# Simulating Human-Like Counseling: A Path- and Scenario-Guided Framework for Psychological Support Dialogue

**Yuanchen Shi[1,2], Longyin Zhang[3], Maodong Li[1,2], Yibin Zheng[1,2], Xiuhong Wang[4], Fang Kong[1,2*]**

[1]School of Computer Science and Technology, Soochow University
[2]Jiangsu Key Lab of Language Computing, Suzhou 215123, China
[3]Aural & Language Intelligence, A*STAR Institute for Infocomm Research
[4]Institute of Science and Technology Information, Jiangsu University
20227927002@stu.suda.edu.cn, zhang_longyin@a-star.edu.sg, 20254027002@stu.suda.edu.cn, 20234227027@stu.suda.edu.cn, xiuhongwang@ujs.edu.cn, kongfang@suda.edu.cn

## Abstract

The growing demand for psychological support underscores the lack of high-quality counseling dialogue datasets, particularly in non-English contexts. We propose PGSim, a **P**ath-**G**uided **Sim**ulation framework that mirrors real counseling processes—symptom description, problem identification, cause analysis, strategy planning, and iterative adjustment. PGSim models each user scenario as a fine-grained quadruple {Group, Psychological Problem, Problem Cause, Support Focus} and guides dialogue generation through expert-annotated strategy paths. Real counseling dialogues and expert-edited samples are used to fine-tune two language models: a Dialog Generator for strategy-aligned dialogue creation and a Dialog Modifier for expert-level refinement. After automated and human verification, we construct the **C**hinese **P**sychological **s**upport **D**ialogue **D**ataset (CPsDD), containing 68K dialogues across 13 groups, 16 problems, 13 causes, and 12 support focuses. We further present the **C**omprehensive **A**gent **D**ialogue **S**upport **S**ystem (CADSS), which integrates profiling, summarization, strategy planning, and empathetic response. Experiments on CPsDD and ESConv demonstrate that CADSS achieves state-of-the-art results on Strategy Prediction and Emotional Support Conversation tasks.

**Code, Prompt, Dataset, and Case Study** — https://github.com/FakerBoom/CPsDD

**Scenario:** Group (Mid school students), Psychological Problems (Anxiety, Low self-esteem), Problem Cause (Exam failure), Support Focus (Innovative thinking, Logical thinking).

| Path | Dialogue |
|---|---|
| (System) Greeting | -Hello, have you been feeling stressed about school lately? Would you like to talk about your feelings? |
| (User) User Response | -A bit. *My exams haven't been good*, and I feel so useless. |
| (System) Comforting and Encouraging | -Exams can indeed bring a lot of pressure, but your grades don't fully define your worth. Have you done your best? |
| (User) User Response | -Yes, I feel like I've been trying really hard, but I still didn't do well. I feel like I'm letting my parents down, and it's really embarrassing. |
| (System) Guiding question | -Can you tell me what was the hardest part for you? Have you noticed anything you could improve on during your revision? |
| ... | ... |
| (System) Movie Recommendations | -I recommend you watch The Pursuit of Happyness. It tells the story of someone who persists through adversity and eventually achieves their dreams. Maybe it will inspire you. |
| (System) Guiding question | -Apart from studying, do you have any hobbies you enjoy? Do you feel more relaxed when engaging in those activities? |
| (User) User Response | -I like drawing, but I just haven't been in the mood to pick up a pencil lately. |
| ... | ... |
| (System) Closing Language | -The result of this exam doesn't define your future achievements. Failure is a necessary step on the path of growth. I hope you can slowly adjust your mindset and believe that you have the ability to face the next challenge. |
| (User) User Response | -Thank you for the recommendations and advice. I'll try to adjust my mindset and stop constantly doubting myself. |

**Severity of psychological problems**: Starting Level: 4 → Ending Level: 2

Figure 1: An example from CPsDD dataset.

## Introduction

Modern societal pressures have sharply increased the demand for psychological support, particularly via dialogue systems (Divya, Valsaraj, and Al Harthy 2022). Such systems show promise in applications like emotional counseling, customer service, and social media (Zhou et al. 2020). The Emotional Support Conversation (ESC) task has thus emerged, aiming to help users manage emotional distress (Peng et al. 2022). However, due to privacy concerns and the high cost of manual data collection, ESC datasets remain extremely scarce (Liu et al. 2021), especially in the Chinese context.

Large Language Models (LLMs) demonstrate strong capabilities in providing empathetic responses (Sorin, Brin et al. 2024). However, in the context of psychological counseling, their replies often remain lengthy, generic, and lacking in contextual empathy or actionable support (Wang et al. 2024). Existing Chinese psychological dialogue datasets are limited in either annotation scope (e.g., only labeling topic or psychological problem) or scenario diversity. To address these limitations, we propose **PGSim**, a **P**ath-**G**uided **Sim**ulation framework that mimics real-world psychological counseling workflows. As shown in Figure 1, each dialogue is guided by a *strategy path*, composed of expert-defined response strategies such as ”Comforting and Encouraging” or ”Guiding Question”, and grounded in a fine-grained user *scenario*—a structured quadruple {Group, Psychological Problem, Prob-

*Corresponding author

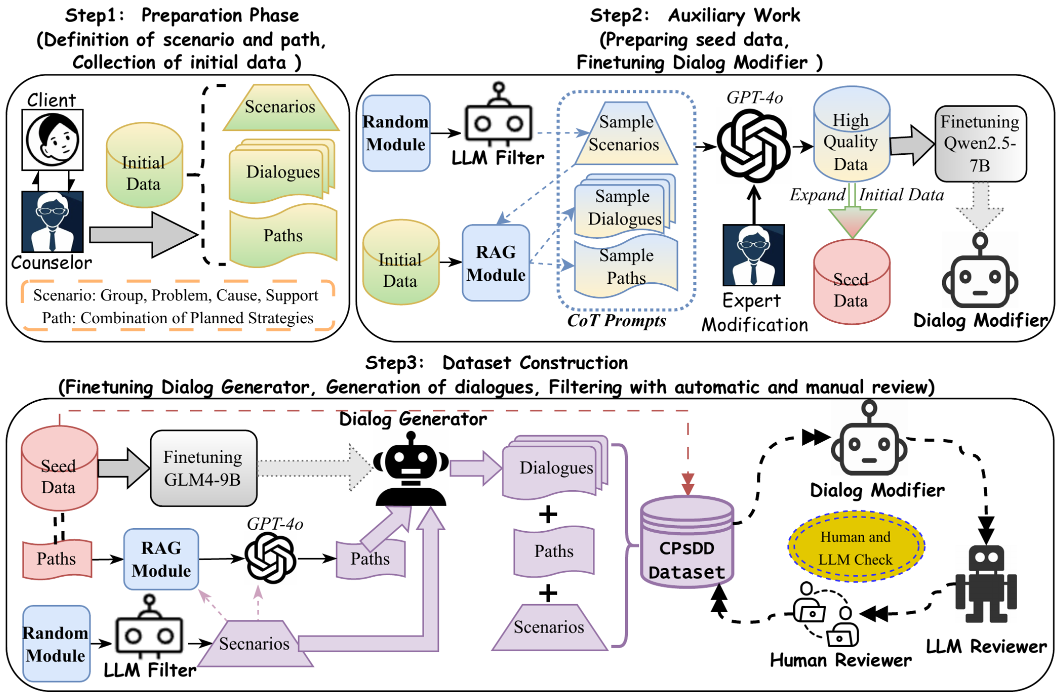



Figure 2: Overview of Path-Guided Simulation framework (PGSim).

lem Cause, Support Focus}.

Based on this design, we collect expert-annotated real counseling dialogues and their strategy paths. Using GPT-4o with Chain-of-Thought (CoT) prompting, we simulate new dialogues aligned with similar scenarios and paths. These are refined by psychological experts and used to fine-tune a Dialog Modifier that learns expert-level revision abilities. Together with the original data, they form seed data for finetuning a Dialog Generator to produce realistic and strategy-aligned dialogues. All generated data are then reviewed by both LLMs and human annotators to ensure quality. Compared with previous simulation frameworks (He et al. 2024; Ye et al. 2024), PGSim achieves structured, strategy-aware generation with scenario-level control, greatly improving controllability and realism.

We construct the **C**hinese **P**sychological **s**upport **D**ialogue **D**ataset (CPsDD), containing 68,136 dialogues across 13 groups, 16 psychological problems, 13 causes, and 12 support focuses. Each dialogue is annotated with its strategy path, user scenario, and the assessed severity change. Building on CPsDD, we develop the **C**omprehensive **A**gent **D**ialogue **S**upport **S**ystem (CADSS), which integrates four agents—Profiler, Summarizer, Planner, and Supporter. Experiments demonstrate that CADSS achieves state-of-the-art (SOTA) performance on both Strategy Prediction (SP) and Emotional Support Conversation (ESC) tasks across Chinese (CPsDD) and English (ESConv) datasets, showing strong bilingual generalization.

**Our main contributions are as follows:**

- We propose **PGSim**, a Path-Guided Simulation framework that mimics real-world psychological counseling by introducing structured strategy paths and fine-grained user scenarios for controllable, high-quality dialogue generation with expert-level refinement.
- We construct **CPsDD**, the first large-scale Chinese psychological support dataset covering comprehensive strategies and scenarios, and present **CADSS**, a multi-agent system integrating profiling, planning, summarization, and empathetic response.
- Extensive experiments show that CADSS achieves SOTA on SP and ESC tasks across CPsDD and ESConv, verifying its effectiveness and cross-lingual generalization.

## Related Work

Psychological support chatbots and datasets have recently garnered growing attention. Early studies explored online mental health consultations using open-domain dialogues (Razzaque and Stockmann 2016), with chatbots integrating natural language understanding and multimodal emotion recognition (Oh et al. 2017), or assisting in the diagnosis and treatment of depression and anxiety (Vaidyam et al. 2019).

With AI advancements, more sophisticated LLM-based systems have emerged. PsyChatbot (Chen et al. 2024) adopted a retrieval-based QA model under the CBT framework. SoulChat (Chen et al. 2023) and EmoLLM (Liu et al. 2024b) focused on enhancing LLMs' empathy and emotional reasoning. Other efforts addressed privacy (Lee, Lee, and Lee 2024) and the feasibility of AI support (Casu et al. 2024).

However, psychological dialogue datasets remain limited due to privacy and collection difficulties. EmpatheticDia-

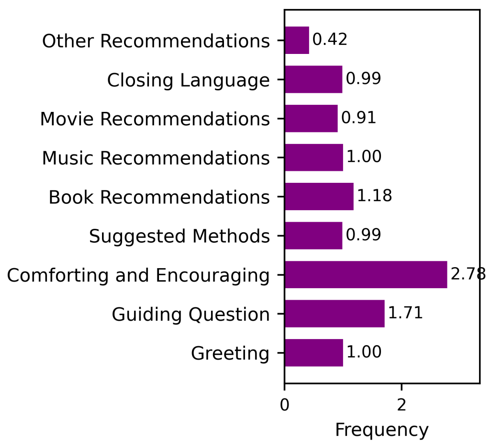



Figure 3: List and average frequency of strategies.

logues (Rashkin 2018), ESConv (Liu et al. 2021), and ES-CoT (Zhang et al. 2024b) focus on emotional support and strategy learning. ServeForEmo (Ye et al. 2024) utilizes LLMs for automatic dialogue generation. In the Chinese domain, while PsyQA (Sun et al. 2021), SmileChat (Qiu, He et al. 2024), CPsyCounD (Zhang et al. 2024a), and Psy-Chat (Li et al. 2023) offer valuable resources, they often lack multi-turn realism or fine-grained annotations.

Building on this, we propose CPsDD, a large-scale Chinese psychological support dialogue dataset with diverse scenarios and strategies, and develop CADSS, a multi-agent system aligned with real-world counseling workflows.

## Path-Guided Simulation Framework

### Preparation Phase

**Definition of Scenario and Path** As illustrated in Figure 2, we define the core components of a psychological support scenario. To mirror real counseling workflows—where clients describe symptoms, counselors identify problems and causes, and propose targeted strategies—we represent each **scenario** as a fine-grained quadruple: {Group, Psychological Problem, Problem Cause, Support Focus}. This structure aligns with actual clinical practice and ensures semantic diversity.

The taxonomy of these components is developed with three certified psychological experts (Shek 2002; Xiao et al. 2013; Pan, Wang, and Derakhshan 2023) experienced in counseling across schools, prisons, and hospitals. We define 13 user groups, 16 problems, 13 causes, and 12 support focuses (see Figure 4). This diversity enables realistic simulations, as identical problems (e.g., depression) may arise from different causes, requiring distinct strategies. Moreover, counseling approaches can vary across user groups (e.g., children vs. prisoners). Our schema captures such subtle variations, allowing flexible and fine-grained scenario generation.

To guide dialogue progression, we define a structured path—an ordered sequence of system response strategies and user replies. Each system turn follows a chosen strategy reflecting its support intention. As shown in Figure 3, we define nine common strategies from expert experience, including comforting, guiding, and multiple recommendation types (e.g., books, movies, music), supported by psychological research on the therapeutic value of artistic media (Shechtman 2008; Young 2012; Bradt and Dileo 2014).

**Collection of Initial Data** To establish the seed data for simulation, we collect 130 real-world Chinese psychological counseling dialogues, provided by licensed experts. All data are anonymized and released with informed consent. Each dialogue is annotated with its scenario quadruple and complete strategy path, ensuring consistent mapping between user conditions and support strategies. These annotated samples not only serve as high-quality training references but also ground the initial fine-tuning of our Generator and Modifier.

### Auxiliary Work

**Preparing Seed Data** After preparing the initial data, we generate a large pool of scenarios to bootstrap large-scale generation. A Random Module samples 200K user scenarios by selecting one user group and randomly combining 1–3 psychological problems, causes, and support focuses from the predefined taxonomy. We employ GLM4-9B (GLM et al. 2024), Qwen2.5-7B (Yang et al. 2024), and DeepSeek-R1 (Liu et al. 2024a) to filter out implausible or inconsistent cases, yielding 135K valid and diverse scenarios.

To ensure coverage across groups, 20 validated scenarios per group are selected, and similar expert-annotated dialogues and paths are retrieved as few-shot examples. These are organized into CoT prompts to guide GPT-4o in generating new dialogues conditioned on both scenario and dialogue path.

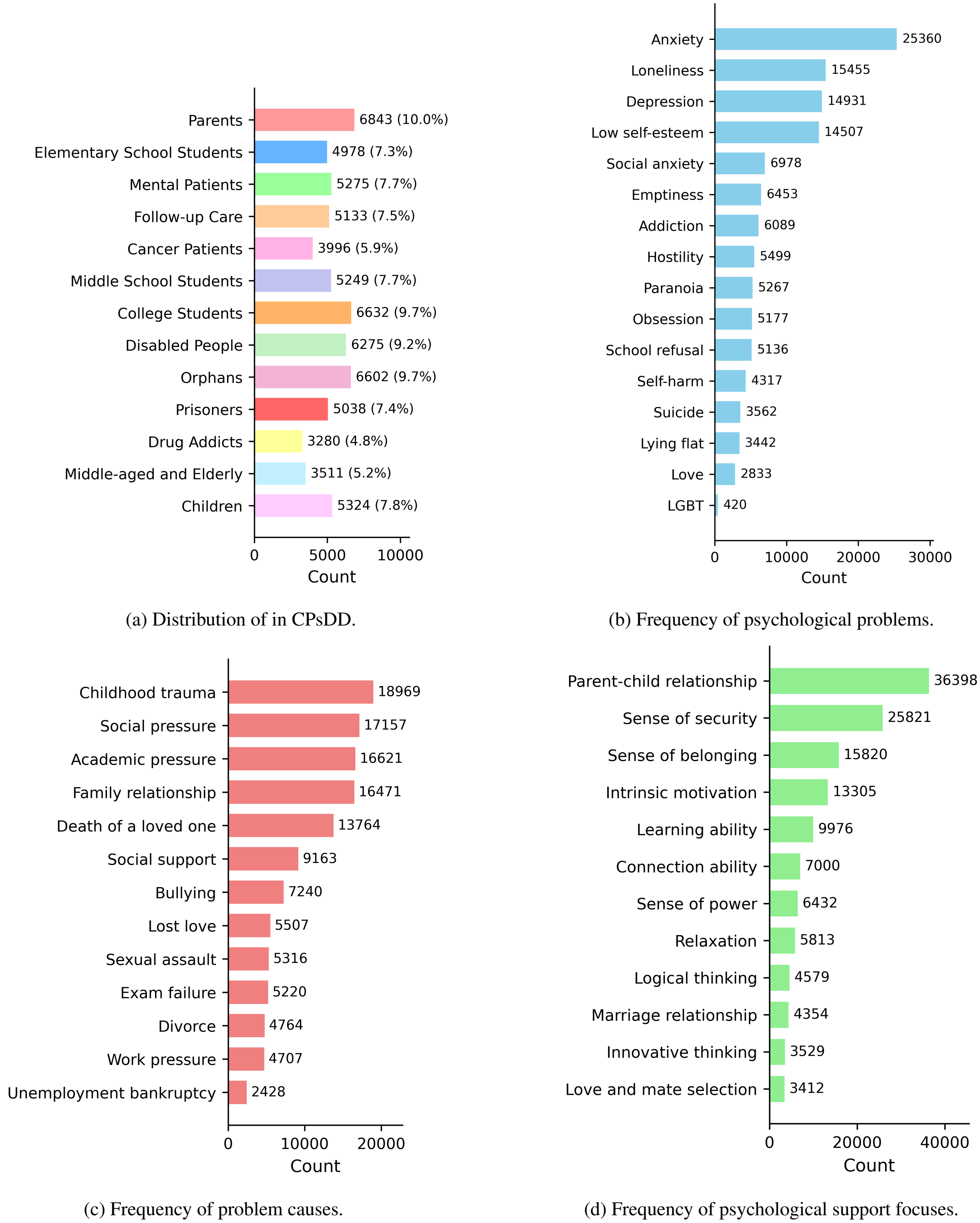
To ensure counseling quality, each GPT-generated dialogue is reviewed and refined by domain experts assigned by context (e.g., school psychologists for student cases, forensic counselors for prison samples). This targeted review maintains contextual accuracy while reducing expert workload. The expert-revised dialogues and initial annotated samples together form the complete Seed Data.

**Finetuning Dialog Modifier** We fine-tune Qwen2.5-7B using Low-Rank Adaptation (LoRA) (Hu et al. 2021) on GPT-generated and expert-revised dialogue pairs. The Modifier refines machine-generated dialogues to enhance relevance and clinical accuracy.

### Dataset Construction

**Finetuning Dialog Generator** We fine-tune GLM4-9B on the Seed Data using LoRA, obtaining a Dialog Generator that produces high-quality psychological support dialogues based on scenarios and strategy paths. Each sample contains a scenario quadruple and its dialogue path, enabling the model to learn both strategic control and semantic coherence.

**Generation of Dialogues** To scale up generation, we randomly sample validated scenarios. For each scenario, similar expert-reviewed samples from the Seed Data are retrieved to construct CoT prompts. GPT-4o simulates a suitable strategy path, which, together with the scenario and retrieved examples, is fed into the Dialog Generator, yielding about 135K raw dialogues aligned with structured scenarios and paths.




(a) Distribution of in CPsDD.

(b) Frequency of psychological problems.

(c) Frequency of problem causes.

(d) Frequency of psychological support focuses.

Figure 4: The frequency distribution of groups, problems, causes, and focuses in CPsDD.

**Filtering with Automatic and Manual Review** A multi-stage review pipeline ensures data quality. First, all dialogues are refined by the Dialog Modifier and those with fewer than 10 utterances are removed. GPT-4o then serves as an LLM Reviewer to score emotional effectiveness, coherence, and scenario consistency (1–10 scale). In the first round, 32K low-

| Datasets | Size | Utts. (System/User) | #Utt.len (System/User) | S | G | P | C | F | Language |
|---|---|---|---|---|---|---|---|---|---|
| ESConv | 1.3K | 29.3K (14.6K/14.6K) | 21.17 (19.90/22.45) | ✓ | × | ✓ | ✓ | × | EN |
| ServeForEmo | 3.8K | 62.9K (30.7K/32.1K) | 17.97 (15.25/20.56) | ✓ | ✓ | ✓ | ✓ | × | EN |
| SmileChat | 55.2K | 628.3K(318.2K/310.1K) | 81.29(56.90/106.32) | × | × | ✓ | × | × | ZH |
| SoulChat | 258.4K | 3.0M(1.5M/1.5M) | 65.61(89.98/41.41) | × | × | ✓ | × | × | ZH |
| CPsyCounD | 3.1K | 48.9K(24.4K/24.4K) | 44.84(56.02/33.67) | × | × | ✓ | × | × | ZH |
| PsyDTCorpus | 5K | 84.0K(42.0K/42.0K) | 45.01(55.72/34.30) | × | × | ✓ | × | × | ZH |
| CPsDD(ours) | 68.1K | 1.3M(0.7M/0.6M) | 61.01(83.25/32.04) | ✓ | ✓ | ✓ | ✓ | ✓ | ZH |

Table 1: Overall statistical comparison. *Utts.* represents utterances, # denotes the average, and *S,G,P,C,F* represents Strategy, Group, Problem, Cause, and Support Focus, respectively.

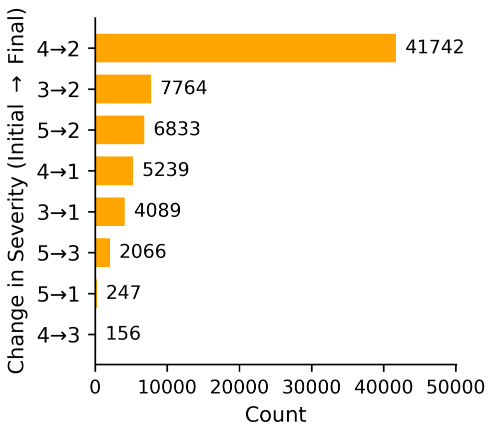



Figure 5: Degrees of relief of psychological problems.

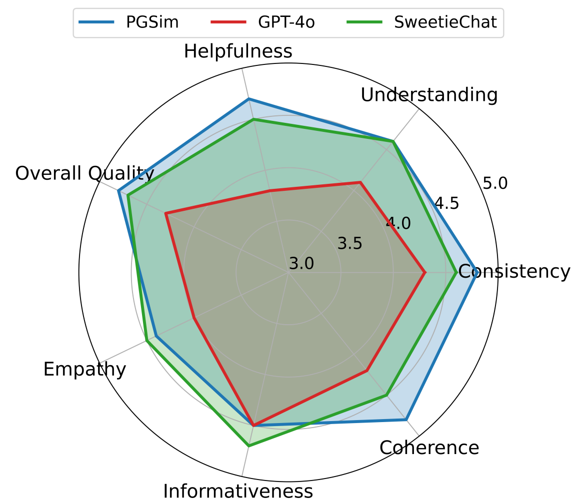



Figure 6: Human evaluations of different methods.

quality samples (score $\leq 6$) are discarded, 38K high-quality ones (score $\geq 9$) retained, and 65K mid-range dialogues (score 7–8) revised using Reviewer feedback and reprocessed by the Modifier. A second round of LLM review follows: samples scoring below 7 are discarded, 9–10 retained, and those scoring 8 sent to domain experts for manual refinement. Expert-revised dialogues are re-evaluated, and only those scoring $\geq 9$ in the final round are preserved.

Combined with the Seed Data, CPsDD contains 68K high-quality, scenario-rich psychological support dialogues.

# CPsDD Dataset

## Data Statistics and Analysis

**Dataset Comparison** Table 1 compares CPsDD with major psychological dialogue datasets in both English and Chinese. CPsDD surpasses English datasets such as ESConv (Liu et al. 2021) and ServeForEmo (Ye et al. 2024) in scale and utterances, offering more comprehensive and emotionally supportive responses. Compared with recent Chinese datasets including SmileChat (Qiu, He et al. 2024), SoulChat (Chen et al. 2023), CPsyCounD (Zhang et al. 2024a), and PsyDTCorpus (Xie et al. 2024), CPsDD achieves competitive scale and richer content. Moreover, it provides the most comprehensive label coverage—*Strategy, Group, Problem, Cause, and Support Focus*—capturing the full spectrum of psychological support strategies and user needs, which makes it particularly suitable for training dialogue systems.

**Data Distribution** CPsDD covers 13 user groups (Figure 4a), with Parents most and Drug Addicts least represented. Figures 4b–4d show the frequency of Problems, Causes, and Support Focuses. Anxiety, Loneliness, Depression, and Low Self-esteem are most prevalent, while Childhood Trauma, Social Pressure, Academic Pressure, and Family Issues dominate as causes. Support mainly centers on Parent–child relationships and Sense of security, reflecting current societal trends. Figure 3 presents the average frequency of response strategies, where Comforting and Encouraging occur most often as key elements of psychological support.

**Psychological Support Effectiveness** We employ GPT-4o-mini as a Judger to evaluate users' psychological severity based on the first three and last user utterances. As shown in Figure 5, most users' negative emotions are alleviated after counseling, with many showing mild improvement and some nearly recovered. On average, emotional severity decreases by at least two levels, indicating that dialogues in CPsDD effectively deliver psychological support.

## Human Evaluation

We randomly select 100 CPsDD samples and use GPT-4o and SweetieChat (Ye et al. 2024) to generate comparable dialogues. 50 Chinese undergraduates rate the dialogues (1–5

| Models | ACC↑ | PPL↓ | B-1↑ | B-2↑ | B-3↑ | B-4↑ | D-1↑ | D-2↑ | D-3↑ | R-L↑ | L-R |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen2.5(Yang et al. 2024) | 26.01 | 26.59 | 20.27 | 7.97 | 4.51 | 2.82 | 81.83 | 98.71 | 99.75 | 19.19 | 1.12 |
| DeepSeek-R1(Liu et al. 2024a) | 27.78 | - | 18.92 | 6.77 | 3.51 | 1.74 | 75.68 | 97.96 | 99.85 | 18.52 | 3.23 |
| GPT-4o | 33.17 | - | 21.26 | 7.33 | 3.44 | 1.87 | 74.31 | 98.69 | 99.89 | 17.93 | 1.83 |
| CKPI(Hao and Kong 2025) | 47.93 | 22.76 | 35.82 | 26.47 | **20.62** | 14.92 | 48.48 | 87.53 | 95.35 | 25.19 | 0.63 |
| SoulChat(Chen et al. 2023) | - | 25.93 | 20.86 | 7.82 | 4.00 | 2.33 | 78.17 | 98.34 | 99.68 | 19.47 | 1.10 |
| MeChat(Qiu, He et al. 2024) | - | 23.59 | 19.86 | 7.42 | 3.99 | 2.43 | 80.35 | 98.60 | 99.78 | 19.25 | 1.03 |
| PsyChat(Qiu et al. 2024) | - | 20.85 | 18.29 | 6.09 | 2.86 | 1.62 | 76.20 | 97.30 | 99.10 | 17.76 | 1.16 |
| MindChat(Xin Yan 2023) | - | **12.79** | 20.32 | 8.18 | 4.53 | 2.77 | 74.42 | 97.05 | 99.14 | 19.64 | 1.43 |
| EmoLLM(Liu et al. 2024b) | - | 14.09 | 20.90 | 9.71 | 5.53 | 3.31 | 72.29 | 97.82 | 99.54 | 21.48 | 2.51 |
| CPsyCounX(Zhang et al. 2024a) | - | 13.17 | 20.59 | 8.51 | 4.57 | 2.73 | 78.54 | 97.87 | 99.49 | 19.41 | 1.34 |
| PsyDTLLM(Xie et al. 2024) | - | 29.93 | 19.62 | 7.10 | 3.51 | 2.06 | **82.49** | 98.91 | 99.87 | 19.14 | 1.13 |
| CADSS(ours) | **80.98** | 21.57 | **37.08** | **26.50** | 20.47 | **16.54** | 81.92 | 98.92 | **99.90** | **37.97** | **1.00** |
| w/o planner | - | 21.05 | 25.82 | 13.80 | 9.66 | 7.25 | 80.89 | 99.05 | 99.87 | 25.71 | 1.08 |
| w/o summarizer | 80.91 | 21.99 | 36.22 | 24.87 | 19.63 | 16.04 | 81.46 | 98.87 | 99.83 | 37.72 | 0.96 |
| w/o profiler | 79.68 | 21.87 | 36.57 | 25.11 | 20.06 | 16.26 | 81.43 | 98.88 | 99.84 | 37.75 | 0.99 |
| w/o planner&summarizer | - | 22.19 | 25.40 | 13.31 | 9.24 | 6.87 | 81.15 | **99.07** | 99.87 | 25.26 | 1.08 |
| w/o planner&profiler | - | 21.87 | 25.69 | 13.79 | 9.67 | 7.29 | 80.54 | 98.97 | 99.84 | 25.79 | 1.12 |
| w/o summarizer&profiler | 79.75 | 21.84 | 36.16 | 24.80 | 19.56 | 15.97 | 81.24 | 98.86 | 99.83 | 37.57 | 0.97 |
| w/o all | - | 23.77 | 25.19 | 13.10 | 9.11 | 6.79 | 80.91 | 99.04 | 99.84 | 24.93 | 1.09 |

Table 2: Overall experimental results of different models on the SP and ESC tasks on CPsDD dataset.

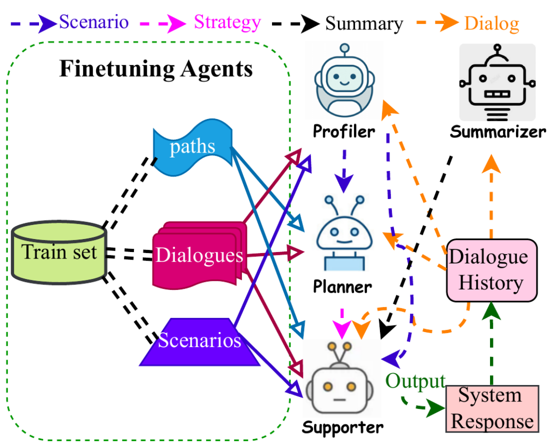



Figure 7: CADSS framework. Solid lines denote data for fine-tuning; dashed lines denote inputs during deployment.

scale) across seven dimensions: Helpfulness, Understanding, Consistency, Coherence, Informativeness, Empathy, and Overall Quality. As shown in Figure 6, PGSim-generated dialogues are more coherent, contextually consistent, and helpful. Although SweetieChat scores higher in informativeness and empathy, CPsDD achieves the best overall rating due to its balanced structure and realistic interaction flow.

## CADSS Multi-Agent System

We design the Comprehensive Agent Dialogue Support System (CADSS), a modular multi-agent framework (Figure 7), to deliver personalized and emotionally attuned psychological support. Inspired by the cognitive workflow of professional counselors, CADSS decomposes the support process into distinct functional roles.

CADSS consists of four collaborative agents: **Profiler** analyzes previous utterances to extract structured features, guiding subsequent agents with personalized information. **Summarizer** uses in-context prompting to condense the dialogue history and infer the client's psychological state (emotion, intent, and distress signals), enabling the system to track longitudinal shifts and contextual nuance. **Planner** predicts the next response strategy from a predefined taxonomy, conditioned on the context and user profile, thus ensuring structured intervention via strategy-path reasoning. **Supporter** generates empathetic, psychologically grounded responses aligned with the predicted strategy and user context, offering comfort, insight, or actionable guidance.

All agents are implemented based on Qwen2.5-7B, with the exception of the Summarizer, which operates via zero-shot prompting for efficiency. This multi-agent design mirrors real-world counseling workflows (diagnosis, planning, support) and facilitates component-wise finetuning, targeted evaluation, and future extensibility (e.g., integrating retrieval).

## Experiments

### Experimental Setup

We evaluate CADSS on both the CPsDD and ESConv datasets for SP and ESC tasks. The data is consistently split into an 8:1:1 ratio (train:dev:test) to fine-tune all agents. This unified multilingual setup promotes robust cross-lingual generalization. We compare CADSS with a range of strong baselines, including specialized mental health LLMs (Table 2).

For evaluation, we report: strategy accuracy (ACC), perplexity (PPL) (Brown et al. 1992), BLEU-n (B-n) (Papineni et al. 2002), ROUGE-L (R-L) (Lin 2004), Distinct-n (D-n) (Li et al. 2015), and Length-Ratio (L-R) (Tian et al. 2017) to assess generation quality, diversity, and verbosity.

### Experimental Results

**Overall Results** As shown in Tables 2 and 3, CADSS achieves SOTA performance across CPsDD and ESConv

| Models | ACC↑ | PPL↓ | B-1↑ | B-2↑ | B-3↑ | B-4↑ | D-1↑ | D-2↑ | R-L↑ |
|---|---|---|---|---|---|---|---|---|---|
| BlenderBot-Joint(Liu et al. 2021) | 17.69 | 17.39 | 18.78 | 7.02 | 3.20 | 1.63 | 2.96 | 17.87 | 14.92 |
| GLHG(Peng et al. 2022) | - | 15.67 | 19.66 | 7.57 | 3.74 | 2.13 | 3.50 | 21.61 | 16.37 |
| MISC(Tu et al. 2022) | 31.67 | 16.27 | 16.31 | 7.31 | 3.26 | 2.20 | 4.62 | 20.17 | 17.51 |
| KEMI(Deng et al. 2023) | - | 15.92 | - | 8.31 | - | 2.51 | - | - | 17.05 |
| TransESC(Zhao et al. 2023) | 34.71 | 15.82 | 17.92 | 7.64 | 4.01 | 2.43 | 4.73 | 20.48 | 17.51 |
| PAL(Cheng et al. 2023) | 34.51 | 15.92 | - | 8.75 | - | 2.66 | 5.00 | 30.27 | 18.06 |
| CKPI(Hao and Kong 2025) | 35.51 | **14.88** | 21.38 | 9.27 | 4.93 | 2.92 | 4.88 | 25.95 | **18.87** |
| Qwen2.5(Yang et al. 2024) | 14.11 | 40.60 | 9.45 | 5.99 | 4.32 | 3.26 | 7.22 | 22.42 | 9.27 |
| DeepSeek-R1(Liu et al. 2024a) | 16.72 | - | 10.96 | 6.13 | 3.86 | 2.47 | 6.69 | 19.47 | 7.13 |
| GPT-4o | 23.72 | - | 15.42 | 7.08 | 5.15 | 3.87 | 8.43 | 29.61 | 9.96 |
| CADSS(ours) | **46.26** | 30.14 | **27.54** | **15.06** | **10.05** | **5.64** | **8.88** | **32.49** | 16.57 |
| w/o planner | - | 31.81 | 26.05 | 13.47 | 8.74 | 4.61 | 8.54 | 31.87 | 15.91 |
| w/o summarizer | 45.64 | 33.11 | 26.99 | 14.73 | 9.78 | 5.40 | 8.65 | 31.28 | 16.31 |
| w/o profiler | 45.57 | 35.51 | 27.17 | 14.80 | 9.84 | 5.45 | 8.71 | 31.81 | 16.30 |
| w/o planner&summarizer | - | 38.65 | 25.77 | 13.51 | 8.79 | 4.62 | 8.54 | 30.00 | 16.59 |
| w/o planner&profiler | - | 39.73 | 25.72 | 13.50 | 8.78 | 4.59 | 8.54 | 29.66 | 16.51 |
| w/o summarizer&profiler | 45.53 | 35.50 | 27.07 | 14.73 | 9.76 | 5.37 | 8.66 | 31.46 | 16.29 |
| w/o all | - | 33.36 | 25.54 | 13.25 | 8.54 | 5.41 | 8.56 | 30.05 | 16.29 |

Table 3: Overall experimental results on the SP and ESC tasks on the ESConv dataset.

| Models | Helpfulness | Understanding | Consistency | Coherence | Informativeness | Empathy | Overall |
|---|---|---|---|---|---|---|---|
| GPT-4o | 18 | 16 | 16 | 8 | **25** | 17 | 17 |
| CKPI | 7 | 16 | 11 | 13 | 9 | 11 | 2 |
| CADSS | **25** | **18** | **23** | **29** | 16 | **22** | **31** |

Table 4: Voting results of different model responses on each evaluation standard.

datasets for both SP and ESC tasks. On CPsDD, CADSS excels in strategy accuracy, BLEU, and ROUGE-L, reflecting superior controllability and response quality. It also exhibits high diversity and well-balanced response lengths (L-R $\approx$ 1), outperforming models like CKPI and DeepSeek, which suffer from repetition or verbosity. On ESConv, CADSS similarly leads in strategy prediction and achieves the best BLEU and Distinct-n scores, indicating strong fluency and generative diversity. While its PPL is slightly higher due to the creative nature of generation, CADSS remains competitive in ROUGE-L. These results emphasize the advantages of CADSS's multi-agent architecture and strategy-guided design in generating empathetic, diverse, and user-aligned psychological dialogues in both Chinese and English settings.

**Ablation Study** We conduct ablation studies on both CPsDD and ESConv to evaluate each agent's contribution to CADSS. As detailed in Tables 2 and 3, removing any single agent results in clear drops in strategy prediction accuracy and response quality (BLEU, ROUGE-L). The Planner is especially critical: its removal causes the sharpest decrease in prediction and generation metrics, yielding responses that are longer and less focused. The Profiler and Summarizer also play important roles in maintaining personalization and coherence. Notably, Distinct scores remain stable due to the inherent diversity of LLM-based models, yet higher PPL in ablated models indicates less confident and controlled generation. These results confirm that all agents are indispensable for achieving robust performance on both Chinese and English psychological support tasks.

**Manual Evaluation** To further evaluate real-world applicability, we deployed CADSS and baselines as API services. We recruited 50 native Chinese-speaking undergraduate students experiencing recent emotional distress (e.g., personal difficulties). Each participant received 30 RMB and initiated a psychological inquiry, interacting with different models via the web interface. Participants rated the model responses across multiple dimensions. As shown in Table 4, CADSS consistently received the highest votes across most metrics, particularly in Overall Quality, demonstrating superior empathetic and professional support capacity. Although predefined strategies may slightly reduce content novelty, CADSS still outperformed alternatives, confirming its effectiveness and user alignment in real psychological scenarios.

## Conclusion

We proposed PGSim, a path-guided simulation framework mimicing real-world psychological counseling by incorporating expert-defined strategy paths and structured user scenarios. Based on PGSim, we constructed CPsDD, the first large-scale Chinese psychological support dialogue dataset with fine-grained strategies, facilitating both Strategy Prediction and Emotional Support Conversation tasks. We further introduced CADSS, a multi-agent support system that achieves state-of-the-art performance on both Chinese and English datasets. Our work effectively demonstrates combining expert knowledge and LLM capabilities to generate realistic, controllable, and high-quality psychological dialogues.

## Ethics Statement

Our work involves sensitive psychological counseling data and the development of a dialogue system for psychological support. To ensure ethical integrity, all data were fully anonymized with informed consent obtained from users, following strict privacy protocols. Certified psychological experts reviewed and refined all dialogues to guarantee factual accuracy and avoid harmful or triggering content. While efforts were made to ensure diversity and fairness, some population or issue biases may remain. The proposed system should not replace professional mental health care. Deployment must include disclaimers, human oversight, and referral mechanisms. All datasets and models will be released for research under ethical guidelines, and we welcome community feedback to enhance safety and inclusivity.

## Acknowledgments

This work was supported by the Project 62276178 under the National Natural Science Foundation of China, the Key Project 23KJA520012 under the Natural Science Foundation of Jiangsu Higher Education Institutions, Project 24BTQ002 under theNational Social Science Foundation of China, the project 22YJCZH091 of Humanities and Social Science Fund of Ministry of Education and the Priority Academic Program Development of Jiangsu Higher Education Institutions. They are also with Jiangsu Key Lab of Language Computing, Suzhou, 215021, P. R. China.

## References

Bradt, J.; and Dileo, C. 2014. Music interventions for mechanically ventilated patients. *Cochrane database of systematic reviews*, (12).

Brown, P. F.; et al. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1): 31–40.

Casu, M.; et al. 2024. AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Appl. Sci*, 14: 5889.

Chen, S. Y., Tiantian; et al. 2024. PsyChatbot: A Psychological Counseling Agent Towards Depressed Chinese Population Based on Cognitive Behavioural Therapy. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Chen, Y.; Xing, X.; Lin, J.; et al. 2023. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of EMNLP 2023*, 1170–1183. Singapore: 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 ACL 2023*, 535–554.

Deng, Y.; Zhang, W.; Yuan, Y.; and Lam, W. 2023. Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st ACL)*, 4079–4095. Toronto, Canada: Association for Computational Linguistics.

Divya, K.; Valsaraj, B. P.; and Al Harthy, M. I. 2022. Suicidality, Psychological distress and Professional help seeking: A cross sectional survey of Omani Student Nurses. *Journal of Positive School Psychology*, 11860–11868.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Hao, J.; and Kong, F. 2025. Enhancing Emotional Support Conversations: A Framework for Dynamic Knowledge Filtering and Persona Extraction. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3193–3202. Abu Dhabi, UAE: Association for Computational Linguistics.

He, T.; Liao, L.; Cao, Y.; et al. 2024. Planning Like Human: A Dual-process Framework for Dialogue Planning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd ACL (Volume 1: Long Papers)*, 4768–4791. Bangkok, Thailand: Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; et al. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lee, J.; Lee, D.; and Lee, J.-g. 2024. Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. *International Journal of Human–Computer Interaction*, 40(7): 1620–1631.

Li, A.; Ma, L.; Mei, Y.; et al. 2023. Understanding Client Reactions in Online Mental Health Counseling. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st ACL (Volume 1: Long Papers)*, 10358–10376. Toronto, Canada: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, A.; Feng, B.; Xue, B.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, S.; Zheng, C.; Demasi, O.; et al. 2021. Towards Emotional Support Dialog Systems. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483. Online: Association for Computational Linguistics.

Liu, Z.; Yang, K.; Xie, Q.; Zhang, T.; and Ananiadou, S. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5487–5496.

Oh, K.-J.; Lee, D.; Ko, B.; and Choi, H.-J. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE international conference on mobile data management (MDM)*, 371–375. IEEE.

Pan, Z.; Wang, Y.; and Derakhshan, A. 2023. Unpacking Chinese EFL students' academic engagement and psychological well-being: The roles of language teachers' affective scaffolding. *Journal of Psycholinguistic Research*, 52(5): 1799–1819.

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 ACL*, 311–318.

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 Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 4324–4330. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Qiu, H.; He, H.; et al. 2024. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of EMNLP 2024*, 615–636. Miami, Florida, USA: Association for Computational Linguistics.

Qiu, H.; Li, A.; Ma, L.; and Lan, Z. 2024. PsyChat: A Client-Centric Dialogue System for Mental Health Support. arXiv:2312.04262.

Rashkin, H. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Razzaque, R.; and Stockmann, T. 2016. An introduction to peer-supported open dialogue in mental healthcare. *BJPsych Advances*, 22(5): 348–356.

Shechtman, Z. 2008. *Treating child and adolescent aggression through bibliotherapy*. Springer Science & Business Media.

Shek, D. T. 2002. Family functioning and psychological well-being, school adjustment, and problem behavior in Chinese adolescents with and without economic disadvantage. *The Journal of genetic psychology*, 163(4): 497–502.

Sorin, V.; Brin, D.; et al. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research*, 26: e52597.

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 Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of ACL-IJCNLP 2021*, 1489–1503. Online: Association for Computational Linguistics.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao, D. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th ACL (Volume 2: Short Papers)*, 231–236.

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

Vaidyam, A. N.; Wisniewski, H.; Halamka, J. D.; Kashavan, M. S.; and Torous, J. B. 2019. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, 64(7): 456–464. PMID: 30897957.

Wang, J.; Xu, C.; Leong, C. T.; Li, W.; and Li, J. 2024. Muffin: Mitigating Unhelpfulness in Emotional Support Conversations with Multifaceted AI Feedback. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of ACL 2024*, 567–585. Bangkok, Thailand: Association for Computational Linguistics.

Xiao, S.; Li, J.; Tang, M.; et al. 2013. Methodology of China's national study on the evaluation, early recognition, and treatment of psychological problems in the elderly: the China Longitudinal Aging Study (CLAS). *Shanghai Archives of Psychiatry*, 25(2): 91.

Xie, H.; Chen, Y.; Xing, X.; Lin, J.; and Xu, X. 2024. PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling. arXiv:2412.13660.

Xin Yan, D. X. 2023. MindChat: Psychological Large Language Model. https://github.com/X-D-Lab/MindChat.

Yang, A.; Yang, B.; Zhang, B.; et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Ye, J.; Xiang, L.; Zhang, Y.; and Zong, C. 2024. SweetieChat: A Strategy-Enhanced Role-playing Framework for Diverse Scenarios Handling Emotional Support Agent. *arXiv preprint arXiv:2412.08389*.

Young, S. D. 2012. *Psychology at the Movies*. Wiley Online Library.

Zhang, C.; Li, R.; Tan, M.; et al. 2024a. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of ACL 2024*, 13947–13966. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, T.; Zhang, X.; Zhao, J.; Zhou, L.; and Jin, Q. 2024b. ESCoT: Towards Interpretable Emotional Support Dialogue Systems. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd ACL (Volume 1: Long Papers)*, 13395–13412. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, W.; Zhao, Y.; Wang, S.; and Qin, B. 2023. TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of ACL 2023*, 6725–6739. Toronto, Canada: Association for Computational Linguistics.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1): 53–93.