SocialSim: Towards Socialized Simulation of Emotional Support Conversation

Zhuang Chen[1,2*], Yaru Cao[3*], Guanqun Bi[2†], Jincenzi Wu[4], Jinfeng Zhou[2], Xiyao Xiao[5], Si Chen[6], Hongning Wang[2], Minlie Huang[2]

[1]School of Computer Science and Engineering, Central South University
[2]CoAI Group, DCST, IAI, BNRIST, Tsinghua University
[3]Northwest Minzu University
[4]The Chinese University of Hong Kong
[5]Lingxin AI
[6]Academy of Arts & Design, Tsinghua University
zhchen18@foxmail.com, biguanqun@mail.tsinghua.edu.cn

Abstract

Emotional support conversation (ESC) helps reduce people's psychological stress and provide emotional value through interactive dialogues. Due to the high cost of crowdsourcing a large ESC corpus, recent attempts use large language models for dialogue augmentation. However, existing approaches largely overlook the social dynamics inherent in ESC, leading to less effective simulations. In this paper, we introduce SocialSim, a novel framework that simulates ESC by integrating key aspects of social interactions: *social disclosure* and *social awareness*. On the seeker side, we facilitate social disclosure by constructing a comprehensive persona bank that captures diverse and authentic help-seeking scenarios. On the supporter side, we enhance social awareness by eliciting cognitive reasoning to generate logical and supportive responses. Building upon SocialSim, we construct SSConv, a large-scale synthetic ESC corpus of which quality can even surpass crowdsourced ESC data. We further train a chatbot on SSConv and demonstrate its state-of-the-art performance in both automatic and human evaluations. We believe SocialSim offers a scalable way to synthesize ESC, making emotional care more accessible and practical.

Introduction

Emotional support conversation (ESC) is a form of communication aimed at providing comfort, understanding, and encouragement to someone who is experiencing emotional distress or facing challenging situations (Beebe et al. 2002). ESC is widely used in various domains, including therapy (Rogers 1995), counseling (Sutton and Stewart 2017), peer support programs (Mead 2014), and online mental health services (Riva 2004). To train machine systems for ESC, Liu et al. (2021a) collect an ESConv corpus by employing crowdsourcing workers to act as seekers and supporters engaging in conversations on predetermined topics. However, the high cost of crowdsourcing limits both the number of conversations and the diversity of topics in such corpora.

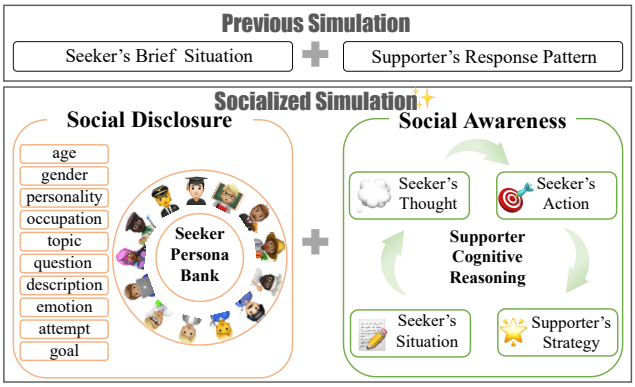

Figure 1: SocialSim recognizes ESC as a social activity and simulates it with both social disclosure and awareness.

With the advent of large language models (LLMs) like GPT-4 (Achiam et al. 2023) and LLaMA (Touvron et al. 2023), recent studies have made progress in using LLMs to augment dialogue data for ESC (Zheng et al. 2023a,b). However, these approaches largely neglect the inherently social nature of ESC, resulting in a noticeable gap between synthetic and crowdsourced corpora. Drawing on theories of social intelligence (Albrecht 2009), as shown in Figure 1, we identify two main areas where this gap arises. **1)** ***Seeker's social disclosure***. Existing methods provide very limited seeker demographics, such as a very brief description of the help-seeking situation, which restricts the specificity and diversity of the dialogues. **2)** ***Supporter's social awareness.*** The current focus is primarily on replicating example dialogues rather than engaging in the thinking process as human supporters should. This reliance on external response patterns, rather than internal cognitive reasoning, reduces the depth and relevance of the simulated dialogues.

To address these challenges, we propose SocialSim, a framework designed for socialized simulation of emotional support conversations. **1)** On the seeker side, we facilitate social disclosure via ***persona realism***. Specifically, we

*These authors contributed equally.
†Corresponding author.

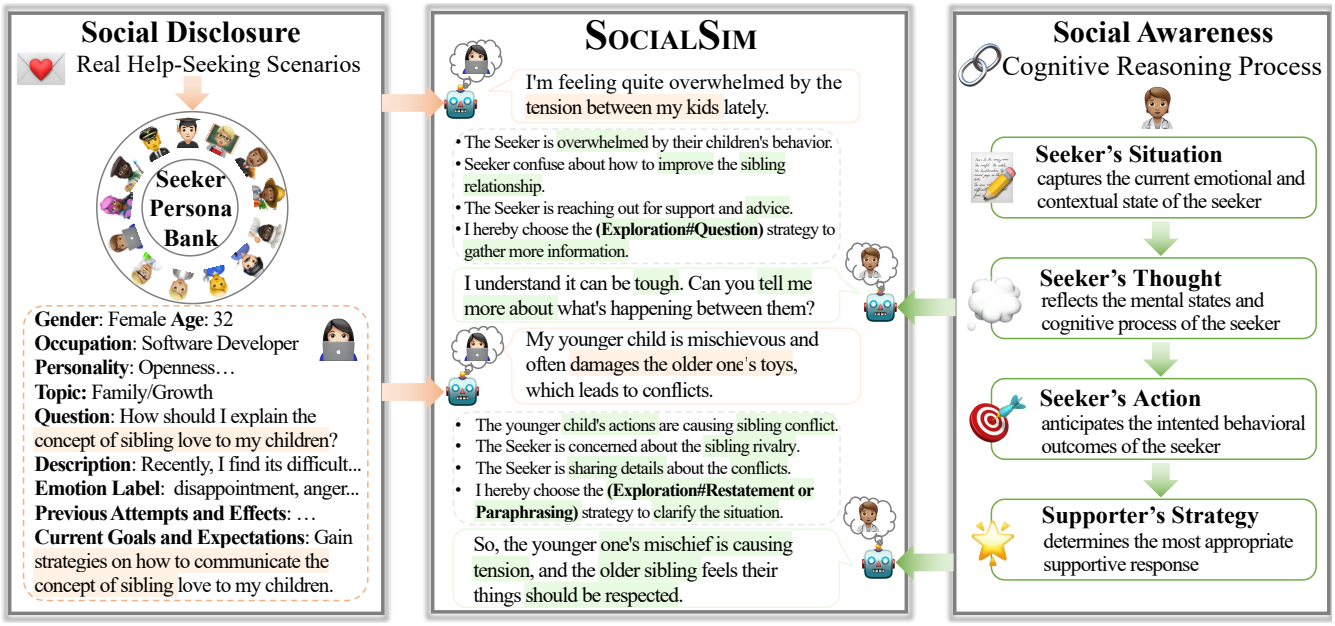

Figure 2: The SocialSim framework. The socialized simulation of emotional support conversation is achieved by conducting persona realism on the seeker side and eliciting cognitive reasoning on the supporter side.

first collect real-world help-seeking scenarios from PsyQA (Sun et al. 2021), a psychological health support dataset. Then drawing on psychological theories (Costa and McCrae 1999), we build a realism pipeline to transform those scenarios into a persona bank that includes comprehensive demographics like gender, age, occupation, personality, and other related details about emotional distress. With the informative and diverse personas, we prepare SocialSim to synthesize realistic and specific dialogues. **2)** On the supporter side, we enhance social awareness via ***cognitive reasoning***. Inspired by related work on cognitive behavioral therapy (Beck 2020) and theory of mind (Beaudoin et al. 2020), we implement a reasoning chain that mimics human supporter's thinking process to obtain a more in-depth understanding about the seeker's dialogue history and mental states, fostering the generation of tailored and supportive responses. **3)** Utilizing the SocialSim framework, we prompt LLMs and generate SSConv, a synthesized ESC dataset of which quality surpasses not only existing synthetic datasets but also the crowdsourced corpus validated by human inspections. We then train a chatbot on SSConv and demonstrate its state-of-the-art performance via both automatic and human evaluations, establishing the effectiveness of SocialSim.

Our key contributions are as follows: 1) We introduce SocialSim, a framework that fosters socialized simulation of ESC by integrating the social disclosure of seekers and social awareness of supporters. 2) We create SSConv, a high-quality synthesized ESC dataset via prompting LLMs under SocialSim, proving its superior quality through rigorous evaluation. 3) We develop and evaluate a chatbot trained on SSConv, demonstrating it outperforms existing methods in delivering meaningful and supportive emotional interactions. We hope SocialSim can bridge the gap between synthetic and real-world emotional support conversations by simulating the social dynamics, thereby making ESC more accessible and benefiting a broader community.

SocialSim: Socialized Simulation Framework

In this section, we introduce SocialSim, a socialized simulation framework for emotional support conversations. As shown in Figure 2, SocialSim is divided into three parts:

- On the seeker side, we conduct persona realism to build an informative and diverse help-seeking scenario bank. This step prepares for simulating authentic conversations by facilitating the seeker's social disclosure.
- On the supporter side, we elicit an explicit cognitive reasoning procedure before responding. This step supports mimicking human supporters' thinking processes, generating more empathetic and supportive responses, thereby enhancing the supporter's social awareness.
- By combining the designs from both the seeker and supporter sides, we instruct the LLMs to synthesize new dialogues to achieve a socialized simulation of ESC.

Persona Realism for Social Disclosure

Social disclosure refers to the extent to which a help-seeker reveals personal and emotional information during a conversation, which is crucial for generating authentic and empathetic responses. Existing methods often lack social disclosure because they provide minimal or generic seeker profiles, limiting the depth and diversity of the generated dialogues. To address this, we propose persona realism, a method that constructs detailed and diverse seeker personas using real-world scenarios. By incorporating attributes such

as personality traits, emotional struggles, and contextual details, persona realism can enhance the authenticity and variety of social disclosure when synthesizing dialogues.

Help-Seeking Scenario Collection To enhance the authenticity and specificity of the ESC simulation, we collect real-world help-seeking scenarios that provide rich and realistic details. We select PsyQA (Sun et al. 2021), a Chinese psychological health support dataset in a Q&A format. Each scenario in PsyQA contains a pair of {*question*, *description*}, which provides a brief and detailed account of emotional struggles, respectively. Additionally, each scenario is labeled with a help-seeking theme, covering 9 major topics and 100 subtopics, including personal growth, interpersonal relationships, family issues, etc. For example, one seeker asks "*How should I explain the concept of sibling love to my children?*" and describes feeling uncertain about how to convey the importance of familial bonds, reflecting complex emotions like a desire to nurture closeness and concern over potential conflicts. Such scenarios offer rich emotional content that is crucial for realistic ESC simulation. To ensure the safety and quality of data, we first filter out irrelevant and sensitive topics like suicide, racial discrimination, and professional medical treatment, then discard descriptions shorter than 65 words to ensure the informativeness of scenarios. We then translate the remaining scenarios into English using GPT-4 and manually validate the translated output to ensure the preservation of the original context and emotional content. Finally, we collect 3,229 scenarios that are rich in information, safe in content, and diverse in topics, making them well-suited for creating realistic help-seeker personas in ESC simulations.

Structured Persona Realism Building on the extensive real-world help-seeking scenarios, we draw upon established psychological theories, particularly the Five-Factor Model of personality (Costa and McCrae 1999), to systematically organize these scenarios into structured seeker personas. Each persona is designed in a key-value format, providing detailed and accurate information crucial for the subsequent generation of dialogues. The key attributes in personas are as follow: {*gender*, *age*, *occupation*, *personality*, *topic*, *question*, *description*, *emotion label*, *previous attempts and effects*, *current goals and expectations*}.

The construction of personas is essentially an information extraction task. We start by selecting a representative help-seeking scenario and manually constructing a corresponding persona to create a demonstration of the task. Following this, we employ GPT-4 to structure all collected scenarios into personas following the constructed demonstrations. Despite the rich detail in the original scenarios, some key attributes may still be missing. In these cases, we allow GPT-4 to make reasonable inferences to complete the persona. For instance, if a seeker's occupation is not explicitly mentioned but can be inferred from the context, GPT-4 is used to make an educated guess. After generation, we manually review and refine all outputs to ensure accuracy and authenticity, which results in a bank containing 3,229 authentic seeker personas. Detailed explanations of key attributes and extraction prompts can be found in the technical appendix.

Cognitive Reasoning for Social Awareness

Social awareness refers to the ability to understand and respond to the emotions, thoughts, and social dynamics of others during interactions. It is crucial for simulating ESC because these conversations inherently involve complex human emotions and social cues that require sensitivity and understanding. Existing methods often lack social awareness, focusing primarily on replicating surface-level response patterns without delving into the deeper cognitive processes that real human supporters use. This gap results in less effective simulations that fail to capture the nuances of real emotional support. To address this, we propose cognitive reasoning as a way to enhance social awareness. Cognitive reasoning involves a structured approach to analyze the seeker's emotional and situational context before generating a response. By integrating cognitive reasoning, we enable the simulation to produce more empathetic and supportive responses, ultimately improving the social awareness and overall effectiveness of ESC simulations.

Our design of cognitive reasoning is inspired by psychological theories emphasizing the importance of understanding and interpreting the mental states of others to provide effective emotional support (Wu et al. 2023; Beck et al. 2024). We also draw on the concept of chain-of-thought prompting (Wei et al. 2022), a method that structures reasoning into sequential, logical steps, enabling more complex and human-like decision-making. In our cognitive reasoning process, we define four types of reasoning nodes: *Situation*, *Thought*, *Action*, and *Strategy*. The *Situation* node captures the current emotional and contextual state of the seeker, such as feeling anxious due to a stressful work environment. The *Thought* node reflects the seeker's internal cognitive processes, like worrying about job performance. The *Action* node anticipates the behavioral outcomes that might result from these thoughts and emotions, such as avoiding work tasks. Finally, the *Strategy* node determines the most appropriate supporting stragety and its purpose, such as offering reassurance and suggesting stress management techniques. By sequentially traversing these nodes in the reasoning process when generation, we ensure that each response is tailored to the seeker's specific needs and is grounded in a deep understanding of their psychological state.

Socialized Simulation for Dialogue Generation

Prepared by the seeker-side personas and supporter-side cognitive reasoning process, we formulate dialogue generation as a task of transforming "`persona` + `reasoning` $\rightarrow$ `dialogue`", and prompt LLMs to accomplish it in an in-context learning way. Specifically, we start by selecting 50 high-quality conversations from ESConv, then manually supplement them with seekers' detailed personas and supporters' reasoning processes to form demonstrations. Next, we randomly select a demonstration and a help-seeking scenario from the persona bank, and prompt GPT-4 to generate a complete dialogue with cognitive reasoning. We here emphasize two rules for the generation: One is that the seeker's utterances should strictly follow the persona information to keep appropriate social disclosure; The other is that the sup-

porter's responses should be generated after a complete cognitive reasoning process to ensure effective social awareness. We then manually inspect the outputs and collect the **SSConv** corpus containing 3,229 synthetic ESC dialogues. Due to the space limitation, we provide the detailed LLM configurations, instruction prompts, and inspection rules in the supplementary technical appendix.

SSConv: Socially Simulated ESC Corpus

In this section, we provide a detailed overview about the characteristics of our constructed dataset, SSConv, including its statistics, quality, topics, strategies, and the role of personalized information.

Statistics SSConv comprises 3,229 dialogues, as detailed in Table 1. Each dialogue in our dataset consists of 18 to 40 utterances, which strikes a balance between providing sufficient substance and avoiding unnecessary verbosity. This range is more refined compared to ESConv, as it minimizes the risk of dialogues either lacking depth or becoming overloaded with irrelevant details. On average, our dialogues have 24 utterances. This range reduces the chance that the generated dialogues lack substance or contain excessive irrelevant information.

Category	SSConv	ESConv
# Sessions	3,229	1,300
# Utterances	77,337	38,365
Avg. # Utterances	24	29.5
Avg. Utterance Length	20.5	16.4
Min. # Max Turn	18	16
Max. # Max Turn	40	120
# Seeker Utterances	38,667	19,989
Avg. # Seeker Utterances	12	15.4
Avg. Seeker Utterance Length	17.8	14.8
# Supporter Utterances	38,670	18,376
Avg. # Supporter Utterances	12	14.1
Avg. Supporter Utterance Length	23.1	18.1

Table 1: Statistics comparison across different datasets.

Quality To evaluate the quality of the generated dataset, we adopt a manual evaluation method. Specifically, we randomly collect 30 dialogues from each of the two variants of SSConv: with social awareness, and strategy only, as well as from other datasets such as ESConv, ExTES, and AugESC. 30 trained workers are employed, with each dialogue assessed by three different workers across the following criteria, scored from 0 to 3: (1) **Informativeness (Inf.)**: Detail in the help-seeker's description of their emotional problems. (2) **Understanding (Und.)**: Supporter's grasp of the help-seeker's experience and feelings. (3) **Helpfulness (Hlp.)**: Effectiveness in alleviating the help-seeker's emotional distress. (4) **Safety (Saf.)**: Ensuring the content of the conversation is safe. (5) **Specificity (Spe.)**: Accuracy in reflecting the help-seeker's specific emotional situation. (6) **Human-likeness (Hlk.)**: Naturalness and anthropomorphism in the dialogue. The results are reported in Table 2. The higher

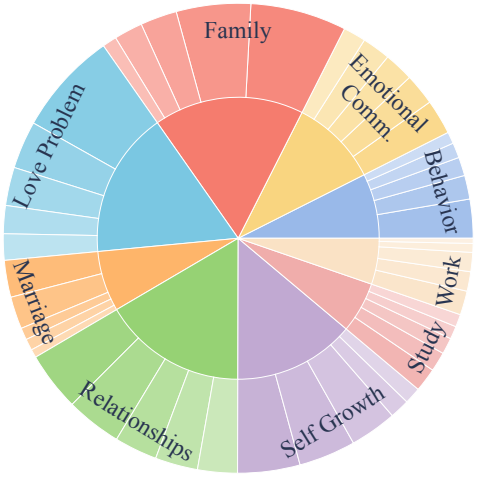

Figure 3: Nine primary topics in SSConv. We further detail the top-5 sub-topics in the appendix.

scores being better for all metrics. With the addition of social awareness, the quality scores are the highest across all dimensions, highlighting the importance of our reasoning process for emotional support simulation. Notably, our SSConv received a full score of 3.00 in the safety criteria, with all workers agreeing that the conversation content was completely free of offensive or sensitive content. Even with only the strategy component, SSConv still outperforms the human-written ESConv, demonstrating the high quality of our approach. The 16-category strategy-based ExTES also achieved strong results, indicating that a more detailed strategy classification is beneficial for ESC tasks. The detailed evaluation method is provided in the supplementary technical appendix.

	Inf.	Und.	Hlp.	Saf.	Spe.	Hlk.
SSConv	**2.76**	**2.86**	**2.79**	**3.00**	**2.75**	**2.57**
SSConv (Stra.)	2.64	2.51	2.60	**3.00**	2.60	2.47
ESConv	2.48	2.49	2.16	2.89	2.17	2.25
ExTES	2.73	2.68	2.50	2.96	2.63	2.55
AugESC	2.04	1.58	1.46	2.83	2.02	1.82

Table 2: Quality assessment with human evaluation. "Stra." means only using the strategy node for dialogue generation.

Topic To ensure the diversity and broad coverage of emotional support dialogues, it is essential to include comprehensive emotional support dialogue scenarios. Drawing on extensive literature related to psychological counseling (Burleson 2003) and insights from previous research on emotional support (Reblin and Uchino 2008; Meng and Dai 2021; Shensa et al. 2020; Graham et al. 2019), we develop a comprehensive taxonomy that covers a wide range of emotional topics and refine it using the keywords extracted from real help-seeking information (Sun et al. 2021). Currently, we identify 9 primary topics and 102 subtopics with emotionally impact. This is nearly three times the amount of

the previous maximum, which consisted of 36 types (Zheng et al. 2023c). These topics cover various aspects of daily life, addressing seekers' diverse emotional needs. Our expanded scope enhances the content and breadth of emotional support scenarios.

Strategy To understand the distribution of response strategies at different dialogue stages, we divide a conversation with N utterances into four equal stages. The k-th utterance by the supporter uses strategy S, represented by its position k/N. We then calculate the proportion of different strategies used within these stages. As shown in Figure 4, the distribution trends of emotional support strategies align closely with those used by real supporters in the crowdsourced ESConv dataset. From a stage perspective, we observe supporters generally follow the helping skills framework sequence: Exploration→Comforting→Action, but also make flexible adjustments. Delving deeper into the use strategies, we find that self-disclosure is relatively limited. This highlights the importance of active listening and aligns with the fact that the LLM supporter lacks a personal setting. There is an increase in questions, providing suggestions, and sharing information, which is consistent with our enriched seeker-side information and reasoning process.

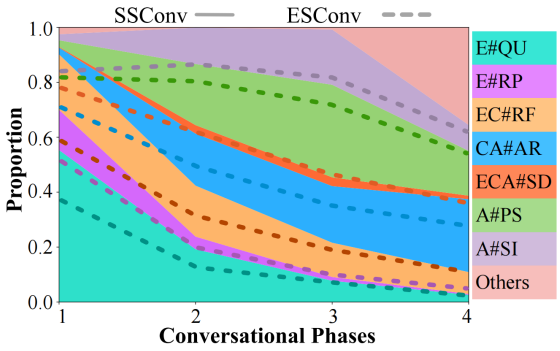

Figure 4: Strategies distribution in SSConv. Details of strategies can be found in the appendix.

To provide further insights into strategy transitions, we visualize these transitions in Figure 5 to illustrate the most common strategy flow patterns among the top five strategies. Several distinct patterns emerge from the visualization. The sequence `E#Qu.→E#Qu.→EC#RF.→A#PS.→CA#AR.` is the most prevalent strategy sequence. These transitions indicate that supporters typically ask multiple questions to identify the seekers' issues and explore the seekers' situation before offering advice. The emotional supporters usually begin by understanding the feelings behind the seekers' distress, then proceed to provide relevant suggestions and reassure their thoughts, which aligns with expected practices.

Personalization The additional seekers' persona bank should have an impact on the utterances in the dialogue. The seeker is expected to share more personal information. Correspondingly, if the supporter provides tailored emotional support and suggestions, their responses should align with

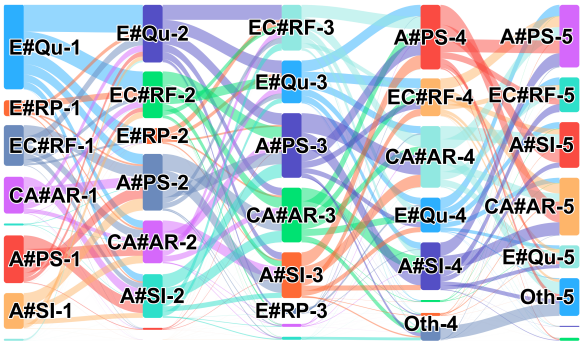

Figure 5: Strategies transition in SSConv. "E, C, A" denote the "Exploration, Comforting, Action" stages.

the seeker's specific information. To verify the impact, we examine whether the utterances contain information in the persona bank, such as the description of the situation, the help-seeking event, and the expectations for the consultation, etc. Specifically, a comparative analysis is conducted to assess the similarity between the utterances and the corresponding persona information, with a randomly sampled persona from another individual used as a comparison. We calculate the proportion of words in the utterance that overlap with the words in the persona information, reporting this as word overlap. Additionally, we compute the cosine similarity between the utterance embedding and the corresponding seeker persona bank embedding, reporting this as an embedding similarity.

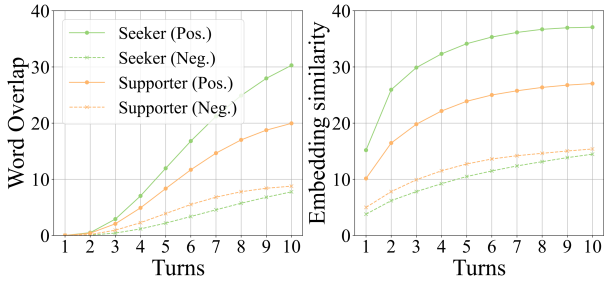

Figure 6: Persona coverage as the conversations progress. Pos.: the corresponding seeker's persona. Neg.: the randomly selected persona from another individual.

The results are shown in Figure 6a and Figure 6b. The initial utterance typically falls within the greeting sentence, which generally doesn't involve much personal information. The conversation then progresses into the Exploration phase of the helping skills framework, where a gradual understanding of personal information begins, leading to a rapid increase in relevance to the persona bank. As the dialogue deepens, the focus shifts from exploring unknown information to utilizing already-known details, causing the curve to level off. The supporter's trajectory resembles that of the seeker, although the proportion of persona-related information is slightly lower due to the need to integrate helping

skills and psychological knowledge. However, it remains higher than that of the comparison persona. This indicates that in our dialogues, the seeker and supporter effectively tailor emotional utterances to the seeker's specific issues.

Experiments

Experiment Settings

Corpora To verify the effectiveness of the proposed SocialSim framework for ESC simulation, we conduct comparative experiments between the generated **SSConv** corpus and existing datasets, including **ESConv** (Liu et al. 2021b), **AugESC** (Zheng et al. 2022), and **ExTES** (Zheng et al. 2023c). Specifically, ESConv is a crowdsourced dataset containing 1,053 dialogues. AugESC is created by first training GPT-J on ESConv's dialogues, then using it to extend the brief situations from EmpatheticDialog (Rashkin et al. 2018) to generate 65,000 new dialogues. ExTES is produced by providing ChatGPT with definitions of support strategies and an example dialogue from ESConv, resulting in 11,177 new dialogues. It is important to note that the two synthetic datasets, AugESC and ExTES, significantly outnumber SSConv, which contains only 3,229 dialogues. We select two test sets: SSConv-test is split from SSConv with the ratio train:test=9:1, representing a diverse range of help-seeking scenarios across various topics. ESConv-test consists of 200 held-out dialogues from ESConv, validating the model's performance on in-domain scenarios with more focused and limited topics.

Models We train a consistent backbone language model, Llama-2-7b (Touvron et al. 2023), across all four datasets. The training is conducted for 5 epochs using the AdamW optimizer (Loshchilov and Hutter 2019) with LoRA (Hu et al. 2021), a learning rate of 5e-5, and a batch size of 8 on one Tesla V100 GPUs. For clarity, we add a "$\circ$" after the corpus name to indicate the model trained on it. Additionally, considering that SSConv also includes cognitive reasoning contents after generation, we train an enhanced model, SSConv$^\bullet$, where the reasoning content is first generated and then followed by the final response during both training and inference stages. We compare the performance of SSConv$^\circ$ and SSConv$^\bullet$ to further investigate the impact of cognitive reasoning on the machine supporter.

Metrics For automatic evaluation, we adopt several standard metrics widely used in existing work, including word overlap-based metrics BLEU-{1, 2} (B-1, B-2) (Papineni et al. 2002), ROUGE-L (R-L) (Lin 2004) and METEOR (Banerjee and Lavie 2005), embedding-based metrics Extrema (Liu et al. 2016), and diversity metrics Distinct-{1, 2} (D-1, D-2) (Li et al. 2016). Additionally, since the scales of different metrics vary, we designed a normalized average (NAvg) metric to more intuitively reflect the model's overall capability. Specifically, we use ESConv results as the reference point, calculate the ratio of the experimental results to the ESConv value for each metric, and then average these ratios across all dimensions.

For human evaluation, we ask the participants to evaluate the models based on the following five aspects: 1) Fluency: the fluency and understandability of the model's responses. 2) Identification: how deeply the model explored the participant's situation and its effectiveness in identifying problems. 3) Comforting: the model's skill in providing comfort. 4) Suggestion: the helpfulness of the model's suggestions. 5) Overall: the participant's overall preference for emotional support. The metrics in 2), 3), and 4) correspond to the three stages in the ESC framework.

Main Results

Automatic Evaluation The automatic evaluation results for all models are presented in Table 3. On the broader-topic SSConv-test set, SSConv$^\circ$ significantly outperforms models trained on other datasets, demonstrating the high quality of dialogues generated by SocialSim's socialized simulation of emotional support conversations. Furthermore, the enhanced version, SSConv$^\bullet$, which explicitly generates the cognitive reasoning process before producing responses, shows even greater improvements. This highlights that reasoning before responding not only benefits synthetic data generation but also contributes significantly to model training, despite the potential increase in inference costs. In contrast, on the ESConv test set, all models exhibit similar performance levels to the original ESConv$^\circ$, with no significant differences observed. This indicates that the synthetic data does not compromise the model's ability to perform on in-domain topics.

Among the baselines, ExTES$^\circ$ demonstrates better performance on both test sets compared to the original ESConv$^\circ$. This advantage may stem from ExTES's use of more refined support strategies in its data synthesis, which enhances the model's ability to generate more effective responses. On the other hand, AugESC$^\circ$ shows relatively weaker performance, likely due to its use of a smaller 6B-level LLM for generation. While the automatic evaluation results provide evidence of SSConv$^\circ$'s superiority, it is important to note that these metrics primarily measure semantic overlap with golden responses and do not fully capture the model's true performance. Therefore, we conduct further human evaluations to gain a more comprehensive understanding.

Interactive Human Evaluation We select the ESConv and ExTES models, which perform well in automatic evaluations, for a manual evaluation against SocialSim. We hire 30 workers, each of whom engages in 2 to 3 sessions with each of the three models, with at least 8 turns per conversation. In total, 216 sessions were collected, and the results are summarized and presented in the Table 5. SocialSim outperforms ESConv across all dimensions and generally performs better than ExTES. This suggests that incorporating personal information and reasoning chains can help seekers feel more respected and attended to, thereby enhancing user experience.

Ablation Study To evaluate the impact of various components on the overall performance, we conducted an ablation study by systematically removing specific nodes in the cognitive reasoning process. The results are summarized in Table 4. The ablation study reveals that the configuration with all components—Situation, Thought, Action, and Strategy—achieves the highest performance, indicating the

Test Set	Models	B-1	B-2	R-L	METEOR	Extrema	D-1	D-2	NAvg
SSConv-test	ESConv$^\circ$	22.05	8.96	18.27	15.67	47.90	3.47	22.15	1.000
	ExTES$^\circ$	28.56	14.54	23.02	23.32	49.04	2.90	19.00	1.198
	AugESC$^\circ$	19.15	7.21	16.50	13.73	47.04	3.47	23.24	0.926
	SSConv$^\circ$	31.44	17.20	24.65	24.75	49.96	3.55	22.77	1.338*
	SSConv$^\bullet$	**32.19**	**18.34**	**26.02**	**25.85**	**50.96**	**3.55**	**23.52**	**1.390***
ESConv-test	ESConv$^\circ$	18.82	7.90	**18.89**	15.94	49.54	4.41	25.39	1.000
	ExTES$^\circ$	**23.32**	**9.19**	18.66	**17.98**	**48.28**	3.79	22.02	**1.031**
	AugESC$^\circ$	15.00	5.88	16.94	13.63	47.91	4.55	**26.34**	0.904
	SSConv$^\circ$	21.47	7.93	17.57	17.10	47.15	4.08	25.08	1.002
	SSConv$^\bullet$	21.73	8.00	17.36	16.36	47.26	**4.57**	25.87	1.017

Table 3: Results of automatic evaluation. The best results are highlighted in **bold**. The second-best results are underlined. Results with $*$ are significantly better than baselines ($p < 0.05$) based on a one-tailed unpaired t-test.

Situation	Thought	Action	Strategy	B-1	B-2	R-L	METEOR	Extrema	D-1	D-2	NAvg
a✗	✗	✗	✗	31.44	17.20	24.65	24.75	49.96	3.55	22.77	1.338
✗	✓	✓	✓	31.97	17.76	25.56	25.69	**50.39**	**3.61**	24.12	1.379
✓	✗	✓	✓	31.96	17.74	25.48	25.47	50.35	3.53	24.10	1.372
✓	✓	✗	✓	32.00	17.71	25.37	25.47	50.44	3.59	24.41	1.376
✓	✓	✓	✗	**32.40**	18.18	25.74	25.80	50.59	3.45	23.01	1.378
✓	✓	✓	✓	32.19	**18.34**	**26.02**	**25.85**	50.96	3.55	**23.52**	**1.390**

Table 4: Ablation study on different nodes in the supporter's cognitive reasoning process.

SSConv$^\circ$ vs.	ESConv$^\circ$			ExTES$^\circ$		
	Win	Loss	Tie	Win	Loss	Tie
Fluency	**58**	3	11	27	7	**38**
Identification	**61**	3	8	**33**	13	26
Comforting	**57**	3	12	**27**	21	24
Suggestion	**60**	3	9	**33**	18	21
Overall	**65**	2	5	**37**	16	19

Table 5: Results of interactive human evaluation.

necessity of a complete logical sequence for optimal model performance. In contrast, the absence of all components results in the poorest performance, underscoring the importance of each node. Omitting only the initial node Situation or final node Strategy leads to a relatively minor decline, suggesting the model's partial ability to infer missing nodes. However, the removal of intermediate components, especially Thought, notably disrupts the logical flow, resulting in a more pronounced drop in performance, highlighting the critical role of maintaining the reasoning coherence.

Related Work

Emotional support conversation (ESC) is a dialogue generation task, where the model acts as the supporter to help the help-seeker alleviate emotional distress. Effective emotional support typically involves skills such as empathy, comfort, and providing advice. Liu et al. (2021a) introduce the ESC task and, through laborious worker training and quality control mechanisms, crowdsource the ESConv dataset. Existing ESC models typically enhance performance through structural improvements. Some approaches inject commonsense knowledge to improve understanding of help-seekers (e.g. MISC (Tu et al. 2022), C3KG (Li et al. 2022), GLHG (Peng et al. 2022)), while others employ cognitive reasoning to gradually infer the mental state of help-seekers (e.g. DialogueCoT(Chae et al. 2023), CueCoT (Wang et al. 2023)). Additionally, some models enhance response relevance by introducing detailed persona information (e.g., PAL (Cheng et al. 2023)). Although these methods make progress, their performance remains limited by existing datasets, which only supplement information without addressing deeper nuances. To address these limitations, some studies attempt to use large language models to expand datasets, such as AugESC (Zheng et al. 2023a) and ExTES (Zheng et al. 2023b). However, these expansions are still based on simple scenarios and dialogue examples. In contrast, our approach enriches the persona information on the help-seeker's side and incorporates cognitive reasoning on the supporter's side. This aims to construct a deeper, more comprehensive dataset that better simulates social interactions.

Conclusion

In this paper, we introduce SocialSim, a socialized simulation framework for emotional support conversations. By constructing a comprehensive persona bank and incorporating cognitive reasoning processes, SocialSim effectively simulates the socially-driven nature of ESC. Building on SocialSim, we generate SSConv, a large-scale synthetic corpus that demonstrates superior quality to existing datasets. We further train a chatbot on SSConv, achieving state-of-the-art performance in both automatic and human evaluations, further demonstrating the effectiveness of SocialSim.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (No. 62125604), the NSFC Key Project (No. 61936010), and the NSFC Project (No. 62441614). This work was also supported by Tsinghua Precision Medicine Foundation, Tsinghua University - Beijing Tsingshang Architectural Decoration Engineering Co., Ltd. Joint Institute for Smart Scene Innovation Design, and China Postdoctoral Science Foundation (No. 2023M741944).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Albrecht, K. 2009. *Social intelligence: The new science of success*. John Wiley & Sons.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Beaudoin, C.; Leblanc, É.; Gagner, C.; and Beauchamp, M. H. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10: 2905.

Beck, A. T.; Rush, A. J.; Shaw, B. F.; Emery, G.; DeRubeis, R. J.; and Hollon, S. D. 2024. *Cognitive therapy of depression*. Guilford Publications.

Beck, J. S. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Beebe, S. A.; Beebe, S. J.; Redmond, M. V.; and Salem-Wiseman, L. 2002. *Interpersonal communication: Relating to others*. Allyn and Bacon Boston.

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

Chae, H.; Song, Y.; Ong, K. T.; Kwon, T.; Kim, M.; Yu, Y.; Lee, D.; Kang, D.; and Yeo, J. 2023. Dialogue Chain-of-Thought Distillation for Commonsense-aware Conversational Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 5606–5632. Association for Computational Linguistics.

Cheng, J.; Sabour, S.; Sun, H.; Chen, Z.; and Huang, M. 2023. PAL: Persona-Augmented Emotional Support Conversation Generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 535–554. Association for Computational Linguistics.

Costa, P.; and McCrae, R. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01): 1999.

Graham, S. A.; Depp, C. A.; Lee, E. E.; Nebeker, C.; Tu, X. M.; Kim, H.-C.; and Jeste, D. V. 2019. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Current Psychiatry Reports*, 21.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Li, D.; Li, Y.; Zhang, J.; Li, K.; Wei, C.; Cui, J.; and Wang, B. 2022. C3KG: A Chinese Commonsense Conversation Knowledge Graph. *CoRR*, abs/2204.02549.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021a. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards Emotional Support Dialog Systems. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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. Online: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*. OpenReview.net.

Mead, S. 2014. *Intentional peer support: An alternative approach*, volume 1. Intentional Peer Support West Chesterfield.

Meng, J.; and Dai, Y. 2021. Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication*, 26(4): 207–222.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Sun, Y.; and Li, Y. 2022. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. In *Proceedings of the Thirty-First International Joint*

Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, 4324–4330. ijcai.org.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Annual Meeting of the Association for Computational Linguistics*.

Reblin, M.; and Uchino, B. N. 2008. Social and emotional support and its implication for health. *Current opinion in psychiatry*, 21(2): 201–205.

Riva, G. 2004. Cybertherapy: Internet and virtual reality as assessment and rehabilitation tools for clinical psychology and neuroscience. *(No Title)*.

Rogers, C. R. 1995. *On becoming a person: A therapist's view of psychotherapy*. Houghton Mifflin Harcourt.

Shensa, A.; Sidani, J. E.; Escobar-Viera, C. G.; Switzer, G. E.; Primack, B. A.; and Choukas-Bradley, S. 2020. Emotional support from social media and face-to-face relationships: Associations with depression risk among young adults. *Journal of affective disorders*, 260: 38–44.

Sun, H.; Lin, Z.; Zheng, C.; Liu, S.; and Huang, M. 2021. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. *ArXiv*, abs/2106.01702.

Sutton, J.; and Stewart, W. 2017. *Learning to counsel: How to develop the skills, insight and knowledge to counsel others*. Robinson.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.; and Yan, R. 2022. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 308–319. Association for Computational Linguistics.

Wang, H.; Wang, R.; Mi, F.; Deng, Y.; Wang, Z.; Liang, B.; Xu, R.; and Wong, K. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 12047–12064. Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, J.; Chen, Z.; Deng, J.; Sabour, S.; and Huang, M. 2023. COKE: A Cognitive Knowledge Graph for Machine Theory of Mind. *arXiv preprint arXiv:2305.05390*.

Zheng, C.; Sabour, S.; Wen, J.; and Huang, M. 2022. AugESC: Large-scale Data Augmentation for Emotional Support Conversation with Pre-trained Language Models.

Zheng, C.; Sabour, S.; Wen, J.; Zhang, Z.; and Huang, M. 2023a. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1552–1568.

Zheng, Z.; Liao, L.; Deng, Y.; and Nie, L. 2023b. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Zheng, Z.; Liao, L.; Deng, Y.; and Nie, L. 2023c. Building Emotional Support Chatbots in the Era of LLMs. *ArXiv*, abs/2308.11584.