ESR-Coach: Leveraging Large Language Models for Training People to Provide Emotionally Supportive Responses

Gongyao Jiang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
gjiang024@connect.hkust-gz.edu.cn

Junze Li
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
jlijj@connect.ust.hk

Xiaojuan Ma
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
mxj@cse.ust.hk

Qiong Luo
The Hong Kong University of Science and Technology
Hong Kong, Hong Kong
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
luo@cse.ust.hk

Abstract

Effectively providing emotional support is a critical yet intricate interpersonal skill. Supporters often lack accessible and practical training opportunities to develop this competency. To address this gap, we introduce ESR-Coach, a Large Language Model (LLM)-based coaching system designed to train individuals in emotionally supportive communication. ESR-Coach leverages multiple AI agents to generate practice scenarios, demonstrate reference responses, and provide assessments on user practice replies. We evaluate the proficiency of our system on these three tasks, demonstrating high-fidelity case generation, helpful exemplary responses, and valid response assessments. In our user study (N=20), ESR-Coach helped participants achieve an average improvement of 17% in response helpfulness. After training, participants also employed more diverse and effective strategies. We further discuss the social intelligence of LLMs and their potential to foster humans' interpersonal skills in real-world scenarios.

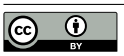
CCS Concepts

• **Human-centered computing** → **User studies**; • **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**; • **Computing methodologies** → *Natural language generation.*

Keywords

Emotional Support, AI Coaching for Humans, Communication Skills Coaching, Human-LLM Interaction

†Corresponding Author.



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789094

ACM Reference Format:
Gongyao Jiang, Junze Li, Xiaojuan Ma, and Qiong Luo. 2026. ESR-Coach: Leveraging Large Language Models for Training People to Provide Emotionally Supportive Responses. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3742413.3789094

1 Introduction

Emotional support is the communicative process that helps alleviate individuals' emotional distress and assists them in working through life challenges [12, 42, 46]. This form of support from social networks like friends, family, and colleagues plays a crucial role in individual mental health and well-being [30]. In everyday social interactions, people naturally turn to their personal networks for comfort and guidance during difficult times [11, 39, 64]. These daily communications serve as a first-line setting where individuals share emotions and provide verbal encouragement. However, despite the importance of these interactions, most people in supportive roles lack training in how to provide effective emotional support [20, 54, 82].

Traditionally, developing proficient emotional support skills requires specialized training typically available to professional counselors and mental health practitioners [48, 57]. This training demands substantial resources in terms of time, financial investment, and access to qualified trainers, creating barriers that make it impractical for the general population [35, 74]. Earlier computational approaches to training support skills employed traditional machine learning techniques, which primarily focused on predicting trainee performance or delivering relatively fixed content [13, 18, 61]. While these systems provided valuable assistance, they often required significant human oversight and struggled to create widely accessible and adaptive learning experiences. The recent emergence of powerful Large Language Models (LLMs) presents a new opportunity to overcome these long-standing limitations [82]. Unlike earlier systems, LLMs possess strong generative capabilities and contextual understanding, enabling them to offer more accessible and adaptable training practices [44, 74] with reduced reliance on human supervision.

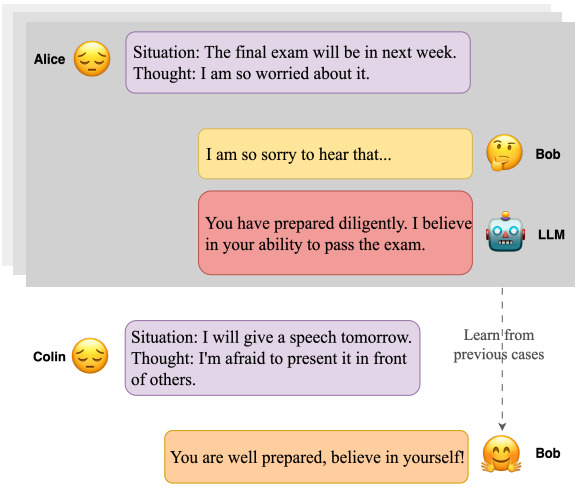


Figure 1: People learn from LLM-generated examples and provide their own support responses to others.

Building on this potential, we introduce ESR-Coach, exploring the use of LLMs to help individuals enhance their supportive language. As illustrated in Figure 1, people can learn to provide supportive responses from the interaction with LLMs. Our system design is inspired by prior work that uses LLMs to simulate practice cases [15, 47, 74], provide example responses [16, 35], and offer feedback [14, 44, 69]. We combine these elements into a unified framework ESR-Coach where multiple AI agents work together to coach the user in emotional support. Specifically, (1) a demonstrator provides high-quality reference responses for users to learn from; (2) an assessor evaluates user inputs and offers immediate feedback; and (3) a generator dynamically creates emotionally challenging scenarios tailored to the user's evolving performance. All three agents are prompted with strategies derived from Hill's Helping Skills Theory [33, 34] and cognitive restructuring [9, 65], aiming to help users learn response strategies that are effective in dealing with others' negativity.

We first conducted an evaluation with four psychology experts to verify the reliability of each module's implementation. Expert ratings showed that the representative LLMs (GPT-3.5 and 4) generated high-fidelity negative situations and thoughts, along with helpful responses, with both dimensions averaging about 4 out of 5. To enhance the assessor LLM, we developed a data augmentation approach that fine-tuned the assessor using expanded expert-annotated data, which substantially improved its alignment with human judgment.

To evaluate the efficacy of our ESR-Coach in enhancing users' emotionally supportive communication skills, we explore the following research questions through a user study:

RQ1: How well does ESR-Coach improve users' performance in providing emotionally supportive responses?

RQ2: How well does ESR-Coach shape users' strategic development in emotional support?

To answer these questions, following previous work [35] involving 15 participants, we recruited 20 participants through a crowdsourcing platform, Prolific, to interact with ESR-Coach. Each participant completed a total of 35 cases, each comprising a help-seeker's situation and their associated negative thoughts. For a refined before-and-after comparison, the coaching process consisted of five sequential phases, comprising three evaluation phases (5 cases each) interleaved with two training phases (10 cases each). During the three evaluation phases, participants were asked to respond to cases without the assistance of ESR-Coach. In the two training phases, we supplied our participants with dynamic cases, reference responses, and feedback generated by ESR-Coach. Additionally, participants were divided into five groups to interact with different parts of ESR-Coach to investigate how ESR-Coach influenced learning outcomes. The assessor LLM in ESR-Coach assessed the helpfulness of users' responses during the coaching process, resulting in a total of 700 data points. Regarding **RQ1**, our analysis suggested that ESR-Coach has the potential to enhance the quality of user responses, with helpfulness scores increasing by 17 percent on average. Users often reached a high performance plateau in the intermediate evaluation, indicating a potential for rapid adaptation of human learners. The group study showed that providing demonstration responses, delivering actionable feedback, and generating dynamic cases all contributed to improving the helpfulness of user responses. We investigated **RQ2** by analyzing dynamics in strategy usage and measuring strategy effectiveness. Results showed that users tended to employ a wider range of support strategies and apply them more effectively after training. Case studies further illustrated instances where participants learned to select and adapt appropriate strategies tailored to specific emotional situations.

We summarize our contributions as follows:

- We propose ESR-Coach, a multi-agent LLM framework that trains and assesses people in providing supportive responses.
- We investigate the feasibility and effectiveness of ESR-Coach, encompassing the generation of training cases, the creation of reference responses, and the assessment of users' performance.
- Our user study presents preliminary findings on the improvement of emotionally supportive responses from users coached by ESR-Coach.

2 Background and Related Work

2.1 Emotional Support

The aim of this study is to explore the use of LLMs in training individuals to provide verbal support in social relationships, such as those among peers, friends, and family members. Building upon previous emotional support conversation systems [46, 71], this study is grounded in Hill's Helping Skills Theory [33, 34]. Our work adapts this established framework for everyday, peer-to-peer support contexts. Mainstream emotional support conversation systems [46, 86] incorporate seven types of response strategies in Hill's Helping Skills Theory, excluding the "challenging" strategy, as it is more suited for professional therapy than for layperson support. Furthermore, cognitive restructuring [9, 50, 65] studies suggest that reframing people's negative thoughts to be positive can be beneficial to their mental health. Therefore, we propose a similar strategy to cognitive restructuring, called "perspective shifting". This strategy aims to help individuals seeking emotional support view situations from a more positive perspective [67]. It is gentler

Strategy	Description	Example
Question	Seeking details about the issue to assist the individual in clarifying their challenges.	Could you elaborate on how you felt during that moment?
Restatement or Paraphrasing	Rephrasing the individual's words in a simpler, more concise manner to aid them in gaining clarity.	It seems as though you feel overlooked by others. Is that accurate?
Reflecting Feelings	Expressing and identifying the emotions the individual is experiencing.	It appears that you're quite anxious about this interview, and it holds significant importance to you.
Self-disclosure	Sharing personal experiences or emotions that are similar to those of the individual to show empathy.	I understand exactly how you feel; I too struggle with speaking to unfamiliar people.
Affirmation and Reassurance	Acknowledging the person's strengths, motivations, and abilities while offering support and confidence.	You've put in great effort, and I'm confident you'll succeed!
Providing Suggestions	Offering potential actions that might help, being cautious not to dictate specific steps.	Trying some deep breathing exercises might help you relax.
Giving Information	Supplying relevant details, facts, opinions, resources, or answers to questions to inform the individual.	Many others face similar situations, so try not to stress too much.
Perspective Shifting	Presenting alternative viewpoints or guiding the person towards a more optimistic outlook.	Perhaps they're hesitant to share the truth because they don't want to upset you.

Table 1: Strategies for providing supportive responses.

than "challenging" and is better suited for everyday conversations rather than professional counseling settings. In summary, we adopt the seven types of response strategies and add "perspective shifting" as the eighth, replacing the "challenging" strategy. All of our prompts provided to both users and LLMs include descriptions of these strategies. All strategies, their descriptions, and examples are listed in Table 1.

Hill's Helping Skills Theory divides the helping process into three stages. To explicitly identify and analyze negativity, we follow a previous cognitive reframing study [65], providing cases consisting of situations and thoughts to LLMs and users. In our adaptation, both the LLM demonstrator and the human user provide supportive responses in a single turn to a given case, enabling more efficient training and focused practice on specific strategies. Additionally, we disclosed multiple strategies to users in each round of training, improving the training efficiency.

2.2 AI for Training People

Artificial intelligence has long been applied to facilitate training and educational processes [7, 58, 75]. Traditional approaches typically rely on machine learning techniques to facilitate the teaching process, primarily by predicting trainee behavior and performance [18, 49, 72, 83] and providing auxiliary support [3, 13, 61, 78]. The rise of generative AI, particularly LLMs, has introduced new capabilities for dynamically creating contextualized content and facilitating interactive practice [1, 2, 38, 79]. This has spurred growing exploration into using LLMs for social and communication skills training [5, 26, 45, 63, 82].

Our study explores how to train individuals to provide verbal support to emotional help-seekers via LLMs. Research related to this topic can be broadly categorized into three categories: communication simulation, response assistance, and real-time feedback. The first category of studies simulates the roles of individuals seeking mental support by generating cases and interactive dialogues, thus providing psychological helpers with practice opportunities [15, 47, 74]. For example, Wang et al. [74] developed a framework that uses LLMs programmed with structured cognitive models to simulate individuals, enabling trainees to practice conversations. Louie et al. [47] introduced a pipeline that leverages expert-defined principles to guide LLMs in creating customized role-playing agents for communication practice. The second category of work has examined how LLMs can assist helpers by providing problem detection or response suggestions in real-time while communicating with individuals [16, 35]. For instance, Hsu et al. [35] introduced an AI tool that diagnoses which counseling strategies are needed in a given context and suggests example responses to counselors during practice sessions. The final group of research employs LLMs to provide feedback for skill enhancement [14, 44, 52, 69]. As exemplified by Lin et al. [44], who created an interactive trainer that delivers just-in-time feedback grounded in psychological theory, significantly improving participants' communication skill mastery. Steenstra et al. [69] developed an LLM-powered training system with a simulated patient that provides turn-by-turn performance

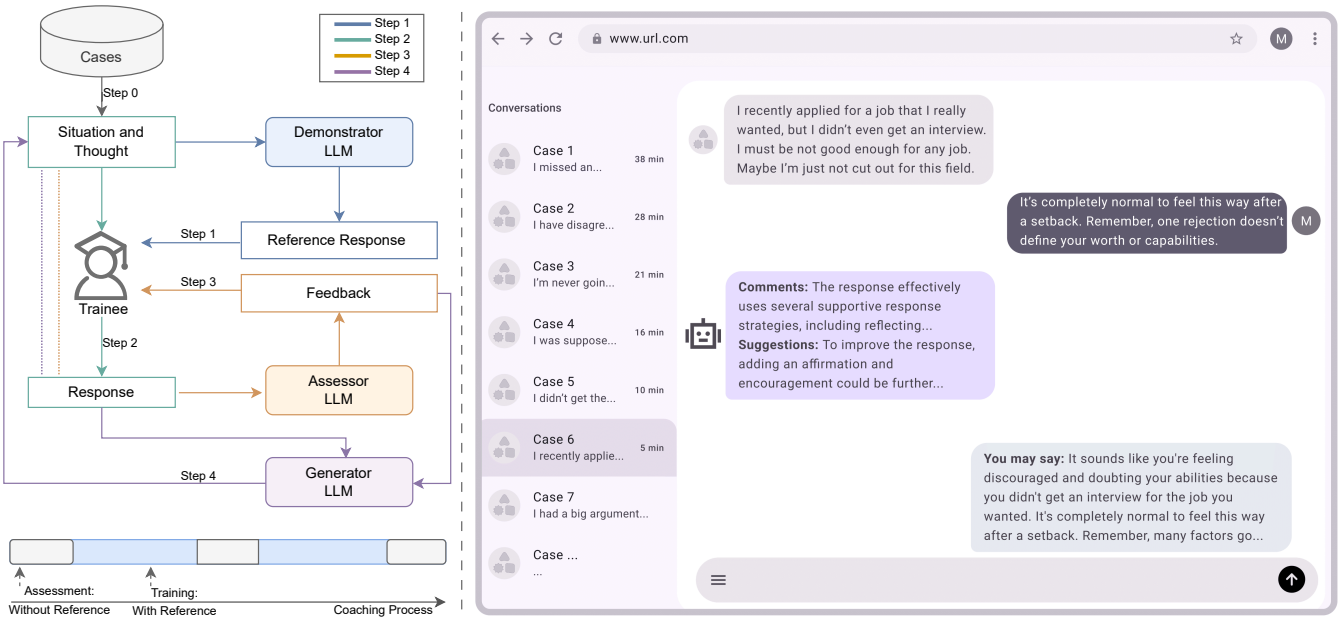


Figure 2: Our ESR-Coach system trains people to provide supportive responses.

visualizations and detailed feedback for counseling skill acquisition. Our study encompasses all these three categories of methods by utilizing LLMs to adaptively create cases for users to practice, generate reference responses, and offer constructive feedback.

3 ESR-Coach

This section details our proposed ESR-Coach training system. We first present the overall system design and its core components. We then describe a pre-study expert evaluation that assessed the capabilities of various LLMs for the system's key tasks. Based on its findings, we introduce a data augmentation method to enhance the assessor LLM. Finally, we outline the complete, integrated coaching process used in our user study.

3.1 System Design

Figure 2 provides an overview of our **ESR-Coach**. It includes a response demonstrator LLM that offers supportive responses as references for human users, an assessor LLM that assesses user responses and provides feedback for improvement, and a case generator LLM that creates training cases based on users' previous performance. All prompts used for the LLMs are listed in the Appendix C.

3.1.1 Case and Response Demonstration. As illustrated in Figure 2, ESR-Coach first randomly selects a case $\boldsymbol{x}$ that includes a situation and a negative thought from the manually annotated dataset [65]. Given the case $\boldsymbol{x}$, the demonstrator LLM $\mathcal{M}$ generates a response $\boldsymbol{y}_m$ for the human user $\mathcal{P}$'s reference, as formulated in Equation 1.

$$\boldsymbol{y}_m = \mathcal{M}(\boldsymbol{x}) \tag{1}$$

Subsequently, the user $\mathcal{P}$ considers both the case $\boldsymbol{x}$ and the LLM-generated reference response $\boldsymbol{y}_m$ to formulate their own response $\boldsymbol{y}_r$, as shown in Equation 2.

$$\boldsymbol{y}_r = \mathcal{P}(\boldsymbol{x}, \boldsymbol{y}_m) \tag{2}$$

3.1.2 Response Feedback. Evaluating user responses is essential for tracking improvements over time [14]. Moreover, existing research has demonstrated that real-time feedback is effective for enhancing users' skill development. Thus, we use an LLM as the assessor $\mathcal{E}$ to produce feedback $\boldsymbol{e}$ for each response $\boldsymbol{y}$ given a case $\boldsymbol{x}$, as formulated in Equation 3.

$$\boldsymbol{e} = \mathcal{E}(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

The most basic form of $\boldsymbol{e}$ is a single score $\boldsymbol{s}$, which the assessor LLM assigns to the target response $\boldsymbol{y}$. Prior research has shown that comments and suggestions can help participants to refine their writing during human-computer interactions [52, 53]. Thus, we prompt the assessor LLM to generate a comment $\boldsymbol{c}$ and a suggestion $\boldsymbol{u}$ before providing the score $\boldsymbol{s}$. This forms a structured feedback tuple $\boldsymbol{e} = \langle \boldsymbol{s}, \boldsymbol{c}, \boldsymbol{u} \rangle$ in natural language.

3.1.3 New Case Generation. Generating training cases adapted to previous performance can enhance the effectiveness of training [8, 22]. Thus, we employ an LLM as the case generator. Given the previous case $\boldsymbol{x}$, response $\boldsymbol{y}$, and feedback $\boldsymbol{e}$, the generator LLM $\mathcal{G}$ will produce a new training case tailored for the user. To minimize latency that could affect the user experience, we input only the current round's data into the generator $\mathcal{G}$ to generate a case for the next round of training. This generation process is formulated in Equation 4.

$$\boldsymbol{x}^{(i+1)} = \mathcal{G}\left(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{e}^{(i)}\right) \tag{4}$$

The entire training process, in which the user receives both a new personalized case and a reference response, crafts their own response, and receives feedback, will iterate until a specified number of rounds (cases) is completed.

3.2 Expert Evaluation of LLMs

Prior to carrying out the user study on the entire ESR-Coach system, we evaluated the capabilities of various LLMs in training emotional support responses. We began by evaluating LLM-generated cases $\boldsymbol{x}_{\mathrm{m}}^{(i)}$ encompassing a pair of situations and thoughts, as well as their corresponding supportive responses $\boldsymbol{y}_{\mathrm{m}}^{(i)}$. Subsequently, we examined the consistency between assessment scores $\boldsymbol{s}_{\mathrm{m}}^{(i)}$ given by an LLM and $\boldsymbol{s}_{\mathrm{e}}^{(i)}$ by experts.

3.2.1 Evaluation Setup. We collected real-world cases from a previous study [65], which curated 300 pairs of situations and thoughts from Thought Record Datasets [10] and Mental Health America. Subsequently, we prompted an LLM with medium performance, namely GPT-3.5 [43, 87], to represent typical LLMs and generate another 300 pairs of situations and thoughts. Then, we divided this dataset of 600 cases into two groups, with each group containing 150 real and 150 generated cases. After that, we employed two representative LLMs, GPT-3.5 and GPT-4, to each provide supportive responses to one of the two groups.[1] This resulted in 300 responses generated by GPT-3.5 and the other 300 generated by GPT-4.

To evaluate the quality of these LLM-generated materials, we engaged four domain experts, all holding or pursuing master's degrees in psychology, to assess LLM performance across three key tasks: case generation, response demonstration, and response assessment. Each expert was assigned 150 samples for evaluation. A two-level stratified sampling approach was employed to ensure an equal number of real and generated cases, as well as an equal number of responses from GPT-3.5 and GPT-4, in each assessment group. Experts evaluated case fidelity and response helpfulness using a 1-5 Likert scale. Additionally, they had the option to modify the LLM-generated responses and were required to label the response strategies of either the original or their modified responses.

Assessment is a crucial component of a training system [82]. While LLMs have shown promising performance in text assessment [40, 87], little research has been done on the use of LLMs in the automatic assessment of emotional responses. Thus, we prompted LLMs to assess the 600 supportive responses generated by GPT-3.5 and GPT-4, using a 1–5 Likert scale. Then we compared these assessment scores to those made by human experts. To accommodate variability, we had the LLMs generate an assessment score for each response five times and report both the mean and standard deviation.

3.2.2 Generation of Cases and Responses. Figure 3 shows that both real and LLM-generated cases were regarded as high fidelity, with only a small difference (0.11 on the scale of 1 to 5) in scores between the generated and real samples. This finding highlights the remarkable capacity of GPT-3.5 in generating situations and thoughts that closely resemble real-world scenarios. Previous studies have demonstrated that a model's capabilities correlate strongly with both the emotional intelligence [59] and the quality of data generation [84]. Given that GPT-3.5 attains medium performance among LLMs [43, 87], we believe LLMs in general can generate socially and emotionally high-fidelity cases.

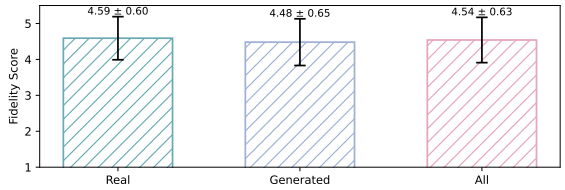


Figure 3: Fidelity scores of real cases and LLM-generated cases.

	Real	Generated	All
GPT-3.5	3.98 (± 0.89)	3.99 (± 0.89)	3.99 (± 0.89)
GPT-4	4.49 (± 0.74)	4.53 (± 0.57)	4.51 (± 0.66)
Average	4.23 (± 0.86)	4.26 (± 0.80)	4.25 (± 0.83)

Table 2: Helpfulness of responses from LLMs, where two models respond to both real and generated cases.

	GPT-3.5	GPT-4	All
# Modified Responses	75	35	110
Avg. Modification Ratio (%)	47.61	21.04	39.15

Table 3: Statistics on the modification of LLM responses by experts. # denotes 'the number of'.

Table 2 shows that, as assessed by our experts, both GPT-3.5 and GPT-4 achieved high scores on response helpfulness, with little difference in scores between real and generated cases. Furthermore, GPT-4 outperformed GPT-3.5, which aligns with general performance differences between the two models. In addition to scoring, revisions made by experts to the LLM-generated responses serve as another measure of generation quality. Intuitively, fewer manual revisions indicate a higher level of confidence in the responses. We compute the word-level minimum editing distance (MED) between each revised response and the original response. Table 3 summarizes the number of modified responses and the average MED of all these modified responses. We see that experts modified fewer responses generated by GPT-4 than those by GPT-3.5, and their modifications were less extensive on GPT-4 responses than on GPT-3.5 responses. This finding aligns with the helpfulness assessments conducted by the experts. In summary, LLMs demonstrate a considerable ability to generate supportive responses.

3.2.3 LLM Assessment v.s. Expert Assessment. As illustrated in Table 4, LLMs exhibited low consistency with experts in response assessment. Even the advanced model GPT-4 obtained a Pearson coefficient of only 0.21. This finding highlights the inadequacy of LLMs in assessing supportive responses. Moreover, we calculated

[1] Including a wider variety of models would strengthen our study, but it would also significantly increase the manpower cost. As such, for an evaluation of LLMs' capabilities before the user study, we selected the most representative series for ease of comparison under limited budget.

	Pearson	Spearman
GPT-3.5	0.15 (± 0.07)	0.10 (± 0.05)
GPT-4	0.21 (± 0.06)	0.19 (± 0.08)

Table 4: Consistency between LLMs' and experts' assessment scores on supportive responses.

the average difference between the scores given by the LLMs and experts. We found that on average, the scores given by GPT-3.5 were 0.73 points higher than those of experts, while the scores given by GPT-4 were 0.42 points higher. This indicates that LLMs tend to provide higher scores than human experts. Consequently, we need to close this gap in using LLMs to assess supportive responses, as our goal is to develop an LLM-driven coaching system.

3.3 Enhancement for Assessor LLM

3.3.1 Data Augmentation. As described in Section 3.2.3, we found that LLM assessments of supportive responses differed considerably from those made by humans. A simple resolution is to use a human-annotated dataset $\boldsymbol{D}_e$ that contains triplets $\langle \boldsymbol{x}, \boldsymbol{y}_m, \boldsymbol{s}_e \rangle$ for training an assessor LLM, where $\boldsymbol{s}_e$ denotes the evaluation by psychology experts on the response $\boldsymbol{y}_m$ to case $\boldsymbol{x}$. However, the responses submitted by the human user in the coaching process should be a mix of responses $\boldsymbol{y}_r$ obtained with given reference responses and the user's own formulations $\boldsymbol{y}_p$. In such a scenario, the assessor LLM trained solely on LLM-generated responses $\boldsymbol{y}_m$ may provide inaccurate assessment scores on the mixed responses.

To address this issue, we propose a data augmentation approach to align the training data used for fine-tuning the assessor LLM with that utilized in the actual coaching process. Our approach is grounded in the observed correlation between LLM capability and response quality [87], which was also evident in our expert evaluation. Specifically, as shown in Table 2, GPT-3.5 received an average helpfulness score that was approximately 0.5 points lower than GPT-4 when responding to the same cases.

To construct a spectrum of response quality that mirrors what might be produced by human trainees at different skill levels, we employ a less powerful LLM $\mathcal{M}_w$ (Qwen-1.5) to generate weaker responses $\boldsymbol{y}_w$. Based on the observed performance gap between GPT-4 and GPT-3.5, we posit a comparable degradation for $\mathcal{M}_w$ relative to GPT-3.5. Given that helpfulness scores are integers on a 1-5 scale, we apply a decrement of 1 point to the original expert scores $\boldsymbol{s}_e$ to assign scores $\boldsymbol{s}_a$ to these weaker responses. Formally, for each case $\boldsymbol{x}$ with its expert-annotated response score $\boldsymbol{s}_e$, we generate:

$$\begin{aligned} \boldsymbol{y}_w &= \mathcal{M}_w(\boldsymbol{x}) \\ \boldsymbol{s}_a &= \max(\boldsymbol{s}_e - 1, 1) \end{aligned} \tag{5}$$

where $\boldsymbol{y}_w$ represents the weak response and $\boldsymbol{s}_a$ is the score assigned to it. The max operation ensures scores remain within the valid range [1, 5]. This approach creates a training set with varied response quality while maintaining consistent relative scoring across responses to the same case.

Through this process, we obtain the augmented fine-tuning data $\boldsymbol{D}_a$ ($|\boldsymbol{D}_a| = |\boldsymbol{D}_e|$), which contains triples $\langle \boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{s}_a \rangle$. We union this augmented data $\boldsymbol{D}_a$ with the expert-annotated data $\boldsymbol{D}_e$ to fine-tune the assessor LLM on this mixed dataset, enabling the LLM to learn how to evaluate both reference responses generated by strong LLMs and those generated by weaker LLMs, thus equipping the assessor LLM to handle a variety of responses from human users.

	Pearson	Spearman
Llama-3.1-8B	0.21 (± 0.06)	0.21 (± 0.06)
GPT-4†††	0.24 (± 0.05)	0.26 (± 0.06)
Our Assessor	**0.67** (± 0.05)	**0.69** (± 0.05)
- Suggestions†	0.66 (± 0.07)	0.68 (± 0.06)
- Comments††	0.66 (± 0.06)	0.68 (± 0.06)
- Sugg & Comm††	0.65 (± 0.06)	0.67 (± 0.05)

$\dagger : p < 0.05$, $\dagger\dagger : p < 0.01$, $\dagger\dagger\dagger : p < 0.001$

Table 5: Performance of our fine-tuned assessor on the evaluation set.

3.3.2 Evaluation of Fine-tuned Assessor LLM. To ensure performance, our assessor LLM $\mathcal{E}$ was fine-tuned on 600 assessment data $\boldsymbol{D}_e$ labeled by human experts and another 600 augmented data $\boldsymbol{D}_a$ generated by using a weaker LLM. We used Llama-3.1-8B as the foundation model for finetuning due to its open-source accessibility and reasonable performance. The fine-tuned assessor LLM exhibits a Pearson coefficient of 0.79 and a Spearman coefficient of 0.78 with the expert assessment data $\boldsymbol{D}_e$, indicating a high correlation between its scores and those of human experts. Nevertheless, the advantage of the weaker LLM-generated data $\boldsymbol{D}_a$, i.e., how accurately the fine-tuned assessor assesses weak responses, remains to be examined. Therefore, we built an additional evaluation dataset to test the assessor's performance.

Evaluation Dataset. To save human labeling effort, we utilized another weaker LLM to construct the evaluation set. To prevent the model from favoring data generated by itself [87], we adopted the same augmentation method as in Section 3.3.1 but used a different model (Mistral-7B) to generate the evaluation set $\boldsymbol{D}_a^*$. The scores for responses $\boldsymbol{y}_w^* \in \boldsymbol{D}_a^*$ were assumed to be lower than those for responses $\boldsymbol{y}_m \in \boldsymbol{D}_e$ generated by stronger LLMs.

Before evaluating the assessor, we verified our assumption on this evaluation set $\boldsymbol{D}_a^*$ by checking whether $\boldsymbol{y}_w^*$ was indeed weaker than $\boldsymbol{y}_m$. For an efficient assessment, we conducted a pair-wise comparison. We randomly sampled 50 $\langle \boldsymbol{x}, \boldsymbol{y}_m \rangle$ tuples from expert-annotated data $\boldsymbol{D}_e$ and retrieved their corresponding responses $\boldsymbol{y}_w^*$ from the evaluation data $\boldsymbol{D}_a^*$, forming triples $\langle \boldsymbol{x}, \boldsymbol{y}_m, \boldsymbol{y}_w^* \rangle$. Due to resource constraints, we invited two PhD students, who learned both response strategies and example reference responses. They were asked to choose the better response from a pair consisting of $\boldsymbol{y}_m$ and $\boldsymbol{y}_w^*$, without knowing which response corresponds to $\boldsymbol{y}_m$ or $\boldsymbol{y}_w^*$. The average Cohen's Kappa Score between the human selections and the assumed better responses is 0.77 (with a human-human agreement score of 0.80), indicating a high level of agreement between the evaluation data and human assessment.

Results. Next, we used the evaluation set $\mathcal{D}_a^*$ to evaluate our fine-tuned assessor LLM. Following previous studies on evaluating assessors [28, 87], we calculated the Pearson and Spearman coefficients

to assess the correlation between the assessor's predictions and the assigned ground truth scores for the evaluation set. We repeated the experiment five times and report both the average and standard deviation. Table 5 lists the correlation coefficients between the predicted scores from our trained assessor and the assigned scores for the responses in the evaluation set. It shows that our fine-tuned assessor's assessment exhibited a high correlation with the evaluation data and outperformed both the Llama model (without fine-tuning) and the GPT-4 model ($p < 0.001$). Additionally, the augmentation of comments and suggestions led to a slight performance improvement, as indicated by minor decreases in the correlation scores.

3.4 The Coaching Process

We selected GPT-3.5 as the demonstrator $\mathcal{M}$ and the case generator $\mathcal{G}$ for its good performance in the previous evaluation. For the assessor $\mathcal{E}$, we chose Llama 3.1-8B fine-tuned on our augmented assessment data. We intentionally chose not to adopt the most advanced models (e.g., GPT-4) for our final system, as we aim to investigate the advantages and disadvantages of using LLMs for coaching within a stable and broadly applicable framework. Our chosen number and models of LLMs for the user study are comparable to those in related work [35, 44, 76], which typically involved one or two LLMs. Finally, we integrated these components into our ESR-Coach system and conducted a user study, examining users' performance in providing supportive responses with and without training using our ESR-Coach.

To ensure effective training and improvement tracking, we design a coaching process in a sequence of five phases, as illustrated in the bottom-left corner of Figure 2. In the first, third, and last phases ("evaluation"), users provide supportive responses to cases without using ESR-Coach, whereas in the second and fourth phases ("training") users provide supportive responses using ESR-Coach. This allows us to compare user performance across the three evaluation phases to measure the impact of the two training phases.

In the first evaluation phase, a case $\boldsymbol{x}$ that includes a situation and a negative thought is randomly selected from a case database [65] per round. In each round, the user $\mathcal{P}$ is required to provide a response $\boldsymbol{y}_p$ to the given case $\boldsymbol{x}$ without any reference response. After completing all practices in the evaluation phase, the user proceeds to the following training phase. In the training phase, the user engages in iterative training using our ESR-Coach. In the first round, the case generator LLM generates a case $\boldsymbol{x}$ based on their performance in the final round of the preceding evaluation phase. Then, the user is given a reference response from the demonstrator LLM and formulates their response. The response to the case $\boldsymbol{x}$ is then evaluated by the assessor LLM $\mathcal{E}$, which provides feedback $\boldsymbol{e}$ to the user to aid in learning. In the next round, the user receives a new case generated by the case generator LLM based on the user's previous performance, and undergoes the same process as the first round. After round-by-round training, the user advances to the next evaluation phase. Upon completion of the three evaluation phases, interleaving with two training phases, the coaching process concludes.

Gender		Age		Race/Ethnicity	
Man	65%	25-30	30%	White	45%
Women	35%	31-40	40%	Asian	15%
		41+	30%	Black	25%
				Mixed	10%
				Other	5%

Table 6: Breakdown of participant demographics by gender, age, and race/ethnicity.

4 User Study

We conducted a user study on our coaching system **ESR-Coach** to evaluate the improvement of trainees' responses throughout the coaching process.

4.1 Setup

We evaluated the effectiveness of our system through a before-and-after controlled experiment. The entire coaching process for each participant consisted of the following five phases: (1) evaluation before training, (2) first training, (3) evaluation post first training, (4) second training, and (5) evaluation post second training. Each evaluation phase contained 5 cases, and each training phase 10 cases. Thus each trainee completed a total of 35 cases in the study.

We recruited 20 participants from Prolific and divided them into five groups. This resulted in 700 data points, which is comparable to prior exploratory studies (258 in Lin et al. [44] and 66 in Wang et al. [74]) that use LLMs to coach humans in the mental health domain. Nevertheless, we acknowledge it limits the statistical power for between-group comparisons. Our primary goal with this design was to conduct an investigation into the learning mechanisms and outcomes, and provide initial evidence for the effect of each component, which can be validated at a larger scale in future work. Table 6 lists the demographic information of our 20 participants. As our objective is not to train professional counselors, we do not require trainees to possess prior psychological knowledge. Nevertheless, there may be disparities in the trainees' initial conversational skills. Therefore, we required that trainees learn from the provided response techniques before beginning their training with our system to mitigate these initial discrepancies. In order to facilitate the tracking of trainee behavior and the customization of the training process, we built a training platform from scratch. We used Flask as the backend and JavaScript as the frontend, keeping this simple and facilitating follow-up work. We provided a guideline that includes the use guide of the system and an introduction and explanations of response strategies for each participant on the front page of our system. Participants can only click to start the coaching process if they confirm that they have read the guideline.

All participants within a group received the same level of system support, while the five groups differed in the specific coaching components they had access to. Specifically, during the evaluation phase, all participants responded to cases $\boldsymbol{x}$ given to them without any reference response. In comparison, in the training phase, all groups received different levels of assistance from the coaching system. The first group received LLM reference responses $\boldsymbol{y}_r$ from the demonstrator $\mathcal{M}$. The second group further received

LLM comments $\boldsymbol{c}$ from the assessor $\mathcal{E}$, whereas the third group received suggestions $\boldsymbol{u}$. The fourth group received both comments and suggestions from the assessor $\mathcal{E}$. Finally, the fifth group not only received reference responses, comments and suggestions, but also training cases from the generator $\mathcal{G}$ based on their previous performance, as described in Section 3.1.3.

4.2 Measurements

Our before-and-after experimental design allows for a direct comparison of user performance, providing clear evidence of the impact of interaction with ESR-Coach. To comprehensively evaluate the effectiveness of our system in training users to provide emotionally supportive responses, we investigated the following dimensions.

Helpfulness of Response. We collected a total of 700 response samples (20 participants × 35 cases each). To assess the quality of these responses at scale, we employed our fine-tuned assessor LLM $\mathcal{E}$. This assessor showed high agreement with human experts (Pearson r=0.79) on standard responses, and also reliable performance (Pearson r=0.67) on a generated assessment dataset with lower-quality responses. We used it to score the helpfulness of each response on a 1-5 scale. This automated assessment enabled two levels of analysis. First, we compared the helpfulness scores from the pre-training and post-training evaluation phases to measure the overall improvement. Second, we examined the performance trajectory across all five phases, including the two training sessions, to provide a more detailed view of the learning process throughout the coaching experience. To complement these quantitative findings, we also conducted case studies to provide qualitative insights into how participants' responses evolved.

Diversity and Effectiveness of Response Strategy. Beyond helpfulness, we examined how participants' use of support strategies evolved. Specifically, we used GPT-4, prompted with ten expert-annotated demonstrations, to automatically label the strategies present in each user response. This allowed us to track the diversity of strategies employed by participants before, during, and after training. We further investigated the correlation between strategy diversity and response helpfulness to understand whether employing a wider range of strategies leads to more effective support. Additionally, we analyzed the situational appropriateness of strategy use through case studies, comparing the strategies, scores, and contextual fit of responses from different phases.

Additionally, our design of multiple user groups enabled us to isolate the contribution of each system component. By comparing the learning outcomes across the five groups, which received different combinations of reference responses, feedback (comments and/or suggestions), and dynamically generated cases, we could assess the individual and combined effects of demonstration, assessment, and personalization on trainee improvement.

5 Results

In this section, we present our findings from the user study. We structure the results around our two research questions, first quantifying the performance improvements achieved with ESR-Coach (RQ1), and then delving into the evolution of users' strategic capabilities (RQ2).

5.1 RQ1: Improvement in Response Performance

5.1.1 Improvement of Helpfulness Score. Table 7 presents the trainee performance across three evaluation phases. Analysis of the overall trends reveals consistent improvement throughout the coaching process. The average helpfulness score across all groups was calculated as $(3.17 + 3.53 + 3.55 + 3.52 + 3.26)/5 = 3.41$ in the initial evaluation, increasing to $(3.27+3.79+3.79+3.91+3.75)/5 = 3.70$ after the first training phase. This represents an overall improvement of $(3.70 - 3.41)/3.41 \times 100\% = 8.5\%$. The improvement was largely sustained in the final evaluation phase, with an average score of $(3.32 + 3.73 + 3.76 + 3.90 + 3.82)/5 = 3.71$.

Participants showed the most substantial gains after the first training phase, with the improvement rate from initial to first post-training evaluation (8.5%) being substantially higher than the additional improvement from first to second post-training evaluation (0.3%). This pattern shows a rapid initial learning curve, where learners efficiently apply supportive response skills from limited exposure to training materials. The sustained performance in the final evaluation suggests that the improvements persisted throughout the duration of our study session.

The observed plateau after the first training phase may reflect the efficient design of our coaching system, which enables users to reach a proficient level of supportive communication with minimal training iterations. This has practical implications for designing accessible and time-efficient emotional support training programs. Together, these results quantify how well ESR-Coach improves performance: users achieved substantial and sustained improvement, with the majority of gains realized rapidly during the initial training phase. Having established that ESR-Coach leads to significant overall improvement, we next dissect these results to understand the contribution of each system component.

5.1.2 Effectiveness of Each Component. To understand how these improvements are achieved, we analyzed the contribution of each system component through our comparison study between user groups. The demonstrator module alone provided a foundation for improvement, achieving a 4.73% gain in helpfulness scores ($p < 0.01$). This establishes that high-quality reference responses serve as an effective starting point for skill development.

The assessment components exhibited distinct but complementary benefits. Comments from the assessor yielded a 7.37% improvement ($p < 0.05$), focusing on qualitative feedback, while suggestions contributed a 6.76% gain ($p < 0.05$), providing actionable guidance. When combined, these assessment elements produced synergistic effects, achieving 10.80% improvement ($p < 0.01$), suggesting that comprehensive feedback covering both evaluative and prescriptive aspects maximizes learning potential.

The condition that integrated all three components of ESR-Coach (Group 5) achieved the highest effectiveness with 17.18% improvement. This integrated approach, combining demonstration, comprehensive feedback, and dynamic case generation, creates a cohesive learning environment where each module reinforces the others. The dynamic case generator appears to amplify the effectiveness of other components by providing appropriately challenging scenarios tailored to individual progress.

Group	Method	Before	Post-1st Tr.	Post-2nd Tr.	Impr.(%)
1	Demonstrator $\mathcal{M}$	3.17±.04	3.27±.04	**3.32**±.12	4.73††
2	+ Assessor's Comments $\mathcal{E}_c$	3.53±.05	**3.79**±.06	3.73±.07	7.37†
3	+ Assessor's Suggestions $\mathcal{E}_u$	3.55±.04	**3.79**±.13	3.76±.11	6.76†
4	+ Assessor's Full Feedback $\mathcal{E}$	3.52±.07	**3.91**±.11	3.90±.12	10.80††
5	+ Full Feedback + Generator (All)	3.26±.08	3.75±.12	**3.82**±.02	**17.18**†

$\dagger : p < 0.05$, $\dagger\dagger : p < 0.01$, $\dagger\dagger\dagger : p < 0.001$

Table 7: The average helpfulness scores of each trainee group in the three evaluation phases. Post-1st Tr. denotes the evaluation phase after the 1st training phase. † represents the statistical significance (paired t-test within group) of the helpfulness improvement before and after training.

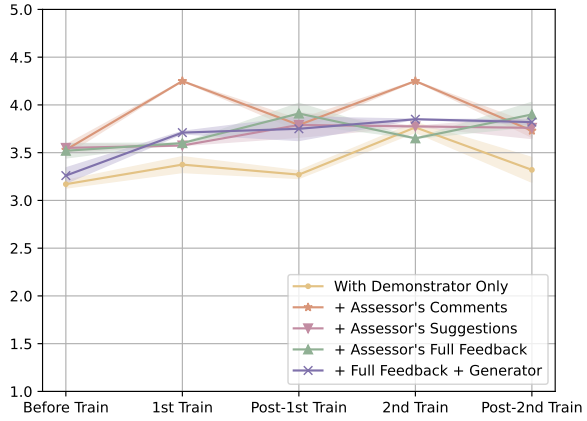


Figure 4: Score variations during the coaching process. Each curve corresponds to a method in Table 7 in order.

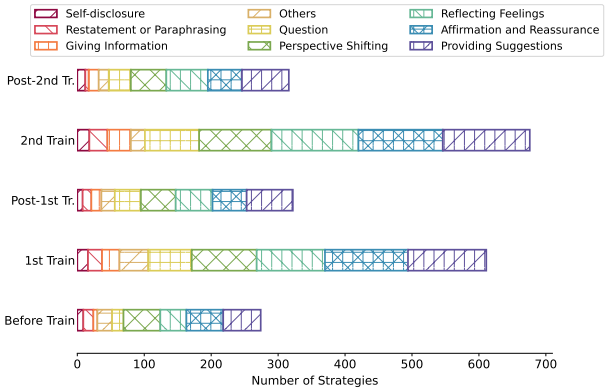


Figure 5: Statistics of strategies employed in trainee responses during the coaching process, divided by phases.

The progressive enhancement from individual components to the fully integrated system underscores the importance of our holistic design approach. Rather than operating in isolation, the demonstrator, assessor, and generator form an interconnected ecosystem that supports different aspects of the learning process, ultimately enabling substantially greater improvement than any single component could achieve independently.

5.1.3 Performance Dynamics. We further analyzed the learning process by examining score variations across all phases, as shown in Figure 4. The trajectories reveal distinct patterns in how different coaching approaches affect skill applications over time.

Groups receiving only reference responses (Group 1) or reference responses with comments (Group 2) exhibited higher scores during training phases compared to evaluation phases. This pattern suggests that trainees may adopt a conservative approach when lacking explicit improvement suggestions, potentially relying more heavily on the provided reference responses during practice sessions.

In contrast, the complete ESR-Coach system (Group 5) demonstrated a smoother and more consistent performance curve throughout the coaching process. The scores for this group progressed steadily from 3.26 in the initial evaluation to 3.82 in the final evaluation, with consistent improvement across both training and evaluation phases. This pattern indicates a more stable and effective performance improvement, where adaptively generated training scenarios appear to facilitate continuous improvement without the performance fluctuations observed in other conditions.

These findings complement the overall performance improvements reported in RQ1 by showing how different coaching components influence the improvement of response performance. While all groups showed improvement, the integrated approach of ESR-Coach supported a more coherent increase, underscoring the value of combining demonstration, feedback, and personalization in emotional support training.

5.2 RQ2: Evolution of Strategy Usage and Effectiveness

While RQ1 established that ESR-Coach improves response performance, RQ2 seeks to understand *how* this improvement is achieved by examining the evolution of users' strategic behaviors. We analyze both the diversity of strategies employed and their contextual effectiveness.

5.2.1 Response Strategy Usage. We analyzed the evolution of strategy usage by counting occurrences in each phase, as shown in Figure 5. Overall, the most frequently employed strategies were

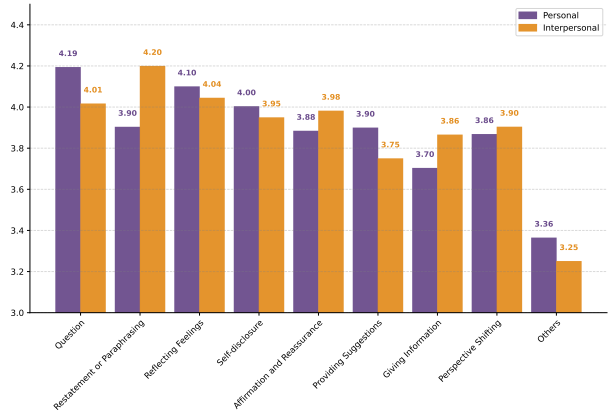


Figure 6: The average helpfulness scores of different strategies when applied to personal versus interpersonal issues.

Providing Suggestions, Affirmation and Reassurance, Reflecting Feelings, and Perspective Shifting. The increase in strategy usage after training sessions, particularly for Question and Reflecting Feelings strategies, indicates the flexible skill application of users. Notably, the Question strategy showed substantial growth, with usage increasing from 17 instances in the initial phase to 81 during the second training phase. This strategy allows trainees to express empathy and provide assistance by actively seeking to understand the help-seeker's situation. The progressive adoption of such inquiry-based approaches throughout the coaching process reflects trainees' improving sophistication in navigating emotional support conversations.

5.2.2 Impact of Response Strategies. To investigate how the effectiveness of support strategies varies across different contexts, we first categorized the emotional dilemmas in our cases into two types: **personal issues** and **interpersonal issues**. This categorization was performed automatically by GPT-4. While automated labeling has limitations, it serves as a consistent proxy for analyzing strategy distribution. We then analyzed the average helpfulness score achieved by each strategy within each context, as summarized in Figure 6.

Our analysis first reveals a fundamental finding: all eight predefined support strategies achieved higher scores (ranging from 3.75 to 4.20) than the "Others" category (3.25-3.36) in both contexts. This result demonstrates the overall effectiveness of the theory-grounded strategies used in ESR-Coach. Furthermore, we observed distinct patterns of effectiveness across the two contexts. Some strategies demonstrated context-independent robustness, showing minimal performance difference (<=0.1). For example, "Reflecting Feelings" and "Self-disclosure" performed similarly well in both contexts, indicating that emotional validation and building rapport through shared experience are universally appreciated regardless of the issue type. Conversely, other strategies exhibited context-dependent effectiveness. The strategy of "Restatement or Paraphrasing" was markedly more effective for interpersonal issues. In situations involving others, accurately paraphrasing the help-seeker's perspective may be particularly valuable for fostering clarity and mutual understanding.

A clear pattern emerged among information-handling strategies. "Question" (seeking information from the help-seeker) and "Providing Suggestions" (offering actionable advice) were both more effective for personal issues. This aligns with the nature of internal struggles, which often benefit from self-reflection and personalized guidance. In contrast, "Giving Information" (providing factual insights) was more effective for interpersonal issues, where external facts and resources can help navigate complex social situations. These findings, in response to RQ2, suggest that users of ESR-Coach not only employed more strategies, but also appeared to adapt their response approaches based on the specific emotional context.

5.2.3 Case Study. This case analysis examines the strategic development of User 2 (in Group 1, receiving reference responses) and User 18 (in Group 5, the complete system), who were selected as representative examples of the dominant learning patterns observed in their respective groups. We show three evaluation phases using an identical case between two groups for a fair assessment. Figure 7 presents the case study. Each color box in the figure presents user responses, key issues identified from the assessor LLM's feedback, and helpfulness scores, corresponding to the cases.

In the initial evaluation, both users leveraged few strategies and showed little empathy. Their responses primarily offered practical suggestions while addressing surface-level issues rather than underlying emotional concerns. Their responses provided suggestions without acknowledging feelings of being perceived as lazy, reflecting a common pattern where novice supporters focus on immediate solutions. These initial responses of both users offered basic utility but lacked emotional depth, resulting in the moderate helpfulness scores of 3/5.

In the intermediate evaluation, User 2 utilized a new strategy Perspective Shifting alongside the used strategy Providing Suggestions. This expansion demonstrates that reference responses serve as a useful learning resource for individuals. By observing reference responses, people are able to acquire basic response strategies accordingly. New strategy utilization was similarly identified in User 18. Notably, there was a consistent inclination among these users to acquire and employ the Perspective Shifting strategy, which may be attributed to the fact that this type of perspective transformation is novel and engaging for ordinary individuals. In addition, both users tend to provide suggestions regarding others' circumstances but neglect emotional engagement. This impacts the demonstration of empathy to some degree, which in turn reduces the helpfulness of their replies.

In the evaluation after second training, User 2 maintained consistent application of previously acquired strategies through habitual use of Perspective Shifting and Providing Suggestions. This pattern highlights the importance of receiving feedback during the training process. Without targeted feedback, users may plateau in their strategic development despite initial gains. With the complete ESR-Coach, User 18 demonstrated strategic integration by combining reflecting feelings with questioning and perspective taking.

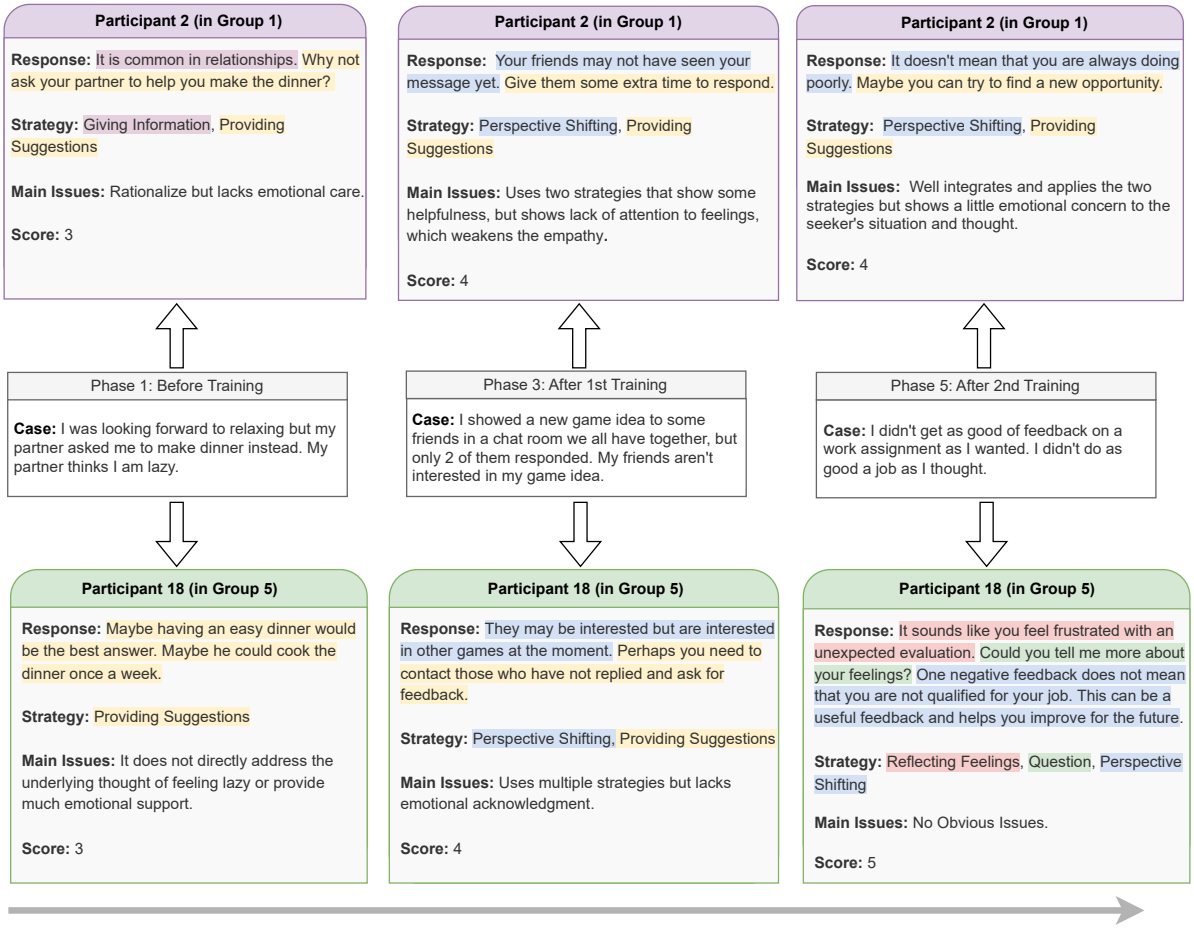


Figure 7: Strategic development of User 2 (Group 1) and User 18 (Group 5) across three evaluation phases. The case study illustrates the progression in strategy application, accompanied by feedback and helpfulness scores.

User 18's response sequence addressed both emotional and cognitive dimensions, creating a more nuanced support response. This progression from isolated strategy use to contextual integration illustrates the potential of ESR-Coach to foster flexible strategy application.

6 Discussion

6.1 From Machine Social Intelligence to Human Skills

Social intelligence, defined as the capacity to understand and effectively navigate human social interactions [70, 80], is a core requirement for systems that coach interpersonal skills [82]. Previous work has investigated the social intelligence of LLMs, showing considerable abilities of LLMs to understand the diverse social context [17, 60, 73]. For LLMs to serve as credible coaches for the improvement of communication skills, they should demonstrate capabilities beyond linguistic fluency, encompassing a functional understanding of human emotions, social contexts, and the nuanced principles of supportive communication [81, 82].

Our user study provides empirical evidence that the LLM-based coaching system effectively facilitates the use of adaptive interpersonal skills in human trainees. The significant improvement in participants' response helpfulness (subsection 5.1), alongside their increased use of diverse and contextually appropriate support strategies (subsection 5.2), suggests that our system guided users toward more effective and context-aware communication. As shown in Section 5.2, results show that participants were able to not only identify what to say but also flexibly adjust their support. As shown in Table 7, the most substantial growth in strategic capability occurred in the group that received the complete coaching system. This suggests that a multifaceted approach combining demonstration, detailed feedback, and adaptive practice is effective for cultivating communication flexibility.

Furthermore, our work highlights the potential of LLM-driven systems as scalable platforms for social intelligence training. Recent explorations into LLMs' capabilities in theory of mind [17] and social reasoning [25] are advancing the frontier of machine social intelligence. As these foundational abilities improve, we anticipate they will enable the creation of even more refined and

adaptive coaches. Such future systems could better understand a trainee's specific emotional state and interpersonal needs, thereby offering more personalized guidance to support the development of sophisticated interpersonal skills across diverse populations and situations.

6.2 Learning of Social Skills

Human learning, particularly the acquisition of complex social and emotional skills, is a profoundly intricate process [21, 36, 37]. It involves the cognitive understanding of strategies, the development of situational awareness, and the cultivation of empathy [23, 29]. The inherent subjectivity of social interactions further complicates both the learning and the assessment of these competencies [32, 35, 44].

Our study provides preliminary insights into this process within a structured coaching environment. As presented in Figure 4, participants showed rapid initial gains after the first training, highlighting the strong learning capabilities of humans. The case study in Section 5.2.3 further investigates this progression, showing how users evolved from offering simple solutions to constructing nuanced responses that combined emotional validation with perspective shifting. Beyond these immediate improvements, the sustainability of learning effects is a critical consideration for real-world application. While this aspect remains under-explored, the stable performance participants maintained in later phases needs future investigation into long-term skill maintenance through long-term post-testing.

Our findings also highlight the challenges in fostering deep learning. In Figure 4, the variance in individual learning trajectories and the plateau effects observed in some groups indicate differing paces of performance improvement. The behavioral measures we employed, while effective for tracking performance, offer a limited view of the underlying cognitive and emotional changes that drive improvement. Ultimately, a deeper understanding of how these social skills are acquired and internalized will likely require cross-disciplinary collaboration [37]. Future work could integrate perspectives from cognitive science [41, 68] and neuroscience [24, 27] to complement behavioral data and illuminate the hidden mechanisms of learning.

6.3 Design Considerations

6.3.1 LLM-based Assessment. A fundamental challenge in our system design involved balancing assessment quality with practical constraints. While human expert evaluation provides the most reliable quality judgments, it cannot provide the real-time feedback required for effective coaching. To address this tension, we developed a hybrid approach using a fine-tuned LLM as the assessor. The solution leveraged expert-annotated data to maintain alignment with human judgment while achieving the scalability needed for immediate feedback during practice sessions. This work establishes a foundation for the future exploration of more advanced training paradigms for assessor LLMs. For instance, leveraging the iterative human-machine interaction inherent in coaching, future systems could employ active learning [62] or reinforcement learning from human feedback (RLHF, [6]) to continuously refine the assessor based on challenging cases and user feedback.

6.3.2 System Design of Multiple Agents. Our three-agent architecture was designed to implement a complete coaching cycle encompassing demonstration, practice, and feedback. For the assessment module, we employed a data augmentation approach aimed to expose the assessor to a broader range of response qualities, better preparing it to evaluate the varied responses produced by human learners. This multi-agent design, evaluated through both expert evaluation in Section 3.2 and user study in Section 5, provides a starting point for exploring richer paradigms of LLM collaboration in coaching. For instance, Qian et al. [55] adopted a hierarchical design on LLM collaboration to finalize tasks. With the increasing number of agents and context length, context engineering [51] paves the way for more adaptive and flexible multi-agent collaboration. For example, the amount and extent of the information exchanged among agents could be curated based on the user responses and scenarios [4].

6.3.3 Generalizability. In this initial system design, we prioritized establishing a clear foundation for emotional support training by focusing on general principles rather than incorporating specific relationship contexts or personality factors. This approach allowed us to investigate core learning mechanisms under controlled conditions. These simplifications in turn provide a scalable training foundation that can be extended with more complex interpersonal dynamics in future work. Moreover, this LLM-based system has the potential to be used in the coaching of other interpersonal skills. As pointed out in Shaikh et al. [63], humans can learn to resolve interpersonal conflict via an LLM-driven simulation system with feedback, which is consistent with our findings in emotional support. Therefore, our framework presents a promising approach that could be adapted for training other interpersonal competencies, such as the development of dialogic leadership [66] and qualitative interviewing [77].

6.4 Ethical Considerations

We describe our procedures for research ethics management, including research ethics review, participant selection and consent, and risk control, in Appendix A. Beyond these standard compliance measures, we considered broader ethical implications in deploying AI for emotional coaching during our system design.

A concern with LLM-based coaching is the potential for misuse. Specifically, users might use the system to automate superficial human connection without genuine empathy, or become over-reliant on AI assistance. Therefore, we positioned ESR-Coach as a learning scaffold rather than a communication substitute. Our system design prioritizes active cognitive engagement over passive consumption, requiring users to formulate their own responses and reflect on AI feedback. Moving forward, it is crucial to investigate how to maintain this balance in long-term usage, ensuring that users apply these skills in their own authentic voices rather than merely mimicking AI patterns.

Deploying generative models in mental health contexts carries the risk of generating emotionally inappropriate cases or feedback. We mitigated this risk not only through the strict filtering and expert review described in our Ethics Statement but also by grounding the AI's behavior in established psychological frameworks. By incorporating Hill's Helping Skills Theory [33] into the system

prompts, we imposed a theoretical soft guardrail that aligns LLM output with standards of professional support. While this theoretical grounding serves as an effective baseline for safety, real-world complexity demands more dynamic safeguards. Future systems should evolve to incorporate continuous expert verification for high-stake emotional scenarios, ensuring that AI-mediated support remains reliable at scale.

6.5 Limitations and Future Work

We summarize several limitations that point to valuable future directions. The sample size of our user study, while comparable to related work [35, 44, 74] in this emerging domain, remains relatively small, which limits the statistical significance of our between-group comparisons. Future research with larger and more diverse participant pools would help validate and extend our results. Similarly, while we expanded the initial dataset through LLM generation, the seed data remained limited. Future work could incorporate broader and more varied sources of emotional scenarios to enhance model robustness.

The coaching paradigm we implemented, while effective for improving users' responses, constitutes a simplified version of real-world emotional support. ESR-Coach currently operates through single-turn exchanges and does not incorporate personal relationships or individual personalities. Extending this framework to support multi-turn conversations and contextual factors would create more realistic practice environments [35, 85]. This could involve developing LLM agents that simulate consistent help-seeker personas across extended conversations.

Regarding the evaluation and supervision, our study employed a sequential coaching design to track individual improvement. To build upon this, future work could incorporate a control group that practices without ESR-Coach to attribute the observed gains to ESR-Coach's coaching and to rigorously quantify the improvement beyond practice effects. Additionally, our evaluation relies on the "LLM-as-a-judge" paradigm [87], which has inherent limitations, such as exhibiting a preference for longer responses, showing a propensity for lenient scoring, and generating feedback with a somewhat uniform style [19, 28]. To narrow this gap, we fine-tuned our assessor on expert-annotated data to better align its judgments with human preferences. In the future, exploring a hybrid system that combines LLM workers with human expert supervision could better capture the subtlety of empathetic communication.

7 Conclusion

We developed an LLM-based coaching system that effectively trains individuals in providing supportive responses to others. Our framework includes a training case generator LLM, a reference response demonstrator LLM, and an assessor LLM that provides feedback to trainees. Through expert evaluation, we found that LLMs could generate highly realistic training materials and helpful responses. Fine-tuned on our extended dataset, the assessor LLM becomes capable of providing reliable feedback. A user study suggested that trainees could improve their supportive responses by interacting with ESR-Coach.

8 GenAI Usage Disclosure

We used ChatGPT and DeepSeek for the purpose of correcting grammar, enhancing expressions, and assisting programming. In our research, we employed generative AI to provide assistance (i.e., practice cases, reference responses, and feedback) to users in learning to deliver emotional support responses, which is our central focus. Additionally, we used LLMs for data augmentation, in order to enhance the LLMs' performance for a fair and valid assessment of user responses. However, it is crucial to clarify that: (1) The method and experiment were designed by us independently. (2) All experimental datasets were derived from empirical results.

Acknowledgments

We thank the anonymous reviewers for providing constructive comments, which help us improve this work significantly. This work was supported by a start-up grant at the Hong Kong University of Science and Technology (Guangzhou).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Hesham Allam, Juan Dempere, Vishwesh Akre, Divya Parakash, Noman Mazher, and Jinesh Ahamed. 2023. Artificial intelligence in education: an argument of Chat-GPT use in education. In *2023 9th International Conference on Information Technology Trends (ITT)*. IEEE, 151–156.

[3] Pengcheng An, Kenneth Holstein, Bernice d'Anjou, Berry Eggen, and Saskia Bakker. 2020. The TA framework: Designing real-time teaching augmentation for K-12 classrooms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–17.

[4] Anthropic. 2025. Effective Context Engineering for AI Agents. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents. Accessed: 2025-10-01.

[5] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120, 41 (2023), e2311627120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2311627120 doi:10.1073/pnas.2311627120

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[7] Joseph Beck, Mia Stern, and Erik Haugsjaa. 1996. Applications of AI in Education. *XRDS: Crossroads, The ACM Magazine for Students* 3, 1 (1996), 11–15.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[9] George M Bodner and THERESA LB McMILLEN. 1986. Cognitive restructuring as an early stage in problem solving. *Journal of Research in Science Teaching* 23, 8 (1986), 727–737.

[10] Franziska Burger, Mark A Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: extracting schemas from thought records. *PloS one* 16, 10 (2021), e0257832.

[11] Moira Burke and Mike Develin. 2016. Once more with feeling: Supportive responses to social sharing on Facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1462–1474.

[12] Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*. Routledge, 569–612.

[13] Rafael A Calvo, Stephen T O'Rourke, Janet Jones, Kalina Yacef, and Peter Reimann. 2010. Collaborative writing support tools on the cloud. *IEEE Transactions on Learning Technologies* 4, 1 (2010), 88–97.

[14] Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4130–4161. doi:10.18653/v1/2024.acl-long.227

[15] Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614* (2023).

[16] Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4295–4304. doi:10.18653/v1/2023.findings-emnlp.284

[17] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024. ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15959–15983.

[18] Evandro B Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior* 73 (2017), 247–256.

[19] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475* (2024).

[20] Carol S Dweck and David S Yeager. 2019. Mindsets: A view from two eras. *Perspectives on Psychological science* 14, 3 (2019), 481–496.

[21] Maurice Elias, Joseph E Zins, and Roger P Weissberg. 1997. *Promoting social and emotional learning: Guidelines for educators*. Ascd.

[22] Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48, 1 (1993), 71–99.

[23] William Estes. 2022. *Handbook of learning and cognitive processes*. Psychology Press.

[24] John DE Gabrieli. 1998. Cognitive neuroscience of human memory. *Annual review of psychology* 49, 1 (1998), 87–115.

[25] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems* 36 (2023), 13518–13529.

[26] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.

[27] Usha Goswami. 2004. Neuroscience and education. *British journal of Educational psychology* 74, 1 (2004), 1–14.

[28] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[29] Hyowon Gweon. 2021. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in cognitive sciences* 25, 10 (2021), 896–910.

[30] Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice* 4, 1 (2008), 189–210.

[31] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=d7KBjmI3GmQ

[32] Cecilia Heyes. 2012. What's social about social learning? *Journal of comparative psychology* 126, 2 (2012), 193.

[33] Clara E Hill. 2020. *Helping skills: Facilitating exploration, insight, and action (5th ed.)*. American Psychological Association.

[34] Clara E Hill and Karen M O'Brien. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association* (1999).

[35] Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2025. Helping the Helper: Supporting Peer Counselors via AI-Empowered Practice and Feedback. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW095 (May 2025), 45 pages. doi:10.1145/3710993

[36] Knud Illeris. 2018. A comprehensive understanding of human learning. In *Contemporary theories of learning*. Routledge, 1–14.

[37] Peter Jarvis. 2012. *Towards a comprehensive theory of human learning*. Routledge.

[38] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

[39] Krzysztof Kaniasty and Fran H Norris. 2000. Help-seeking comfort and receiving social support: The role of ethnicity and context of need. *American journal of community psychology* 28, 4 (2000), 545–581.

[40] Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=8euJaTveKw

[41] Piet AM Kommers, David H Jonassen, and J Terry Mayes. 1992. *Cognitive tools for learning*. Springer.

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

[43] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval. *GitHub repository* (5 2023).

[44] Inna Lin, Ashish Sharma, Christopher Rytting, Adam Miner, Jina Suh, and Tim Althoff. 2024. IMBUE: Improving Interpersonal Effectiveness through Simulation and Just-in-time Feedback with Human-Language Model Interaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 810–840. https://aclanthology.org/2024.acl-long.47

[45] Ryan Liu, Howard Yen, Raja Marjieh, Thomas L. Griffiths, and Ranjay Krishna. 2023. Improving Interpersonal Communication by Simulating Audiences with Language Models. *arXiv preprint arXiv:2311.00687* (2023). https://arxiv.org/abs/2311.00687

[46] 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)*. 3469–3483.

[47] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10570–10603. doi:10.18653/v1/2024.emnlp-main.591

[48] Aaron R Lyon, Shannon Wiltsey Stirman, Suzanne EU Kerns, and Eric J Bruns. 2011. Developing the mental health workforce: Review and application of training approaches from multiple disciplines. *Administration and Policy in Mental Health and Mental Health Services Research* 38, 4 (2011), 238–253.

[49] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 313, 25 pages. doi:10.1145/3491102.3517482

[50] Isaac Marks, Karina Lovell, Homa Noshirvani, Maria Livanou, and Sian Thrasher. 1998. Treatment of posttraumatic stress disorder by exposure and/or cognitive restructuring: A controlled study. *Archives of general psychiatry* 55, 4 (1998), 317–325.

[51] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334* (2025).

[52] Emma O'neil, João Sedoc, Diyi Yang, Haiyi Zhu, and Lyle Ungar. 2023. Automatic Reflection Generation for Peer-to-Peer Counseling. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. 62–75.

[53] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[54] Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300294

[55] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15174–15186. https://aclanthology.org/2024.acl-long.810

[56] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.

[57] William N. Robiner. 2006. The mental health professions: Workforce supply and demand, issues, and challenges. *Clinical Psychology Review* 26, 5 (2006), 600–625. doi:10.1016/j.cpr.2006.05.002

[58] Ido Roll and Ruth Wylie. 2016. Evolution and revolution in artificial intelligence in education. *International journal of artificial intelligence in education* 26, 2 (2016), 582–599.
[59] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 5986–6004. https://aclanthology.org/2024.acl-long.326
[60] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4463–4473.
[61] Silvia Schiaffino, Patricio Garcia, and Analia Amandi. 2008. eTeacher: Providing personalized assistance to e-learning students. *Computers & Education* 51, 4 (2008), 1744–1754.
[62] Burr Settles. 2009. Active learning literature survey. (2009).
[63] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
[64] Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 614–625.
[65] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9977–10000.
[66] Carolyn M Shields. 2004. Dialogic leadership for social justice: Overcoming pathologies of silence. *Educational administration quarterly* 40, 1 (2004), 109–132.
[67] Michelle N Shiota and Robert W Levenson. 2012. Turn down the volume or change the channel? Emotional effects of detached versus positive reappraisal. *Journal of personality and social psychology* 103, 3 (2012), 416.
[68] Thomas J Shuell. 1986. Cognitive conceptions of learning. *Review of educational research* 56, 4 (1986), 411–436.
[69] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–22.
[70] Robert J Sternberg and Avery Siying Li. 2020. Social Intelligence: What It Is and Why. *Social Intelligence and Nonverbal Communication* (2020).
[71] 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)*. 308–319.
[72] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. 2020. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior* 104 (2020), 106189.
[73] Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024. Towards Objectively Benchmarking Social Intelligence of Language Agents at the Action Level. In *Findings of the Association for Computational Linguistics ACL 2024*. 8885–8897.
[74] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 12772–12797. doi:10.18653/v1/2024.emnlp-main.711
[75] Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications* 252 (2024), 124167.
[76] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105* (2024).
[77] Carol AB Warren. 2002. Qualitative interviewing. *Handbook of interview research: Context and method* 839101 (2002), 103–116.
[78] Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. 2020. An interaction design for machine teaching to develop AI tutors. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.
[79] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems* 37 (2024), 15674–15729.
[80] Toshio Yamagishi. 2001. Social Intelligence. *Trust in society* 2 (2001), 121.
[81] Bufang Yang, Yunqi Guo, Lilin Xu, Zhenyu Yan, Hongkai Chen, Guoliang Xing, and Xiaofan Jiang. 2025. Socialmind: Llm-based proactive ar social assistive system with human-like perception for in-situ live interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 1 (2025), 1–30.
[82] Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204* (2024).
[83] Fan Yang and Frederick WB Li. 2018. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & education* 123 (2018), 97–108.
[84] Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. Llm–assisted data augmentation for chinese dialogue–level dependency parsing. *Computational Linguistics* (2024), 1–24.
[85] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.
[86] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 1552–1568.
[87] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

A Ethics Statement

IRB Approval. This research was ethically reviewed and approved by the Institutional Review Board (IRB) of our organization.

Research Target. This study presented a system to train lay people's communication skills. In particular, the proposed system uses LLMs to coach people in providing supportive responses to others around them who express negative thoughts. Our system was not targeted for professional or psychological counselors, and this study did not evaluate any clinical outcomes.

Participants. All participants were aged 18 above and provided informed consent, regarding the study purpose, risks, and data collection. Before the study began, we informed all potential participants that people 1) with mental health issues or 2) who have a traumatic experience that may cause severe distress when recalled would not be allowed to participate. Furthermore, we requested the experts to provide their results from a mental health self-assessment test. As our objective is not to train professional counselors, we do not require trainees to possess prior psychological knowledge. Nevertheless, there may be disparities in the trainees' initial conversational skills. Therefore, we mandate that trainees learn from the provided response techniques before commencing their experience with our system to mitigate these initial discrepancies. In return for their work, all participants were paid a reasonable amount of money.

Data Collection and Usage. No private information was collected and all data were manually filtered to ensure anonymity of people and locations. All data and software developed in this undertaking were utilized solely for research purposes.

Risk Control. We first required the psychology experts to report risky cases during their annotation process. Before the user study, all cases were manually checked, and those that potentially posed risks were removed. We set filters on mental health issues on the crowdsourcing platform and invited only those users who declared to have no mental health issues and be under no risk. Nonetheless, AI dynamically generated content may make participants uncomfortable in some specific conditions. Therefore, on one hand, we

	Count	Avg. # Words
Case		
Real	300	27.63
Generated	300	31.33
Total	600	29.49
Response		
GPT-3.5	300	31.73
GPT-4	300	31.24
Total	600	31.49
Feedback		
Comments $\in \mathcal{D}$	600	80.40
Suggestions $\in \mathcal{D}$		42.81
Comments $\in \mathcal{D}_a$	600	59.08
Suggestions $\in \mathcal{D}_a$		56.66

Table 8: Statistics of training materials, including cases (input), responses (output) and feedback.

Strategy	Count
Question	215
Restatement or Paraphrasing	175
Reflecting Feelings	416
Self-disclosure	26
Affirmation and Reassurance	461
Providing Suggestions	253
Giving Information	85
Perspective Shifting	308

Table 9: Statistics of each strategy.

added response strategies based on Hill's Helping Theory [33] to the prompts for LLMs as guidance. On the other hand, we informed the participants that they have the right to terminate the study at any time without incurring any negative consequences. This risk assessment and control have been incorporated into our IRB review and approval and provided in the informed consent form for the participants. No risks were reported during the entire experiment.

B Data Statistics

Table 8 lists the statistical information of expert-annotated data. It shows that LLM-generated cases are, on average, a little longer than real cases. The responses of GPT-3.5 and GPT-4 are of roughly the same considerable length. The LLM-generated comments and suggestions are longer than cases and responses. Considering the workload of experts, we did not require them to do a fine-grained annotation to label what strategy was used in each sentence of responses. Thus, we list the total number of occurrences of each strategy in all responses. In Table 9, we observe that the most commonly-used strategy is "Affirmation and Reassurance", which acknowledges the help-seeker's strengths, motivations and abilities. In contrast, the strategy "Self-disclosure" is used least, because LLMs have no actual experience as humans.

C Prompts for LLMs

This study involved prompting LLMs to generate cases, responses, evaluations and so on. Our prompts followed a structured pattern, including backgrounds, instructions, constraints and input. Figures 8–12 list all prompts.

For case generation, an LLM is first prompted to generate 30 roles in the real world, such as students, teachers, parents. Then, for each role, we repeatedly prompt the LLM to generate a total of 10 corresponding cases. For other generation tasks, we have provided the table of strategies (Table 1) as references for generating theory-grounded responses or feedback.

D Implementation Details

We utilized GPT-3.5-turbo-0125 and GPT-4o-2024-05-13[2] in our study. We used the default temperature setting and empirically set top_p to 0.4 in all cases, ensuring the stability of the LLMs' output while retaining diversity. We used GPT-3.5-turbo-0125 as the case generator and the response demonstrator in SR-Coach. For fine-tuning the assessor LLM, we adopted the full fine-tune strategy, using Llama-3.1-8B-Instruct as the backbone. We used 4 × A6000 GPUs for fine-tuning, with a batch size of 8 samples per GPU. To mitigate the GPU memory consumption, we utilized ZeRO stage 3 [56] and allowed offloading some memory to the CPU. The fine-tuning epoch was 5 and the initial learning rate was 2e-5 with a 0.1 warmup ratio. We chose the last checkpoint as the well-trained assessor for the evaluation and use in the human coaching process. All automated experiments with LLMs were run 5 times, and the mean and standard deviation are reported. Due to the constraints of labor costs, LLM experiments involving manual annotation such as manual evaluation of LLM responses, example generation, and response assessment were run once with a fixed seed (42). To avoid the preference problem that arises between identical models [43], we chose Qwen1.5-7B-Instruct to generate augmented training data D_a and Mistral-7B-Instruct-v0.1 to generate the evaluation set D_a^*, both less proficient than GPT-3.5-turbo and GPT-4 in various benchmarks [31, 43, 87]. We utilized vllm[3] to accelerate the LLM inference.

E Expert Annotation Interface

We used Label Studio[4] to collect the annotations from experts. The GUI for expert annotations is presented in Figure 13. The experts assessed and labelled the data in the form of questionnaires, illustrated in Table 10.

F Guideline and Consent Form of User Study

We provided a guideline and consent for each participant on the front page of our system, presented as plain text in Table 11.

[2]https://platform.openai.com/docs/models
[3]https://docs.vllm.ai
[4]https://labelstud.io

Prompt for Case Generation

You are an assistant with empathy.
Instructions:
Please generate 10 cases including a situation that {group_name} may face and corresponding negative thoughts that might arise (due to cognitive traps) in the situation.
Constraints:
1. In the first person (for example, I...).
2. Keep the JSON format with keys "group", "situation", "negative_thoughts."

Figure 8: Prompt for generating cases with situations and negative thoughts for the construction of data.

Prompt for Response Demonstrator

You are an assistant with empathy. Given someone who faced a situation and had a negative thought, please provide a supportive response. Below are some response strategies that may be helpful: {table_of_strategies}
Instructions:
1. Choose one or multiple strategies from the provided list to address negativity.
2. Explain the selected strategies and why they are beneficial.
3. Craft a brief, supportive response based on these strategies.
Constraints:
1. Consider a suggested reframed thought, but feel free to innovate.
2. Mimic an empathetic companion's tone in your response.
3. Format your output as JSON with keys "strategy_type", "strategy_description", and "response."
Input:
{situation_and_thought}

Figure 9: Prompt for generating supportive responses to negative thoughts.

Prompt for Assessor

You are an assistant with empathy. Given a context (including both a situation and a thought) and a supportive response, you should generate an evaluation and provide suggestions for improvement.
You should assess the response using the following 1-5 Likert scale:
[1] Not at all helpful
[2] Slightly helpful
[3] Neutral
[4] Somewhat helpful
[5] Very helpful
Instructions:
1. Offer the evaluation and output the score.
2. Provide suggestions for improving the response, with fewer suggestions for higher scores (responses with a 5 rating may not need to be revised).
Constraints:
Output must be in JSON format, with keys "comment", "score", and "suggestion."
Input:
{case_and_response}

Figure 10: Prompt for assessing supportive responses and providing feedback.

Prompt for Comment and Suggestion Generation

You are an assistant with empathy. Given a provided context (including both a situation and a thought), a supportive response, and a score assigned by psychological experts, generate a rationale for the score and provide suggestions for improvement.
Psychological experts assess the response using the following 1-5 Likert scale:
[1] Not at all helpful
[2] Slightly helpful
[3] Neutral
[4] Somewhat helpful
[5] Very helpful
Below are some response strategies for supportive responses that may be helpful:
{table_of_strategies}
Instructions:
1. Offer a rationale for the assigned score, don't mention the score.
2. Provide suggestions for improving the response, with less suggestions for higher scores (responses with 5 rating may not need to be revised).
Constraints:
Output must be in JSON format, with keys "comment" and "suggestion."
Input:
{case_response_and_score}

Figure 11: Prompt for generating comments and suggestions based on expert scores.

Prompt for Dynamic Case Generator

You are an assistant with empathy. Given a context, a response, and a feedback from the psychological expert, you should generate a new training case that includes a new real-world situation and a thought that a help-seeker may have had.
Below are some response strategies that may be helpful:
{table_of_strategies}
Instructions:
1. Generate a real-world situation that a help-seeker faced.
2. Generate a corresponding negative thought that the help-seeker had.
Constraints:
1. In the first person (for example, I...).
2. Format your output as JSON with keys "situation", "thought."
Input:
{case_response_and_feedback}

Figure 12: Prompt for dynamically generating new training cases based on feedback.

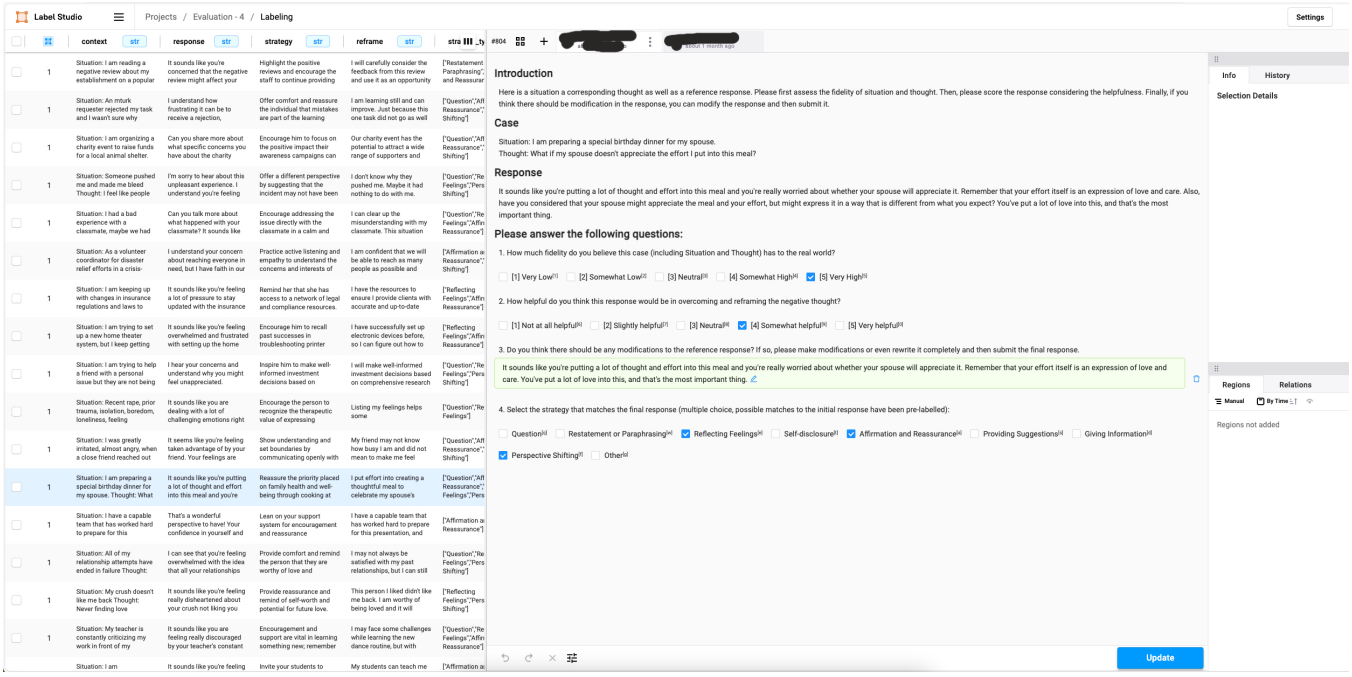


Figure 13: GUI for expert annotation.

Introduction
Here is a situation, a corresponding thought as well as a reference response. Please first assess the fidelity of situation and thought. Then, please score the response considering the helpfulness. Finally, if you think there should be modification in the response, you can modify the response and then submit it.
Case
Situation: I am trying to help a friend with a personal issue but they are not being receptive to my advice.
Thought: My friend doesn't value my opinion and support, which makes me feel unappreciated.
Response
I hear your concerns and understand why you might feel unappreciated. However, have you considered that your friend might be going through a tough time and is struggling to accept help? It's not necessarily a reflection of the value they see in your support. It's possible that they highly value your support, but they just need some time to process their own feelings.
Please answer the following questions:
1. How much fidelity do you believe this case (including Situation and Thought) has to the real world?
[1] Very Low
[2] Somewhat Low
[3] Neutral
[4] Somewhat High
[5] Very High
2. How helpful do you think this response would be in overcoming and reframing the negative thought?
[1] Not at all helpful
[2] Slightly helpful
[3] Neutral
[4] Somewhat helpful
[5] Very helpful
3. Do you think there should be any modifications to the reference response? If so, please make modifications or even rewrite it completely and then submit the final response:

_____
4. Select the strategies that match the final response:
[ ] Question
[ ] Restatement or Paraphrasing
[ ] Reflecting Feelings
[ ] Self-disclosure
[ ] Affirmation and Reassurance
[ ] Providing Suggestions
[ ] Giving Information
[ ] Perspective Shifting

Table 10: An example of questionnaire for experts.

[IMPORTANT] Before you start experiencing our system, please read the consent form and guideline below.

Purpose
This study aims to investigate whether large language models can help train humans to provide supportive responses toward others' negative thoughts.
Procedures
In this study, assuming others have some situations and negative thoughts, you are assigned to write appropriate supportive responses to verbally support them. There are 35 samples in total. Among these samples, 20 samples have AI-generated responses for reference. You can consider them. Then you can write on your own, or you can revise or directly submit the generated response. You should learn from the reference and make appropriate adjustments to your ideas, rather than just submitting them for reference. 15 samples are without references at the beginning, mid-term, and end, so you should write on your own in these cases. Note that you should provide complete sentences, not short phrases.
Participant Requirements
1. Each participant must be 18 years and older.
2. Each participant must be familiar with English.
3. Each participant must not have a history of and must be currently free of mental illness.
4. If you have had a traumatic experience that may cause severe distress when recalled, you are not allowed to participate.
5. Ideally, each participant hopes to gain from the experience an increased ability to cope with negativity from friends or family members. This is not mandatory, but if you are not interested, we do not recommend that you participate in this experiment.
Benefits
Participants may learn some communication skills for helping others by giving supportive responses. The publication of this research can benefit the research community.
Compensation
You will receive 9 GBP for the whole experience. The entire process may take you approximately 1 hour.
Risks
There is negativity in the data. Although we have manually filtered potentially harmful information, negativity can be mildly psychologically taxing for you.
Rights
Your participation is voluntary and can be freely stopped at any point by you.
Confidentiality Assurance
This study will not collect any private information of you.
Consent Confirmation
You must carefully check the following statements, which indicate your consent:
1. I confirm I am over 18 years old.
2. I confirm I don't and haven't had mental illness or a traumatic experience that may cause severe distress when recalled.
3. I have read and understood this consent form.
4. I agree to participate in this study.
As an online system, we collect your consent through your registration. Our registration system requires you to enter your Prolific ID for your credentials. If you sign up for an account and confirm the information with our system, that indicates your informed consent.

How to respond?
Providing appropriate responses can be challenging. Here we provide several useful strategies to respond to others' negativity for your reference. You can use one or several strategies in your response. Before experiencing our system, please read these strategies below. You should read them anytime if you feel unsure how to respond.
{Table of Response Strategies}
How to use the system?
You will see the log-in window at first. You should enter your Prolific ID into both Username and Password to register and sign in. Make sure you enter the correct ID. Then, you will enter the system and should click on "Case xx" on the sidebar to begin. At the beginning, the mid-term and the end of this training, you will not receive reference responses and feedback. This means you will need to complete your response independently. Please respond carefully.

Table 11: Consent form and guideline for participants.