

# STAMPsy: Towards SpatioTemporal-Aware Mixed-Type Dialogues for Psychological Counseling

Jieyi Wang<sup>1</sup>, Yue Huang<sup>1</sup>, Zeming Liu<sup>2\*</sup>, Dexuan Xu<sup>1</sup>, Chuan Wang<sup>3</sup>, Xiaoming Shi<sup>4</sup>,  
Ruiyuan Guan<sup>1</sup>, Hongxing Wang<sup>5</sup>, Weihua Yue<sup>6</sup>, Yu Huang<sup>1\*</sup>

<sup>1</sup>Peking University, Beijing, China

<sup>2</sup>School of Computer Science and Engineering, Beihang University, Beijing, China

<sup>3</sup>Beijing Jiaotong University, Beijing, China

<sup>4</sup>East China Normal University, Shanghai, China

<sup>5</sup>Xuanwu Hospital Capital Medical University, Beijing, China

<sup>6</sup>Peking University Sixth Hospital, Beijing, China

joysw@stu.pku.edu.cn, zmliu@buaa.edu.cn, hy@pku.edu.cn

## Abstract

Online psychological counseling dialogue systems are trending, offering a convenient and accessible alternative to traditional in-person therapy. However, existing psychological counseling dialogue systems mainly focus on basic empathetic dialogue or QA with minimal professional knowledge and without goal guidance. In many real-world counseling scenarios, clients often seek multi-type help, such as diagnosis, consultation, therapy, console, and common questions, but existing dialogue systems struggle to combine different dialogue types naturally. In this paper, we identify this challenge as how to construct mixed-type dialogue systems for psychological counseling that enable clients to clarify their goals before proceeding with counseling. To mitigate the challenge, we collect a mixed-type counseling dialogues corpus termed STAMPsy, covering five dialogue types, task-oriented dialogue for diagnosis, knowledge-grounded dialogue, conversational recommendation, empathetic dialogue, and question answering, over 5,000 conversations. Moreover, spatiotemporal-aware knowledge enables systems to have world awareness and has been proven to affect one's mental health. Therefore, we link dialogues in STAMPsy to spatiotemporal state and propose a spatiotemporal-aware mixed-type psychological counseling dataset. Additionally, we build baselines on STAMPsy and develop an iterative self-feedback psychological dialogue generation framework, named Self-STAMPsy. Results indicate that clarifying dialogue goals in advance and utilizing spatiotemporal states are effective.

**Datasets** — <https://github.com/JOY-SWang/STAMPsy>

## Introduction

One in eight people worldwide is living with mental health conditions, yet the growing demand for mental health care is facing significant challenges due to insufficient resources of existing consultants (World Health Organization 2023). The uneven distribution of face-to-face counseling resources and the high cost of therapy have exacerbated the situation of mental disorders (Ierardi, Bottini, and Riva Crugnola 2022).

\*Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Thanks to the development of generative AI tools, such as Large Language Models (LLMs), online counseling is gradually becoming a good alternative to traditional face-to-face counseling (Olawade et al. 2024; Song et al. 2023).

Recent studies have leveraged LLM to aid clients in obtaining better online psychological counseling (Ke et al. 2024; Wang et al. 2023), falling into two categories: ① task-oriented dialogues or recent LLM-based simulations in psychology theory assume that clients have explicit goals (comfort-seeking, therapy querying, etc.) (Wang et al. 2024c; Chen et al. 2023); ② non-task dialogue-modeling approaches with pre-trained models for single-turn QA or multi-turn empathetic dialogue (Liu et al. 2021a; Zheng et al. 2023). Almost none of these studies naturally mix different types of dialogue (called *mixed-type dialogues*). However, in real-world application scenarios, there are multiple dialogue types in psychological counseling. For example, as shown in Fig.1, the psychological counseling bots could proactively make QA or conversational recommendations after task-oriented dialogue for diagnosis to improve user experience. However, to the best of our knowledge, no previous research has been done on this challenge. To mitigate this challenge, we present a novel task, mixed-type dialogues for psychological counseling, as shown in Fig.1.

Besides, beyond verbal conversations and body language, time and environment are two important nonverbal cues that affect counseling (Sherry, Paula, and Cynthia 2009) and mental health (Woo and Postolache 2008; Levitan et al. 2019). Therefore, we delve into the spatiotemporal states for this task, leading the model to generate more world-aware answers. For recommending therapy, therapists generally give different advice for late-night insomnia and early morning awakening, relaxation exercises before bed, and avoiding long naps.

To facilitate the study of this task, we collate the first SpatioTemporal-Aware Mixed-type dialogues dataset for Psychological counseling (STAMPsy), stamped with spatiotemporal state and psychological helping skills, including five dialogue types: task-oriented dialogues for diagnosis, knowledge-grounded dialogue, conversational recommendation, empathetic dialogue, and QA, containing 5,006

Dialog Type	Task-oriented Dialogue for Diagnosis	QA	Recommendation	Knowledge-grounded Dialog	Empathetic Dialog
Goal	<b>Dialogue #1: Period: late at night 24:00 p.m.; Location: Home</b>		<b>Dialogue #2: Period: morning 6 a.m.; Location: Outdoor</b>		
Diagnosis	<b>User#1:</b> I'm especially anxious... It's <b>late</b> but I can't sleep at home at all. <b>Bot:</b> I'm sorry to hear about your recent anxiety and insomnia issues. .... not being able to fall asleep. Let's start by filling out an anxiety scale to get your mental state.		<b>User#2:</b> I'm especially anxious ... It's <b>6 a.m.</b> , and I didn't sleep at all last night. <b>Bot:</b> I'm sorry to hear about your recent anxiety and insomnia issues. Senior year is really stressful and the mind tends to be sensitive. It's normal for you to feel this way.		
Recommend Places	<b>User#1:</b> Can you recommend an outdoor place to take a break? <b>Bot:</b> Going out late at night may not be the safest option, especially considering you are a student. If your home has a patio or balcony, this may be a safer option.		<b>User#2:</b> ...recommend an outdoor place ...? <b>Bot:</b> Many places are still closed at 6 a.m., but it's a good time to enjoy the peace of the <b>early morning</b> . Recommend nearby <b>parks</b> within 500m.		
Empathetic Dialog	<b>User#1:</b> Thank you. <b>Bot:</b> Have faith that you can fall into sweet <b>dreams</b> and don't be anxious.		<b>User#2:</b> Thank you. <b>Bot:</b> You may feel tired today, but try to give yourself some grace. Wish you a good <b>day!</b>		
Knowledge Grounded Dialogue	<b>User#1:</b> Is there any other treatment other than what is recommended above? <b>Bot:</b> .....I suggest we use Cognitive Behavioral Therapy to help you cope better with these dilemmas. <b>User#1:</b> What is Cognitive Behavioral Therapy? <b>Bot:</b> Simply put, cognitive behavioral therapy is a therapy that works by adjusting an individual's thinking patterns and behavioral habits, ..., it helps you better cope with the challenges of reality.				
Reference Knowledge Graph					

Figure 1: An example of STAMPsy with spatiotemporal state and reference knowledge.

dialogues and 61,832 utterances, with at least three dialogue types in each dialogue of STAMPsy. It should be noted that data processing is guided by psychological professionals. Specifically, we propose organizing the data according to the professional psychological theory of “Case Conceptualization”, including modules to gain the client’s profile and personal information, spatiotemporal state, and goal, checked by professionals. As shown in Figure 1, to help clarify the client’s goal, the therapist proactively gives recommendations like scale testing and therapies, interspersed with emotional solace. Furthermore, the counselor’s reply **changes with time and environment**. Although both suffer from insomnia, client #1 hopes to fall asleep, while client #2 hopes to stay alert during the day. Therefore, the counselor offers different suggestions, relaxation ways, and coffee respectively, which vividly demonstrates the effectiveness of spatiotemporal state on mixed-type dialogue.

To promote research on mixed-type dialogues for psychological counseling, we conduct bench-marking experiments on STAMPsy for the three sub-tasks: Helping Skills Selection, SpatialTemporal State Processing(STSP), and Adaptive Retrieval Augmented Generation. Moreover, inspired by Asai et al. (2023), we propose a novel Iterative Self-Feedback framework termed **Self-STAMPsy** to self-reflect on the consultation process and optimize model, imitating the supervision report and fine-tuning pre-trained dialogue models for comparison with ① different instruction texts as input and ② adaptive retrieval-augmented generation from knowledge graph. Experimental results also demonstrate that utilizing the above four modules improves the answers.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to identify the new challenge that there are multiple dialogue types in psychological counseling for online clients.
- To mitigate this challenge, we propose the first Chi-

nese mixed-type dialogue dataset stamped with the spatiotemporal state and goals for psychological counseling (STAMPsy), where the spatiotemporal state of counseling is stamped to better connect the psychological dialog system with reality and explore clients’ potential emotional responses to their surroundings.

- We build baselines on STAMPsy and propose a novel iterative self-feedback framework Self-STAMPsy simulating the supervision process. Extensive experiments demonstrate that helping skill, spatiotemporal state and self-feedback retrieval can improve performance.

## Related Work

### Applications of LLMs in Psychology

Recently, large language models(LLMs) with professional expertise and much better inferential capability have the potential to converse like a real person(Wang et al. 2024a; Tao et al. 2024), leveraging their capabilities for tasks like reasoning and interaction. Especially in psychology, LLMs are proven capable of replacing human participants in experiments(APA 2024), like emulating human social dynamics(Park et al. 2023; Wang et al. 2024b). Therefore, we make LLMs serve as therapists and clients to generate data similar to counseling dialogues. However, previous research has mainly focused on simply establishing an identity for LLMs, with limited relative information and without considering complex scenarios in real-world counseling(Cho et al. 2023). We include the spatiotemporal state during counseling and the goal sequence of the client-counselor dialogue, which influences the client’s emotional state, and is crucial for assisting LLMs in better solving clients’ problems.

### Datasets for Mental Health Support

Research on mental health support typically relies on high-quality psychologically pertinent datasets, such as ESConv

Datasets	S.T.-aware	Mixed-type	Psy.	Dialogue Types
ESConv(Liu et al. 2021a)	×	×	✓	Emotional response generation
PsyQA(Sun et al. 2021)	×	×	✓	single-turn conversations for mental health support
AugESC(Zheng et al. 2023)	×	×	✓	Large emotional response generation
CPsyCoun(Zhang et al. 2024)	×	×	✓	Multi-turn dialogues with different consultation topics
BlendedSkillTalk(Smith 2020)	×	✓	×	Know., empathetic dialogue, chitchat
DuRecDial 2.0(Liu et al. 2021b)	×	✓	×	Rec., chitchat, QA, task-oriented dialogue
DuClarifyDial(Liu et al. 2022)	×	✓	×	Rec., know. chitchat, QA, task-oriented dialogue
MidMed(Shi et al. 2023)	×	✓	×	Rec., empathetic dialogue, know., QA, diagnosis-oriented dialogue
DuSinc(Zhou et al. 2022)	✓	×	×	Service information augmented dialogue
STAMPsy (Ours)	✓	✓	✓	Rec., chitchat, know., QA, diagnosis-oriented dialogue

Table 1: Comparison of STAMPsy with other datasets. “Psy.”, “S.T.-aware”, “know.”, and “rec.” stand for psychological counseling dialogue, spatiotemporal-aware knowledge, knowledge-grounded dialogue, and recommendation, respectively.

(Liu et al. 2021a), AugESC(Zheng et al. 2023), CPsyCoun(Zhang et al. 2024), PsyQA(Sun et al. 2021). Although these datasets enhance the performance of LLMs in psychology, most of them mix all types of dialogues, leaving the model unable to give more targeted answers. Table 1 lists high-quality psychological and mixed-type dialogue datasets. However, they primarily focus on pure psychological counseling or cannot capture and perceive other mixed information that might be used in the counseling process. Faced with the absence of a dataset for mixed psychological counseling conversations, we proposed STAMPsy, added mixed dialogue goals to conversations, and conducted in-depth research on the positive impact of the spatiotemporal state on counseling conversations. This could be the initial step towards a spatiotemporal perception of mixed dialogue similar to real-world psychological counseling.

### Mixed-type Dialogue Systems

Recent research on mixed-type dialogue has witnessed a marked upsurge, especially in open-source datasets. Some researchers train a unified, comprehensive conversation model(Madotto et al. 2020; Roller et al. 2020), which integrates multiple dialogue skills into a singular framework by combining various single-skill conversation datasets, such as persona-chat and task-oriented dialogue. Others focus on mixed-type dialogue datasets and models, designed to aggregate diverse dialogue skills to meet specific needs, like recommending music and places, but unable to solve psychological problems(Liu et al. 2020, 2021a). Compared with them, we compiled the first mixed-type spatiotemporal dialogue dataset for psychology, which examines the effectiveness of integrating counselor helping skills and fusing spatiotemporal state in online psychological counseling.

### Dataset Collection

This paper aims to construct a spatiotemporal-aware mixed-type dialogue dataset for online psychological counseling. To more authentically reflect the patterns of real-world psychological counseling, we devise multiple modules to mimic practical counseling case report patterns. The construction flowchart is presented in Figure 2.

**Data Source** Given privacy concerns, it is difficult to obtain real-world consultation records. Therefore, we analyzed

public cases and in-depth descriptions from popular counseling books(KAMA 2015; Sherry, Paula, and Cynthia 2009; Hill 2009). To better fit the online context, we also selected conversations from PsyQA(Sun et al. 2021), a Chinese dataset of psychological health support with the Helping Skills Theory. After collection, we filter private information, divide long cases into several meetings manually, and ultimately collect 4,000 single texts as the origin input.

### Reference Knowledge Extraction

Psychological counseling tends to be a long, multi-turn dialogue, and response quality hinges on communication coherence and personal information consistency. Inspired by Park et al. (2023), we use short-term memory to update the context. For long-term memory, continuous interactions, personal information, psychological knowledge, and spatiotemporal stamps help maintain persona consistency. We utilize LLMs serving as structured psychological assistants to capture information from original case text and collate them into triples in the form of  $[Domain|Slot|Value]$ . “Domain” includes “Personal Information”, “Spatial-Temporal Information” and “Psychological Knowledge”.

**Personal Information** Directed by 9-BOX Case Conceptualization Model(Meichenbaum 2009), we collate “Personal Information” with 9 slots step by step, including ①Personal Profile and Background, ②Problem Presentation, ③Comorbidity, ④Stressors, ⑤Efficacy, Adherence, and Satisfaction of treatments received, ⑥Strengths, ⑦Summary Risk & Protective Factors, ⑧Outcomes and ⑨Barriers.

**Spatial-Temporal Information** The increasingly popular online text psychological support enables counseling to take place anytime, anywhere(Liu and Gao 2021). However, previous online counseling dialogue systems cannot comprehend complex scenarios with limited information. Therefore, we focus on the spatiotemporal state to better connect the reality. Moreover, spatial effects, environment, and time are three important nonverbal behaviors in communication because people have emotional responses to their surroundings(Sherry, Paula, and Cynthia 2009). Accordingly, for each round of conversation, we extract and label non-private spatiotemporal state  $ST_i$  such as time, location, or weather based on regular expressions, and conclude their influence on mental health to spatiotemporal stamps.

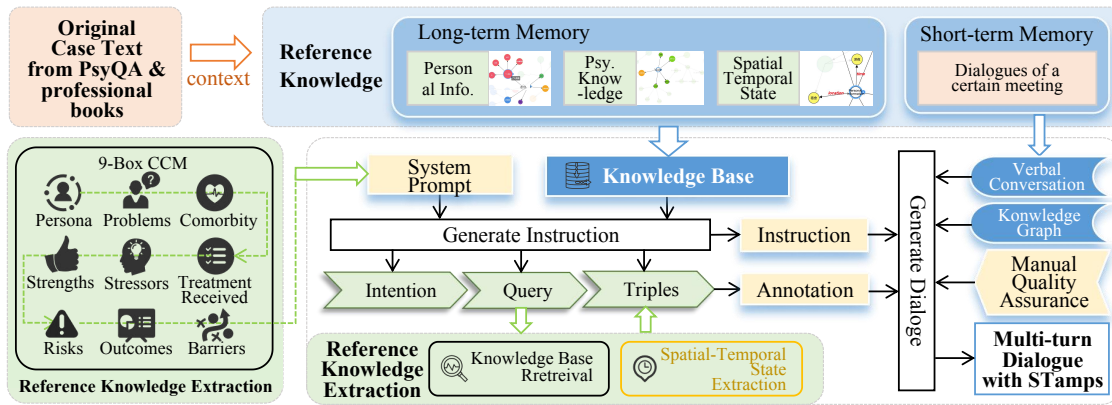


Figure 2: The collection workflow of STAMPsy. We extract reference knowledge from context under 9-Box CCM and then gain multi-turn dialogues with a multipart instruction. All the dialogues are annotated and revised by psychological experts.

**Psychological Knowledge** The collection of knowledge from the dialogues and the original knowledge base is converted to triples. Inspired by the construction and application of the temporal knowledge graphs (TKGs)(Trivedi et al. 2017, 2019), where the fact extends from a triple  $(s, p, o)$  (respectively representing the subject, predicate as a relation type, and object, and saved as memory after a manual check by psychological experts) to a quadruple with a timestamp  $t$ . We collect an event and redefine it as a spatiotemporal state stamped event represented as a quadruple,  $[Domain|Slot|Value|Stamp]$ . For “Relaxing Method Recommendation” slot, the value contains “drink coffee” with the timestamp “morning” but “read a book” to “night”.

Based on reference knowledge quadruples, we conclude a multipart system prompt shaped by instructions based on the CCM and expand single-turn text to multi-turn dialogues. What’s more, we invite clinical psychologists to check the consistency and correctness of the dialogues.

### Dataset Annotation

Dataset annotation involves labeling dialogue goals and exploring knowledge graphs, done by 8 psychology experts. The goals are split into counselor’s helping skill(Hill 2009) and client’s behaviors(Hill et al. 1992)(right of Fig. 2 in parentheses), with detailed guidance and knowledge triples provided to assist annotation.

**Target Helping Skills Sequence** During trial annotation, we extensively analyze the generated conversations above and find that the goals sequence patterns in the dialogue are similar to Counselor Helping Skills commonly used in psychology(Bickmore, Schulman, and Sidner 2011; Hill 2009; Lent, Hill, and Hoffman 2003). Thus, we assume that each counseling can be realized through an organized helping skill sequence, which may reveal the common the layout of high-quality mixed-type conversation in psychological counseling, simplifying the process of data annotation.

Inspired by MultiWOZ(Budzianowski et al. 2018), for each dialogue session, we provide a targeted dialogue helping skill sequence assigned by at least three of eight help-

ing skills(Hill 1986), which are “Immediacy”(Imme.), “Interpretations”(Inptn.), “Self-disclosures”, “Open questions”, “Feeling Reflection”(Feel.), “Restatements”(Rest.), “Information giving”(Info.), “Direct guidance”(Guid.) and “Others”. Direct guidance is classified to recommending places, relaxing ways, lifestyles, therapies, and music. Specifically, thanks to spatial stamps, the “Direct guidance: Recommended Place” label aims to guide users to a private place while counseling outdoors or in a crowded environment. These helping skills can also be mapped to the five types of conversation(see Tab. 2). The whole dialogue sequences with the first five distinct helping skills are shown in Fig. 3.

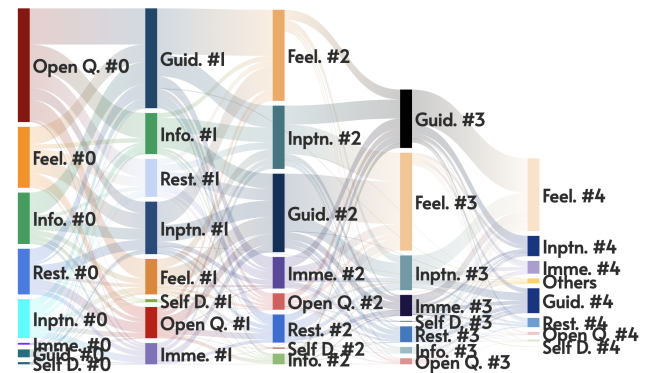


Figure 3: Sequence visualization of the common dialogue goal flow patterns within the first 5 counselor-helping skills.

**Knowledge Graph Construction** We incorporate knowledge graphs, archiving extensive knowledge in the form of triples, into psychological counseling to provide more accurate interactive questions and answers(Yang, Zhang, and Erfani 2020). After experts recheck the correctness and completeness of triples, annotators relate entities about disorder, symptom, therapy, etc., and relationships including disorder-symptom relation, disorder-therapy relation, etc, to construct a psychological knowledge graph  $K$ . For each dialogue, we

add a subset including persona and spatiotemporal stamps.

Goal Type	# Num	Avg.Len.	Helping Skill
Diagnosis	3745	85.71	Imme., etc.
QA	4500	89.82	Intpn., etc.
Knowledge-grounded dialog	4857	67.06	Info., etc.
Place	89	75.15	
Relaxation	1128	82.72	
Recommend	2210	<b>97.48</b>	Direct
Lifestyle	1777	91.69	Guidance
Therapy	137	74.46	
Music	5330	89.3	
Total	<b>7229</b>	66.90	Feel., etc.
Empathetic dialog			

Table 2: The number and the average length of goal types and counselor helping skills.

## Data Quality Audit and Analysis

All annotated samples are thoroughly reviewed to identify low-quality dialogues and remove contentious illusions or hazardous information. Substandard samples are re-annotated. Following Liu et al. (2020), we employ two senior psychologists for data quality evaluation on 1,000 randomly sampled dialogues. Specifically, they assign “1” for dialogues following annotation guidance, and “0” for the rest. The final average evaluation score of our dataset is “0.91”, with a Kappa value(Tinsley and Weiss 1975) of “0.84”, indicating that the dialogues are of high quality. Finally, we propose a Chinese SpatioTemporal-Aware Mixed-type multi-turn dialogues dataset for Psychological counseling. As shown in the bottom right corner of Fig.2, each meeting consists of multiple rounds of dialogues with at least three distinct Helping Skills.

Table 2 shows the number and the average length of goal types and counselor helping skills. Empathetic dialogue is most common because emotional solaces such as encouragement, validation, and consolation are important in counseling(Lee, Seo, and Kim 2007). In contrast, diagnosis is relatively rare, where external knowledge and backgrounds are extra required. Furthermore, we found that Diagnosis, QA, Know., and Rec. these four types of activities are more inclined to contain psychological knowledge, which means helping skills like Imme., Intpn., Info., and Guid. tend to involve external knowledge, and that answer can be enhanced with RAG. Then we optimized our generation template.

## Methodology

### Overview

Recently, the methodologies for different types of dialogues have gradually shifted to generative and end-to-end modeling. Following these trends, we propose a pre-trained mixed-type dialogue model **Self-STAMPsy** based on Qwen2(Bai et al. 2023). During training, a dialogue with a sequence of utterances between a client and a counselor is given. Then the dialogue is processed into a set of samples  $\{(c_i, t_i)\}$ , where  $t_i$  is  $i_{th}$  target counselor response,  $c_i$  is the concatenation of all former utterances before  $t_i$ , and  $D$  is the train set. Dialogue generation is formulated as an end-to-end di-

alogue generation(E2EDG) task, which aims to generate  $t_i$  conditioned on  $c_i$  as a query into LLMs.

Accordingly, Self-STAMPsy has four modules, helping skills selection, spatiotemporal state extraction, adaptive retrieval augmented generation, and iterative self-feedback. Helping skills selection and spatiotemporal state extraction aim to obtain dialogue goals  $\tilde{g}_i$  and spatiotemporal stamps  $\widetilde{ST}_i$ , respectively. Moreover, by integrating a novel iterative self-feedback mechanism, we get the final response  $t_i$ . The framework of Self-STAMPsy is plotted in Figure 4.

### Helping Skills Selection

The Helping Skills Selection sub-task takes dialogue context and current dialogue goal sequence as input, then outputs predicted counselor helping skill. Formally, this task is regarded as a multi-class classification problem. Specifically, this module is divided into two baseline models, the dialogue goal prediction with or without context. The input of the prediction module is a dialogue context  $c_i$ . With context  $c_i$ , it outputs the counselor helping skill  $h_i$  and predicts recommended goal  $g_i$ . For the baseline model without contextual information, The classification process is formulated,

$$p_i = f(c_i),$$

where  $f$  is the classification function BERT(Kenton and Toutanova 2019) and  $p_i \in |\mathcal{R}|^{|\mathcal{G}|}$  is the predicted probability value,  $\mathcal{G}$  is the predefined category set. The dialogue activity  $g_i$  is selected as the predicted dialogue activity if the value of the dimension is the highest probability value in  $p_i$ . For the model with context, we input multiple consecutive sentences  $c_1, c_2, c_3, \dots$  to BERT in the form of [CLS] $c_1$ [SEP][CLS] $c_2$ [SEP][CLS] $c_3$ ... and compute the mean loss of [CLS] located at the beginning of each sentence. After Helping Skills Selection, to promote the effectiveness, we convert dialogue goal  $g_i$  into natural language with predefined templates, represented as  $\tilde{g}_i$ . If the predicted goal is “Recommend Therapy”, the converted instruction will be “The therapist will then design a therapy”.

### SpatioTemporal Stamp Processing

SpatioTemporal State Processing(STSP) contains two steps, spatiotemporal state extraction and stamp generation. The state  $ST_i$  can be obtained from the Memory Module. Subsequently, we utilize the position embeddings of  $c_i$  and  $ST_i$  as inputs and derive emotional state  $\widetilde{ST}_i$  by employing the mechanism previously analyzed for the goal. During training, we minimize the negative log-likelihood(NLL) loss:

$$\begin{aligned} \mathcal{L}_{NLL_{\widetilde{ST}_i}} &= -\mathbb{E} \log p(\widetilde{ST}_i | ST_i, c_i) \\ &= -\mathbb{E} \sum_{i=1}^n \log p(\widetilde{ST}_i | ST_i, c_i, \widetilde{ST}_{i < i}) \end{aligned}$$

where  $\widetilde{ST}_i < i$  denotes previously generated emotional state in query  $Q$ . Then based on the impact of time and place on one’s emotions, we convert  $S$  into emotional state  $\widetilde{ST}_i$ .

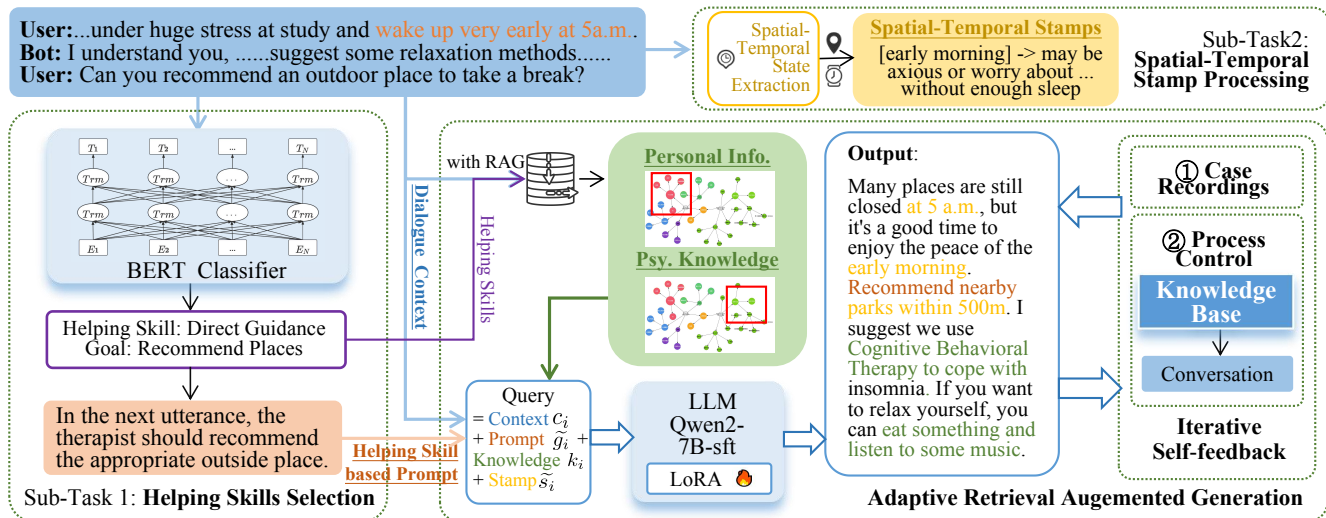


Figure 4: The framework of the proposed Self-STAMPsy. A detailed prompt template is open-sourced.

### Adaptive Retrieval Augmented Generation

RAG methods enhance the input of LLMs with pertinent retrieved passages, thereby minimizing factual inaccuracies in knowledge-intensive tasks (Gua et al. 2020; Lewis et al. 2020). Graph-RAG (Edge et al. 2024) has claimed the use of graphs in connection with LLMs and RAG is powerful. Therefore, we combine knowledge graph  $K$  with LangChain, a RAG tool, and get augmented response  $t_i$ .

**Iterative Self-feedback** To ensure interaction quality and dialogue consistency, we add the Iterative Self-feedback mechanism. We design a multipart prompt with two main modules: ① **Case Recordings** for quality assurance. In counseling, the counselor is required to reflect on their client between every two sessions, trying to get the causes of their problems, the underlying themes behind the problems, and the interventions that apply to helping the client. Referring to Hill (2009), we designed the model to self-reflect each turn, including six parts: explicit content, implicit content, barriers to defense and change, distortions, countertransference, and personal assessment. ② **Process Control** utilizes historical interactions to determine the optimal timing to end dialogues. If the conversation isn't over ( $i < n$ ), the output  $t_i$  will be used for iterative model optimization.

## Experiments

### Baselines

Following PsyCoun (Hu et al. 2024), we carefully select five fine-tuned robust LLMs, GPT4, and two Chinese psychological models as baselines to explore and evaluate the performance of various models. To optimize network performance, we perform LORA (Hu et al. 2022) and use LLaMa-Factory<sup>1</sup> based on STAMPsy and under the same parameters.

<sup>1</sup><https://github.com/hiyouga/LLaMA-Factory>.

Helping Skill	Prec.	Recall	F1
Immediacy	<b>64.84</b>	<b>26.92</b>	<b>38.05</b>
	59.14	22.58	28.09
Interpretations	<b>77.92</b>	<b>87.57</b>	<b>82.46</b>
	71.68	81.71	75.79
Open questions	93.97	93.80	94.38
	<b>96.80</b>	<b>97.39</b>	<b>96.08</b>
Feeling reflection	84.85	80.31	82.52
	<b>87.31</b>	<b>82.81</b>	<b>85.00</b>
Restatements	<b>70.03</b>	31.49	43.44
	68.95	<b>41.87</b>	<b>48.96</b>
Information giving	58.44	48.75	53.16
	<b>69.68</b>	<b>60.66</b>	<b>64.42</b>
Direct guidance	<b>83.34</b>	78.48	80.84
	82.80	<b>87.42</b>	<b>85.05</b>
Weighted avg.	78.89	70.47	73.13
	<b>79.40</b>	<b>74.73</b>	<b>75.97</b>

Table 3: Sub-Task1: Helping Skills Selection. The BERT strategy classification results for each helping skill. We compare the performance between the models without (the upper row) or with (the lower row) context.

### Experimental Setting

For the BERT classifier, we use a mini-batch size of 128 and the Adam optimizer with default parameters (fixed learning rate 0.001,  $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times e^{-8}$ ) (Kingma and Ba 2014). For finetuning, we add a dense output layer on top of the model with a cross-entropy loss function. Our experiments are conducted on the workstation with Linux 6.1.0-10-amd64, four NVIDIA A800 GPUs, and Debian 10.

### Evaluation

For Helping Skills Selection, we calculate the precision, recall, and f1-score in the test set as a classification task. In addition, to assess whether the model truly understands the

Baselines	Automatic Metrics					Manual Metrics				
	GHSC	STSP Acc.	BLEU-1/2	ROUGE-L	BertSim	Rel.	Info.	Human.	Help.	Empa.
ChatGLM4-sft	25.45	42.50	31.37/15.16	37.49	83.15	1.66	1.47	1.42	1.45	1.28
Baichuan2-sft	24.54	48.02	28.78/14.72	35.46	83.10	1.24	1.22	1.38	0.87	1.12
BLOOMZ-sft	9.10	51.13	16.91/6.58	27.76	76.84	0.24	0.10	0.94	0.11	0.10
LLaMa3-zh-sft	21.82	35.37	21.07/14.17	31.24	76.50	1.05	1.07	1.42	0.84	1.36
GPT4	69.09	52.50	38.63/21.46	43.86	87.51	1.85	1.70	1.60	1.85	1.80
MindChat	20.91	40.39	20.33/12.96	20.65	83.19	1.79	1.18	1.68	0.92	1.10
EmoLLM	23.64	41.15	23.80/19.35	27.52	78.69	1.68	0.55	1.55	0.82	1.57
Self-STAMPsy	<b>70.91</b>	<b>56.08</b>	<b>42.48/28.94</b>	<b>44.75</b>	<b>87.63</b>	<b>1.85</b>	<b>1.72</b>	<b>1.88</b>	<b>1.85</b>	<b>1.82</b>
-w/o Helping Skills	32.73	55.82	41.54/27.63	39.51	86.62	1.81	1.56	1.84	1.77	1.73
-w/o $ST_i$	65.45	48.32	40.06/26.86	38.09	86.19	1.60	1.50	1.83	1.70	1.31
-w/o Self-feedback	36.36	55.93	42.22/27.76	40.86	87.42	1.67	1.55	1.69	1.48	1.64
-w/o all modules	29.09	48.13	39.22/23.76	37.86	83.42	1.57	1.47	1.39	1.44	1.54

Table 4: Evaluation results of different models on GHSC and STAMPsy test set, the best scores are shown in bold. STSP Acc. means the result of sub-task 2 on accuracy.  $ST_i$  means the SpatioTmeporal state information. Other metrics are for generation.

helping skills, we devise a benchmark, called GHSC, following Hill Counselor Verbal Response Category System(Hill et al. 1992) and evaluate all models on GHSC. For STSP, we calculate the accuracy. As mentioned above, the problem of response generation is formulated as an E2EDG task. The input to the generation model is the concatenation of the dialogue context  $c_i$ , the predicted dialogue helping skill prompt  $\hat{g}_i$ , the reference knowledge  $k_i$  and the spatiotemporal stamp  $\widehat{ST}_i$ . The output is the counselor’s response  $t_i$ . To evaluate the performance of E2EDG, we utilize both automatic metrics and manual metrics. For automatic evaluation, we use several classical metrics, including BLEU- $n$  (Papineni et al. 2002), ROUGE(Lin 2004), and BertSim(Su 2020). For manual evaluation, we asked four psychotherapists to evaluate sessions from each end-to-end conversation with consensus judgment, encompassing five feature indicators: Relevance(Rel.), Informativeness(Info.), Human-likeness(Human.), Helpfulness(Help.), Empathy(Emp.).

## Results

The outcome of **Helping Skills Selection** is listed above in Table 3. Obviously, context information helps the classifier perform better and get a higher weighted F1-score, except “Immediacy” and “Interpretations” are relatively worse with context. This is reasonable because these two helping skills are generally at the very beginning of the conversation and lack preceding text. Thus, we add the knowledge base and use RAG to improve the results in the subsequent task.

In Table 4, the results of **GHSC** show that current LLMs are unable to accurately clarify goals and select the appropriate helping skills. With helping skills selected ahead, Self-STAMPsy performs better than the base model for nearly 40%. Additionally, the evaluation results of **STSP** and **E2EDG** show that we can effectively extract spatiotemporal information through all models but Self-STAMPsy outperforms all the baselines in all metrics and all tasks.

**Ablation Study** Table 4 shows the ablation results, where “w/o Helping Skills”, “w/o SpatialTemporal Info.”, “w/o Self-feedback” respectively means removing dialogue goal instructions, spatiotemporal state instructions, and iterative

feedback from our model. Results show that reducing any module of Self-STAMPsy leads to poorer results, illustrating the effectiveness of each module in Self-STAMPsy.

Take “a woman worrying about her kid’s grades” as an example, other models without the guidance of helping skills tend to explore how to improve children’s learning, which appears more like communication about education rather than addressing the emotional concerns of the client, quite different from golden professional replies. Subsequent conversations should focus on the client’s issues, so it is better to come into play the “emotional reflection”. Our model first correctly categorizes the helping skill, and based on the goal of “Reflection of Feeling”, Self-STAMPsy can achieve better effects and return to the client’s concerns, substantiating the significance of clarifying the goals of counseling in advance, which helps to prevent the direction from deviating from the original request. In addition, focusing on spatiotemporal states can lead to more realistic answers. While MindChat and EmoLLM tend to give single short emotional soothing, other LLMs often list “pale” suggestions, quite different from real counseling. Self-STAMPsy can better match practical scenarios, such as concentrating the dialogue’s focus on the visitor, emotions due to time and environment. Therefore, our benchmark also has impactful potential in online counseling and emotional dialogue.

## Conclusion

In this paper, we first identify the challenge of how to develop mixed-type dialogue systems for psychological counseling for clients to articulate their goals ahead of the counseling process. Then, we take a step forward by collecting STAMPsy, a mixed-type dialogue dataset for psychological health support with 8 counselor helping skills annotated, which contains 5k mixed-type conversations and 62K utterances. Furthermore, we propose Self-STAMPsy, a spatiotemporal-aware mixed-type dialogue generation model with adaptive RAG and a novel iterative self-feedback mechanism. Finally, experimental results demonstrate the effectiveness of Self-STAMPsy, showcasing the potential of LLMs as supplemental tools that can boost the accessibility and efficiency of online counseling services.

## Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFC2503903), the National Natural Science Foundation of China (62436006 and 62406015), the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (CI2023C062YLL) and Beijing Nova Program.

## References

- APA. 2024. What psychologists need to know about the evolution of generative AI. *Monitor on Psychology*, 55: 1.
- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *ArXiv*, abs/2310.11511.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Bickmore, T. W.; Schulman, D.; and Sidner, C. L. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of biomedical informatics*, 44(2): 183–197.
- Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5016–5026.
- Chen, S.; Wu, M.; Zhu, K. Q.; Lan, K.; Zhang, Z.; and Cui, L. 2023. LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. *arXiv:2305.13614*.
- Cho, Y.; Kim, M.; Kim, S.; Kwon, O.; Kwon, R. D.; Lee, Y.; and Lim, D. 2023. Evaluating the efficacy of interactive language therapy based on LLM for high-functioning autistic adolescent psychological counseling. *arXiv preprint arXiv:2311.09243*.
- Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv:2404.16130*.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org.
- Hill, C. E. 1986. An overview of the Hill counselor and client verbal response modes category systems. In Greenberg, L. S.; and Pinsof, W. M., eds., *The psychotherapeutic process: A research handbook*, 131–159. Guilford Press.
- Hill, C. E. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.
- Hill, C. E.; Corbett, M. M.; Kanitz, B.; Rios, P.; Lightsey, R.; and Gomez, M. 1992. Client behavior in counseling and therapy sessions: Development of a pantheoretical measure. *Journal of Counseling Psychology*, 39(4): 539–549.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.
- Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; and Wang, M. 2024. PsychoLLM: Enhancing LLM for Psychological Understanding and Evaluation. *arXiv:2407.05721*.
- Ierardi, E.; Bottini, M.; and Riva Crugnola, C. 2022. Effectiveness of an online versus face-to-face psychodynamic counselling intervention for university students before and during the COVID-19 period. *BMC psychology*, 10(1): 35.
- KAMA. 2015. *Psychological Cases*. China Fortune Press.
- Ke, L.; Tong, S.; Cheng, P.; and Peng, K. 2024. Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review. *arXiv:2401.01519*.
- Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Lee, S.; Seo, Y.; and Kim, D. 2007. Validation of the counselor activity self-efficacy scales. *Korean J Couns Psychother*, 19: 655–73.
- Lent, R.; Hill, C.; and Hoffman, M. A. 2003. Development and Validation of the Counselor Activity Self-Efficacy Scales. *Journal of Counseling Psychology*, 50: 97–108.
- Levitán, R. D.; Klein, R.; Bakshi, N.; Laposa, J.; Hill, S.; Kloiber, S.; et al. 2019. Morningness-eveningness scores predict outcomes differentially for depressed patients attending morning vs. afternoon day treatment streams. *Chronobiology International*, 36(11): 1581–1591.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, 9459–9474.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*.
- Liu, J.; and Gao, L. 2021. Analysis of topics and characteristics of user reviews on different online psychological counseling methods. *International Journal of Medical Informatics*, 147: 104367.
- Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021a. 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. Online: Association for Computational Linguistics.
- Liu, Z.; Wang, H.; Niu, Z.-Y.; Wu, H.; and Che, W. 2021b. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Pro-*

- cessing, 4335–4347. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Liu, Z.; Wang, H.; Niu, Z.-Y.; Wu, H.; Che, W.; and Liu, T. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1036–1049.
- Liu, Z.; Xu, J.; Lei, Z.; Wang, H.; Niu, Z.-Y.; and Wu, H. 2022. Where to Go for the Holidays: Towards Mixed-Type Dialogs for Clarification of User Goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1024–1034.
- Madotto, A.; Lin, Z.; Bang, Y.; and Fung, P. 2020. The Adapter-Bot: All-In-One Controllable Conversational Model. *arXiv:2008.12579*.
- Meichenbaum, D. 2009. Psycho-cultural assessment and interventions: The need for a case conceptualization model. In *13th annual meeting of the Melissa Institute, Florida*.
- Olawade, D. B.; Wada, O. Z.; Odetayo, A.; and et al. 2024. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, 3: 100099.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Shuster, K.; Smith, E. M.; et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Sherry, C.; Paula, N.; and Cynthia, O. 2009. *Interviewing and Change Strategies for Helpers: Fundamental Skills and Cognitive-Behavioral Interventions*. Brooks/Cole, 6th edition. Instructor's Edition.
- Shi, X.; Liu, Z.; Wang, C.; Leng, H.; Xue, K.; Zhang, X.; and Zhang, S. 2023. MidMed: Towards Mixed-Type Dialogues for Medical Consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8145–8157.
- Smith, E. M. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.
- Song, J.; Litvin, B.; Allred, R.; Chen, S.; Hull, T. D.; and Areán, P. A. 2023. Comparing Message-Based Psychotherapy to Once-Weekly, Video-Based Psychotherapy for Moderate Depression: Randomized Controlled Trial. *J Med Internet Res*, 25: e46052.
- Su, J. 2020. SimBERT: Fusion retrieval and generation. *spaces.ac.cn*.
- 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. In *Findings of the Association for Computational Linguistics*, 1489–1503.
- Tao, Z.; Chen, X.; Jin, Z.; Bai, X.; Zhao, H.; and Lou, Y. 2024. EVIT: Event-Oriented Instruction Tuning for Event Reasoning. *arXiv preprint arXiv:2404.11978*.
- Tinsley, H. E.; and Weiss, D. J. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4): 358–376.
- Trivedi, R.; Dai, H.; Wang, Y.; and Song, L. 2017. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *international conference on machine learning*, 3462–3471. PMLR.
- Trivedi, R.; Farajtabar, M.; Biswal, P.; and Zha, H. 2019. Dyrep: Learning representations over dynamic graphs. In *International conference on learning representations*.
- Wang, H.; Du, X.; Yu, W.; Chen, Q.; Zhu, K.; Chu, Z.; Yan, L.; and Guan, Y. 2024a. Learning to Break: Knowledge-Enhanced Reasoning in Multi-Agent Debate System. *arXiv:2312.04854*.
- Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
- Wang, L.; Zhang, J.; Yang, H.; Chen, Z.; Tang, J.; Zhang, Z.; Chen, X.; Lin, Y.; Song, R.; Zhao, W. X.; Xu, J.; Dou, Z.; Wang, J.; and Wen, J.-R. 2024b. User Behavior Simulation with Large Language Model based Agents. *arXiv:2306.02552*.
- Wang, R.; Milani, S.; Chiu, J. C.; Zhi, J.; Eack, S. M.; Labrum, T.; Murphy, S. M.; Jones, N.; Hardy, K.; Shen, H.; Fang, F.; and Chen, Z. Z. 2024c. PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. *arXiv:2405.19660*.
- Woo, J.-M.; and Postolache, T. T. 2008. The impact of work environment on mood disorders and suicide: Evidence and implications. *International Journal on Disability and Human Development*, 7(2): 185–200.
- World Health Organization. 2023. World Mental Health Day 2023. Web page. [Online; accessed: 2024-08-12].
- Yang, S.; Zhang, R.; and Erfani, S. 2020. GraphDialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. *arXiv:2405.16433*.
- Zheng, C.; Sabour, S.; Wen, J.; Zhang, Z.; and Huang, M. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of ACL*.
- Zhou, H.; Xu, X.; Wu, W.; Niu, Z.; Wu, H.; Bao, S.; Wang, F.; and Wang, H. 2022. Link the world: Improving open-domain conversation with dynamic spatiotemporalaware knowledge. *arXiv preprint arXiv:2206.14000*.