Self-chats from Large Language Models Make Small Emotional Support Chatbot Better

Zhonghua Zheng[1*] **Lizi Liao**[2] **Yang Deng**[3] **Libo Qin**[4] **Liqiang Nie**[1]

[1] Harbin Institute of Technology, Shenzhen [2] Singapore Management University
[3] National University of Singapore [4] Central South University

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

Abstract

Large Language Models (LLMs) have shown strong generalization abilities to excel in various tasks, including emotion support conversations. However, deploying such LLMs like GPT-3 (175B parameters) is resource-intensive and challenging at scale. In this study, we utilize LLMs as "Counseling Teacher" to enhance smaller models' emotion support response abilities, significantly reducing the necessity of scaling up model size. To this end, we first introduce an iterative expansion framework, aiming to prompt the large teacher model to curate an expansive emotion support dialogue dataset. This curated dataset, termed ExTES, encompasses a broad spectrum of scenarios and is crafted with meticulous strategies to ensure its quality and comprehensiveness. Based on this, we then devise a *Diverse Response Inpainting* (DRI) mechanism to harness the teacher model to produce multiple diverse responses by filling in the masked conversation context. This richness and variety serve as instructive examples, providing a robust foundation for fine-tuning smaller student models. Experiments across varied scenarios reveal that the teacher-student scheme with DRI notably improves the response abilities of smaller models, even outperforming the teacher model in some cases. The dataset and codes are available[1].

1 Introduction

The recent rise of Large Language Models (LLMs) has underscored their aptitude in generalization by adeptly performing tasks through mere conditioning on a scant number of in-context exemplars or straightforward task descriptions in natural language (Brown et al., 2020; Bahrini et al., 2023; Qin et al., 2023). Moreover, the exceptional ability of LLMs to assimilate and retain a broad spectrum of knowledge (Sap et al., 2020; Biswas, 2023), encompassing factual and commonsense realms, has

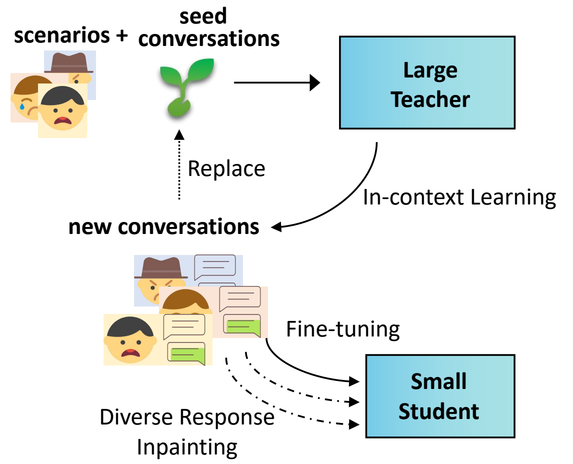


Figure 1: We use teacher-generated conversations with diverse response inpainting to better teach the student.

been notably impactful. This prowess has notably reshaped numerous arenas, including the domain of Emotional Support Conversations (ESC), enriching both dataset development and model construction.

Previous compilation of ESC datasets relied heavily on methods such as psychotherapy video transcripts (Shen et al., 2020), online repositories (Medeiros and Bosse, 2018), and questionnaires Liu et al. (2021). While these sources offer high-quality data, they come with significant costs. To this end, recent works (Zheng et al., 2023b; Liao et al., 2023) highlight how the rise of LLMs has revolutionized this space. The intrinsic generalization capabilities and vast knowledge pools of LLMs now facilitate the expansion and enrichment of ESC datasets. However, these datasets generated still lack diversity in ES scenarios and fail to provide fine-grained guidance from ES strategies.

Transitioning to the realm of ESC model (or ChatPal model) construction, the era preceding LLMs saw a reliance on predefined templates and meticulously crafted rules (van der Zwaan et al., 2012), which were beleaguered by a lack of generality. However, with the proliferation of datasets, a shift towards data-driven models has been ob-

*Work was done during an internship at SMU.
[1]https://github.com/pandazzh2020/ExTES

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11325–11345
August 11-16, 2024 ©2024 Association for Computational Linguistics

served (Cheng et al., 2022), deploying a myriad of techniques ranging from hierarchical graph networks (Peng et al., 2022) to relatively diminutive Transformer models (Tu et al., 2022) or even pre-trained language models (Sharma et al., 2021; Deng et al., 2023). Despite their advancements, a glaring deficit of these models is their inefficacy in adeptly navigating unfamiliar scenarios. Contrarily, LLMs, with their expansive knowledge and robust generality, have been utilized as sagacious experts in response generation (Zhang et al., 2023a), yielding superior performance results.

Nevertheless, a critical limitation shadowing such prompt-based ChatPal model (Zhang et al., 2023a) is its dependency on exceedingly large models, encapsulating hundreds of billions of parameters (Kojima et al., 2022; Wei et al., 2022). The deployment of these behemoths on a large scale is deterred by their exorbitant computational demands and inference costs. Hoffmann et al. (2022) shows that, for a given compute budget, the best performances are not achieved by the largest models but by smaller models trained on more data. Our endeavor is thus channeled towards empowering smaller models to generate emotional support responses, thereby making large-scale deployment a viable proposition.

In light of this, we propose to engage LLMs as "counseling teacher" to augment the emotional support response adeptness of smaller models, thereby significantly reducing the need for large model sizes. Starting with a carefully crafted set of dialogues encapsulating a variety of scenarios and fine-grained strategies, we engage a large teacher model to iteratively generate a large number of generalized and high-quality emotional support conversations. The ensuing curated dialogues are then employed to fine-tune a compact, agile student model to exhibit emotional support response proficiency. By leveraging the large model as a teacher, we unlock the potential for *Diverse Response Inpainting* (DRI), enabling the generation of multiple unique and consistent responses through filling in the masked conversation context, thereby enriching the fine-tuning dataset and encapsulating a flexible response spectrum. This maneuver significantly elevates the performance of student models without additional human annotation.

In summary, our contributions are threefold:

- We leverage LLMs as "counseling teacher" to enhance the emotional support response capabilities of smaller models, thereby alleviating the requirement for large model sizes.
- Our methodology enables *diverse responses* for each conversation context via a novel Diverse Response Inpainting approach, enriching the fine-tuning data and mirroring the flexible response spectrum inherent in ESC.
- Experiments show that our method not only contributes a high-quality and large-scale ExTES dataset, covering a wide range of emotional support scenarios and strategies but also yields a compact ChatPal that rivals the performance of much larger models.

2 Related Work

Emotional Support ChatBots. Emotional Support (ES) ChatBots in real-world have been largely hindered by the glaring lack of large-scale well-annotated datasets (Sun et al., 2021). Most existing studies in emotional support conversations prioritize dataset collection from psychotherapy video transcripts (Shen et al., 2020) or online sources (Medeiros and Bosse, 2018), such as stress-related Twitter interactions (Medeiros and Bosse, 2018), mental health reddits (Sharma et al., 2020), and online support groups (Hosseini and Caragea, 2021; Li et al., 2021b). 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. Hence, they further construct AUGESC with LLMs, an augmented dataset, which largely extends the scale and topic coverage of ESConv (Zheng et al., 2023b).

Other than datasets, there have been various ways to build ES conversation models. Early works mainly rely on predefined templates and handcrafted rules (van der Zwaan et al., 2012), which suffer from limited generality. Recent works explored data-driven models (Cheng et al., 2022), such as by leveraging hierarchical graph network (Peng et al., 2022) or relatively small Transformer models (Tu et al., 2022). More recently, researchers resort to pre-trained language models (Sharma et al., 2021; Deng et al., 2023) or LLMs (Zhang et al., 2023a). In our work, besides contributing a new dataset, we further investigate an effective way on learning from large model to finetune a smaller

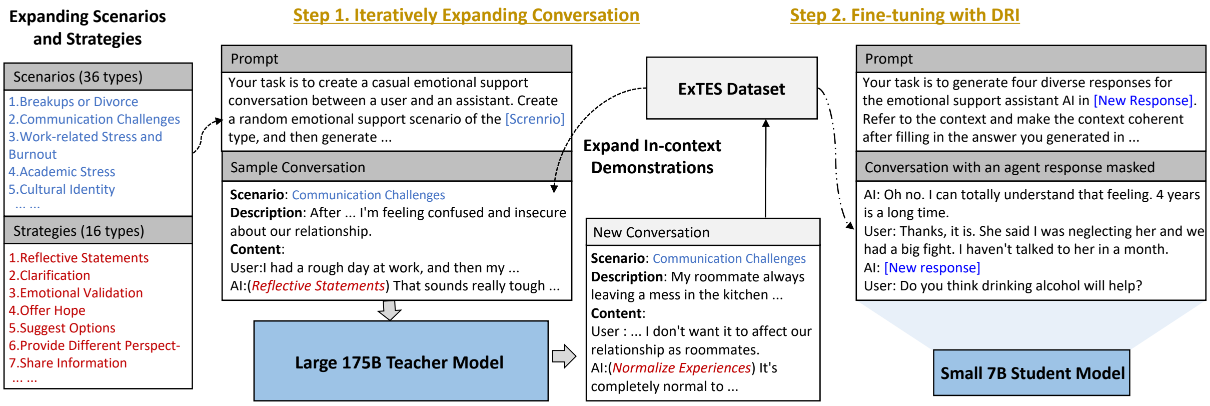


Figure 2: Detailed overview of our proposed method. Initiated with a meticulously designed set of dialogues spanning diverse scenarios with comprehensive strategies, it is followed by two steps: **Step 1**: a very large teacher model is prompted to generate emotional support conversations in an iterative expansion fashion. **Step 2**: the curated conversation samples are used to fine-tune a small, lightweight student to exhibit emotion support response capabilities. The LM-based teacher further enables **Diverse Response Inpainting (DRI)**—generating multiple distinct responses for each conversation context to enrich the fine-tuning data and capture the nature of flexible response space. This boosts the performance of student models without any additional human annotation.

ChatPal with compatible performance.

Knowledge Distillation. Knowledge distillation (KD) is a technique where a smaller "Student" model learns from a larger "Teacher" model, aiming to reduce size and latency without compromising accuracy (Gou et al., 2021; Hinton et al., 2015). KD has found extensive application across various domains (Cheng et al., 2020, 2018). Our research can be perceived as a nuanced variant of KD, aligning with efforts to enhance the performance of smaller models through leveraging LLMs. Similar endeavors have been undertaken, where LLMs have been distilled or employed for data augmentation purposes (Wang et al., 2021; Ding et al., 2022; Kang et al., 2023). A notable strand within this realm involves utilizing LLMs for generating both task labels and task-related descriptions, aimed at training smaller models on various tasks (Shridhar et al., 2022; Li et al., 2022; Ho et al., 2022; Hsieh et al., 2023). Unlike traditional setups, the teacher model in our framework is designed to generate a variety of emotional support responses via diverse response inpainting. This unique configuration aims at enriching the student model's capacity with comprehensive guidance, thereby distinguishing our method from previously established ones.

3 Teacher-Student Framework

In this section, we elucidate how the teacher-student framework functions. As illustrated in Figure 2, we curate a meticulously designed set of dialogues as our starting point with diverse scenarios and comprehensive strategies. Then, in a two-step fashion, we first iteratively expand these conversations using a large teacher model and then fine-tune a small student ChatPal with DRI.

3.1 Comprehensive Scenarios and Strategies

To create diverse emotional support conversations with broad coverage, we developed a comprehensive set of 36 emotional support scenarios (detailed in Appendix D), drawing from 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). This is a significant expansion from the five scenarios in ESConv (Liu et al., 2021), catering to diverse life situations and user emotional needs. Similarly, based on references (Hill, 1999; Organization et al., 2020), we compiled 16 emotional support strategies in Table 1. This represents a two-fold increase compared to the eight strategies in ESConv, enabling teacher models to provide more targeted suggestions and broadening the scope of emotional support.

3.2 Iterative Expansion via Teacher

Building on (Brown et al., 2020; Bahrini et al., 2023), we harness the capabilities of the ChatGPT teacher model to iteratively produce new dialogues, utilizing both complete dialogue exemplars and new scenarios enriched task descriptions.

Data collection initialization: We began with the creation of 100 seed dialogues, derived from reputable emotion support datasets such as ESConv

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)	3,887	3.9%
Overall	97,893	100%

Table 1: Statistics of response strategies used in ExTES.

(Liu et al., 2021), ETMHS (Sharma et al., 2020), and Reddit (Yeh et al., 2015). These dialogues underwent manual correction and strategic response labeling. Their quality is ensured via rigorous human evaluations, as highlighted in Appendix F.

Iterative data expansion: As depicted in Figure 2, the large teacher model uses the initial 100 seed dialogues as examplars paired with new scenarios enriched task descriptions to generate new conversations. These new dialogues, guided by our prompt template in Appendix E, both extend the dataset and serve as the next iteration's seeds. The LLM produces these dialogues while marking them with suitable emotional support strategies. With this iterative method, the initial dialogues were soon superseded by 1k dialogues from diverse scenarios, allowing for a scalable process that can easily incorporate new seeds and scenarios.

Quality assurance: Although our template specifies the desired dialogue format and criteria, inconsistencies occasionally arise, such as data format errors, duplications, omitted response strategies and non-compliance to scenarios *etc.* We prioritize data integrity; hence, we engage in human reviews and enact manual corrections. It's noteworthy that our approach requires substantially less human intervention than traditional methods like questionnaires (Liu et al., 2021) or crowd-sourcing (Budzianowski et al., 2018), with a mere 10% of the generated dialogues necessitating adjustments. Any dialogue requiring substantial modification is promptly discarded. After screening and adjustments, we consolidate approximately 11k dialogues, resulting in the ExTES dataset.

3.3 Fine-tune Small ChatPal Student

After collecting the ExTES dataset, we fine-tune small student models on generated conversations. In order to obtain a better small ChatPal model, selecting an efficient fine-tuning method is critical. Hence, we explored three fine-tuning methods: conventional DialoGPT Fine-Tuning (DialoGPT-FT), LLaMA Adapter-Tuning (7B-Adapter), and LLaMA LoRA-Tuning (7B-LoRA). Based on our preliminary results, the 7B-LoRA version performed the best (see Table 8 and Appendix I for more details). Therefore, we focus on this setting for further building our small ChatPal model.

Specifically, suppose $P_\Phi(y|x)$ is the learner of LLaMA-7B, where Φ is the set of network parameters initialized with pre-trained weights Φ_0. In conventional full fine-tuning, the model is updated to $\Phi_0 + \Delta\Phi$ by following the gradient to maximize the conditional language modeling objective:

$$\max_{\Phi} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log P_\Phi(y_t|x, y_{<t}),$$

where x is the conversation context, y is the response by supporter and $y_{<t}$ is the part decoded before step t. Z refers to the whole training set.

To overcome the challenge in large size of $\Delta\Phi$, the LoRA-Tuning adopts a parameter-efficient approach, where the task-specific parameter increment $\Delta\Phi = \Delta\Phi(\Theta)$ is further encoded by a much smaller-sized set of parameters Θ with $|\Theta| \ll |\Phi_0|$. Hence, the objective becomes optimizing over Θ:

$$\max_{\Theta} \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log P_{\Phi_0+\Delta\Phi(\Theta)}(y_t|x, y_{<t}).$$

3.4 Diverse Response Inpainting

To further enhance the student model's performance, we introduce the diverse response inpainting (DRI) mechanism. This mechanism prompts the larger teacher model to fill in the masked response position with a range of diverse responses given the same conversation context, offering a broader learning scope for the student. Specifically, DRI works by completing partial dialogues—those missing an agent's response turn—using predictions from the teacher model. Notably, in emotional support conversations, each response can be approached with a variety of strategies, leading to diverse output. Leveraging the teacher model's vast generative capacity and inherent randomness, we capitalize on this diversity. This results in richer fine-tuning guidance signals in an enlarged dataset, capturing a wide range of potential responses.

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

Table 2: The statistics of our ExTES vs. ESConv.

Specifically, a complete dialogue d is a sequence of utterances, $d = (u_1, r_1, u_2, r_2, \ldots, u_t, r_t, \ldots, u_T, r_T)$. We use the same notation for partial dialogues, denoting the unobserved utterance with the $\diamond$ symbol. For example, $(u_1, r_1, u_2, r_2, u_3, \diamond, u_4, r_4)$ is a partial dialogue where utterance r_3 is unobserved. We refer to it as "masked" response. We also use the shorthand $d_{m(r_3)}$ to denote a dialogue d with r_3 masked. To complete the partial dialogue $d_{m(r_3)}$, we generate replacement for r_3, denoted $\hat{r}_3$. The inpainted dialogue is then:

$$DRI(d_{m(r_3)}) = (u_1, r_1, u_2, r_2, u_3, \hat{r}_3, u_4, r_4).$$

An example is shown in Appendix H, we use ChatGPT to generate multiple diverse and consistent responses to capture a flexible response space.

4 Dataset Characteristics and Quality

General Statistics. Our compiled dataset, named ExTES, encompasses a total of 11,177 dialogues. Detailed breakdowns are presented in Table 2. Each dialogue averages 18.2 utterances. Notably, while user utterances tend to exhibit negative sentiments, assistant responses predominantly exude positive tones, underscoring their role in providing emotional support. An illustrative dialogue from our dataset can be found in Appendix A.

The average dialogue length in ExTES, at 18.2 utterances, emphasizes the iterative exchanges often needed for effective emotional support. This length surpasses that of earlier datasets on emotional chatting (Zhou and Wang, 2018) and empathetic dialogue (Rashkin et al., 2019). While our dialogues are shorter than ESConv's, they exhibit a denser average utterance length (26.0 words), indicating richer content. Further annotation specifics are in Table 1 and Table 10. Dominant emotional challenges are rooted in communication issues and work stresses, possibly heightened by recent global economic trends.

Dialogue Quality Evaluation. The fine-tuning data's quality is paramount for optimizing our smaller model's performance. To ensure the excellence of the ExTES dataset, we conducted a thorough human evaluation and benchmarked it against ESConv, a crowdsourced dataset. Our evaluation framework, inspired by (Li et al., 2021a; Zheng et al., 2023b), comes with a set of guidelines provided in Appendix L. Our evaluation focuses on the following key metrics: **Informativeness** measures how well the individual seeking support articulates their emotional challenges. **Understanding** gauges the supporter's grasp of the individual's experiences and emotions. **Helpfulness** evaluates the effectiveness of the supporter's efforts in mitigating the individual's emotional distress. **Consistency** ensures participants consistently adhere to their roles and exhibit non-contradictory behavior. **Coherence** checks if conversations have seamless topic transitions. All metrics employ a four-level Likert scale (Allen and Seaman, 2007), ranging from 0 to 3, where a higher score indicates superior quality. For this evaluation, we engaged five master's students as annotators, assessing 50 randomly selected dialogues from both ExTES and ESConv for a comprehensive comparison.

	ESConv	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 3: Human evaluation of ExTES quality (scores from 0 to 3). κ denotes Fleiss' Kappa (Fleiss, 1971), indicating fair to moderate inter-annotator agreement ($0.2 < \kappa < 0.6$).

As shown in Table 3, it demonstrates that the large teacher model can generate high-quality emotional support dialogues with proper demonstrations and ES scenario guidance. Dialogues collected by our method show similar evaluation scores compared to crowdsourced ESConv. It is even better than crowdsourced dialogues in terms of Informativeness and Helpfulness. According to our observation, this might be because the answers generated by large teacher model tend to have more substantial and complete content.

Strategy Distribution. In this analysis, we aim to show whether the large teacher model annotated response strategies show reasonable patterns across different stages of a conversation. To do this, we considered a conversation with N responses in total, where the k-th response r_k adopts the strategy S.

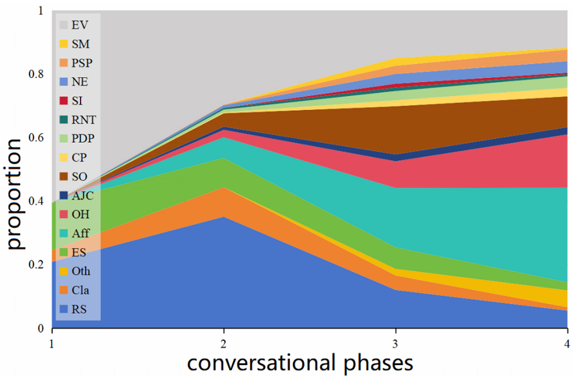


Figure 3: Distribution of strategies at different phases.

The position of it 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. To gain insight into strategy distribution across these phases, we scrutinized every dialogue in our dataset, cataloging the frequencies of strategies within each phase. The gathered data offers a snapshot of how strategies are employed throughout the progression of a conversation. As depicted in Figure 3, distinct but reasonable trends emerge regarding the utilization of ES strategies over the conversation's course. For instance, *Emotional Validation* is predominantly used in the initial phases to convey understanding to the help-seeker, while in the concluding stages, *Affirmation* is favored to offer encouragement.

Toxicity Assessment To assess potential toxicity in our ExTES dataset, we employed the Perspective API[2], a widely recognized tool for toxicity detection (Zheng et al., 2023a). This API evaluates utterances for toxicity based on six distinct attributes. Table 4 reveals that our dataset demonstrates minimal toxicity, even lower than the manually curated ESConv dataset. We consider the level of toxicity to be normal. Actually, further reductions in toxicity scores may affect the quality of emotional support conversations. Because users seeking emotional support might express some hateful or aggressive content, which will increase toxicity levels. Significantly, the Severe Toxicity score, which tracks intensely hateful or aggressive comments, stands at a mere 0.0016, likely reflecting the safety features of ChatGPT. Moreover, the ChatPal model, fine-tuned using ExTES, shows further reduced toxicity levels, especially in categories like Toxicity, Severe Toxicity, Insult, and Profanity. This trend aligns with our goal of creating a bot to interact in a compassionate and respectful manner.

[2] https://perspectiveapi.com/

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

Table 4: Results of toxicity assessment using Perspective API. Lower scores are better. ChatPal Responses are generated by LoRA finetuning on ExTES dataset.

5 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) How effective is our small ChatPal for providing emotional support? (Q2) How is the effect of using large teacher model to capture comprehensive scenarios and strategies? (Q3) What is the effect of diverse response inpainting?

5.1 Baselines

We will compare our model with the following baselines (detailed in Appendix J):
LLaMA (Touvron et al., 2023). It is the vanilla open and efficient large-scale language model.
ChatGPT (Ouyang et al., 2022). ChatGPT is known for its language understanding and text generation capabilities.
Ask-Expert (Zhang et al., 2023a). Ask-Expert is a framework for emotional support with structured expert conversations.
AUGESC (Zheng et al., 2023b). AUGESC augments dialogues and utilizes the ExTES dataset to fine-tune GPT-J model. We also fine-tune AUGESC with DRI to demonstrate DRI's robustness, which is denoted as **AUGESC+DRI**.
ChatPal / DRI. A variant fine-tuned on ExTES dataset without diverse response inpainting DRI.

5.2 Evaluation Metrics

The automated evaluation metrics we used comprised of METEOR (Banerjee and Lavie, 2005), BLEU-4 (B-4), ROUGE-L (R-L) (Lin, 2004), Vector Extrema (Forgues et al., 2014) and the Distinct-2/3 (Li et al., 2016). The responses were tokenized using the NLTK (Loper and Bird, 2002). For human evaluation, we use similar metrics as introduced in Section 4 but focus on evaluating the generated responses. We use Informativeness **(Inf.)** of the supporter responses, Understanding **(Und.)**,

Methods	METEOR	B-4	R-L	Extrema	D-2	D-3
ChatGPT	21.86	2.048	13.76	60.76	**75.88**	**95.29**
Ask-Expert	29.85	2.126	17.10	60.33	72.18	94.50
LLaMA	16.27	1.175	9.834	50.86	29.21	50.56
AUGESC	29.62	2.390	21.89	60.38	64.23	84.21
AUGESC+DRI	32.98	2.315	**22.62**	62.24	69.75	93.54
ChatPal / DRI	30.67	**2.491**	20.85	63.73	61.94	82.80
ChatPal	**33.12**	2.437	21.09	**65.44**	66.93	90.71

Table 5: Results of automatic evaluation demonstrate the advantages of our teacher-student framework.

Methods	Inf.	Und.	Hel.	Con.	Coh.	Ove.
ChatGPT	2.47	2.07	2.34	**2.41**	**2.55**	2.40
Ask-Expert	2.15	1.34	1.78	1.94	1.84	1.84
LLaMA	1.59	1.21	1.68	1.44	1.58	1.71
AUGESC	2.16	1.83	2.09	1.85	2.35	2.23
AUGESC+DRI	2.32	2.20	2.46	2.12	2.40	2.43
ChatPal / DRI	2.31	2.04	2.19	2.36	2.37	2.33
ChatPal	**2.49**	**2.31**	**2.51**	2.39	2.41	**2.48**
κ	0.42	0.33	0.37	0.35	0.40	0.41

Table 6: Human evaluation results. The scores (from 0 to 3) are averaged over all the samples rated by three annotators. κ denotes Fleiss' Kappa (Fleiss, 1971), indicating fair or moderate inter-annotator agreement ($0.2 < \kappa < 0.6$).

Helpfulness **(Hel.)**, Consistency **(Con.)**, Coherence **(Coh.)**, and a new **Overall (Ove.)** which evaluates how good the emotion support model is in general.

5.3 Overall Evaluation (Q1)

5.3.1 Automatic Evaluation Results

To demonstrate the effectiveness of our teacher-student framework, we compare our ChatPal with other methods and report results in Table 5. See Appendix F for other supplementary experiments.

Firstly, regarding the content-based metrics (*incl*, METEOR, B-4, R-L, and Extrema), it is evident that our ChatPal consistently outperforms other baselines. Among them, ChatGPT exhibits a significant superiority over LLaMA. Ask-Expert further improves the performance by excelling in offering more specific advice than the vanilla ChatGPT. Built upon a small language model, AUGESC can achieve competitive performance as Ask-Expert, indicating the advantages of distilling the knowledge from large models. Overall, our method integrates a broader range of emotional support strategies and scenarios that are distilled from the large teacher, allowing for a more generalizable ChatPal model.

Secondly, when assessing diversity-based metrics (namely, *incl*, D-2, and D-3), it's evident that methods rooted in ChatGPT naturally generate responses that are both lengthier and richer in content compared to others. The Ask-Expert method, with its fixed guiding prompts, somewhat restricts ChatGPT's response diversity. Yet, extreme diversity isn't always advantageous. By tailoring ChatGPT to specific emotional support scenarios, our student model not only elicits a range of responses for its own education but also strikes a balance in diversity. This makes it more diverse than the original LLaMA and more measured than Ask-Expert. Overall, our teacher-student framework delivers dual benefits: it produces a sizable, high-quality ESC dataset and refines a smaller ChatPal that rivals the performance of its larger counterparts.

5.3.2 Human Evaluation Results

To complement automatic evaluation (Ye et al., 2022a,b; Wan et al., 2022), we further conduct human evaluation on the generated responses with five annotators. We randomly sample 50 conversations from ExTES's test data for comparison. The annotators were asked to rate the performance of different models. The outcomes of comparison (as shown in Table 6) demonstrate the following findings. (1) It reveals that our final ChatPal (student model) trained on our ExTES dataset achieves better performances than the vanilla ChatGPT (teacher model) on most metrics. It also confirms the high quality and practicality of our ExTES dataset in enhancing emotional support capabilities. (2) We find that Ask-Expert, due to its reliance on fixed formats, is only suitable for providing specific actionable advice and cannot offer comprehensive emotional support, hence it received lower scores. On the other hand, AUGESC may provide unhelpful responses to unfamiliar questions, resulting in lower scores on the Understanding and Helpfulness metrics. This is potentially due to LLaMA's larger generation space and better comprehension compared to GPT-J. Based on our expanded wide-ranging scenarios and comprehensive strategies, our ChatPal outperforms other models in almost all metrics. In general, the results show the effectiveness of our teacher-student framework, enhancing the ability of smaller models to provide emotional support.

5.4 Advantages of ExTES Dataset (Q2)

5.4.1 Performance on New Scenarios

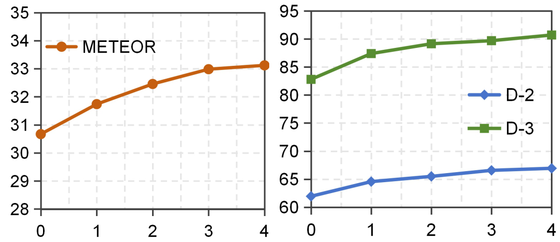
To affirm the efficacy of our expanded scenarios, we only choose new scenarios from the ExTES dataset (31 out of 36 scenarios that are different

Methods	METEOR	B-4	R-L	Extrema	D-2	D-3
ChatGPT	22.29	2.114	12.52	60.56	**74.96**	**94.13**
Ask-Expert	24.61	2.190	17.13	59.85	72.10	93.38
LLaMA	14.46	1.256	10.24	50.11	27.76	48.04
AUGESC+DRI	31.96	2.289	20.57	28.09	66.51	91.07
ChatPal	**32.56**	**2.425**	**20.98**	**61.63**	68.07	92.25

Table 7: Automatic evaluation results in new scenarios. It reveals that our student model outperforms other methods on most metrics.

Method	Stra?	METEOR	B-4	R-L	Extrema	D-2	D-3
DialoGPT-FT	✘	26.03	1.721	13.37	53.27	49.29	62.92
	✔	26.82	1.966	13.23	55.71	53.11	77.47
7B-Adapter	✘	28.48	1.944	16.95	64.47	60.43	82.62
	✔	29.71	1.987	16.39	62.73	60.83	82.24
7B-LoRA (ChatPal / DRI)	✘	30.31	2.333	19.60	65.06	63.64	84.90
	✔	30.67	**2.491**	20.85	**65.44**	61.94	82.80
ChatPal	✘	31.05	2.402	20.94	64.51	**69.88**	**91.96**
	✔	**33.12**	2.437	**21.09**	63.73	66.93	90.71

Table 8: Comparison of fine-tuning methods. We compare the no-strategy (✘) and with-strategy (✔) variants.



(a) METEOR (b) D-2/3

Figure 4: The impact of the number of diverse responses k, ranging from k=0 (w/o DRI) to 4.

from ESConv) for testing. The automatic and human evaluation of various methods in new scenarios are shown in Table 7 and Table 13 (Appendix F.3). For large language models, ChatGPT and Ask-Expert are less sensitive to varying scenarios, thanks to ChatGPT's generation capabilities while Ask-Expert further instructs ChatGPT to respond by using tailored prompts. On the other hand, vanilla LLaMA and AUGESC+DRI struggle to provide specific advice in unseen scenarios, due to limited generation capabilities of relatively small models. While AUGESC+DRI achieves a similar performance to ChatPal, which illustrates the generality of our approach in new scenarios. Our approach ChatPal, which involves venturing into new scenarios and fine-tuning from high-quality datasets, equips it to address a wide range of user emotional issues with greater empathy and provide more detailed guidance.

5.4.2 Effect of Strategy Guidance

To show the effect of fine-grained strategies in ExTES for helping fintuning, we conduct an ablation study on all three fine-tuning schemes. Results are presented in Table 8. We observe that the variants with strategies are generally better than those without strategies in all schemes, except for their performance on D-2/3 metrics. This is because, under the guidance of specific strategies, the response generation space becomes more constrained, reducing the diversity of responses in certain extent. Therefore, we refer our final ChatPal model as the version trained with strategy annotation and enhanced with DRI.

5.5 Diverse Response Inpainting Effect (Q3)

Table 5 and 6 also show the comparison of performance between our student model and its variant w/o DRI. Additionally, Figure 4 demonstrates the impact of generating varying numbers of diverse responses during DRI for later finetuning. Compared to the variants w/o DRI, the student models exhibit a significant performance improvement. But our ChatPal and AUGESC+DRI scores lower on B-4 metric than the variants w/o DRI. This is understandable, under the support of diverse responses, the student model can provide a wider range of emotional support replies. Additionally, generating diverse responses further expands the data scale based on our ExTES dataset, which effectively enhances the quantity of high-quality data. Overall, leveraging the teacher model to generate diverse responses, the performance of small student model can further elevate the performance and help building a more powerful and versatile emotional support chatbot.

6 Conclusion

In this paper, we proposed a teacher-student framework and demonstrated the potential of LLMs as "counseling teacher" in enhancing the emotional support response-abilities of smaller models. By leveraging the in-context generalization and extensive knowledge reservoirs of LLMs, we curated a large-scale emotional support conversation dataset (ExTES) and deliberately fine-tuned smaller models with diverse response inpainting mechanism to exhibit proficiency in providing emotional support. Extensive experiments validate the advantages of the ExTES dataset as well as the superiority of the proposed teacher-student framework.

Limitations

Our proposed approach relies heavily on LLMs and is subject to the same limitations, namely, known biases in the training data and the ability to hallucinate incorrect information. Since our student model (ChatPal) is trained on conversations generated by LLM, it is possible that such characteristics of the teacher model can get passed along to the student. Additionally, it is known that for different cultures, the emotional support strategies can be very diverse which requires cultural background knowledge and reasoning processes (Gibson et al., 2016). And our fine-tuning data is only available in English and cannot provide support for other languages at this moment.

On the other hand, our method currently cannot run on small devices such as mobile phones, but we're concentrating on utilizing LLMs (ChatGPT 175B) as a "counseling teacher" to enhance the emotional support capacities of smaller models (LLaMA 7B). Based on our approach, running the fine-tuned student model (LLaMA-7B) on machines with normal computational power (such as RTX 3090) is also a form of progress. Furthermore, we believe that this will be addressed in the future with advancements in model compression and optimization techniques.

Ethical Considerations

Working in the field of emotional support requires additional ethical considerations. Regarding safety, we acknowledge the limitations of the current framework proposed and the potential risks associated with deploying them directly for emotionally vulnerable individuals. We do not recommend the direct deployment of the fine-tuned models from this work into real-life situations; currently, they are only suitable for academic research. While we intend to develop models for the greater good of society, it is crucial to recognize that the dataset contains potentially problematic content, including toxic or biased material that could be used to generate negative or offensive content. We openly provide the dataset collected for this work to assist in supporting future improvements in ESC.

On the other hand, our proposed system relies heavily on large language models and therefore inherits their well-known problems centered around societal biases learned through pretraining, hallucinations, and expensive use of resources (Weidinger et al., 2021). Various controls are included to constrain the LLMs to the emotional support task, but these are unlikely to fully wash away their inherent issues. Significant further progress needs to be made in areas like debiasing, grounding in actuality, and efficient serving before we can safely deploy this type of system in a production setting.

Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SIS-SMU-010).

References

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.

I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. Quality progress, 40(7):64–65.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. ChatGPT: Applications, Opportunities, and Threats. arXiv preprint arXiv:2304.09103.

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.

Som S Biswas. 2023. Role of chat gpt in public health. Annals of biomedical engineering, 51(5):868–869.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

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.

Jian Cheng, Peisong Wang, Gang Li, Qinghao Hu, and Hanqing Lu. 2018. Recent advances in efficient computation of deep convolutional neural networks. Frontiers of Information Technology & Electronic Engineering, 19:64–77.

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.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. 2020. A survey of model compression and acceleration for deep neural networks.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. 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.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? ArXiv, abs/2212.10450.

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.

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.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. arXiv preprint arXiv:2212.10071.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

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.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. ArXiv, abs/2305.02301.

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.

Junmo Kang, Wei Xu, and Alan Ritter. 2023. Distill or annotate? cost-efficient fine-tuning of compact models. ArXiv, abs/2305.01645.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Chunyuan Li, Heerad Farkhoor, 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.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. ArXiv, abs/2210.06726.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

Xinmeng Li, Wansen Wu, Long Qin, and Quanjun Yin. 2021a. How to evaluate your dialogue models: A review of approaches.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021b. 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.

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.

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. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit.

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.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5925–5941.

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. Association for Computational Linguistics.

Maija Reblin and Bert N Uchino. 2008. Social and emotional support and its implication for health. Current opinion in psychiatry, 21(2):201.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 27–33.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In Proceedings of the Web Conference 2021, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. 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. Association for Computational Linguistics.

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.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling reasoning capabilities into smaller language models. ArXiv.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

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.

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.

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.

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022. A unified dialogue user simulator for few-shot data augmentation. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3788–3799.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

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.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023a. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. 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. 2023a. Click: Controllable text generation with sequence likelihood contrastive learning.

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

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. Association for Computational Linguistics.

A Data Example from our dataset

An detailed example can be seen in Figure 5. Each conversation is labeled with 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.

Scene: Communication Challenges

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.

User : Hey.
AI : Hi, how are you?
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.
AI (*Reflective Statements*): That sounds really irritating. It must be tough to constantly clean up after someone else.
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.
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.
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.
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.
User : You're right. I'll try to approach it from a calm and understanding perspective. Hopefully, we can come to a resolution.
AI (*Affirmation*): I believe in your ability to handle this situation maturely. Remember to take care of yourself throughout this process.
User : Thank you for the support. It means a lot to me. I'll do my best to handle it positively.
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.
User : I really appreciate your positivity and encouragement. It definitely gives me hope. Thank you again for being there.
AI : You're welcome! I'm here whenever you need to talk. Take care and hang in there!
User : Thank you. I'll keep that in mind. Take care too. Goodbye!

Figure 5: Data example from our ExTES dataset.

B Definitions of Strategies

Below are 16 emotional support strategies and examples that we have compiled. In general, emotional support strategies help in two aspects: (1) During dataset collection, the role of the emotional support strategy primarily assists the Teacher model (ChatGPT) in generating more targeted and plausible conversations. (2) During model training and inference, the strategy and corresponding response are predicted sequentially. The stragegy helps the generation model to narrow down their generation space. During inference, the strategy are also predicted.

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 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.

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

C Strategy Transition

We present the top-5 most frequent strategy transitions with 3-5 hops in Table 9. These transitions indicate that supporters usually ask questions and explore the user’s situation before comforting the user. Emotional support supporters 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.

D Details of Scenarios

Below are 36 emotional support scenarios and examples that we have compiled. And Table 10 is the statistics of all ES 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

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

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, and 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 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: Navigating 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 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 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.

E Template of Expanding Conversation

The template for ChatGPT to iteratively expand conversations (Figure 2) 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. ... ... 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 ...*

F Other Experiments

F.1 The quality of Seed Dialogues

Table 11 shows the results of human evaluation on seed dialogues and ExTES.

	Seeds	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
Overall	2.38	**2.45**	0.52

Table 11: Human evaluation of seed dialogues quality and 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 ($0.2 < \kappa < 0.6$).

F.2 Experiments Across Datasets

Assessing the generalizability of knowledge from the synthesized dataset to a human-annotated dataset is crucial. We conducted experiments across datasets (as shown in the table 12). We finetune ChatPal separately using ESConv and ExTES. The resulting models are then tested on the Test Set of both datasets for automatic evaluation. We find that 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. In addition, 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 appear in ExTES test set.

Test Set	Train Set	METEOR	B-4	R-L	Extrema	D-2	D-3
ESConv	ESConv	24.23	1.670	17.19	**58.57**	44.09	60.78
	ExTES	**27.07**	**2.312**	**20.57**	55.56	**63.83**	**83.93**
ExTES	ESConv	24.08	1.687	16.70	53.41	46.83	65.94
	ExTES	**33.12**	**2.437**	**21.09**	**63.73**	**66.93**	**90.71**

Table 12: Experiments across datasets. We finetune ChatPal separately using ESConv and ExTES. The resulting models are then tested on the test set of both datasets.

F.3 Human Evaluation in New Scenarios

Table 13 shows the results of human evaluation in new scenarios

Methods	Inf.	Und.	Hel.	Con.	Coh.	Ove.
ChatGPT	**2.41**	2.04	2.36	2.42	2.37	2.39
Ask-Expert	1.80	1.65	1.79	1.52	1.89	1.93
LLaMA	1.24	1.22	1.14	1.86	1.65	1.55
AUGESC+DRI	1.68	1.74	1.72	2.03	1.82	1.92
ChatPal	2.37	**2.38**	**2.42**	**2.47**	**2.39**	**2.46**
κ	0.45	0.31	0.35	0.33	0.47	0.42

Table 13: The human evaluation in new scenarios (scores from 0 to 3). The Fleiss' Kappa is a fair or moderate inter-annotator agreement ($0.2 < \kappa < 0.6$).

F.4 Comparative Analysis: DRI vs. Data Scaling

we conduct additional experiments to compare the performance of our DRI technique by simply increasing the data scale by incorporating new examples (shown in Table 14). We compare the performance of the student model using different proportions of the training dataset (10%, 50%, and 100%) and DRI.

Compared to the version without DRI, we find that the student model shows significant performance improvement, especially in D-2/3 scores. This indicates that DRI focuses on generating diverse results, making the responses produced by the student model more varied. Additionally, when using 10% of the training data with DRI, the performance is slightly better than the version using only 50% of the data. This is partly because DRI expands the data scale, bringing the 10% data volume to a level comparable to 50%, and on the other hand, it reduces the similarity between the data while expanding the data scale, effectively increasing the quantity of high-quality data. Therefore, we believe that the gains brought by DRI are different from simply increasing the data scale. In situations where the data scale is equivalent, leveraging the teacher model to generate diversified responses can further enhance the performance of the student model, contributing to the development of a more robust and universally applicable emotion-supportive chatbot.

Proportions	METEOR	B-4	R-L	Extrema	D-2	D-3
10%	22.65	1.662	16.89	51.87	42.08	59.61
10%+DRI	28.32	2.012	18.57	55.36	51.71	73.44
50%	27.39	2.170	17.38	55.21	48.25	69.45
50%+DRI	31.49	2.205	20.15	64.04	63.33	85.51
100%	30.67	**2.491**	20.85	63.73	61.94	82.80
100%+DRI	**33.12**	2.437	**21.09**	**65.44**	**66.93**	**90.71**

Table 14: The automatic evaluation of comparing the DRI by simply increasing the data scale.

G Details of Data Collection

As shown in Appendix E our template specifies the desired dialogue format and criteria, but inconsistencies occasionally arise, such as data format errors, duplications, omitted response strategies and non-compliance to scenarios *etc.* We employed manual inspection and corrections to ensure compliance within the dataset's 11,000 dialogues. Five master's students conducted the manual review, dedicating approximately 8 days (each screening around 350 dialogues per day). Despite the manual review process, our method requires significantly less human intervention compared to traditional methods like questionnaires (Liu et al., 2021) or crowd-sourcing (Budzianowski et al., 2018), with less than 10% of the generated dialogues necessitating adjustments. Any dialogues requiring extensive modifications were promptly discarded. After screening and adjustments, we consolidate approximately 11k dialogues, resulting in the ExTES dataset. This process costs about $210 to call OpenAI's API[3] and use the gpt-3.5-turbo-0613 version. And we clarify that stating ChatGPT's parameter count as 175B is a rough estimate, as the exact details have not been officially disclosed to avoid potential misinformation.

[3]https://platform.openai.com/docs/api-reference

H Diverse Response Inpainting Example

Figure 6 shows the process of diverse response inpainting. This method further improves the student model without any additional manual annotation. Specifically, given the same dialogue context, ChatGPT model will infill extra four responses into the same turn (K=4). When fine-tuning the LLaMA model, the format of the fine-tuning data comprises input-output pairs (for instance, breaking down a conversation into 10 input-output pairs). Therefore, utilizing DRI (K=4) allows for a fivefold expansion of a single round of dialogue response (one input-output pair), generating four new input-output pairs. This significantly enriches the fine-tuning data and provide more guidance to the student model under same situation.

I Fine-tune Methods

I.1 Fine-tune Methods

We explore the following three methods to fine-tune our ChatPal (student model):

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-2[4] to model a multiturn dialogue session as a long text and frame the generation task as language modeling.

LLaMA Adapter-Tuning LLaMA-Adapter (Zhang et al., 2023b) is a form of prefix-tuning that prepends a learnable adaption-prompt to the inputs of the attention blocks in LLaMA. 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 by prefix tuning (Li and Liang, 2021) and the original adapter method (Houlsby et al., 2019), Adapter-Tuning introduces some new sublayers (i.e., adapter layers) acting as low-rank bottlenecks within each Transformer layer. Generally, instead of tuning all parameters, Adapter-Tuning focuses

[4]https://huggingface.co/docs/transformers/model_doc/gpt2

Below is an emotional support dialog. Your task is to generate four diverse responses for the emotional support assistant AI in [New Response]. Refer to the context and make the context coherent after filling in the answer you generated in [New Response].

User: Hi I am feeling blue, things aren't going well at all
AI: Hello. I am sorry to hear that. Can I ask you more about it?
User: My girlfriend and I split after 4 years!
AI: Oh no! I am sorry. How do you feel? Sad? Joyful?
User: Very sad and lonely. I loved her. What should I do?
AI: Oh no. I can totally understand that feeling. 4 years is a long time.
User: Thanks, it is. She said I was neglecting her and we had a big fight. I haven't talked to her in a month.
AI: **[New response]**
User: Do you think drinking alcohol will help?
AI: Ohhhh. Well, I never ditch my partner, always my partner ditches me, so I do not know why they do that. I think it will worsen the situation.
User: You had a similar experience? "I want to drink till I forget about her.
AI: Yes, I quit alcohol 3 years ago. It took me 20 years.
User: How did you do it?
AI: It took me 1 whole year to forget about my ex. My ex told me I am a party girl because I drink too much, so I stopped. I was trying to get him to like me more. Since then I have not been drinking.
User: Did you find someone else?
AI: Nope. I am cool without a man.
User: That's the way to be. But I am too lonely. I have a secret account so I can still see her on Facebook. Do you think that's a good idea?
AI: Haha, smart, yes. No need to force yourself to forget about her. I heard men cannot do without a woman but women don't really need a man.
User: Yeah, that way maybe I can get her back at the right time. I feel a little better - thanks.

Diverse responses:
1.Focus on taking care of yourself physically and emotionally. Make sure you're eating well, getting enough sleep, and engaging in activities that bring you joy.
2.Share your feelings with people you trust. They can provide emotional support and be there for you during this tough time.
3.Take some time to think about what you've learned from this relationship and how you've grown. It can be a valuable opportunity for personal growth.
4.If you find it challenging to cope with your emotions, consider talking to a therapist or counselor who can provide guidance and support.

Figure 6: An example of generation diverse responses. The DRI task description and the conversation context are given in ChatGPT to generate multiple diverse responses. The square below is the four different responses generated in [New Response].

on tuning mainly the adapter layers.
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.

I.2 Experimental Setup

We select LLaMA LoRA-Tuning to build our small ChatPal model. During the fine-tuning phase, we set the maximum input sequence length to 256 and the rank K in LoRA to 8. We use LLaMA-7B model and initialize the checkpoints with the 8-bit integer format (int8) parameters released by Touvron et al. (2023). These parameters remain fixed throughout training, reducing GPU memory consumption and improving training speed. We use the Adam optimizer to update LoRA parameters with a batch size of 128 and learning rates of 3e-4, respectively. The trainable LoRA parameters are fine-tuned on NVIDIA A100-40GB GPUs, and the training duration is approximately 15 hours. Finally, we yield a small ChatPal with compatible performance to much larger models, thereby significantly alleviating the requirement for large model sizes.

J Baselines

We will compare our model with five different baselines:
LLaMA (Touvron et al., 2023). LLaMA is an open and efficient large-scale base language model that sources publicly available datasets. This model is trained on a large amount of unlabeled data, making it well suited for fine-tuning a variety of tasks, and can be run on a single V100 GPU[5].

[5]We chose the LLaMA-7B version based on the needs of the emotional support task.

ChatGPT (Ouyang et al., 2022). ChatGPT is a model for processing sequential data with amazing language understanding and text generation capabilities, and in particular, it trains the model by connecting it to a large corpus of real-world conversations. ChatGPT can be used for a wide range of domains, including emotional support tasks.
Ask-Expert (Zhang et al., 2023a). Ask-Expert is a framework in emotional support domain, where the structure of expert conversation is outlined by prespecified prompts which reflect a reasoning strategy taught to practitioners in the field. Blenderbot model (Shuster et al., 2022) utilizing "Ask-Expert" shows quality improvements across all expert sizes.
AUGESC (Zheng et al., 2023b). Zheng et al. (2023b) prompt a fine-tuned LLM to complete full dialogues from available dialogue posts of various topics, which are then postprocessed based on heuristics. They proposed AugESC dataset and then fine-tuned GPT-J model, which is superior to strong baselines of dialogue augmentation.
Our Chatpal w/o DRI We only fine-tune LLaMA on our ExTES dataset w/o diverse response inpainting, which is an original variant of our small ChatPal and can help us understand the influence of diverse responses in Section 5.5.

K Why Synthesized Dataset Is Essential

In this section, we discuss why the synthesized dataset is essential for the emotional support conversational task. Firstly, previous compilation of ESC datasets relied heavily on methods such as psychotherapy video transcripts (Shen et al., 2020), online repositories (Medeiros and Bosse, 2018), and questionnaires Liu et al. (2021). While these sources offer high-quality data, they come with significant costs. As language models advanced, conventional data collection methods became insufficient to meet the demands of training models. Secondly, the intrinsic generalization capabilities and vast knowledge pools of LLMs now facilitate the expansion and enrichment of ESC datasets. Based on this, we further address the problem that existing small-scale datasets still lack diversity in ES scenarios and cannot provide fine-grained emotional support strategy guidance. 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. Lastly, we investigate different fine-tuning strategies to endow LLaMA with effective and flexible emotional support capabilities. 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.

L Guideline of Human Evaluation

We present the guideline of human evaluation in Figure 7. Before showing them the final evaluation materials, we first train our human evaluators by providing them this form, together with detailed instructions on how to carefully do the evaluations, what these metrics and corresponding scores mean *etc*.

Guideline of Human Evaluation	
You need to score the conversation between the help seeker (User) and the emotional support assistant (AI). Read the definitions and examples of evaluation metrics below to rate the results generated by different models. These examples illustrate how each metric can be applied to evaluate an emotional support conversation.	
Scores	3 (Excellent) , 2 (Good) , 1 (Accepted) , 0 (Unsatisfactory)
(1) Informativeness	
Definition	Informativeness measures how well the individual seeking support articulates their emotional challenges.
Examples	1. Low Informativeness: "I'm feeling really bad today." 2. High Informativeness: "I've been feeling overwhelmed because of work. I have tight deadlines, and my boss has been giving me extra tasks. I don't have much time for myself, and it's really stressing me out."
(2) Understanding	
Definition	Understanding gauges the supporter's grasp of the individual's experiences and emotions.
Examples	1. Low Understanding: "That sucks." 2. High Understanding: "I can imagine how stressful it must be to have such a heavy workload and demanding boss. It sounds like you're going through a tough time right now."
(3) Helpfulness	
Definition	Helpfulness evaluates the effectiveness of the supporter's efforts in mitigating the individual's emotional distress.
Examples	1. Low Helpfulness: "I'm sorry to hear that. I hope you feel better soon." 2. High Helpfulness: "It sounds like you could use some time management strategies to handle your workload more effectively. Have you considered talking to your boss about your workload or seeking support from colleagues?"
(4) Consistency	
Definition	Consistency ensures participants consistently adhere to their roles and exhibit non-contradictory behavior.
Examples	1. Inconsistent Behavior: Initially providing empathetic responses and later becoming dismissive or indifferent about the person's feelings. 2. Consistent Behavior: Maintaining a supportive and empathetic tone throughout the conversation, showing genuine care and concern.
(5) Coherence	
Definition	Coherence checks if conversations have seamless topic transitions.
Examples	1. Low Coherence: Frequent topic changes without exploring any of them in depth. For example, discussing work stress, then suddenly switching to talking about hobbies without any connection. 2. High Coherence: A focused conversation that explores a specific issue thoroughly before transitioning to a related topic. For instance, discussing work stress and then gradually shifting the conversation to coping mechanisms or self-care strategies.

Figure 7: Guideline of human evaluation for dialogue quality.