → ES → RS	5.27 %
	EV → RS → SO → Cla	4.36 %
5-Hop	EV → RS → EV → Aff → SO	1.97 %
	EV → RS → SO → Aff → RS	1.34 %
	RS → EV → SO → OH → SO	0.89 %
	EV → RS → ES → SO → Aff	0.45 %
	EV → ES → RS → NE → Cla	0.27 %

Table 5: Proportions of top-5 strategy transitions in responses. The adjacent same strategies are merged. Abbreviations are consistent with the Appendix B.

across the different phases of the conversation.

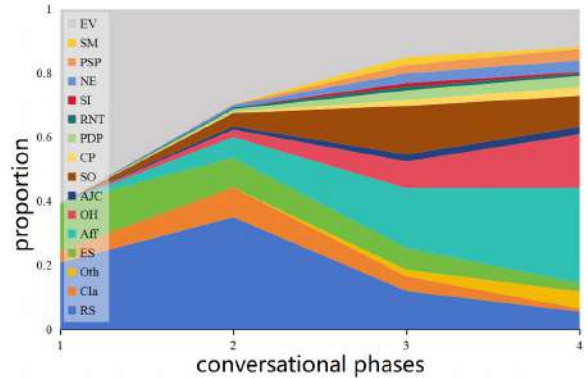


Figure 2: The distribution of strategies at different phases.

Strategy Transition We present the top-5 most frequent strategy transitions with 3-5 hops in Table 5. These transitions indicate that assistants usually ask questions and explore the help-seekers situations before comforting the help-seek users. Emotional support assistants usually first understand the cause of the user’s distress and then say some words of comfort or express sympathy for the user’s experience. This is generally as expected. It also might not be wise enough to make actionable suggestions at the beginning of the whole dialogue.

5 Training Techniques Exploration

In the rapidly evolving field of artificial intelligence, utilizing large language models in an effi-

cient and effective manner has become increasingly important. Parameter-efficient finetuning stands at the forefront of this pursuit, allowing researchers and practitioners to reuse pre-trained models while minimizing their computational and resource footprints. We explore the following methods to adapt large pre-trained language models on our dataset. We used DialoGPT and LLaMa-7B as the backbones of the compared models:

5.1 DialoGPT Fine Tuning

DialoGPT (Zhang et al., 2020) is a medium-sized GPT2 Model trained on 147M conversation-like exchanges extracted from Reddit. It was trained with a causal language modeling (CLM) objective on conversational data and is therefore powerful at response generation in open-domain dialogue systems. In order to fine-tune DialoGPT, we use CLM training. We follow the OpenAI GPT-2³ to model a multiturn dialogue session as a long text and frame the generation task as language modeling.

5.2 LLaMA Adapter-Tuning

LLaMA-Adapter (Zhang et al., 2023) is a form of prefix-tuning that prepends a learnable adaption-prompt to the inputs of the attention blocks in LLaMA. In total, there are only 1.2M parameters to update during finetuning, which significantly reduces the memory footprint and speeds up training. Recently, LLaMA-Adapter v2 (Gao et al., 2023) is developed to further include more trainable parameters. We use LLaMA-Adapter v2 to demonstrate instruction-tuning LLaMA 7B on our dataset. Inspired from prefix tuning (Li and Liang, 2021) and the original adapter method (Houlsby et al., 2019), Adapter-Tuning introduces some new sub-layers (i.e., adapter layers) acting as low-rank bottlenecks within each Transformer layer. Generally, instead of tuning all parameters, Adapter-Tuning focuses on tuning mainly the adapter layers.

5.3 LLaMA LoRA-Tuning

Low-rank adaption (LoRA) (Hu et al., 2021) is a technique to approximate the update to the linear layers in a LLM with a low-rank matrix factorization. This significantly reduces the number of trainable parameters and speeds up training with little impact on the final performance of the model. We demonstrate this method by instruction-tuning LLaMA 7B on our dataset. The authors

take inspiration from (Li et al., 2018; Aghajanyan et al., 2021) which show that the learned over-parametrized models in fact reside on a low intrinsic dimension. Based on the inherent low-rank characteristics of the large model, the bypass matrix is added to simulate the fine-tuning of the full model parameters. LoRA achieves the purpose of lightweight fine-tuning through a simple and effective solution. It turns various large models into professional models in different fields through light fine-tuning.

6 Experiments

In this section, building upon the validation of our ExTES dataset’s quality from prior sections, our experiments concentrate on three critical facets: (Q1) Which is the most effective fine-tuning technique for constructing an emotional support chatbot? (Q2) Whether the new dataset adheres to specific criteria, such as toxicity levels, and if it offers broad applicability? (Q3) How does the insights drawn from human assessments complement with automatic evaluations?

6.1 Automatic Evaluation

6.1.1 Experiments on ExTES (Q1)

To assess the implications of employing support strategies on model performance when using either DialoGPT or LLaMA as the backbone framework, we evaluated the performance of the no-strategies and strategies versions. The automated evaluation criteria we used comprised of METEOR (Banerjee and Lavie, 2005), BLEU-2/4 (B-2/4), ROUGE-L (R-L) (Lin, 2004), Vector Extrema (Forgues et al., 2014) and the Distinct-1/2/3 (Li et al., 2016). The responses were tokenized using the NLTK (Loper and Bird, 2002) for this purpose.

To demonstrate the effectiveness of the model, we compare it with no-strategies and strategies variants. Results for nine metrics are reported in Table 6. Overall, we have the following observations and discussions: (1) The models based on LoRA-Tuning and Adapter-Tuning consistently outperform the DialoGPT variant on all metrics, suggesting that LLaMA, with larger parameter size, is more suitable for emotional support chatbot. (2) The LoRA-Tuning model is slightly better than the Adapter-Tuning model on most metrics (except B-2), indicating that our dataset can play the best effect on LLaMA-LoRA. Therefore, in subsequent experiments, we will use LLaMA-LoRA

³https://huggingface.co/docs/transformers/model_doc/gpt2

Backbones	Variants	PPL	METEOR	B-2	B-4	R-L	Extrema	D-1	D-2	D-3
DialoGPT	no-strategies	13.11	26.03	4.438	1.721	13.37	53.27	19.13	49.29	62.92
	strategies	13.71	26.82	4.773	1.966	13.23	55.71	16.70	53.11	77.47
LLaMA-Adapter	no-strategies	15.25	28.48	6.751	1.944	16.95	64.47	23.23	60.43	82.62
	strategies	15.82	29.71	6.317	1.987	16.39	62.73	22.90	60.83	82.24
LLaMA-LoRA	no-strategies	15.67	30.31	6.105	2.333	21.60	65.06	21.73	63.64	84.90
	strategies	16.02	30.67	6.416	2.491	20.85	65.44	21.81	61.94	82.80

Table 6: Results of automatic evaluation. The results in bold indicate significant superiority over all the competitors.

Attributes	ExTES	ESConv	LoRA-Responses
Toxicity	0.0501	0.0760	0.0358
Severe Toxicity	0.0016	0.0036	0.0016
Identify Attack	0.0047	0.0095	0.0048
Insult	0.0219	0.0183	0.0137
Profanity	0.0251	0.0401	0.0222
Threat	0.0073	0.0098	0.0078

Table 7: Results of our toxicity assessment using Perspective API. Lower scores are better. The LoRA-Responses are generated by LLaMA-LoRA finetuned on our ExTES dataset.

for evaluation. (3) The variant with strategies is generally better than the variant without strategies, but slightly lower than the variant without strategies in R-L and D-1/2/3. This is because, under the guidance of strategies, the generation space of replies will be smaller, which reduces the diversity of prediction and generates more targeted replies.

6.1.2 Toxicity Assessment (Q2)

It has always been an important yet challenging problem to control language models to avoid generating texts with undesirable attributes, such as toxic language and unnatural repetition (Zheng et al., 2023). We hence use the Perspective API⁴ (a toxicity detection API widely used in online discussions) to evaluate the potential toxicity in our dataset ExTES. All utterances are scored for toxicity across six production attributes via the Perspective API. For each attribute, we report the average score (from 0 to 1, with lower scores being safer) across all utterances.

As shown in Table 7, our dataset scored very low and exhibited little toxicity. It is generally lower than that of ESConv, the manually curated dataset. We consider the level of toxicity to be normal. Ac-

tually, further reductions in toxicity scores may affect the quality of emotional support conversations. Because users seeking emotional support might express some hateful or aggressive contents, which will increase toxicity level in data. It’s noteworthy that the score for Severe Toxicity—referring to extremely hateful, aggressive, or disrespectful comments—is as low as 0.0016. This low score can likely be credited to the inherent safety mechanisms within the ChatGPT model. Furthermore, when examining the LLaMA-LoRA model finetuned on our ExTES dataset, we observe even reduced levels of toxicity, particularly in categories such as Toxicity, Severe Toxicity, Insult, and Profanity. Such a reduction is advantageous, as our aim is for the emotional support bot to engage users with kindness and courtesy.

6.1.3 Cross-dataset Experiments (Q2)

To validate the generality of the collected ExTES dataset, we apply LoRA-Tuning on LLaMA separately using the ESConv and ExTES training set. The resulting models are then tested on the test set of both datasets for comparison. As shown in Table 8, the performance of the model trained on ESConv is generally worse than the one trained on ExTES across various indicators. Firstly, the model trained on ExTES showcases remarkable performance on the ESConv test set, which demonstrates that ExTES possesses remarkable generality to be adapted into various emotional support applications. Secondly, the performance gap between the model trained on ExTES and ESConv on the ExTES test set is more substantial than that on the ESConv test set. This is mainly because the total amount of ESConv data is small, and there are many unseen scenarios that ESConv does not cover but appears in ExTES test set. Naturally, it becomes more difficult to provide an appropriate

⁴<https://perspectiveapi.com/>

Test Set	Train Set	PPL	METEOR	B-2	B-4	R-L	Extrema	D-1	D-2	D-3
ESConv	ESConv	15.36	24.23	6.310	1.670	17.19	58.57	25.72	44.09	60.78
	ExTES	15.64	27.07	6.325	2.312	20.57	55.56	24.88	63.83	83.93
ExTES	ESConv	15.23	24.08	6.333	1.687	16.70	53.41	24.79	46.83	65.94
	ExTES	16.02	30.67	6.410	2.491	20.85	65.44	21.81	61.94	82.80

Table 8: Experiments across datasets. We apply LoRA-Tuning on LLaMA separately using ESConv and ExTES. The resulting models are then tested on the test set of both datasets for automatic evaluation.

response for the unseen scenarios. Overall, the generality of ExTES can contribute to a more robust and generalizable ES conversation system.

6.2 Human Evaluation (Q3)

We performed human interaction evaluation. We recruited 50 students from different majors and collected a total of 100 interactive conversations and ratings (each participant contributed two). Each participant was asked to talk about the same emotional issue with three models. Each conversation lasts at least 10 turns, after which participants can continue or end the conversation. Participants are only allowed to talk about emotional issues, but the conversation is open as long as it stays on topic. After the dialogue, participants were asked to rate the performance of the three models according to the following aspects, which followed the evaluation protocol of (Liu et al., 2021). To evaluate the models, participants were prompted with the following questions: (1) **Fluency**: which model’s responses were more coherent and easily comprehensible? (2) **Identification**: which model delved deeper into your situation, effectively identifying your problems? (3) **Comforting**: which model displayed more adeptness in providing comfort and support? (4) **Suggestion**: which model offered more helpful suggestions for addressing your issues? (5) **Overall**: which model’s emotional support did you prefer in general? The study encompassed three models: (1) LLaMA-ESConv, (2) LLaMA-ExTES-no strategies and (3) LLaMA-ExTES-strategies.

The outcomes of comparison demonstrate the following findings: (a) It reveals that LoRA-Tuning LLaMa on our ExTES significantly enhanced its capability to provide emotional support across all metrics. (b) The strategies version slightly outperforms the non-strategies version overall, especially on Suggestion. This indicates that the strategic

ExTES vs.	ESConv			ExTES (no-strategies)		
	Win	Tie	Loss	Win	Tie	Loss
Flu.	28%	50%	22%	43%	48%	9%
Ide.	27%	53%	20%	28%	58%	14%
Com.	49%	37%	14%	25%	28%	47%
Sug.	57%	22%	21%	42%	36%	22%
Ove.	51%	28%	21%	42%	44%	14%

Table 9: Human evaluation results. The results revealed that LoRA-Tuning LLaMa on our ExTES significantly enhanced its capability to provide emotional support across all metrics.

guidance is effective and worth further exploration. These findings also underscore the considerable impact of dataset quantity on model performance. In general, the fine-tuning of pre-trained models on our dataset rendered them more preferable to users, affirming ExTES’s high quality and utility in bolstering emotional support capabilities.

7 Conclusion and Future Work

In this paper, we address the under-researched integration of emotional support conversation bots with Large Language Models. By overcoming data scarcity and training challenges for building emotion support conversation models, we leveraged human expertise and Large Language Models’ computational strength to establish a comprehensive and expansive emotional support dialogue dataset. Then we fine-tuned LLaMA to a specialized emotional support chat model. Our method based on extensive emotional support conversation datasets and parameter-efficient tuning, showed promising results in providing emotional support across various scenarios. This study represents a significant advancement in emotional support dialogue systems, laying a strong foundation for future applications and exploration.

References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.
- Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. [Intrinsic dimensionality explains the effectiveness of language model fine-tuning](#). 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 7319–7328, Online. Association for Computational Linguistics.
- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 337–371. PMLR.
- I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. *Quality progress*, 40(7):64–65.
- 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.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.
- Brant Burleson. 2003. Emotional support skills. *Handbook of communication and social interaction skills*, pages 180–399.
- Tomer Cagan, Stefan L Frank, and Reut Tsarfaty. 2017. Data-driven broad-coverage grammars for opinionated natural language generation (onlg). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1331–1341.
- Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Jishnu Chowdhury, Yong Zhuang, and Shuyi Wang. 2022. Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning.
- Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.
- Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023b. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4079–4095.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model.
- Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-Im: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642.
- James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Interspeech 2016*, pages 1447–1451.
- Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21:1–18.
- Clara E. Hill. 1999. Helping skills: Facilitating exploration, insight, and action. In *American Psychological Association*.
- Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *AAAI Conference on Artificial Intelligence*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems.
- Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

- Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.
- Chunyuan Li, Heerad Farkhor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.
- Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021a. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.
- Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021b. Mmconv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 675–684.
- Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents in the post-chatgpt world. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3452–3455.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.
- Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In *Practical Applications of Agents and Multi-Agent Systems*.
- Jingbo Meng and Yue (Nancy) Dai. 2021. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication*, 26(4):207–222.
- World Health Organization et al. 2020. Mental health and psychosocial considerations during the covid-19 outbreak, 18 march 2020. Technical report, World Health Organization.
- Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019a. [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.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019b. 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.
- Maija Reblin and Bert N Uchino. 2008. Social and emotional support and its implication for health. *Current opinion in psychiatry*, 21(2):201.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020a. [A computational approach to understanding empathy expressed in text-based mental health support](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

- Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020c. A computational approach to understanding empathy expressed in text-based mental health support.
- Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020d. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*.
- Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *SIGdial 2020*, pages 10–20.
- Ariel Shensa, Jaime E. Sidani, César G. Escobar-Viera, Galen E. Switzer, Brian A. Primack, and Sophia Choukas-Bradley. 2020. Emotional support from social media and face-to-face relationships: Associations with depression risk among young adults. *Journal of Affective Disorders*, 260:38–44.
- 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.
- Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of gpt family llms’ question answering performance. *arXiv preprint arXiv:2303.07992*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint arXiv:2201.08239*.
- 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. 2023. Llama: Open and efficient foundation language models.
- 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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 308–319.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.
- 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*.
- Chenchen Ye, Lizi Liao, Fuli Feng, Wei Ji, and Tat-Seng Chua. 2022a. Structured and natural responses co-generation for conversational search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 155–164.
- Chenchen Ye, Lizi Liao, Suyu Liu, and Tat-Seng Chua. 2022b. Reflecting on experiences for response generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5265–5273.
- Chun-Hung Yeh, Anuradha Welivita, and Pearl Pu Faltings. 2015. A dialogue dataset containing emotional support for people in distress. *arXiv preprint arXiv:1503.08895*.
- Rohola Zandie and Mohammad H. Mahoor. 2020. Emptansfo: A multi-head transformer architecture for creating empathetic dialog systems.
- Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, AoJun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation.
- Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning.
- Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Xianda Zhou and William Yang Wang. 2018. *MojiTalk: Generating emotional responses at scale*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Data Example from our dataset

Here we detail the conversation that Figure 3 demonstrates to show details that our dataset contains. The detailed example can be seen in Figure 4. Each conversation is labeled its scene category and a brief of description of the user. In the context of each conversation, the strategies used by the assistant are labeled red.

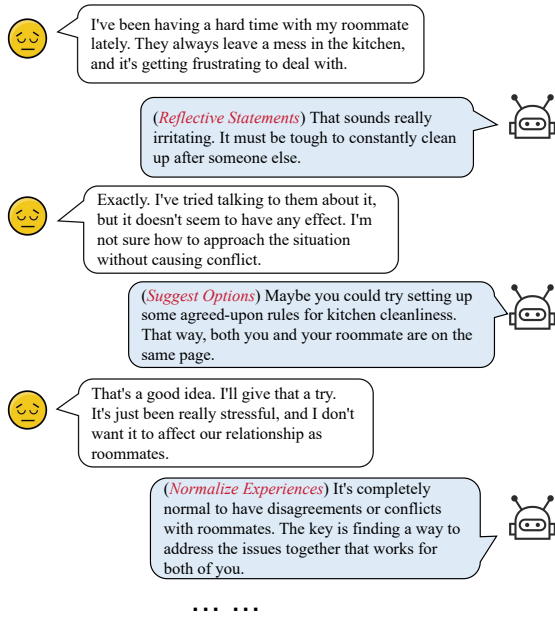


Figure 3: An example chat showing effective emotional support (adapted from our dataset) being provided to the user (left) by the assistant (right). The support strategies used by the assistant are marked in parentheses before the utterances.

B Definitions of Strategies

Reflective Statements (RS) Repeat or rephrase what the User has expressed to show that you're actively listening. Example: User: "I'm feeling really overwhelmed with all the work I have to do." Assistant: "It sounds like you're feeling overwhelmed with your workload."

Clarification (Cla) Seek clarification to ensure a clear understanding of the User's emotions and experiences. Example: User: "I just can't shake off this feeling of sadness." Assistant: "Could you help me understand what might have triggered this feeling of sadness?"

Emotional Validation (EV) Acknowledge and validate the User's emotions without judgment. Example: User: "I'm so frustrated with myself for making the same mistake again." Assistant: "It's

Scene: Communication Challenges
<p>Description: With my roommate always leaving a mess in the kitchen, it's becoming increasingly hard to address the issue and maintain a harmonious living environment.</p>
<p>User: Hey.</p> <p>AI: Hi, how are you?</p> <p>User: I'm alright, thanks. But I've been having a hard time with my roommate lately. They always leave a mess in the kitchen, and it's getting frustrating to deal with.</p> <p>AI (Reflective Statements): That sounds really irritating. It must be tough to constantly clean up after someone else.</p> <p>User: Exactly. I've tried talking to them about it, but it doesn't seem to have any effect. I'm not sure how to approach the situation without causing conflict.</p> <p>AI (Suggest Options): Maybe you could try setting up some agreed-upon rules for kitchen cleanliness. That way, both you and your roommate are on the same page.</p> <p>User: That's a good idea. I'll give that a try. It's just been really stressful, and I don't want it to affect our relationship as roommates.</p> <p>AI (Normalize Experiences): It's completely normal to have disagreements or conflicts with roommates. The key is finding a way to address the issues together that works for both of you.</p> <p>User: You're right. I'll try to approach it from a calm and understanding perspective. Hopefully, we can come to a resolution.</p> <p>AI (Affirmation): I believe in your ability to handle this situation maturely. Remember to take care of yourself throughout this process.</p> <p>User: Thank you for the support. It means a lot to me. I'll do my best to handle it positively.</p> <p>AI (Offer Hope): I'm confident that with open communication and a willingness to find a solution, you'll be able to improve your living situation. Don't hesitate to reach out for support if you need it.</p> <p>User: I really appreciate your positivity and encouragement. It definitely gives me hope. Thank you again for being there.</p> <p>AI: You're welcome! I'm here whenever you need to talk. Take care and hang in there!</p> <p>User: Thank you. I'll keep that in mind. Take care too. Goodbye!</p>

Figure 4: Data example from our dataset.

completely understandable to feel frustrated when you make a mistake."

Empathetic Statements (ES) Express understanding and empathy towards the User's experiences. Example: User: "I'm really struggling with my self-confidence right now." Assistant: "I can imagine how challenging it must be to navigate through situations that affect your self-confidence."

Affirmation (Aff) Provide positive reinforcement and encouragement to uplift the User's spirits. Example: User: "I feel like I'm not good enough." Assistant: "You've accomplished so much already, and your abilities speak for themselves. Don't underestimate your capabilities."

Offer Hope (OH) Share optimistic perspectives or possibilities to instill hope. Example: User: "I don't know if things will ever get better." Assistant: "Remember that change is constant, and there are always opportunities for growth and positive change."

Avoid Judgment and Criticism (AJC) It's important to create a non-judgmental and safe space for the User to express their emotions without fear of criticism. Refrain from passing judgment or being overly critical of their experiences or choices. Example: User: "I'm feeling so guilty for taking time off work to focus on my mental health." Assistant: "Taking care of your mental health is crucial, and it's not something to feel guilty about. Your well-being should always be a priority, and I'm glad you recognized that. Is there anything I can do to support you during this time?"

Suggest Options (SO) Offer practical suggestions or alternative perspectives for addressing the issue at hand. Example: User: "I'm having trouble managing my stress." Assistant: "Have you considered trying relaxation techniques like deep breathing or mindfulness exercises?"

Collaborative Planning (CP) Work together with the User to develop an action plan. Example: User: "I want to improve my time management skills." Assistant: "Let's brainstorm some strategies together. How about breaking tasks into smaller, more manageable chunks?"

Provide Different Perspectives (PDP) Offer alternative ways of looking at the situation to help the User gain new insights. Example: User: "I'm devastated that my project didn't succeed." Assistant: "Sometimes setbacks can lead to unexpected opportunities for learning and growth. It's a chance to reassess and try again."

Reframe Negative Thoughts (RNT) Help the User reframe negative thoughts into more positive or realistic ones. Example: User: "I'm such a failure." Assistant: "Instead of thinking that way, let's focus on what you've learned from this experience and how you can apply it moving forward."

Share Information (SI) Provide educational or factual information about emotions, coping mechanisms, or self-care practices. Example: User: "I'm struggling to manage my anxiety." Assistant: "Did you know that deep breathing exercises and grounding techniques can help reduce anxiety symptoms? Would you like me to explain how to practice them?"

Normalize Experiences (NE) Explain that certain emotions or reactions are common and part of the human experience. Example: User: "I feel so guilty for taking time for myself." Assistant: "It's common to feel guilty about self-care, but it's essential for your well-being. Remember, you deserve

to prioritize your needs too."

Promote Self-Care Practices (PSP) Advocate for engaging in activities that promote well-being and self-care. Example: "Make sure to take some time for yourself and do something that brings you joy and relaxation."

Stress Management (SM) Provide suggestions for stress management techniques like exercise, meditation, or spending time in nature. Example: "Engaging in regular physical activity can help reduce stress and improve mood."

Others (Oth) Interact with friendly greetings and employ additional supportive techniques that are not covered by the previously mentioned categories.

C ChatGPT Self-Chat Template

The template for ChatGPT self-chat (Figure 1) is as follows:

Remember here is a comprehensive list of typical strategies for responding in conversations for emotional support, along with examples for each: 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. 3. Emotional Validation: Acknowledge and validate the person's emotions without judgment. 4. Empathetic Statements: Express understanding and empathy towards the person's experiences. 5. Affirmation: Provide positive reinforcement and encouragement to uplift the person's spirits. 6. Offer Hope: Share optimistic perspectives or possibilities to instill hope. 7. Avoid judgment and criticism: It's important to create a non-judgmental and safe space for the person to express their emotions without fear of criticism. Refrain from passing judgment or being overly critical of their experiences or choices. 8. Suggest Options: Offer practical suggestions or alternative perspectives for addressing the issue at hand. 9. Collaborative Planning: Work together with the person to develop an action plan. 10. Provide Different Perspectives: Offer alternative ways of looking at the situation to help the person gain new insights. 11. Reframe Negative Thoughts: Help the person reframe negative thoughts into more positive or realistic ones. 12. Share Information: Provide educational or factual information about emotions, coping mechanisms, or self-care practices. 13. Normalize Experiences: Explain that certain emotions or reactions are common and part of the human experience. 14. Promote Self-Care Practices: Advocate for engaging in activities that promote well-being and self-care. 15. Stress Management: Provide suggestions for stress management techniques like exercise, meditation, or spending time in nature. 16. Others: Other strategies. Example:

\$(SEED EXAMPLE)

Your task is to create a casual emotional support conversation between a user and an assistant. Create a random emotional support scenario of the '\$[SCENE]' type, write it in the Description, and then generate a complete set of dialogue. Make the conversation more like a real-life chat and be specific. Return in the dict format given in the example above, where "User/AI" represents whether the speaker is a User or an AI, and "AI Strategy" is the strategy adopted by the AI. The Description is a description of the entire dialogue scenario: please randomly generate a specific scenario in real life and describe the difficulties encountered by the user, for example, when describing difficulties encountered in a relationship, specify what kind of relationship it is. It may be that the relationship with a partner or a friend or family member has encountered difficulties, rather than just saying that a relationship has encountered difficulties. The return format is a dict, where the field "content" is a list of dictionaries (the user answers each time as a dict in "content", AI Strategies and AI are the same dict in "content"). The "scene" is the same as in the above example, do not change.

D Examples of Scenarios

Breakups or Divorce Example 1: Processing the emotions and grief following the end of a long-term relationship. Example 2: Seeking guidance on how to navigate a recent breakup and move forward.

Conflicts or Communication Problems Example 1: Dealing with a misunderstanding or disagreement with a close friend or family member. Example 2: Seeking advice on resolving conflicts with a romantic partner and improving communication.

Communication Challenges Example: Helping a person find effective ways to express their needs and concerns to their partner, fostering open and constructive communication.

Coping with the Death of a Loved One Example 1: Navigating the stages of grief and finding ways to honor the memory of the deceased. Seeking support in managing the emotional impact of losing a close family member or friend.

Dealing with the Loss of a Pet Example 1: Processing the deep sadness and emptiness after the death of a beloved pet. Example 2: Seeking understanding and comfort while grieving the loss of a long-time companion animal.

Work-related Stress and Burnout Example 1: Coping with excessive workload, pressure, and a demanding work environment. Example 2: Seeking strategies to manage stress and achieve a healthier work-life balance.

Financial Worries and Uncertainty Example 1: Navigating financial challenges such as debt, job loss, or unexpected expenses. Example 2: Seeking emotional support and practical advice to alleviate financial stress and regain stability.

Unemployment-related stress Example: Encouraging someone who is about to lose their job due to poor company performance, discussing the possibility of changing jobs, prioritizing self-care, and staying positive.

Academic Stress Example: Offering guidance and study tips to a student feeling overwhelmed by their workload, helping them create a study plan and adopt healthy stress management techniques.

Depression and Low Mood Example 1: Dealing with feelings of sadness, loss of interest, and lack of motivation. Example 2: Seeking guidance on coping mechanisms and professional help for managing depression symptoms.

Managing Bipolar Disorder Example 1: Finding support and strategies to navigate the highs and lows of bipolar disorder. Example 2: Seeking advice on maintaining stability, managing medication, and recognizing warning signs.

Anxiety and Panic Example: Providing guidance and techniques for someone who experiences social anxiety, helping them gradually face their

fears and build confidence in social situations.

Depression and Low Mood Example: Being there for a person experiencing depression, actively listening to their struggles, and encouraging them to seek professional help and engage in self-care activities.

Adjusting to a New Job or Role Example 1: Coping with the challenges and expectations of a new job or promotion. Example 2: Seeking guidance on adapting to a new work environment and building professional relationships.

Moving to a New City or Country Example 1: Dealing with feelings of homesickness, cultural adjustment, and building a new social network. Example 2: Seeking support in navigating the practical and emotional aspects of relocating to a different city or country.

Career Transitions Example: Assisting someone who is considering a career change, helping them explore their passions, transferable skills, and develop a plan for transitioning into a new field.

Parenthood and Parenting Challenges Example: Supporting a new parent who is feeling overwhelmed and sleep-deprived, offering reassurance, and sharing tips for self-care and coping strategies for the demands of parenthood.

Low Self-Esteem or Lack of Confidence Example 1: Addressing negative self-perceptions and building self-worth. Example 2: Seeking techniques for cultivating self-compassion and improving self-esteem.

Body Image Concerns and Eating Disorders Example 1: Dealing with body dissatisfaction and the impact it has on self-image and overall well-being. Example 2: Seeking support in recovering from an eating disorder and developing a healthy relationship with food and body.

LGBTQ+ Identity Example: Assisting someone in the process of coming out as gay, offering support, connecting them with LGBTQ+ community resources, and being a source of understanding.

Cultural Identity and Belonging Example: Engaging in discussions with someone who is exploring their mixed-race identity and helping them embrace and celebrate their diverse heritage.

Academic Stress or Pressure Example 1: Coping with academic expectations, exam anxiety, or perfectionism. Example 2: Seeking strategies for time management, study techniques, and reducing academic stress.

Job Loss or Career Setbacks Example 1: Navi-

gating the emotions and challenges of losing a job or facing career setbacks. Example 2: Seeking guidance and encouragement for career transitions or exploring new professional opportunities.

Parenting Challenges and Parental Guilt Example 1: Managing parental responsibilities, parenting styles, and dealing with parental guilt. Example 2: Seeking advice on effective communication with children and finding a balance between work and family.

Sibling Rivalry or Family Conflict Example 1: Resolving conflicts and improving relationships with siblings or other family members. Example 2: Seeking guidance on navigating family dynamics, establishing healthy boundaries, and fostering understanding.

Surviving and Recovering from Physical or Emotional Abuse Example 1: Processing the trauma of past abuse and seeking support for healing and recovery. Example 2: Finding resources and coping strategies for managing the emotional impact of abuse.

Healing from Sexual Assault or Domestic Violence Example 1: Navigating the complex emotions, seeking support, and developing coping mechanisms after experiencing sexual assault or domestic violence. Example 2: Accessing information on trauma-informed therapy and support networks for survivors of assault or violence.

Post-Traumatic Stress Disorder (PTSD) Example: Creating a safe and non-judgmental space for a military veteran with PTSD to share their experiences and providing resources for trauma-focused therapy and support groups.

Healing from Abuse Example: Assisting someone who has recently left an abusive relationship, connecting them with local support services, and offering encouragement as they rebuild their life.

Navigating Gender Identity and Transitioning Example 1: Seeking support and resources while exploring gender identity and considering transitioning. Example 2: Accessing guidance on navigating social, medical, and legal aspects of transitioning.

Chronic Illness or Pain Management Example 1: Coping with the emotional impact of a chronic illness, including pain, limitations, and lifestyle adjustments. Example 2: Seeking support in managing daily challenges, finding self-care strategies, and connecting with others facing similar health issues.

Coping with a Diagnosis or Medical Treatment

Example 1: Processing the emotions surrounding a new medical diagnosis and navigating treatment options. Example 2: Seeking emotional support and practical guidance to cope with medical procedures, side effects, and lifestyle changes.

Caregiver Support Example: Offering guidance and resources to a caregiver of an elderly parent, discussing techniques for managing caregiver stress and suggesting respite care options.

Finding Meaning and Purpose in Life Example 1: Exploring questions related to the meaning of life, personal values, and finding purpose. Example 2: Assisting someone who is questioning their life's purpose and exploring different avenues for finding meaning, discussing their values and interests, and encouraging self-reflection.

Spirituality and Faith Example: Offering guidance and resources to someone who is questioning their faith or seeking spiritual fulfillment, providing support as they explore their beliefs and values.

Addiction and Recovery Example: Offering empathy and understanding to someone battling addiction, discussing treatment options, and providing emotional support during their journey to recovery.

Support for Loved Ones or Friends Example: Supporting a parent who has a child dealing with addiction, offering a listening ear, and connecting them with support groups and counseling services.