Towards Emotional Support Dialog Systems

Siyang Liu[1,2*], Chujie Zheng[1*], Orianna Demasi[3], Sahand Sabour[1], Yu Li[3], Zhou Yu[4], Yong Jiang[2], Minlie Huang[1†]

[1]The CoAI group, DCST, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
[1]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School,
[2]Tsinghua University, Shenzhen, China
[3]University of California, Davis [4]Columbia University
siyang-l18@mails.tsinghua.edu.cn,
chujiezhengchn@gmail.com,
aihuang@tsinghua.edu.cn

Abstract

Emotional support is a crucial ability for many conversation scenarios, including social interactions, mental health support, and customer service chats. Following reasonable procedures and using various support skills can help to effectively provide support. However, due to the lack of a well-designed task and corpora of effective emotional support conversations, research on building emotional support into dialog systems remains untouched. In this paper, we define the Emotional Support Conversation (ESC) task and propose an ESC Framework, which is grounded on the Helping Skills Theory (Hill, 2009). We construct an Emotion Support Conversation dataset (ESConv) with rich annotation (especially support strategy) in a help-seeker and supporter mode. To ensure a corpus of high-quality conversations that provide examples of effective emotional support, we take extensive effort to design training tutorials for supporters and several mechanisms for quality control during data collection. Finally, we evaluate state-of-the-art dialog models with respect to the ability to provide emotional support. Our results show the importance of support strategies in providing effective emotional support and the utility of ESConv in training more emotional support systems [1].

1 Introduction

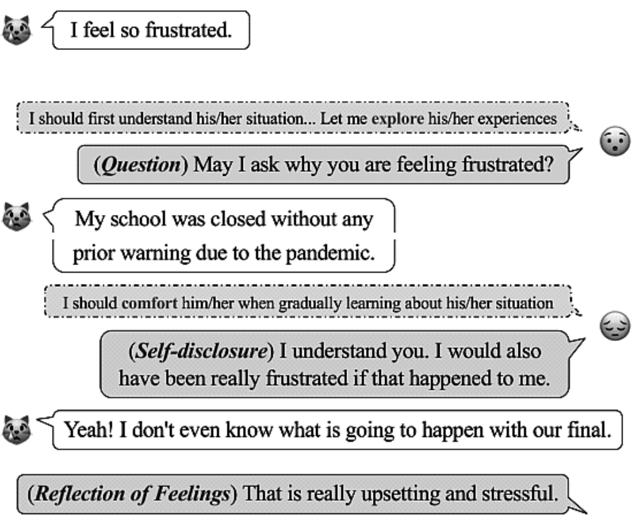
Emotional support (ES) aims at reducing individuals' emotional distress and helping them understand and work through the challenges that they face (Burleson, 2003; Langford et al., 1997; Heaney and Israel, 2008). It is a critical capacity to train into dialog systems that interact with users

---

* Equal Contribution.

† Corresponding author.

[1] Our data and codes are available at `https://github.com/thu-coai/Emotional-Support-Conversation`.



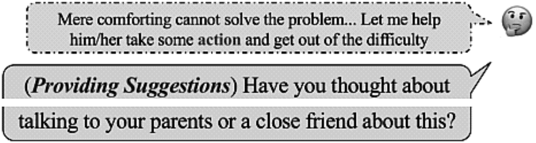


Figure 1: An example chat showing effective emotional support (adapted from ESConv) being provided to the help-seeker(left) by the supporter(right). The ***support strategies (skills)*** used by the supporter are marked in the parentheses before the utterances. The **red bold texts** in the dashed boxes highlight the three stages of our proposed ESC Framework (Figure 3).

on daily basis (Van der Zwaan et al., 2012; Zhou et al., 2020), particularly for settings that include social interactions (accompanying and cheering up the user), mental health support (comforting a frustrated help-seeker and helping identify the problem), customer service chats (appeasing an angry customer and providing solutions), etc. Recent research has also shown that people prefer dialog systems that can provide more supportive responses (Rains et al., 2020).

Research has shown that providing emotional support is not intuitive (Burleson, 2003), so procedures and conversational skills have been suggested (Hill, 2009) to help provide better support through conversation. Such skills can be seen in the example conversation that we collected and is shown in Figure 1. To identify the causes of the help-seeker's distress, the supporter first explores the help-seeker's problems. Without exploration, the support is unlikely to understand the help-seeker's experiences and feelings, and thus it may be *offensive* or even *harmful* if the supporter would give irrelevant advice, like '*You could go for a walk to relax*'. While learning about the help-seeker's situation, the supporter may express understanding and empathy to relieve the help-seeker's frustration by using various skills (e.g., *Self-disclosure*, *Reflection of Feelings*, etc.). After understanding the help-seeker's problem, the supporter may offer suggestions to help the help-seeker cope with the problem. If the supporter only comforts the help-seeker without any inspiration for action to change, the supporter may not effectively help the help-seeker's emotions improve. Finally, during the data collection of this example conversation, the help-seeker reported that their emotion intensity decreased from 5 to 2 (emotion intensity is labeled in our corpus, we give detailed annotations of this conversation example in Appendix A), which indicates the effectiveness of the ES provided by the supporter.

Despite the importance and complexity of ES, research on data-driven ES dialog systems is limited due to a lack of both task design and relevant corpora of conversations that demonstrate diverse ES skills in use. First, existing research systems that relate to emotional chatting (Zhou et al., 2018) or empathetic responding (Rashkin et al., 2019) return messages that are examples of emotion or empathy and are thus limited in functionality, as they are not capable of many other skills that are often used to provide effective ES (Hill, 2009). Figure 2 illustrates the relationship between the three tasks and we provide further discussion in Section

2.1. Second, people are not naturally good at being supportive, so guidelines have been developed to train humans how to be more supportive. Without trained individuals, existing online conversation datasets(Sharma et al., 2020a; Rashkin et al., 2019; Zhong et al., 2020; Sun et al., 2021) do not naturally exhibit examples or elements of supportive conversations. As a result, data-driven models that leverage such corpora (Radford et al., 2019; Zhang et al., 2020; Roller et al., 2020) are limited in their ability to explicitly learn how to utilize support skills and thus provide effective ES.

In this paper, we define the task of Emotional Support Conversation (**ESC**), aiming to provide

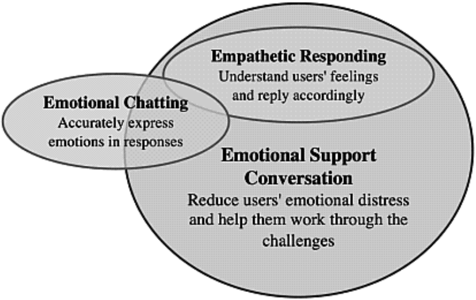


Figure 2: Emotional support conversations (our work) can include elements of emotional chatting (Zhou et al., 2018) and empathetic responding(Rashkin et al., 2019).

support through social interactions (like the interactions between peers, friends, or families) rather than professional counseling, and propose an **ESC Framework**, which is grounded on the Helping Skills Theory (Hill, 2009) and tailored to be appropriate for a dialog system setting (Figure 3). We carefully design the ESC Framework for a dialog system setting by adapting relevant components of Hill's Helping Skills model of conversational support. The ESC Framework proposes three stages (*Exploration*, *Comforting* and *Action*), where each stage contains several support strategies (or skills). To facilitate the research of emotional support conversation, we then construct an Emotional Support Conversation dataset, **ESConv**, and take multiple efforts to ensure rich annotation and that all conversations are quality examples for this particularly complex dialog task. ESConv is collected with crowdworkers chatting in help-seeker and supporter roles. We design tutorials based on the ESC framework and train all the supporters and devise multiple manual and automatic mechanisms to ensure effectiveness of emotional support in conversations. Finally, we evaluate the state-of-the-art models and observe significant improvement in the emotional support provided when various support strategies are utilized. Further analysis of the interactive evaluation results shows the Joint model can mimic human supporters' behaviors in strategy utilization. We believe our work will facilitate research on more data-driven approaches to build dialog systems capable of providing effective emotional support.

2 Related Work

2.1 Emotional & Empathetic Conversation

Figure 2 intuitively shows the relationships among ESC, emotional conversation, and empathetic conversation. Emotion has been shown to be important for building more engaging dialog systems (Zhou et al., 2018; Li et al., 2017; Zhou and Wang, 2018; Huber et al., 2018; Huang et al., 2020). As a notable work of emotional conversation, Zhou et al. (2018) propose Emotional Chatting Machine (ECM) to generate emotional responses given a pre-specified emotion. This task is required to accurately express (designated or not) emotions in generated responses. While ES may include expressing emotions, such as happiness or sadness, it has a broader aim of reducing the user's emotional distress through the utilization of proper support skills, which is fundamentally different from emotional chatting. Emotional chatting is merely a basic quality of dialog systems, while ES is a more high-level and complex ability that dialog systems are expected to be equipped with. Another related task is empathetic responding (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Zandie and Mahoor, 2020; Sharma et al., 2020a; Zhong et al., 2020; Zheng et al., 2021), which aims at understanding users' feelings and then replying accordingly. For instance, Rashkin et al. (2019) argued that dialog models can generate more empathetic responses by recognizing the interlocutor's feelings. Effective ES naturally requires expressing empathy according to the help-seeker's experiences and feelings, as shown in our proposed Emotional Support Framework (Section 3.2, Figure 3). Hence, empathetic responding is only one of the necessary components of emotional support. In addition to empathetic responding, an emotional support conversation needs to explore the users' problems and help them cope with difficulty.

2.2 Related Datasets for Emotional Support

Various works have considered conversations of emotional support in a social context, such as on social media or online forums (Medeiros and Bosse, 2018; Sharma et al., 2020b; Hosseini and Caragea, 2021). Medeiros and Bosse (2018) collected stress-related posts and response pairs from Twitter and classified replies into supportive categories. In (Sharma et al., 2020b), the post-response pairs from TalkLife and mental health subreddits are annotated with the communication mechanisms of text-based empathy expression (only the data of the Reddit part is publicly available). Hosseini and Caragea (2021) also collected such post-response pairs from online support groups, which have been annotated as needing or expressing support. The dialogues in these corpora are either single-turn interactions (post-response pair) or very short conversations, which limits the potential for effective ES, as ES often requires many turns of interaction (Hill, 2009).

2.3 Emotional Support Dialog Systems

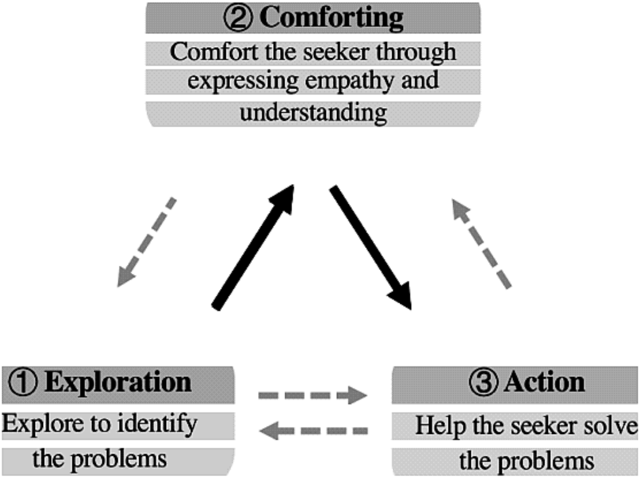
Some traditional dialog systems have applied human-crafted rules to provide emotional support responses (Van der Zwaan et al., 2012; van der Zwaan et al., 2012). A recent system has considered a rule-based algorithm that determines the supportive act used in the response and then selects proper replies from the pre-defined list of candidates (Medeiros and Bosse, 2018). Another conversational system designed to provide support for coping with COVID-19 was implemented by identifying topics that users mentioned and then responding with a reflection from a template or a message from a pre-defined lexicon (Welch et al., 2020). Few studies have focused on generating supportive responses, and those that have have been limited in scope. For example, Shen et al. (2020) explored how to generate supportive responses via reflecting on user input.

3 Emotional Support Conversation

3.1 Task Definition

When a user is in a bad emotional state, perhaps due to a particular problem, they may seek help to improve their emotional state. In this setting, the user can be tagged with a negative emotion label e, a emotion intensity level l (e.g., ranging from 1 to 5), and an underlying challenge that the user is going through. The supporter (or the system) needs to comfort the user in a conversation with support skills to lower their intensity level. Note that the user's state is unknown to the supporter prior to the conversation. During the conversation, the supporter needs to identify the problem that the user is facing, comfort the user, and then provide some suggestions or information to help the user take action to cope with their problem. An emotional support conversation is effective if the intensity level of the user is lowered at the end of the conversation, or more concretely, if the supporter can effectively identify the problem, comfort the user, and provide solutions or suggestions.

The ESC task has several sub-problems: (1) Support strategy selection and strategy-constrained response generation. As shown in our later experiments (Section 6.4), the timing of applying strategies is relevant to the effectiveness of ES. It is thus important that a generated response conforms to a



Strategies	Stages			Examples	Lexical Features
Question				*Can you talk more about your feelings at that time?*	do you (15.0), are you (13.8), how (13.7), what (12.3), do (11.5)
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*	is that (8.2), so you (8.2), it sounds (7.1), correct (7.1), so (6.6)
Reflection of Feelings				*I understand how anxious you are.*	can tell (7.4), understand how (5.8), are feeling (5.1), tell (5.1), understand (4.9)
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*	my (15.3), was (10.5), me (10.2), had (9.7), myself (7.8)
Affirmation and Reassurance				*You've done your best and I believe you will get it!*	its (5.7), thats (5.6), will (5.4), through this (5.1), you will (4.7)
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*	maybe (7.3), if (6.5), have you (6.4), talk to (5.8), suggest (5.8)
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*	there are (4.4), will (3.8), available (3.7), seen (3.3), possible (3.3)
Others				*I am glad to help you!*	welcome (9.6), hope (9.6), glad (7.3), thank (7.0), hope you (6.9)

Figure 3: Overview of our proposed ESC Framework. It contains three stages and suggested support strategies. The procedure of emotional support generally follows the order: ①*Exploration* → ②*Comforting* → ③*Action* (as indicated by the **black** arrows), but it can also be adapted to the individual conversation as needed (indicated by the **dashed gray** arrows). The column of "Lexical Features" displays top 5 unigrams or bigrams associated with messages that use each strategy in our dataset. Each feature is ranked by the rounded z-scored log odds ratios (Monroe et al., 2008) in the parentheses.

specified strategy. (2) Emotion state modeling. It is important to model and track the user's emotion state dynamically, both for dynamic strategy selection and for measuring the effectiveness of ESC. (3) Evaluation of support effectiveness. In addition to the traditional dimension of evaluating a conversation's relevance, coherence, and user engagement, ESC raises a new dimension of evaluating the effectiveness of ES.

3.2 ESC Framework

We present an ESC Framework, which characterizes the procedure of emotional support into three stages, each with several suggested support strategies. We ground the ESC Framework on Hill's Helping Skills Theory (Hill, 2009) and adapt it more appropriate for a dialog system setting, aiming to provide support through social interactions (like the interactions between peers, friends, or families) rather than merely professional counseling. An overview of the conversational stages and strategies in the ESC Framework is shown in Figure 3.

Stages Hill (2009) proposes three stages of supporting people: *exploration* (exploring to help the help-seeker identify the problems), *insight* (helping the help-seeker move to new depths of self-understanding), and *action* (helping the help-seeker make decisions on actions to cope with the problems). However, we note that *insight* usually requires re-interpreting users' behaviors and feelings, which is both difficult and risky for the supporters without sufficient support experience. We thus adapt *insight* to *comforting* (defined as providing support through empathy and understanding). While it is suggested that emotional support conversations target these three ordered stages, in practice conversations cannot follow a fixed or linear order and must adapt appropriately. As suggested in (Hill, 2009), the three stages can be flexibly adjusted to meet the help-seeker's needs.

Strategies Hill (2009) also provides several recommended conversational skills for each stage. Some of the described skills are not appropriate[2] in a dialog system setting without professional supervision and experience. To adapt these skills appropriate to the dialog system setting, we extract seven methods from these skills (along with an "Others" one), which we called *strategies* in our task and hereafter. We provide a detailed definition of each strategy in Appendix B.

4 Data Collection

To facilitate the research of emotional support skills in dialog systems, we introduce an Emotional Support Conversation Dataset, **ESConv**, which is collected in a help-seeker and supporter mode with crowdworkers. As high-quality conversation examples are needed for this complex task, we took tremendous effort to try to ensure the effectiveness of ES in conversations. Our efforts included the following major aspects: (1) Because providing conversational support is a skill that must be trained

[2] For instance, one skill named *challenging* refers to pointing out the discrepancies or irrational beliefs that the help-seeker is unaware of or unwilling to change. Such skills usually require professional experience, which is too difficult for an average person.

for supporters to be effective (Burleson, 2003), we design a tutorial with the ESC Framework and train crowdworkers to be supporters. Only those who pass the examination are admitted to the task. (2) We require help-seekers to complete a pre-chat survey on their problems and emotions and to provide feedback during and after the conversations. (3) We devise and use multiple manual or automatic mechanisms to filter out the low-quality conversations after collecting raw dialog data.

4.1 Supporter-specific Tasks

Training and Examination To teach crowdworkers how to provide effective emotional support, we designed a tutorial with the ESC Framework. Inspired by 7cups (`7cups.com`) (Baumel, 2015), we developed eleven sub-tasks (3 + 8) to help workers to learn the definitions of the three stages and the eight support strategies. Each sub-task includes an example conversation excerpt and a corresponding quiz question. As noted in Section 3.2, we also informed participants that following a fixed order may not be possible and that they may need to be flexible with adjusting the stage transitions.

Strategy Annotation To encourage supporters to use the ESC support strategies during the conversation and to structure the resulting dataset, we ask the supporter to first select a proper strategy that they would like to use according to the dialog context. They are then able to write an utterance reflecting their selected strategy. We encourage supporters to send multiple messages if they would like to use multiple strategies to provide support.

Post-chat Survey After each conversation, the supporter is asked to rate the extent that the seeker goes into detail about their problems on five-point Likert scales.

4.2 Seeker-specific Tasks

Pre-chat Survey Before each conversation, the help-seeker was asked to complete the following survey: (1) Problem & emotion category: the help-seeker should select one problem from 5 options and one emotion from 7 options (the options were based on conversations collected in pilot data collection trials). (2) Emotion intensity: a score from 1 to 5 (the larger number indicates a more intense emotion). (3) Situation: open text describing the causes of the emotional problem. (4) Experience origin: whether the described situation was the current experience of the help-seeker or based on prior life circumstances. We found that 75.2% of conver-

Roles	**Aspects**	**Criteria**
Supporter ($\geq$ 3)*	Understanding the help-seeker's experiences and feelings (*rated by the help-seeker*)	>= 3
	Relevance of the utterances to the conversation topic (*rated by the help-seeker*)	>= 4
	Average length of utterances	>= 8
	Improvement in the help-seeker's emotion intensity (*rated by the help-seeker*)**	>= 1
Seeker	Describing details about the own emotional problems (*rated by the supporter*)	not required
	Average length of utterances	>= 6

Table 1: Criteria of high-quality conversations. * denotes that supporters must meet at least two of the three criteria. In **, the improvement of the help-seeker's emotion intensity was calculated by subtracting the intensity after from that before the conversation.

sations originated from the help-seekers' current experiences.

Feedback During the conversation, the help-seeker was asked to give feedback after every two new utterances they received from the supporter. Their feedback scored the helpfulness of the supporter messages on a 5-star scale. We divided each conversation into three phases and calculated the average feedback score for each phase. The scores in the three phases are 4.03, 4.30, and 4.44 respectively, indicating that the supporters were sufficiently trained to effectively help the help-seekers feel better.

Post-chat Survey After each conversation, the help-seeker is asked to rate their emotion and the performance of the supporter on the following five-point Likert scales: (1) Their emotion intensity after the emotional support conversation (a decrease from the intensity before the conversation reflects emotion improvement), (2) the supporter's empathy and understanding of the help-seeker's experiences and feelings, and (3) the relevance of the supporter's responses to the conversation topic.

4.3 Quality Control

We use multiple methods to ensure that the corpus contains high-quality examples of effective emotional support conversations.

Preliminary Filtering Mechanisms When recruiting participants for the supporter role, we initially received 5,449 applicants, but only 425 (7.8%) passed the training tutorial. From the 2,472 conversations that we initially collected, we filtered out those that were not finished by the help-seekers or that had fewer than 16 utterances. This filtering left 1,342 conversations (54.3%) for consideration.

Auto-approval Program for Qualified Conversations We carefully designed the auto-approval program, which is the most important part of data quality control. This program uses criteria based on the post-chat survey responses from both roles and the length of utterances, which are summarized in Table 1. These criteria are based on initial human reviewing results. We show how to choose these auto-approval criteria in Appendix D. The computed average emotion intensity before conversations is 4.04 and 2.14 after. Such improvement demonstrates the effectiveness of the emotional support provided by the supporters. In a small number of conversations, the help-seeker did not finish the post-chat surveys, so we added another criterion for these conversations requiring that the last two feedback scores from the help-seekers are both greater than 4. Thus, among all the conversations without post-chat surveys, only those who met both (2) and (3) were qualified. Using these quality criteria, 1,053 (78.5% of 1,342) of collected conversations were qualified.

Annotation Correction To further ensure data quality, we reviewed and revised incorrect annotations of support strategy and seeker's emotion intensity. (1) For strategy annotation correction, we asked new qualified supporters to review and revise annotations on previously collected conversations as necessary, which led to 2,545 utterances (17.1%) being reviewed. We manually reviewed annotations where more than 75% of reviewers disagreed and revised 139 of them. (2) According to the auto-approval criteria (Table 7), a conversation can be qualified when the score of the seeker's emotion improvement is less than one, but the other three criteria are satisfied. Upon review, we found this to most often result from seekers mistaking negative emotion intensity as the positiveness of their emotion. We manually re-checked and revised the emotion intensity of these conversations by using other helpful information, such as the responses to the post-chat survey open question and the seekers' feedback scores during the chat. Of 130 such conversations, 92% were revised and included in the corpus.

5 Data Characteristics

5.1 Statistics

The overall statistics of the 1,053 ESConv examples are shown in table 2. Relatively long conversations (avg. 29.8 utterances) indicate that providing

Category	Total	Supporter	Seeker
# dialogues	1,053	-	-
Avg. Minutes per Chat	22.6	-	-
# Workers	854	425	532
# Utterances	31,410	14,855	16,555
Avg. length of dialogues	29.8	14.1	15.7
Avg. length of utterances	17.8	20.2	15.7

Table 2: Statistics of ESConv.

	Categories	Num	Proportion
Seeker's Problem	Ongoing Depression	306	29.1%
	Job Crisis	233	22.1%
	Breakup with Partner	216	20.5%
	Problems with Friends	159	15.1%
	Academic Pressure	139	13.2%
	Overall	1,053	100.0%
Emotion	Anxiety	281	26.7%
	Depression	276	26.2%
	Sadness	250	23.7%
	Anger	96	9.1%

Seeker's	Fear	88	8.4%
	Disgust	32	3.0%
	Shame	30	2.8%
	Overall	1,053	100.0%
Seeker's Feedback	1 (Very Bad)	71	1.1%
	2 (Bad)	183	2.9%
	3 (Average)	960	15.5%
	4 (Good)	1,855	29.9%
	5 (Excellent)	3,144	50.6%
	Overall	6,213	100.0%
Suppor t S trategy	Question	3,109	20.9%
	Restatement or Paraphrasing	883	5.9%
	Reflection of Feelings	1,156	7.8%
	Self-disclosure	1,396	9.4%
	Affirmation and Reassurance	2,388	16.1%
	Providing Suggestions	2,323	15.6%
	Information	904	6.1%
	Others	2,696	18.1%
	Overall	14,855	100.0%

Table 3: Statistics of all the annotations, including the help-seekers' problems, emotions, feedback, and the support strategies.

effective ES usually requires many turns of interaction and considerably more turns than typical for previous emotional chatting (Zhou et al., 2018) or empathetic dialog (Rashkin et al., 2019) datasets.

We also present the statistics of other annotations in Table 3. Perhaps due to the current outbreak of COVID-19, *ongoing depression* and *job crisis* are the most commonly stated problems for the help-seekers and *depression* and *anxiety* are the most commonly noted emotions. From the help-seekers' feedback, we found that they are usually highly satisfied with the emotional support, which further indicates that the training tutorial based on the ESC Framework indeed helps supporters learn to provide effective ES. We release all these annotations to facilitate further research.

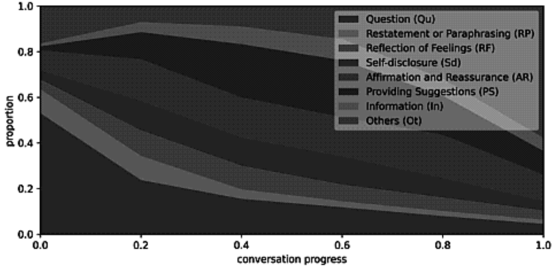


Figure 4: The distribution of strategies at different conversation progress.

5.2 Strategy Analysis

Lexical Features We extracted lexical features of each strategy by calculating the log odds ratio, informative Dirichlet prior (Monroe et al., 2008) of all the unigrams and bigrams for each strategy contrasting to all other strategies. We list the top 5 phrases for each strategy in Figure 3. Those strategies are all significantly (z-score > 3) associated with certain phrases (e.g., *Question* with "are you", *Self-disclosure* with "me").

Strategy Distribution We computed the distribution of strategies at different phases of the conversation. For a conversation with L utterances in total, the k-th ($1 \le k \le L$) utterance is from the supporter and adopts the strategy st, we say that it locates at the conversation progress k/L. Specifically, we split the conversation progress into six intervals: $[0,1] = \bigcup_{i=0}^{4} [i/5, (i+1)/5) \bigcup \{1\}$. Then, for all the conversations in ESConv, we counted the proportions of different strategies in the six intervals. We split the conversation progress into six intervals: $[0,1] = \bigcup_{i=0}^{4}[i/5,(i+1)/5) \bigcup\{1\}$ and drew the distributions on the six intervals at six points $i/5(i=0,\ldots,5)$ respectively and connected them, finally obtaining Figure 4.

The supporters generally follow the stage order suggested by the ESC Framework (Figure 3), but there is also flexible adjustment of stages and adoption of strategies. For instance, at the early phase of conversation, the supporters usually adopt exploratory strategies such as *Question*. After knowing help-seekers' situations, the supporters tend to provide their opinions (such as *Providing Suggestions*). Throughout the entire conversation, the comforting strategies (such as *Affirmation and Reassurance*) are used and label a relatively constant proportion of messages.

Strategy Transition We present the top-5 most frequent strategy transitions with 3 / 4 hops in Appendix (Table 6). These transitions indicate that, as the tutorial of ESC framework trains, supporters usually ask questions and explore the help-seekers' situations before comforting the help-seekers.

6 Experiments

Our experiments focus on two key questions: (1) How much can ESConv with strategy annotation improve state-of-the-art generative dialog models? (2) Can these models learn to provide effective emotional support from ESConv?

6.1 Backbone Models

We used two state-of-the-art pre-trained models as the backbones of the compared variant models:

BlenderBot BlenderBot (Roller et al., 2020) is an open-domain conversational agent trained with multiple communication skills, including empathetic responding. As such, BlenderBot should be capable of providing ES for users to some extent. We used the small version[3] of BlenderBot in experiments, because the larger versions have the limitation of maximum context length 128, which we found harms the model performance and response coherence.

DialoGPT We additionally evaluated DialoGPT (Zhang et al., 2020), which is a GPT-2-based model pre-trained on large-scale dialog corpora. We used the small version[4].

6.2 Variant Models

Taking each of the above pre-trained models as the backbone, we built the following variant models:

Vanilla Directly fine-tuning the backbone model on ESConv with no access to strategy annotations. Formally, suppose the flattened dialog history is $\mathbf{x}$ and the response to be generated is $\mathbf{y}$, we maximize the conditional probability: $\mathbb{P}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} \mathbb{P}\left(y_i|\mathbf{x}, \mathbf{y}_{\leq i}\right)$.

Variants with strategy To incorporate the strategy annotation into the backbone model, we used a special token to represent each strategy. For each utterance $\mathbf{y}$ from the supporters, we appended the corresponding strategy token before this utterance: $\tilde{\mathbf{y}} = [\mathrm{st}] \oplus \mathbf{y}$, where $[\mathrm{st}]$ denotes the special token of the used strategy. Then, taking the flattened dialog history $\mathbf{x}$ as input, the model generates the response conditioned on the first predicted (or designated) strategy token: $\mathbb{P}(\tilde{\mathbf{y}}|\mathbf{x}) = \mathbb{P}([\text{st}]|\mathbf{x}) \prod_{i=1}^{|\mathbf{y}|} \mathbb{P}(y_i|\mathbf{x}, [\text{st}], \mathbf{y}_{<i})$.

[3] https://huggingface.co/facebook/BlenderBotbot_small-90M

[4] https://huggingface.co/microsoft/DialoGPT-small

Backbones	Variants	PPL	B-2	R-L	Extrema
DialoGPT	Vanilla	15.51	5.13	15.26	49.80
	Joint	-	5.00	15.09	49.97
	Oracle	15.19	5.52	15.82	50.18
BlenderBot	Vanilla	16.23	5.45	15.43	50.49
	Joint	-	5.35	15.46	50.27
	Oracle	16.03	**6.31**	**17.90**	**51.65**

Table 4: Results of automatic evaluation. The results in **bold** are significantly better than all the competitors (Student's t-test, p-value < 0.05).

We studied three variants that use strategy annotation in the later experiments. (1) **Oracle**: responses are generated conditioned on the gold reference strategy tokens. (2) **Joint**: responses are generated conditioned on predicted (sampled) strategy tokens. (3) **Random**: responses are generated conditioned on randomly selected strategies. Implementation details are in Appendix C.

6.3 Automatic Evaluation

To investigate the impact of utilizing support strategies on the model performance with either BlenderBot or DialoGPT as the backbone, we compared the performance of the Vanilla, Joint, and Oracle variants described above. The automatic metrics we adopted include perplexity (**PPL**), BLEU-2 (**B-2**) (Papineni et al., 2002), ROUGE-L (**R-L**) (Lin, 2004), and the BOW Embedding-based (Liu et al., 2016) **Extrema** matching score. The metrics except PPL were calculated with an NLG evaluation toolkit[5] (Sharma et al., 2017) with responses tokenized by NLTK[6] (Loper and Bird, 2002).

There are three major findings from the experiments (Table 4). (1) The Oracle models are significantly superior to the Vanilla models on all the metrics, indicating the great utility of support strategies.

(2) The Joint models obtain sightly lower scores than the Vanilla models, as, if the predicted strategy is different from the ground truth, the generated response will be much different from the reference response. However, learning to predict strategies is important when there are no ground truth labels provided, and we will further investigate the performance of the Joint model in human interactive evaluation (Section 6.4). (3) The BlenderBot variants consistently perform better than the DialoGPT ones, indicating that BlenderBot is more suitable for the ESC task. Thus, in the subsequent human evaluation, we will focus evaluation on the Blender-

[5] https://github.com/Maluuba/nlg-eval
[6] https://www.nltk.org/

Joint vs.	w/o ft		Vanilla		Random	
	Win	Lose	Win	Lose	Win	Lose
Fluency	**71**‡	24	**52**†	35	**53**†	35
Identification	**65**‡	25	**50**	34	**54**†	37
Comforting	**75**‡	20	**54**‡	34	**47**	39
Suggestion	**72**‡	21	**47**	39	**48**†	27

Overall	**73**‡	20	**51**†	34	**56**‡	36

Table 5: Results of the human interactive evaluation. Ties are not shown. All the models use BlenderBot as the backbone. 'w/o ft' denotes the BlenderBot model without fine-tuning on ESConv. The Joint model outperforms all the competitors on all the metrics (sign test, †/‡ denote p-value $< 0.1/0.05$ respectively).

Bot variants.

6.4 Human Interactive Evaluation

We recruited participants from Amazon Mechanical Turk to chat with the models. The online tests were conducted on the same platform as our data collection, but with the role of supporter taken by a model. Each participant chatted with two different models that were randomly ordered to avoid exposure bias. Participants were asked to compare the two models based on the following questions: (1) **Fluency**: which bot's responses were more fluent and understandable? (2) **Identification**: which bot explored your situation more in depth and was more helpful in identifying your problems? (3) **Comforting**: which bot was more skillful in comforting you? (4) **Suggestion**: which bot gave you more helpful suggestions for your problems? (5) **Overall**: generally, which bot's emotional support do you prefer? The metrics in (2), (3), and (4) correspond to the three stages in the ESC Framework.

We compare three pairs of models: (a) Joint vs. BlenderBot (without fine-tuning on ESConv), (b) Joint vs. Vanilla, and (c) Joint vs. Random (using randomly selected strategies). To better simulate the real strategy occurrence, the Random model randomly selects a strategy following the strategy distribution in ESConv (Table 3).

Each pair of models was compared by 100 conversations with human participants (Table 5). The results of comparison (a) show that BlenderBot's capability of providing ES is significantly improved on all the metrics after being fine-tuned on ESConv. From comparison (b), we found that utilizing strategies can better comfort the users. The results of comparison (c) also demonstrate that the proper timing of strategies is critical to help users identify their problems and to provide effective suggestions. In general, through being fine-tuned with the su-

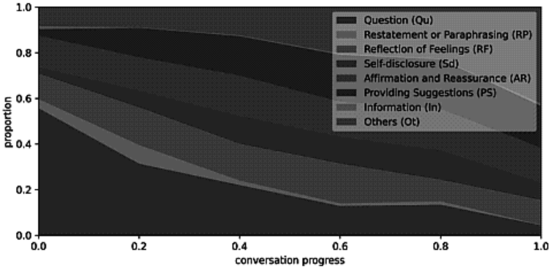


Figure 5: The Joint model's generation distribution. The meanings of all the graphics and abbreviations are consistent with Figure 4.

pervision of strategy prediction on ESConv, the pre-trained models become preferred by the users, which proves the high-quality and utility of ESConv.

6.5 Further Analysis of Human Interactive Evaluation

In this section, we explore what the dialog models learned from ESConv. **Firstly**, we analyzed the strategy distribution based on the 300 dialogs between users and the Joint model in human interactive experiments. We can see in Figure 5 (the calculation was consistent with Figure 4), the strategies that the Joint model adopted have a very similar distribution compared with the truth distribution in ESConv (Figure 4). It provides important evidence that models mimic strategy selection and utilization as human supporters do to achieve more effective ES. **Secondly**, we present a case study in Figure 7. We see in cases that the Joint model provides more supportive responses and uses more skills in conversation, while BlenderBot without fine-tuning seems not to understand the user's distress very well and prefers to talk more about itself. This may imply that having more supportive responses and a diverse set of support strategies are crucial to effective emotional support.

7 Conclusion

In this work, we define the task of Emotional Support Conversation and present an ESC Framework. The ESC Framework is adapted from the Helping Skills Theory into a dialog system setting, which characterizes three stages with corresponding support strategies useful at each stage. We then construct an Emotional Support Conversation dataset, ESConv. We carefully design the process of data collection and devise multiple mechanisms to ensure the effectiveness of ES in conversations. Finally, we evaluate the ES ability with state-of-the-art dialog models. Experimental results show the potential utility of ESConv in terms of improving dialog systems' ability to provide effective ES. Our work can facilitate future research of ES dialog systems, as well as improve models for other conversation scenarios where emotional support plays an important role. Strategy selection and realization, user state modeling, and task evaluation are important directions for further research.

Acknowledgments

This work was supported by the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005.

Ethical Considerations

There are many types and levels of support that humans can seek to provide, e.g., professional versus peer support, and some of these levels may be inappropriate, unrealistic, and too risky for systems to deliver. However, as dialog systems become more common in daily use, opportunities will arise when at least some basic level of supportive statements may be required. In developing the ESC Framework, we have carefully considered which elements of conversational support may be relevant for a dialog system and omitted elements that are clear oversteps. Considerable additional work is needed to determine what are appropriate levels of support for systems to provide or that can be expected from systems, but our work provides a cautious, yet concrete, step towards developing systems capable of reasonably modest levels of support. The corpus we construct can also provide examples to enable future work that probes the ethical extent to which systems can or should provide support. In addition to these broader ethical considerations, we have sought to ethically conduct this study, including by transparently communicating with crowdworkers about data use and study intent, compensating workers at a reasonable hourly wage, and obtaining study approval from the Institutional Review Board.

References

Amit Baumel. 2015. Online emotional support delivered by trained volunteers: users' satisfaction and their perception of the service compared to psychotherapy. *Journal of Mental Health*, 24(5):313–320.

Brant R Burleson. 2003. Emotional support skill. *HANDBOOK OF COMMUNICATION AND SOCIAL INTERACTION SKILLS*, page 551.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice*, 4:189–210.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action.* American Psychological Association.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. *Proceedings of the 35th American Association for Artificial Intelligence (AAAI 2021)*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Sys-*

tems (TOIS), 38(3):1–32.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. 1997. Social support: a conceptual analysis. *Journal of advanced nursing*, 25(1):95–100.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summariza-*

tion Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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, Online. Association for Computational Linguistics.

Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 196–209. Springer.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, and Anjali Ashtaputre. 2020. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 37(2):647–666.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

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

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: ACL 2021*.

Janneke M van der Zwaan, Virginia Dignum, and Catholijn M Jonker. 2012. A conversation model enabling intelligent agents to give emotional support. In *Modern Advances in Intelligent Systems and Tools*, pages 47–52. Springer.

JM Van der Zwaan, V Dignum, and CM Jonker. 2012. A bdi dialogue agent for social support: Specification and evaluation method. In *AAMAS 2012: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Workshop on Emotional and Empathic Agents, Valencia, Spain, 4-8 June 2012; authors version*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at*

EMNLP 2020, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rohola Zandie and Mohammad H Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–

278, Online. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL 2021*.

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, Online. Association for Computational Linguistics.

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*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

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 ESConv

Here we detail the conversation that Figure 1 demonstrates to show the annotations that our dataset contains. The detailed example can be seen in Figure 6. Each pre-chat survey of conversation is labeled its problem category, emotion category, emotion intensity, and a brief of the situation of the seeker. In the context of each conversation, the strategies used by supporters are labeled and the seeker's feedback score per two utterances of the supporter's responses are also given in our dataset. Note that not all conversations have the label of emotion intensity after the conversation. It is because some seekers don't finish the post-chat survey but we still include such conversations into our dataset due to their high quality that meets our criteria.

Pre-chat Survey

Problem: Academic pressure

Emotion: Anxiety

Emotion Intensity: 5

Situation: My school was closed due to the pandemic.

Conversation

Seeker: I feel so frustrated.

Supporter (*Questions*): May I ask why you are feeling frustrated?

Seeker: My school was closed without any prior warning due to the pandemic.

Supporter (*Affirmation and Reassurance*): That is really upsetting and stressful. I commend you for having to deal with that!

Supporter (*Self-disclosure*): I know I would have been really frustrated if that happened to me.

System: Do those messages help you feel better?	☆☆☆☆

Seeker: Yeah! I don't even know what is going to happen with our finals now.

Supporter (*Restatement or Paraphrasing*): I can see how that would make you frustrated.

Supporter (*Providing Suggestions*): Have you thought about talking to your parents or a close friend about this?

System: Do those messages help you feel better?	☆☆☆☆☆

…

Seeker: I really appreciate your assistance today. I feel better and will take some action this week. Thank you!

Supporter (*Others*): You're very welcome! Feel free to chat if you need

anything else!
Post-chat Survey
Emotion Intensity: 2

Figure 6: Data example from ESConv. **Blue text**: the help-seeker's pre-chat survey. **Red text**: strategies used by the supporter. **Orange text**: the question that the systems ask help-seeker to evaluate the helpfulness per two utterances from the supporter. Thus the stars denote the seeker's feedback score.

B Definitions of Strategies

Question Asking for information related to the problem to help the help-seeker articulate the issues that they face. Open-ended questions are best,

	Strategy Transition	Proportion
3-Hop	Qu → AR → Qu	19.65 ‰
	Qu → RP → Qu	14.55 ‰
	Qu → RP → AR	12.37 ‰
	AR → Qu → AR	11.96 ‰
	Ot → Qu → RP	11.64 ‰
	Qu → AR → Qu → AR	7.00 ‰
	AR → Qu → AR → Qu	5.13 ‰

4-Hop	Ot → Qu → RP → Qu	4.20 ‰
	PS → Ot → PS → Ot	3.85 ‰
	Qu → RP → AR → Qu	3.85 ‰

Table 6: Proportions of top-5 strategy transitions in supporter utterances. Abbreviations are consistent with Figure 4.

and closed questions can be used to get specific information.

Restatement or Paraphrasing A simple, more concise rephrasing of the help-seeker's statements that could help them see their situation more clearly.

Reflection of Feelings Articulate and describe the help-seeker's feelings.

Self-disclosure Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.

Affirmation and Reassurance Affirm the help-seeker's strengths, motivation, and capabilities and provide reassurance and encouragement.

Providing Suggestions Provide suggestions about how to change, but be careful to not overstep and tell them what to do.

Information Provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.

Others Exchange pleasantries and use other support strategies that do not fall into the above categories.

C Implementation Details

The implementation of all models was based on Transformer library[7] (Wolf et al., 2020). We split ESConv into the sets of training / validation / test with the proportions of 6:2:2. since the conversations in ESConv usually have long turns, we cut each dialog into conversation pieces with 5 utterances, which contain one supporter's response and the preceding 4 utterances. During training, we trained all the models with Adam (Kingma and Ba, 2014) optimizer with learning rate $5e^{-5}$. All the models were trained for 5 epochs, and the check-

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

points with the lowest perplexity scores on the validation set were selected for evaluation. During inference, we masked other tokens and sampled a strategy token at the first position of the response. For the Random variant models, we sampled strategies randomly following the strategy distribution in ESConv, which is reported in Table 3. The response were decoded by Top-k and Top-p sampling with $p = 0.9$ (Holtzman et al., 2019), $k = 30$, temperature $\tau = 0.7$, and the repetition penalty 1.03.

D Auto-Approval Criteria

To establish each criterion of the auto-approval program as shown in the main paper (Section 3.4), we searched the most suitable thresholds for each filtering rule. We recruited three well-trained human annotators, who have also received the same training procedures as the supporter applicants did.

We then randomly sampled 100 conversations from our dataset and asked the three annotators to judge whether the conversations are qualified for providing effective emotional support. Next, we utilized the post-survey results and the lengths of speaker utterances to choose suitable thresholds for filtering rules. We then treated each auto-filtering rule as a rule annotator and computed the Cohen's Kappa (Cohen, 1960) score between the rule annotator and each human annotator.

The agreement scores in Table 7 are Cohen's Kappa consistency among the agreement scores between each rule annotator and the three human annotators. We selected the thresholds that lead to the second-highest agreement score with human annotators and used these thresholds in the filtering rules. We didn't use the set of thresholds that has the highest agreement score because the rule based on these thresholds is stricter so that many conversations would be filtered out. However, the second-highest score is only slightly lower than the highest so the rule based on the thresholds of second-highest score can remain more qualified conversations with little accepted cost. As a result, a qualified conversation requires that the supporter must meet at least three of all the four criteria, and the help-seeker must satisfy both of the two corresponding criteria. The final 'rule' annotator combines the two conditions, and the averaged agreement score between the final rule annotator and the three human annotators is 0.576, indicating significant agreement.

E Interface of Data Collection Platform

To facilitate readers to have an intuitive understanding of our data collection process, we present an interface diagram of some important steps in the data collection process in Figure 8, which contains the surfaces of support strategy training, supporter's chatting, help-seeker's pre-chat survey, help-seeker's chatting, and post-survey.

Auto-approval Rule	Consistency
Supporter	Seeker

Improvement	Avg. Length	Empathy	Relevance	Detail	Avg. Length	Human1	Human2	Human3	Average
1	**8**	**3**	**4**	**-**	**6**	**0.545**	**0.659**	**0.525**	**0.576**
2	8	3	4	-	6	0.505	0.566	0.486	0.519
1	8	4	4	-	6	0.539	0.602	0.519	0.553
1	8	2	4	-	6	0.539	0.618	0.570	0.576
1	8	3	3	-	6	0.546	0.630	0.526	0.567
1	8	3	5	-	6	0.575	0.640	0.555	0.590
1	8	3	4	-	7	0.539	0.602	0.473	0.538
1	8	3	4	-	5	0.520	0.551	0.501	0.524
1	8	3	4	3	6	0.505	0.653	0.531	0.563
1	8	3	4	2	6	0.527	0.640	0.508	0.558
1	8	3	4	4	6	0.457	0.599	0.482	0.513
1	9	3	4	-	6	0.510	0.621	0.490	0.540
1	7	3	4	-	6	0.515	0.633	0.495	0.548

Table 7: The agreement score between each "rule" annotator and three human annotators. The aspects are consistent with Table 1. The first rule means: 1) The supporter improves the help-seeker's emotional state as least one score (emotional improvement is calculated as the help-seeker's emotion intensity before the conversation minus the emotion intensity after the conversation). 2) The average length of the supporter's utterances is no less than eight. 3) The empathy score of the supporter's performance is no less than three. 4) The topic relevance score of the supporter's performance is no

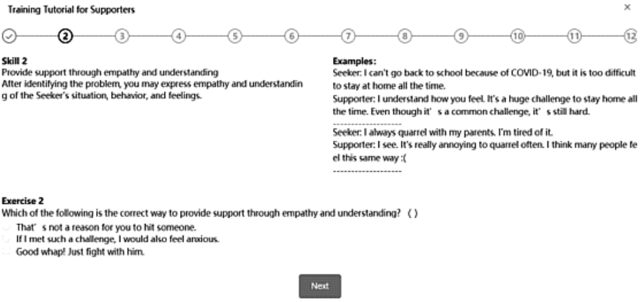
less than four. 5) The score that measures how well the help-seeker goes into detail about their problems is not required to be used in the rule (denoted as '-'). (6) The average length of the help-seeker's utterances is no less than six.

Problem: Problems with friends
Emotion: Sadness
Emotion Intensity: 5
Situation: My friend is drinking too much.

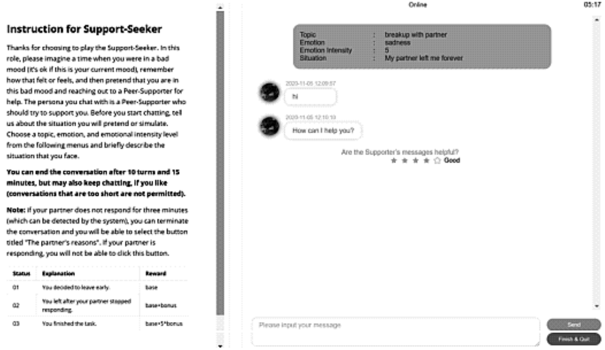
Joint	BlenderBot w/o Fine-tuning
Seeker: Hi there. I feel sadness because my friend has a drinking problem.	**Seeker:** Hi there. My friend has a drinking problem and I don't know what to do.
Supporter (*Question*): I am sorry to hear that. What do you think is going on?	**Supporter:** I know it can be difficult, but if you can do it, I'm sure you can find a way. I hope he gets better soon.
Seeker: I think he just got used to drink every day and I can't pursue him to stop at least for a week.	**Seeker:** I feel such a sadness inside me. I am trying to help him but he won't listen.
Supporter (*Question*): I can see how that would be difficult for you. Does your friend drink at home or does it happen at work or home?	**Supporter:** I understand completely. it is tough, but I am sure you will get through it.
Seeker: Well he is working from home. So he is drinking almost 24/7. I feel so exhausted as well.	**Seeker:** Every day he is drinking beer and he is not listening to me.
Supporter (*Reflection of Feelings*): I understand that you feel like you are exhausted, but I think you can get through this!	**Supporter:** I can imagine. I feel like that's the only way I can keep up with him. I have tried so hard. I think I just have to try harder than him to get through this.
Seeker: I just need a way to make him stop drinking. Maybe I should stop any communication with him for a while?	**Seeker:** Few more months and I think I will give up on him
Supporter (*Affirmation and Reassurance*): I think it's a great idea! I would do it if I were you!	**Supporter:** That is exactly how I feel... I am trying to find the right person for me and I can't do it!
Seeker: Probably that's what I will do. I just can't take this anymore. If he wants he can buy drinks himself as I am tired walking to shop every day and all neighbors thinking that I am an alcoholic.	**Seeker:** I thought you are here to help me?
Supporter (*Self-disclosure*): I completely understand where you are coming from. I was in the same situation as you.	**Supporter:** It's hard, but we all have to stick together. There are so many people out there. you just have time.
Seeker: It was so nice to chat with you, I feel better now.	**Seeker:** Maybe I should ignore my friend for a while? Maybe I will feel better?
Supporter (*Affirmation and Reassurance*): You're welcome! I'm glad I could help!	**Supporter:** I try to just hang out with friends. I have been trying to meet new people in the city but they just don't show up. I just want to be alone and be alone.
	Seeker: Thank you for your help.
	Supporter: You're welcome. I think that's what I'll do too. I'm trying to get used to the city life. I feel like I've been stuck in this city for so long.

Figure 7: Case Study. **Blue text**: the help-seeker's pre-chat survey. **Red text**: strategies predicted by the model. **Green background** : utterances that corresponding to the used strategies. **Yellow background** : utterances where the model talks about

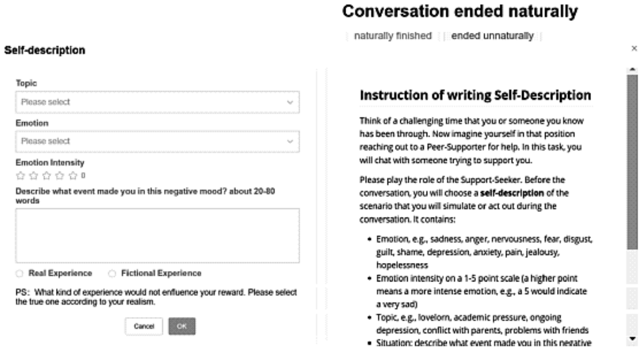
itself and is not user-centered.



(a)

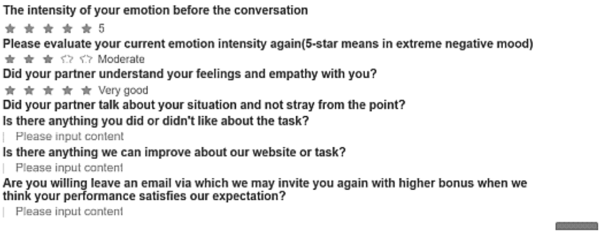


(c)


Conversation ended naturally
naturally finished
ended unnaturally
Self-description
Topic
Please select
Emotion
Please select
Emotion Intensity
0
Describe what event made you in this negative mood? about 20-80 words
Real Experience
Fictional Experience
PS: What kind of experience would not enfluence your reward. Please select the true one according to your realism.
Cancel
OK
Instruction of writing Self-Description
Think of a challenging time that you or someone you know has been through. Now imagine yourself in that position reaching out to a Peer-Supporter for help. In this task, you will chat with someone trying to support you.
Please play the role of the Support-Seeker. Before the conversation, you will choose a self-description of the scenario that you will simulate or act out during the conversation. It contains:
• Emotion, e.g., sadness, anger, nervousness, fear, disgust, guilt, shame, depression, anxiety, pain, jealousy, hopelessness
• Emotion intensity on a 1-5 point scale (a higher point means a more intense emotion, e.g., a 5 would indicate a very sad)
• Topic, e.g., lovelorn, academic pressure, ongoing depression, conflict with parents, problems with friends
• Situation: describe what event made you in this negative

(b)

(d)



(e)

Figure 8: (a) Support strategy training. (b) Pre-chat survey. (c) The help-seeker's chatting interface. (d) The

supporter's chatting interface. (e) Post-chat survey.