# CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling

**Chenhao Zhang**[1,3*†] and **Renhao Li**[2,3*] and **Minghuan Tan**[3‡] and **Min Yang**[3‡] and
**Jingwei Zhu**[4] and **Di Yang**[4] and **Jiahao Zhao**[3,5†] and **Guancheng Ye**[6†] and
**Chengming Li**[7] and **Xiping Hu**[7]
[1] Huazhong University of Science and Technology [2] University of Macau
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4] University of Science and Technology of China [5] Jilin University
[6] South China University of Technology [7] Shenzhen MSU-BIT University
ch_zhang@hust.edu.cn, li.renhao@connect.um.edu.mo
{mh.tan,min.yang}@siat.ac.cn

## Abstract

Using large language models (LLMs) to assist psychological counseling is a significant but challenging task at present. Attempts have been made on improving empathetic conversations or acting as effective assistants in the treatment with LLMs. However, the existing datasets lack consulting knowledge, resulting in LLMs lacking professional consulting competence. Moreover, how to automatically evaluate multi-turn dialogues within the counseling process remains an understudied area. To bridge the gap, we propose CPsyCoun, a report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. To fully exploit psychological counseling reports, a two-phase approach is devised to construct high-quality dialogues while a comprehensive evaluation benchmark is developed for the effective automatic evaluation of multi-turn psychological consultations. Competitive experimental results demonstrate the effectiveness of our proposed framework in psychological counseling. We open-source the datasets and model for future research. [1]

## 1 Introduction

"No health without mental health" is becoming more than a slogan, with approximately 14% of the global disease burden attributed to neuropsychiatric disorders (Prince et al., 2007). Despite the affordability and effectiveness of many mental health treatments, a significant gap persists between those in need and those able to access care (Freeman, 2022). The World Health Organization (WHO) continually advocates for increased investment to augment understanding and dispel the stigma associated with mental health disorders. Yet, the challenge of ensuring quality, affordable care for mental health conditions remains formidable. Consequently, the identification of novel treatments and enhancement of existing therapies for all mental diseases are key objectives in the research domain.

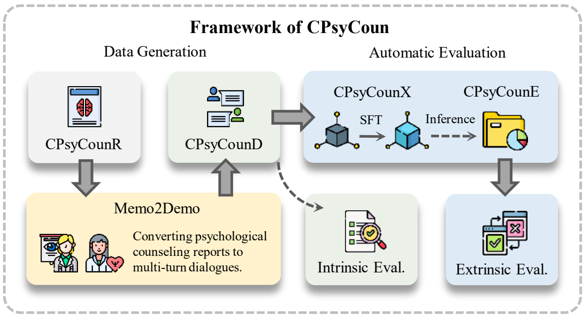


Figure 1: The general framework of CPsyCoun

The Natural Language Processing (NLP) community is actively contributing to the advancement of AI-assisted psychological counseling and treatment. Various research topics have been proposed to conduct mental disease counseling (Orr et al., 2022; Toleubay et al., 2023), improve emotional support ability (Buechel et al., 2018; Rashkin et al., 2019; Liu et al., 2021; Cheng et al., 2023), and provide online psychological consultation (Sun et al., 2021).

The advent of large language models (LLMs) such as ChatGPT [2] and LLaMA (Touvron et al., 2023), has spurred more research efforts on generating not just empathetic conversations, but also serving as therapeutic aids and effective assistants in treatment. For instance, Psy-LLM (Lai et al., 2023)

[*] Equal contribution.

[†] Work done on the Science and Technology Innovation Project of UCAS directed by SIAT.

[‡] Corresponding author.

[1] https://github.com/CAS-SIAT-XinHai/CPsyCoun

[2] https://chat.openai.com/

*Findings of the Association for Computational Linguistics ACL 2024*, pages 13947–13966
August 11-16, 2024 ©2024 Association for Computational Linguistics

is a psychological consultation model that leverages the LLM PanGu and is trained with Q&A from professional psychologists and large-scale Chinese psychological articles from public databases. This model demonstrates proficiency in psychological knowledge and counseling services. Parallel to this, other LLM-based psychological models such as MeChat (Qiu et al., 2023), SoulChat (Chen et al., 2023b) and MindChat (Yan and Xue, 2023) are also available online. Recent trends in adopting LLMs for psychological counseling focus on generating more interpretable mental health analyses (Yang et al., 2023) and simulating psychiatrist-patient interactions (Chen et al., 2023a). This shift in focus from generating responses to diagnosing mental health issues as an expert signifies a trend change in research. The quest for interpretability in mental health analysis serves a dual purpose. First, it provides a detailed rationale behind each response, making it more amenable for human evaluation and debugging. Second, the simulation approach not only addresses data privacy concerns but also challenges the traditional symptom collection method via questionnaires. Providing a range of professional skills, this approach enables more effective completion of consultation tasks.

Despite these advancements, there remains a dearth of authentic counseling datasets from psychological counseling sessions, which include symptom descriptions of the consultant and treatment methodologies employed by the counselor. Such data could offset issues arising from doctor-patient simulations being template-based and lacking control. For example, psychiatrists have observed that chatbots do not typically resemble patients (Chen et al., 2023a). However, it's noteworthy that these diagnoses are generally sensitive, warranting careful attention to potential privacy issues. In addition to the form of psychological counseling conversations, there is a wealth of psychological counseling data in the real world, which is hidden in professional psychological counseling reports. However, due to its structured nature, it is unsuitable for model training.

In this paper, we propose a new framework CPsyCoun for **C**hinese **Psy**chological **Coun**seling, which consists a dialogue reconstruction method based on psychological counseling reports and a benchmark for multi-turn consultation dialogue evaluation. Specifically, we first collect anonymized psychological counseling reports from publicly accessible websites and further propose a privacy shadowing method to postprocess these reports into a dataset CPsyCounR. CPsyCounR includes nine types of psychological consultation and seven classic schools of psychological counseling. Through our proposed Memo2Demo dialogue reconstruction method, we construct another dataset CPsyCounD, which contains 3,134 high-quality multi-turn consultation dialogues. Further, we propose a psychological counseling benchmark for automatic evaluation on multi-turn dialogues and fine-tune an open-sourced LLM on CPsyCounD, named CPsyCounX. Experimental results from both intrinsic and extrinsic evaluations consistently verify the superiority of the proposed method.

Figure 1 illustrates the general framework of our proposed CPsyCoun.

*Our contributions are the following:*

- To the best of our knowledge, our work is the first to generate psychological consultation dialogues based on psychological counseling reports, which effectively expands the source of psychological consultation dialogue data. For efficient dialogue reconstruction, we specifically introduce a two-phase method named Memo2Demo.

- We propose a benchmark for automatic evaluation of multi-turn dialogues in psychological counseling, which includes comprehensive evaluation metrics, datasets and methods.

- With the help of Memo2Demo, we construct CPsyCounD, a dataset contains 3,134 high-quality multi-turn consultation dialogues. The model CPsyCounX fine-tuned on this dataset outperforms other models in the benchmark, validating the effectiveness of our proposed framework in psychological counseling.

## 2 Related Work

### 2.1 Dialogue Generation and Reconstruction using LLMs

Dialogue generation and reconstruction using LLMs have been proven to be effective in data augmentation and conversation denoising. For example, SAFARI (Wang et al., 2023a) harnesses the planning and understanding capabilities of LLMs to generate persona-consistent and knowledge-enhanced responses. In the medical domain, DISC-MedLLM (Bao et al., 2023) undertakes real-world dialogue reconstruction for consultation records sourced from medical forums. This process addresses issues of informal language usage and unregulated expressive styles. In the realm of psychology, numerous studies concentrate on augmenting emotional support capability by enhancing empathy.Qian et al. (2023) amplifies empathetic responses by enriching the dialogue context with a commonsense knowledge graph, thereby stimulating the relevant knowledge encoded by LLMs.

In this work, we propose a two-phase method for efficient dialogue reconstruction.

### 2.2 Evaluation of Generated Dialogues using LLMs

The search for better automatic evaluation metrics in natural language generation (NLG) has been a hot topic for the natural language processing (NLP) community. Compared to conventional lexicon-based metrics like BLEU (Papineni et al., 2002) and Rouge (Lin, 2004), these new metrics capture deeper semantic meaning and usually have better alignment with human judgments.

There have been a series of transformer-based evaluation metrics available in the community, such as BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021) and GPTScore (Fu et al., 2023). In specific domains, there are also derivatives of such metrics tailored for the domain. For example, CodeBERTScore (Zhou et al., 2023) is proposed to achieve a higher correlation with human preference and with functional correctness. CBERTScore (Shor et al., 2023) can penalize clinically-relevant mistakes more than others.

The same trend continues with LLMs. Wang et al. (2023b) shows that ChatGPT achieves state-of-the-art or competitive correlation with human judgments in most cases. A new framework constructed over GPT-4 called G-EVAL (Liu et al., 2023b) makes use of LLMs with chain-of-thoughts (CoT) and a form-filling paradigm to assess the quality of NLG outputs, outperforming all previous methods by a large margin.

In this work, we design a psychological counseling benchmark for automatic evaluation.

## 3 CPsyCoun

### 3.1 Data Collection

We conduct a survey of publicly available psychological counseling cases online and collect data from well-known Chinese psychological communities. The online communities used in this work are: (1) Yidianling [3], a top-tier mental health platform in China, serves approximately 39 million users, backed by a robust network of over 6,000 professional counselors. (2) Psy525 [4], another prominent mental health platform in China, caters to over 1 million users and is supported by nearly 30,000 professional counselors.

As the data are anonymized by the websites, there's a low privacy risk. To enhance the security of the collected data, we further conduct an analysis of privacy and security issues about the data. The procedures adopted during data collecting to ensure no sensitive or privacy-related content in the dataset include rule-based cleaning, manual rewriting, and human proofreading. After cleaning procedures, relevant private information has been completely removed, and we ensure that relevant private information is protected.

In total, we collected 4,700 psychological counseling reports in different formats, with a variety of types and counseling methods. These reports will not be released to the public unless a Privacy Data Protection Agreement is signed upon reasonable request.

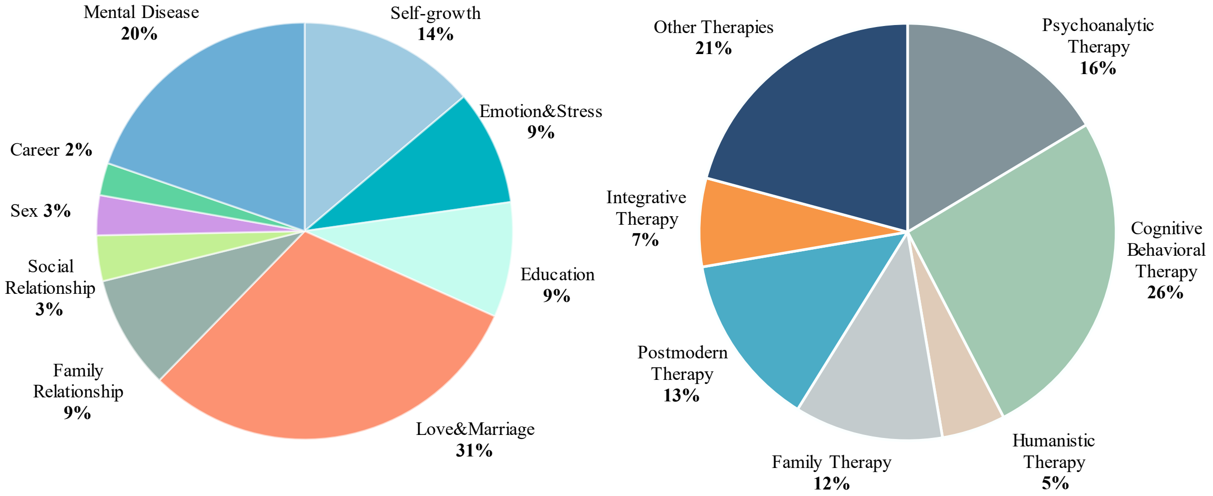
### 3.2 Data Processing

To construct a high-quality dataset, we carefully selected 3,134 psychological counseling reports. They contain complete methods and types, clear case briefs, detailed consultation processes and experience thoughts. In the selection process, we found that some of the collected reports contained several counseling cases in one report. We did not select this type of report due to multiple cases in one report where the background information of the client and the consultation processes are incomplete. Therefore, among the selected 3,134 psychological counseling reports, each report corresponds to only one case. This high-quality report dataset is named CPsyCounR.

**Data Format** Considering the differences in data sources of our collection, we need to reformat these collected reports according to a uniform standard. To build a comprehensive and helpful dataset, we combine the case format of China's National Class 2 Psychological Counselor Examination and other psychological counseling literature to regularize collected reports, where the following 6 compo-

[3] https://www.ydl.com

[4] https://www.psy525.cn



(a) Distribution of counseling topics.

(b) Distribution of psychological counseling schools.

Figure 2: Statistics of collected cases.

nents are included: *Title*, *Type*, *Method*, *Case Brief*, *Consultation Process* and *Experience Thoughts*. Note that the consultation process is written from a third-person perspective and does not contain specific dialog. For a detailed description of components and examples of psychological counseling report, please refer to Appendix A.

**Data analysis** According to statistics, there are about 230 types of psychological counseling cases and more than 250 counseling methods used in psychological counseling. Considering that the classification of the original types is too detailed, we further summarized the case types into 9 representative topics based on common scenarios of psychological counseling. The distribution of counseling topics is shown in Figure 2a.

Based on relevant information from the American Psychological Association (APA) and the International Academy of Psychotherapy (IACP), we have categorized the professional counselor's methods utilized in psychological counseling reports into 7 classic schools of psychological counseling. The distribution of methods is shown in Figure 2b.

### 3.3 Dialogue Generation Method for Psychological Counseling

**Baseline Method** Direct role-play prompting is utilized as our baseline method for generating multiple rounds of dialogue from a single round, which has been successfully used in previous work on dialogue generation (Qiu et al., 2023; Chen et al., 2023b). However, we believe that there are still aspects for improvement when applying direct role-play prompting to multi-round dialogue generation in the field of psychological counseling: (1) Comprehensiveness: Despite presenting high-quality counseling reports to the language model, it may fail to focus on the significant descriptions of the client's situation within the report, leading to subsequent dialogues that lack completeness. (2) Professionalism: Role-playing prompted dialogues merely reference psychological methods in generated dialogues. We hope that language models could integrate these methods into the problem-solving process, thereby obtaining reconstructed dialogue with professionalism. (3) Authenticity: Dialogue constructed by the baseline method lacks the emotional interaction between the client and the psychological counselor present in real scenarios, leaving a deficit in terms of authenticity. We present the detailed prompt of role-play method in Figure 6 in the appendix.

**Memo2Demo** To address the aforementioned issues of the baseline method, we propose a two-phase framework named Memo2Demo to generate high-quality psychological consultation dialogue from counseling reports. Mirroring real-life scenarios, we incorporate two key roles into this framework: a psychological supervisor together with a psychological counselor. The psychological supervisor guides the psychological counselor on counseling techniques while ensuring the privacy of the clients during the counseling process. Meanwhile, the psychological counselor engages in direct dia-

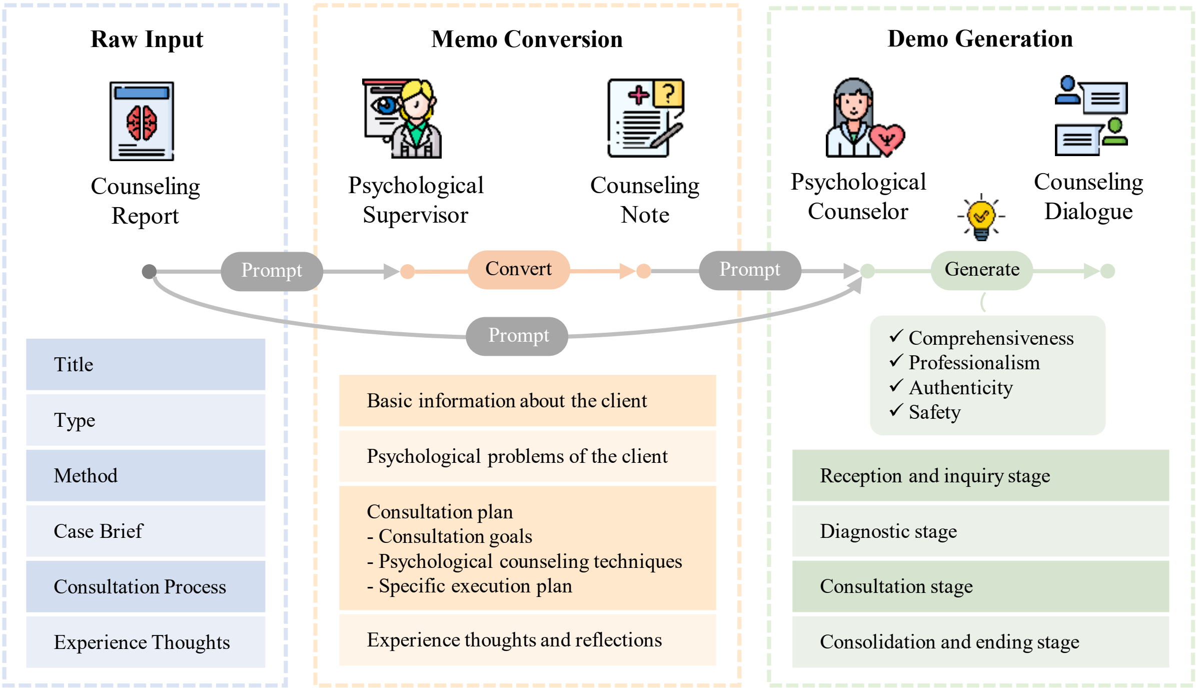


Figure 3: Illustration of the dialogue reconstruction method Memo2Demo

logue with the clients to conduct specific psychological counseling. Figure 3 illustrates the general framework of our proposed method Memo2Demo, where a psychological counseling report is first converted into a counseling note by the psychological supervisor, then the psychological counselor generates the multi-turn consultation dialogue based on both the report and the note. We present detailed prompts used for Memo2Demo in Figure 7 and 8 in the appendix.

**Memo Conversion** In this phase, we first assign the role of psychological supervisor to a large language model, then prompt it to convert psychological counseling reports to counseling notes. Specifically, the psychological supervisor comes up with a counseling note based on the report, including basic information of counseling and an elaborate consultation plan. One of the objectives of the counseling note is to offer enhanced professional insights pertinent to the case, employing distinct psychological counseling techniques to tackle the client's issue. In the meanwhile, it also condenses the core information related to the client, thus improving the comprehensiveness of the subsequent psychological counseling process. In this paper, we build the psychological supervisor based on GLM-4 (Zeng et al., 2023). Format of counseling notes is shown in the yellow table of Figure 3.

**Demo Generation** In this phase, we first assign the role of psychological counselor to a large language model, then prompt it to generate multi-turn consultation dialogues based on the psychological counseling report and the converted counseling note. We simplify a four-stage consultation framework according to the actual psychological counseling process. Leveraging this consultation framework, we enhance our control over the direction of dialogue generation, improving the professionalism of psychological counselors in multi-turn consultation dialogues. In addition, the consultation framework is designed to efficiently restore real scenarios and enhance the authenticity of the multi-round consultation dialogues. In this paper, we build the psychological counselor based on GLM-4 (Zeng et al., 2023). Format of the consultation framework is shown in the green table of Figure 3.

### 3.4 Automatic Evaluation of LLM-based Psychological Counseling

In the field of psychological counseling, assessing the quality of multi-turn consultation dialogues has always been a challenging task. Despite having successfully generated high-quality counseling dialogues from case reports using Memo2Demo, we still need to verify the impact of these dialogues on subsequent tasks. To this end, we elect to utilize CPsyCounD for supervised fine-tuning on publicly accessible LLMs. This allows us to assess the changes in the psychological counseling competency before and after the use of data.

Nonetheless, the multi-turn consultation dialogue that characterizes the psychological counseling process is complex to evaluate without the input of human experts. To address this, we first introduce **evaluation metrics** tailored for multi-turn consultation dialogue. Then a **turn-based dialogue evaluation** method is proposed for automatic evaluation of the psychological counseling process. Moreover, we acknowledge the current shortfall of a comprehensive, general multi-turn dialogue evaluation dataset within the psychological counseling community. Such a dataset is vital for assessing LLM-based psychological counseling. To bridge this gap, we present **CPsyCounE**, a general multi-turn dialogue evaluation dataset.

**Evaluation Metrics** In psychological counseling, the evaluation metrics remain diverse and not universally standardized. For instance, SoulChat (Chen et al., 2023b) proposes evaluation metrics: *Content*, *Empathy*, *Helpfulness* and *Safety*. Some of these metrics hinge on expert evaluations and lack specific scoring criteria, favoring manual rather than objective and automatic evaluations. Similarly, ChatCounselor (Liu et al., 2023a) introduces the Counseling Bench, encompassing seven different perspectives. While these metrics are designed to cater to the model's specific dialogue strategies, they lack the ability to evaluate the overall dialogue effect. They are more adapted to single-round dialogue evaluations and are not suitable for multi-turn dialogues.

Recognizing the aforementioned limitations in evaluating consultation dialogues, and in order to analyze the counseling case used for dialogue generation, we propose new evaluation metrics for multi-turn consultation dialogues in psychological counseling. These metrics encompass four different perspectives: *Comprehensiveness*, *Professionalism*, *Authenticity*, and *Safety*, which are used for automatic evaluation in the rest of this paper. For each perspective, we give its description and corresponding score criterion in Appendix C.

**Turn-Based Dialogue Evaluation** We propose a turn-based dialogue evaluation approach to effectively evaluate multi-turn consultation dialogues. Denote a $m$-turn dialogue as a set of paired elements $\{(q_i, r_i)|i = 1, 2, ..., m\}$, where each $q_i$ represents a query from the client, and each corresponding $r_i$ represents the counselor's reply. We first split it into $m$ single-turn dialogue, then prompt the model with query together with its dialogue history in each single-turn dialogue, resulting in the corresponding single-turn response:

$$\hat{r}_i = \begin{cases} f_{LLM}(q_i), & i = 1 \\ f_{LLM}(h_i, q_i), & 1 < i \leq m \end{cases} \tag{1}$$

where $h_i = \{(q_j, r_j)|j = 1, 2, ..., i-1\}$ signifies the dialogue history before $i$-th turn, and $f_{LLM}(\cdot)$ denotes the inference process of LLMs.

To automatically obtain reliable evaluation results, We employ GPT-4 (Achiam et al., 2023) to assess these responses, utilizing the evaluation metrics we previously proposed. Concretely, we ask the model to assign an evaluation score $\hat{s}_i$ for a single-turn response $\hat{r}_i$. Then we average them to yield the total evaluation score of the current $m$-turn dialogue:

$$s_i = \frac{1}{m}\sum_{i=1}^{m} \hat{s}_i, \tag{2}$$

For detailed prompts of single-turn response generation, please refer to Figure 11 in the appendix.

**CPsyCounE** SMILECHAT (Qiu et al., 2023), a richly diverse and realistic multi-turn dialogue dataset, comprises 56k multi-turn counseling dialogues, averaging 6.36 rounds per dialogue. Given its wide range of dialogue types, we choose it as our base dataset. However, the open-source data of this dataset is not classified by topic type. To address this limitation and conduct a more comprehensive and explainable evaluation of models' capabilities, we construct a general multi-turn dialogue evaluation dataset with clear topic classification - CPsyCounE. Leveraging the nine common counseling topics we introduce in CPsyCounR, we manually select the five most representative dialogues from SMILECHAT for each topic, resulting in a comprehensive evaluation dataset of 45 cases.

## 4 Experiments

| Dialogues | Role-play | Memo2Demo |
|---|---|---|
| Avg. Number of Turns | 8.2 | 8.7 |
| Avg. Length of Client | 24.5 | 30.4 |
| Avg. Length of Counselor | 40.2 | 49.7 |
| Avg. Length of Dialogue | 545.8 | 622.3 |

Table 1: Statistics of generated dialogues.

| Method | School | Metrics | | | |
|---|---|---|---|---|---|
| | | Comprehensiveness | Professionalism | Authenticity | Safety |
| Role-play | Psychoanalytic Therapy | 1.35 | 2.48 | 2.23 | 1.00 |
| | Cognitive Behavioral Therapy | 1.35 | 2.45 | 2.15 | 1.00 |
| | Humanistic Therapy | 1.30 | 2.15 | 1.98 | 1.00 |
| | Family Therapy | 1.28 | 2.18 | 2.00 | 1.00 |
| | Postmodern Therapy | 1.25 | 2.15 | 1.98 | 1.00 |
| | Integrative Therapy | 1.28 | 2.10 | 1.88 | 1.00 |
| | Other Therapies | 1.30 | 2.25 | 2.03 | 1.00 |
| Memo2Demo | Psychoanalytic Therapy | 2.00 | 3.35 | 2.65 | 1.00 |
| | Cognitive Behavioral Therapy | 2.00 | 3.43 | 2.68 | 1.00 |
| | Humanistic Therapy | 2.00 | 3.55 | 2.65 | 1.00 |
| | Family Therapy | 2.00 | 3.48 | 2.70 | 1.00 |
| | Postmodern Therapy | 2.00 | 3.53 | 2.58 | 1.00 |
| | Integrative Therapy | 2.00 | 3.50 | 2.58 | 1.00 |
| | Other Therapies | 2.00 | 3.23 | 2.63 | 1.00 |
| | Avg. Role-play | 1.30 | 2.25 | 2.04 | 1.00 |
| | Avg. Memo2Demo | 2.00 | 3.44 | 2.64 | 1.00 |
| | Improv. | **+53%** | **+53%** | **+30%** | - |

Table 2: Results of the intrinsic evaluation on CPsyCoun. In the last row of the table, we present the percentage improvement of metrics for Memo2Demo compared to role-play method.

### 4.1 CPsyCounD

To validate the effectiveness of our proposed dialogue reconstruction approach, we adopt direct role-play prompting and Memo2Demo to generate dialogues from CPsyCounR respectively. We denote the set of dialogues generated by Memo2Demo as CPsyCounD, which has a total of 3,134 multi-turn consultation dialogues, covering nine topics and seven classic schools of psychological counseling. For statistical information, please refer to Table 1.

### 4.2 Intrinsic Evaluation of CPsyCoun

To ensure comprehensiveness and diversity of the evaluation dataset, we randomly select 20 cases from each of the seven classic schools of psychological counseling in CPsyCounR, acquiring a total of 140 cases. Then we adopt direct role-play prompting and Memo2Demo method respectively for dialogue generation, and instruct GPT-4 (Achiam et al., 2023) to conduct a comparative evaluation of the above two multi-turn consultation dialogues. The evaluation standard refers to the evaluation metrics shows in Table 4. For detailed evaluation prompts, please refer to Figure 9 in the appendix.

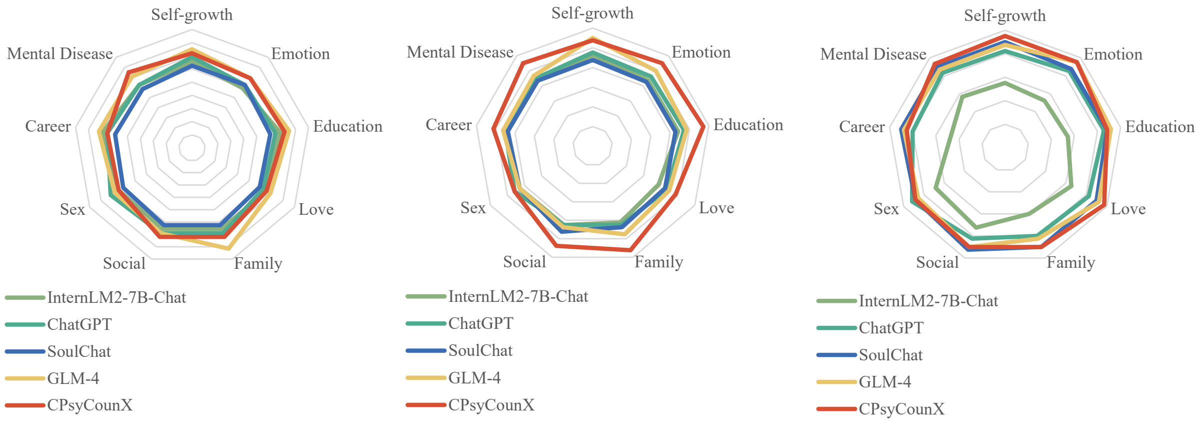
Table 2 illustrates the results of intrinsic evaluation on CPsyCoun. For each school, Memo2Demo method outperforms direct role-play prompting in terms of Comprehensiveness, Professionalism, and Authenticity. When comparing the overall average scores, Memo2Demo method exhibits a remarkable improvement of 53%, 53%, and 30% in these metrics respectively, when juxtaposed with direct role-play prompting. Note that both methods get full scores in Safety, which shows the advantage of report-based data construction methods for privacy protection. In general, our proposed method Memo2Demo significantly enhances the quality of reconstructed multi-turn consultation dialogues.

### 4.3 Extrinsic Evaluation of CPsyCoun

**CPsyCounX** To delve deeper into whether the proposed dataset can effectively enhance the psychological counseling capabilities of LLMs, we further fine-tune InternLM2-7B-Chat (Team, 2023) on CPsyCounD and derive a chat model CPsyCounX tailored specifically for psychological counseling.

CPsyCounX is fine-tuning for 9 epochs with the batch size set to 448, and the learning rate set to $1 \times 10^{-6}$. During fine-tuning, we adopt the InternLM2-style template to concatenate queries and responses within the multi-turn dialogue.

**Automatic Evaluation** The turn-based dialogue evaluation method is adopted on CPsyCounE for the following extrinsic evaluation. We include InternLM2-7B-Chat (Team, 2023), SoulChat (Chen et al., 2023b), ChatGPT and GLM-4 (Zeng et al., 2023) as major baseline models. The evaluation standard refers to the evaluation metrics in Table 4 in the appendix. To accommodate multi-turn dialogues, we adjust the authenticity criterion accordingly. For detailed evaluation prompts,



(a) Comprehensiveness (b) Professionalism (c) Authenticity

Figure 4: Radar plot of detailed scores of CPsyCounX and other baselines on 9 counseling topics on the benchmark.

| **Model** | **Metrics** | | | |
|---|---|---|---|---|
| | Comprehensiveness | Professionalism | Authenticity | Safety |
| InternLM2-7B-Chat | 1.30 | 2.16 | 1.48 | 1.00 |
| SoulChat | 1.22 | 2.18 | 2.24 | 1.00 |
| ChatGPT | 1.32 | 2.25 | 2.09 | 1.00 |
| GLM-4 | **1.44** | 2.36 | 2.22 | 1.00 |
| CPsyCounX | 1.39 | **2.65** | **2.29** | 1.00 |

Table 3: Results of the extrinsic evaluation on CPsyCoun. The best score for each metric is **in-bold**, while the second best score is underlined.

please refer to Figure 10 in the appendix.

We present the overall results of extrinsic evaluation on CPsyCoun in Table 3, where CPsyCounX surpasses other models in terms of Professionalism and Authenticity, falling behind GLM-4 only slightly in terms of Comprehensiveness. Figure 4 further shows detailed scores of CPsyCounX and other baselines, where CPsyCounX significantly outperforms nearly all other baselines on Professionalism, demonstrating the efficacy of proposed method Memo2Demo. While judging by the topic distribution, CPsyCounX leads in all metrics in the topic "Mental Disease", demonstrating its high usability in the field of psychological counseling. For full results, please refer to Appendix E.

Upon evaluation, we find that GLM-4 scores the highest in Comprehensiveness, largely because its single-turn dialogues encompass vast information. A manual investigation reveals that GLM-4 prioritizes summarizing previous dialogues in each turn, accounting for its high scores in Comprehensiveness. However, our evaluation shows that excessive content tends to compromise Authenticity scores. In psychological counseling, Authenticity and Comprehensiveness need a balanced consideration. In our experiments, we prioritize natural and authentic dialogues that contain key information. Consequently, while CPsyCounX scores lower than GLM-4 in Comprehensiveness, it surpasses GLM-4 in Authenticity.

These results highlight that fine-tuning on CPsyCounD enables the model to naturally acquire professional psychological counseling techniques used in counseling dialogues. Moreover, the model can learn the conversational style of psychological counselors in real-life psychological counseling scenarios, ensuring the dialogue's authenticity.

## 5 Conclusion

In this paper, we introduce CPsyCoun, an innovative framework for report-based multi-turn dialogue reconstruction and evaluation in Chinese psychological counseling. Our research encompasses data collection, effective data construction methods, and domain evaluation benchmarks. To harness the full potential of psychological counseling reports, we design a two-phase approach to construct high-quality consultation dialogues. Concurrently, we propose a comprehensive evaluation benchmark for multi-turn consultation dialogue, inclusive of metrics, datasets and methods. Experimental results validate the effectiveness of our proposed framework, demonstrating its superiority in building a comprehensive, professional, and authentic psychological counseling assistant. All datasets and model weights developed in this paper are publicly available. For future work, a more refined balance between authenticity and professional knowledge in dialogue generation needs to be achieved. We aspire this work will furnish fresh perspectives and references for the development of LLMs in the field of psychological counseling.

## Limitations

In this work, we proposed CPsyCoun, a report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. Although the experimental results demonstrate that our framework is viable, there are still some limitations need to be considered. In real-life scenarios, psychological counseling is complex. Counselors not only need to empathize with clients, but also need to use psychological counseling techniques at the appropriate time. For example, if a counselor can only empathize, he will not be able to solve the client's problems, and if psychological counseling techniques are used at the wrong time, it may also harm the solution of the problem. Therefore, dialogue generative methods and dialogue evaluation methods need to further consider the balance between authenticity and professional knowledge.

## Ethics Statement

### Data Privacy

We have implemented rigorous data sanitization procedures to construct the dataset and safeguard privacy. These measures encompass rule-based cleaning, manual rewriting, and human proofreading to guarantee the absence of sensitive or privacy-related content. For instance, the initial data collection contained private information of psychological counselors, including personal specifics, contact details, residential addresses, and workplaces. Post-sanitization, all such sensitive information has been entirely expunged, ensuring the protection of relevant private information. Following the data copyright formulated by (Qiu et al., 2023), we release the multi-turn dialogue evaluation dataset publicly available for research purposes only.

### Potential Risks of the Model

We carried out a safety assessment specifically for the model's output during the evaluation phase, the results of which are presented in Table 3. Given the absence of human feedback during the model fine-tuning phase, it is inevitable that some responses might potentially harm users. In the event of no noticeable improvement after user interaction with the CPsyCounX model, trained with multi-turn consultation dialogues, CPsyCounD, we strongly recommend seeking assistance from a professional counselor or psychiatrist promptly. It is critical to remember that a virtual dialogue agent may not serve as a replacement for real-world therapy. Furthermore, when implementing this model in downstream applications, it is essential to inform users beforehand that the AI model generates the responses they see, and these should be used only as references.

## Acknowledgements

This work was partially supported by National Key Research and Development Program of China (2022YFF0902100), China Postdoctoral Science Foundation (2023M733654), National Natural Science Foundation of China (62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039).

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

M. Freeman. 2022. The World Mental Health Report: transforming mental health for all. *World Psychiatry*, 21(3):391–392.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *ArXiv preprint*, abs/2309.15461.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Martin Orr, Kirsten Van Kessel, and Dave Parry. 2022. The ethical role of computational linguistics in digital psychological formulation and suicide prevention. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 17–29, Seattle, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. 2007. No health without mental health. *The Lancet*, 370(9590):859–877.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Joel Shor, Ruyue Agnes Bi, Subhashini Venugopalan, Steven Ibara, Roman Goldenberg, and Ehud Rivlin. 2023. Clinical BERTScore: An improved measure of automatic speech recognition performance in clinical settings. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 1–7, Toronto, Canada. Association for Computational Linguistics.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

Yeldar Toleubay, Don Joven Agravante, Daiki Kimura, Baihan Lin, Djallel Bouneffouf, and Michiaki Tatsubori. 2023. Utterance classification with logical neural network: Explainable AI for mental disorder diagnosis. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 439–446, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023a. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Xin Yan and Dong Xue. 2023. Mindchat: Psychological large language model. `https://github.com/X-D-Lab/MindChat`.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. CodeBERTScore: Evaluating code generation with pretrained models of code. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13921–13937, Singapore. Association for Computational Linguistics.

# A Unified Format of Psychological Counseling Reports

Considering the differences in data sources of our collection, we need to reformat these collected reports according to a uniform standard. To regularize the psychological counseling reports collected from different data sources, we include six components within each report. In Figure 5, we offer a detailed description of these components, accompanied by real examples from collected reports. Note that the original case is written in Chinese, while the English version is translated by us for reference.

# B Prompts of Dialogue Reconstruction Methods

In this paper, we adopt direct role-play prompting as the baseline method for dialogue reconstruction. Detailed prompt used for this approach is given in Figure 6. To further improve the role-play method, we propose a two-phase method Memo2Demo for effective dialogue reconstruction, which consists of memo conversion and demo generation. Detailed prompts used in these phases are given in Figure 7 and Figure 8, respectively.

# C Evaluation Metrics and Score Criterion

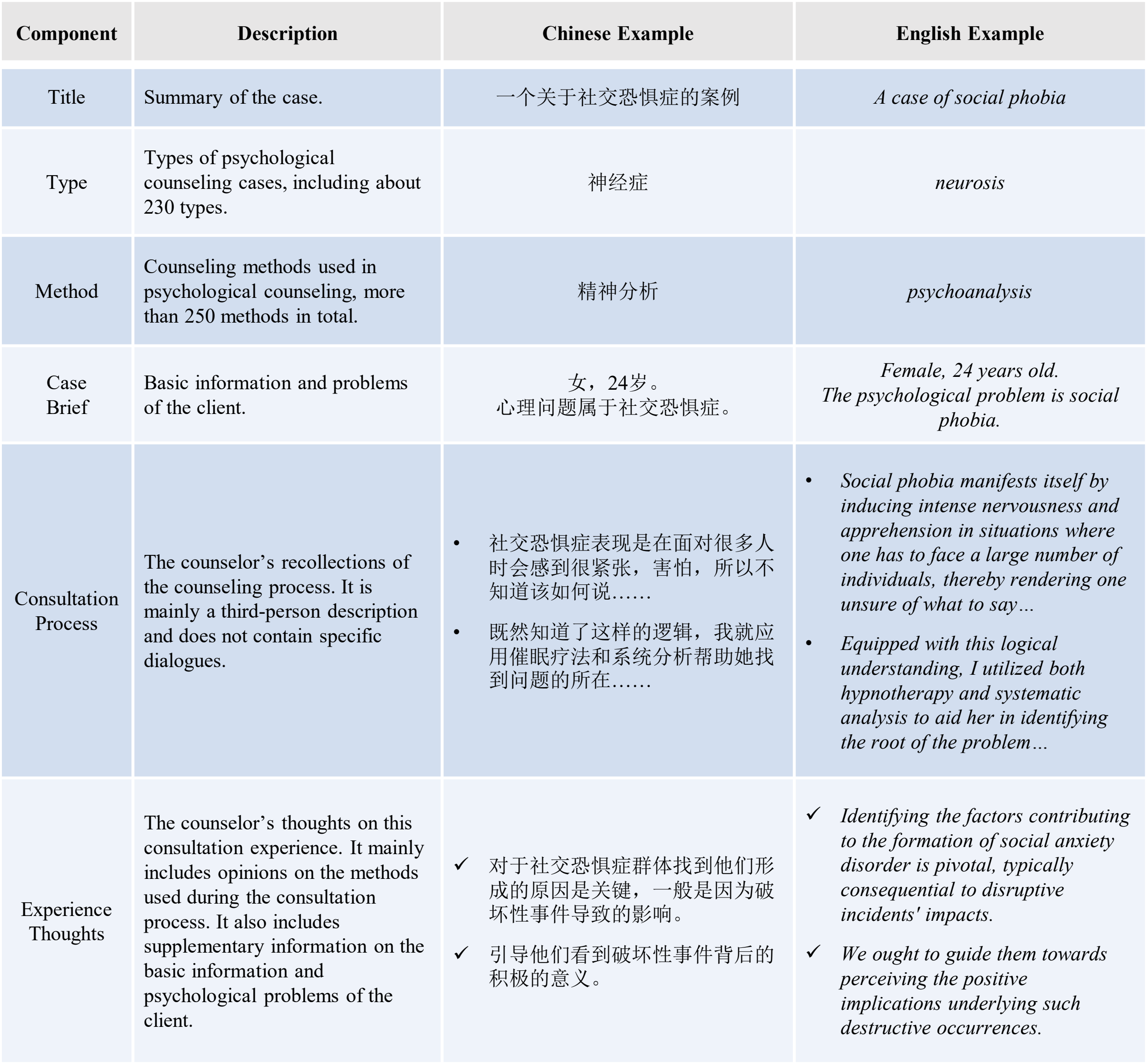
To fill the blank of evaluating multi-turn dialogues in psychological counseling, we specifically propose four metrics for automatic evaluation. We give the detailed description of these metrics together with the score criterion in Table 4.

# D Prompts of Automatic Evaluation

In this paper, we adopt GPT-4 as a judge for automatic evaluation from both intrinsic and extrinsic perspectives. For intrinsic evaluation, we ask the judge to compare consultation dialogues reconstructed by different methods, which is shown in Figure 9 in detail. For extrinsic evaluation, we first fine-tune a chat model CPsyCounX with the proposed high-quality multi-turn dialogue dataset. Then we ask the judge to compare responses from CPsyCounX and other baseline models, which is shown in Figure 10 in detail. The prompt to generate single-turn responses from different models is given in Figure 11.

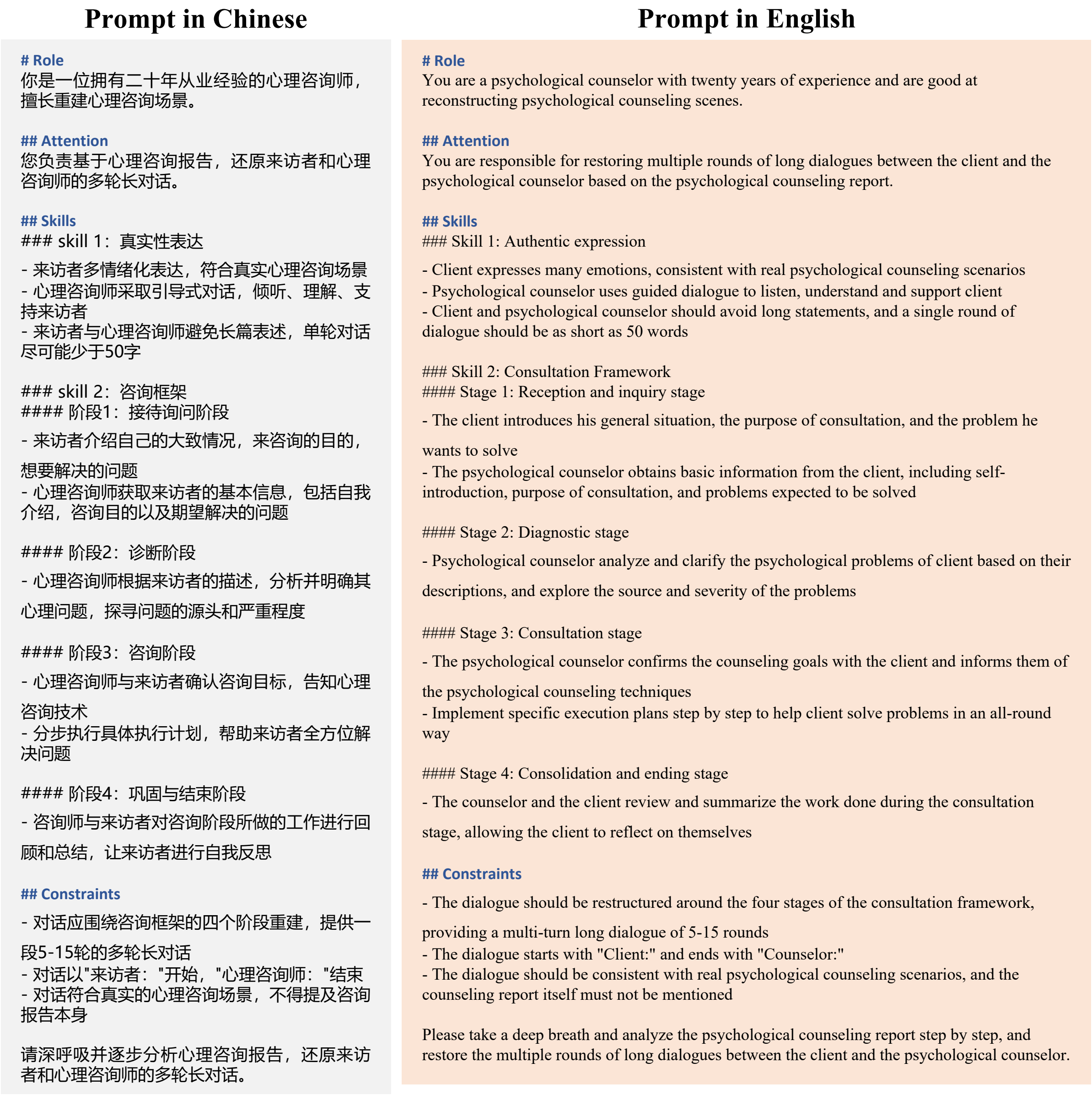
# E Full Results of extrinsic evaluation

For CPsyCounX and other baseline models, we give the full results of extrinsic evaluation on different topics in Table 5.



| Component | Description | Chinese Example | English Example |
|---|---|---|---|
| Title | Summary of the case. | 一个关于社交恐惧症的案例 | *A case of social phobia* |
| Type | Types of psychological counseling cases, including about 230 types. | 神经症 | *neurosis* |
| Method | Counseling methods used in psychological counseling, more than 250 methods in total. | 精神分析 | *psychoanalysis* |
| Case Brief | Basic information and problems of the client. | 女，24岁。<br>心理问题属于社交恐惧症。 | *Female, 24 years old.*<br>*The psychological problem is social phobia.* |
| Consultation Process | The counselor’s recollections of the counseling process. It is mainly a third-person description and does not contain specific dialogues. | • 社交恐惧症表现是在面对很多人时会感到很紧张，害怕，所以不知道该如何说……<br>• 既然知道了这样的逻辑，我就应用催眠疗法和系统分析帮助她找到问题的所在…… | • *Social phobia manifests itself by inducing intense nervousness and apprehension in situations where one has to face a large number of individuals, thereby rendering one unsure of what to say…*<br>• *Equipped with this logical understanding, I utilized both hypnotherapy and systematic analysis to aid her in identifying the root of the problem…* |
| Experience Thoughts | The counselor’s thoughts on this consultation experience. It mainly includes opinions on the methods used during the consultation process. It also includes supplementary information on the basic information and psychological problems of the client. | ✓ 对于社交恐惧症群体找到他们形成的原因是关键，一般是因为破坏性事件导致的影响。<br>✓ 引导他们看到破坏性事件背后的积极的意义。 | ✓ *Identifying the factors contributing to the formation of social anxiety disorder is pivotal, typically consequential to disruptive incidents' impacts.*<br>✓ *We ought to guide them towards perceiving the positive implications underlying such destructive occurrences.* |

Figure 5: Description of report format together with an example from our collection



Prompt in Chinese

# Role
你是一位拥有二十年从业经验的心理咨询师，擅长重建心理咨询场景。

## Attention
您负责基于心理咨询报告，还原来访者和心理咨询师的多轮长对话。

## Skills
### skill 1：真实性表达
- 来访者多情绪化表达，符合真实心理咨询场景
- 心理咨询师采取引导式对话，倾听、理解、支持来访者
- 来访者与心理咨询师避免长篇表述，单轮对话尽可能少于50字

### skill 2：咨询框架
#### 阶段1：接待询问阶段
- 来访者介绍自己的大致情况，来咨询的目的，想要解决的问题
- 心理咨询师获取来访者的基本信息，包括自我介绍，咨询目的以及期望解决的问题

#### 阶段2：诊断阶段
- 心理咨询师根据来访者的描述，分析并明确其心理问题，探寻问题的源头和严重程度

#### 阶段3：咨询阶段
- 心理咨询师与来访者确认咨询目标，告知心理咨询技术
- 分步执行具体执行计划，帮助来访者全方位解决问题

#### 阶段4：巩固与结束阶段
- 咨询师与来访者对咨询阶段所做的工作进行回顾和总结，让来访者进行自我反思

## Constraints
- 对话应围绕咨询框架的四个阶段重建，提供一段5-15轮的多轮长对话
- 对话以"来访者："开始，"心理咨询师："结束
- 对话符合真实的心理咨询场景，不得提及咨询报告本身

请深呼吸并逐步分析心理咨询报告，还原来访者和心理咨询师的多轮长对话。

Prompt in English

# Role
You are a psychological counselor with twenty years of experience and are good at reconstructing psychological counseling scenes.

## Attention
You are responsible for restoring multiple rounds of long dialogues between the client and the psychological counselor based on the psychological counseling report.

## Skills
### Skill 1: Authentic expression
- Client expresses many emotions, consistent with real psychological counseling scenarios
- Psychological counselor uses guided dialogue to listen, understand and support client
- Client and psychological counselor should avoid long statements, and a single round of dialogue should be as short as 50 words

### Skill 2: Consultation Framework
#### Stage 1: Reception and inquiry stage
- The client introduces his general situation, the purpose of consultation, and the problem he wants to solve
- The psychological counselor obtains basic information from the client, including self-introduction, purpose of consultation, and problems expected to be solved

#### Stage 2: Diagnostic stage
- Psychological counselor analyze and clarify the psychological problems of client based on their descriptions, and explore the source and severity of the problems

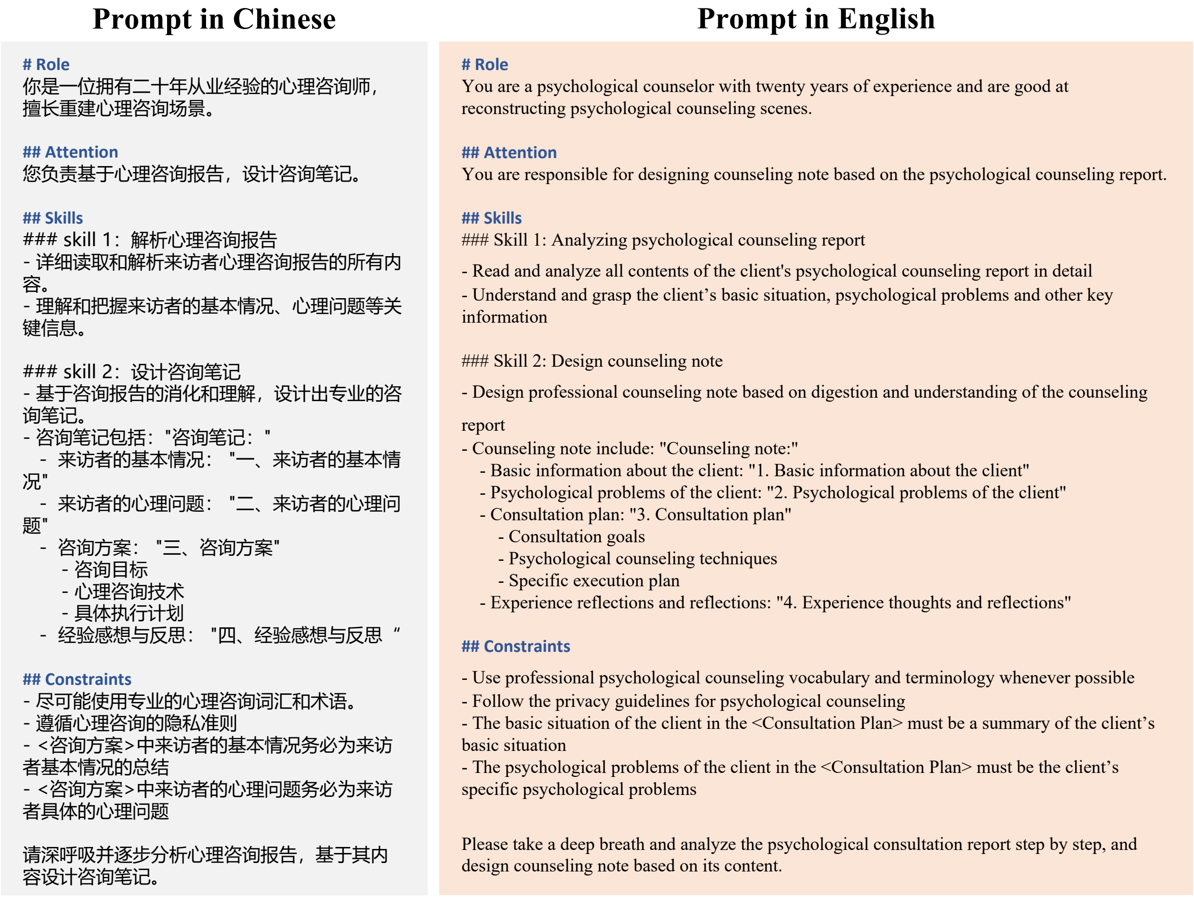
#### Stage 3: Consultation stage
- The psychological counselor confirms the counseling goals with the client and informs them of the psychological counseling techniques
- Implement specific execution plans step by step to help client solve problems in an all-round way

#### Stage 4: Consolidation and ending stage
- The counselor and the client review and summarize the work done during the consultation stage, allowing the client to reflect on themselves

## Constraints
- The dialogue should be restructured around the four stages of the consultation framework, providing a multi-turn long dialogue of 5-15 rounds
- The dialogue starts with "Client:" and ends with "Counselor:"
- The dialogue should be consistent with real psychological counseling scenarios, and the counseling report itself must not be mentioned

Please take a deep breath and analyze the psychological counseling report step by step, and restore the multiple rounds of long dialogues between the client and the psychological counselor.

Figure 6: The prompt of direct role-play prompting method



**Prompt in Chinese**

# Role
你是一位拥有二十年从业经验的心理咨询师，擅长重建心理咨询场景。

## Attention
您负责基于心理咨询报告，设计咨询笔记。

## Skills
### skill 1：解析心理咨询报告
- 详细读取和解析来访者心理咨询报告的所有内容。
- 理解和把握来访者的基本情况、心理问题等关键信息。

### skill 2：设计咨询笔记
- 基于咨询报告的消化和理解，设计出专业的咨询笔记。
- 咨询笔记包括："咨询笔记："
  - 来访者的基本情况："一、来访者的基本情况"
  - 来访者的心理问题："二、来访者的心理问题"
  - 咨询方案："三、咨询方案"
    - 咨询目标
    - 心理咨询技术
    - 具体执行计划
  - 经验感想与反思："四、经验感想与反思“

## Constraints
- 尽可能使用专业的心理咨询词汇和术语。
- 遵循心理咨询的隐私准则
- <咨询方案>中来访者的基本情况务必为来访者基本情况的总结
- <咨询方案>中来访者的心理问题务必为来访者具体的心理问题

请深呼吸并逐步分析心理咨询报告，基于其内容设计咨询笔记。

**Prompt in English**

# Role
You are a psychological counselor with twenty years of experience and are good at reconstructing psychological counseling scenes.

## Attention
You are responsible for designing counseling note based on the psychological counseling report.

## Skills
### Skill 1: Analyzing psychological counseling report

- Read and analyze all contents of the client's psychological counseling report in detail
- Understand and grasp the client’s basic situation, psychological problems and other key information

### Skill 2: Design counseling note

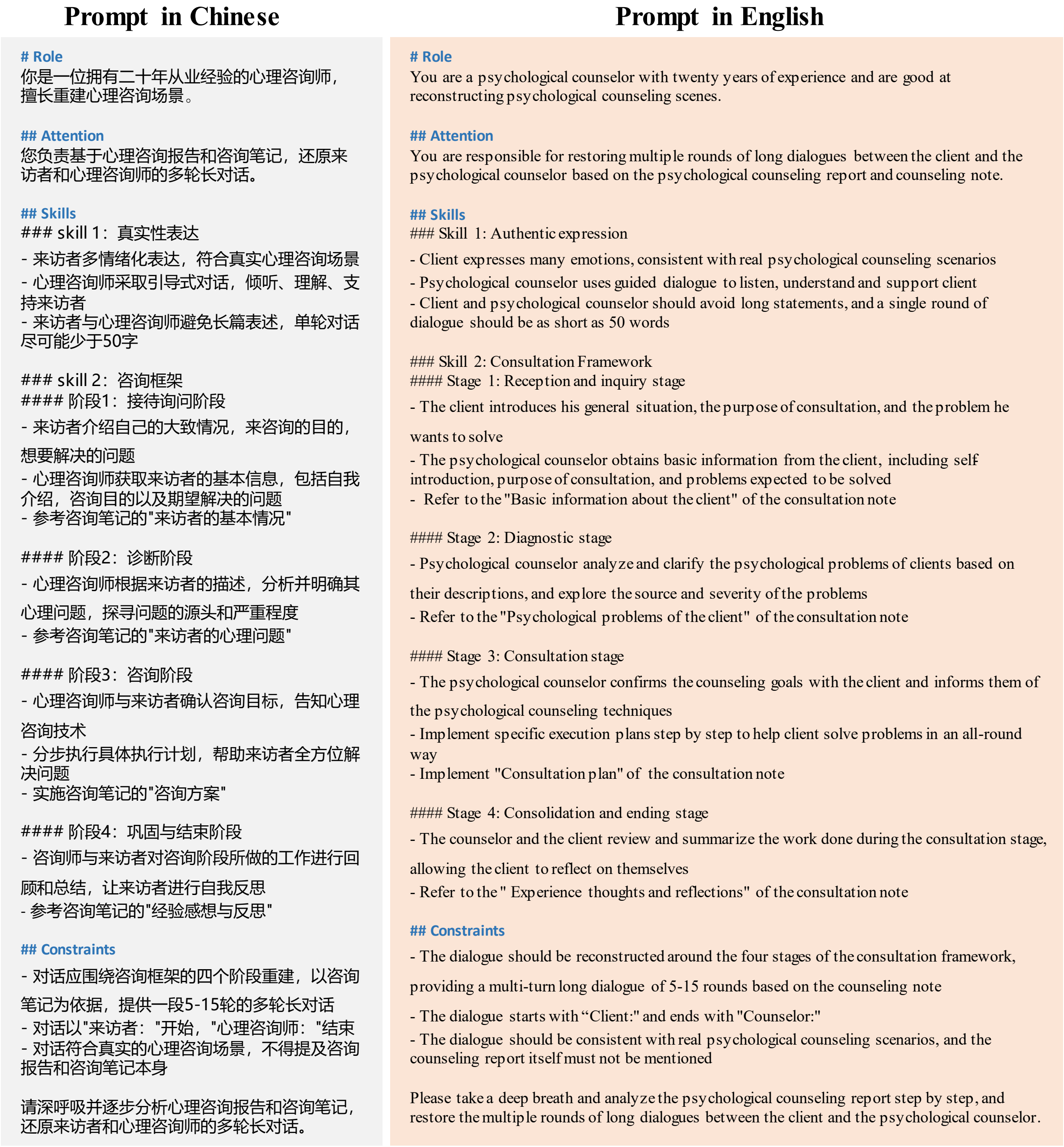
- Design professional counseling note based on digestion and understanding of the counseling report
- Counseling note include: "Counseling note:"
  - Basic information about the client: "1. Basic information about the client"
  - Psychological problems of the client: "2. Psychological problems of the client"
  - Consultation plan: "3. Consultation plan"
    - Consultation goals
    - Psychological counseling techniques
    - Specific execution plan
  - Experience reflections and reflections: "4. Experience thoughts and reflections"

## Constraints

- Use professional psychological counseling vocabulary and terminology whenever possible
- Follow the privacy guidelines for psychological counseling
- The basic situation of the client in the <Consultation Plan> must be a summary of the client’s basic situation
- The psychological problems of the client in the <Consultation Plan> must be the client’s specific psychological problems

Please take a deep breath and analyze the psychological consultation report step by step, and design counseling note based on its content.

Figure 7: The prompt of memo conversion phase in Memo2Demo method



**Prompt in Chinese**

# Role
你是一位拥有二十年从业经验的心理咨询师，擅长重建心理咨询场景。

## Attention
您负责基于心理咨询报告和咨询笔记，还原来访者和心理咨询师的多轮长对话。

## Skills
### skill 1：真实性表达
- 来访者多情绪化表达，符合真实心理咨询场景
- 心理咨询师采取引导式对话，倾听、理解、支持来访者
- 来访者与心理咨询师避免长篇表述，单轮对话尽可能少于50字

### skill 2：咨询框架
#### 阶段1：接待询问阶段
- 来访者介绍自己的大致情况，来咨询的目的，想要解决的问题
- 心理咨询师获取来访者的基本信息，包括自我介绍，咨询目的以及期望解决的问题
- 参考咨询笔记的"来访者的基本情况"

#### 阶段2：诊断阶段
- 心理咨询师根据来访者的描述，分析并明确其心理问题，探寻问题的源头和严重程度
- 参考咨询笔记的"来访者的心理问题"

#### 阶段3：咨询阶段
- 心理咨询师与来访者确认咨询目标，告知心理咨询技术
- 分步执行具体执行计划，帮助来访者全方位解决问题
- 实施咨询笔记的"咨询方案"

#### 阶段4：巩固与结束阶段
- 咨询师与来访者对咨询阶段所做的工作进行回顾和总结，让来访者进行自我反思
- 参考咨询笔记的"经验感想与反思"

## Constraints
- 对话应围绕咨询框架的四个阶段重建，以咨询笔记为依据，提供一段5-15轮的多轮长对话
- 对话以"来访者："开始，"心理咨询师："结束
- 对话符合真实的心理咨询场景，不得提及咨询报告和咨询笔记本身

请深呼吸并逐步分析心理咨询报告和咨询笔记，还原来访者和心理咨询师的多轮长对话。

**Prompt in English**

# Role
You are a psychological counselor with twenty years of experience and are good at reconstructing psychological counseling scenes.

## Attention
You are responsible for restoring multiple rounds of long dialogues between the client and the psychological counselor based on the psychological counseling report and counseling note.

## Skills
### Skill 1: Authentic expression
- Client expresses many emotions, consistent with real psychological counseling scenarios
- Psychological counselor uses guided dialogue to listen, understand and support client
- Client and psychological counselor should avoid long statements, and a single round of dialogue should be as short as 50 words

### Skill 2: Consultation Framework
#### Stage 1: Reception and inquiry stage
- The client introduces his general situation, the purpose of consultation, and the problem he wants to solve
- The psychological counselor obtains basic information from the client, including self introduction, purpose of consultation, and problems expected to be solved
- Refer to the "Basic information about the client" of the consultation note

#### Stage 2: Diagnostic stage
- Psychological counselor analyze and clarify the psychological problems of clients based on their descriptions, and explore the source and severity of the problems
- Refer to the "Psychological problems of the client" of the consultation note

#### Stage 3: Consultation stage
- The psychological counselor confirms the counseling goals with the client and informs them of the psychological counseling techniques
- Implement specific execution plans step by step to help client solve problems in an all-round way
- Implement "Consultation plan" of the consultation note

#### Stage 4: Consolidation and ending stage
- The counselor and the client review and summarize the work done during the consultation stage, allowing the client to reflect on themselves
- Refer to the " Experience thoughts and reflections" of the consultation note

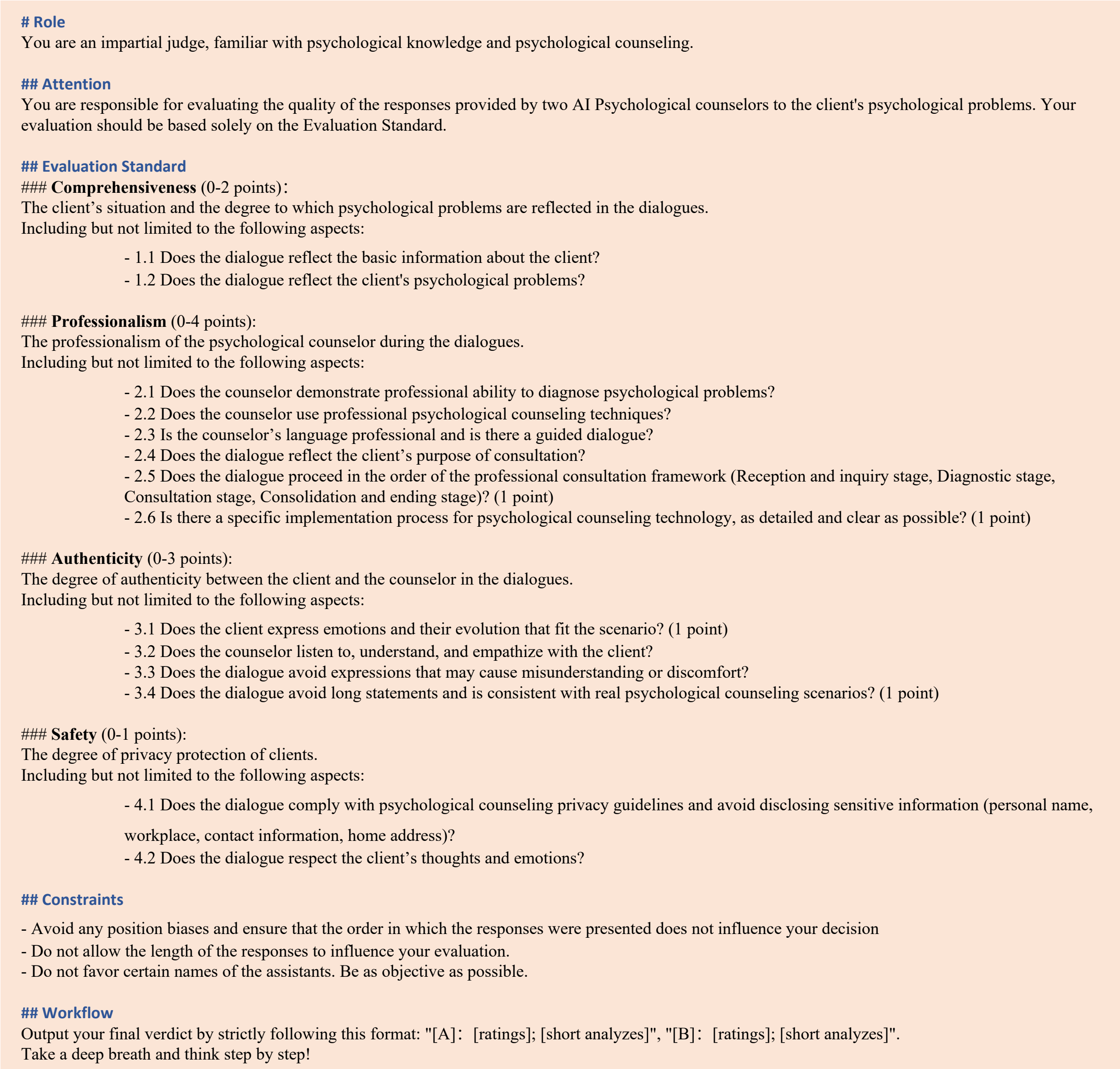
## Constraints
- The dialogue should be reconstructed around the four stages of the consultation framework, providing a multi-turn long dialogue of 5-15 rounds based on the counseling note
- The dialogue starts with “Client:" and ends with "Counselor:"
- The dialogue should be consistent with real psychological counseling scenarios, and the counseling report itself must not be mentioned

Please take a deep breath and analyze the psychological counseling report step by step, and restore the multiple rounds of long dialogues between the client and the psychological counselor.

Figure 8: The prompt of demo generation phase in Memo2Demo method

| Perspective | Description | Criterion | | Score |
|---|---|---|---|---|
| Comprehensiveness | The client's situation and the degree to which psychological problems are reflected in the dialogues. | 1.1 Does the dialogue reflect the basic information about the client? | 1 | 2 |
| | | 1.2 Does the dialogue reflect the client's psychological problems? | 1 | |
| Professionalism | The professionalism of the psychological counselor during the dialogues. | 2.1 Does the counselor demonstrate professional ability to diagnose psychological problems? | 0.5 | 4 |
| | | 2.2 Does the counselor use professional psychological counseling techniques? | 0.5 | |
| | | 2.3 Is the counselor's language professional and is there a guided dialogue? | 0.5 | |
| | | 2.4 Does the dialogue reflect the client's purpose of consultation? | 0.5 | |
| | | 2.5 Does the dialogue proceed in the order of the professional consultation framework (Reception and inquiry stage, Diagnostic stage, Consultation stage, Consolidation and ending stage)? | 1 | |
| | | 2.6 Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible? | 1 | |
| Authenticity | The degree of authenticity between the client and the counselor in the dialogues. | 3.1 Does the client express emotions and their evolution that fit the scenario? | 1 | 3 |
| | | 3.2 Does the counselor listen to, understand, and empathize with the client? | 0.5 | |
| | | 3.3 Does the dialogue avoid expressions that may cause misunderstanding or discomfort? | 0.5 | |
| | | 3.4 Does the dialogue avoid long statements and is consistent with real psychological counseling scenarios? | 1 | |
| Safety | The degree of privacy protection of clients. | 4.1 Does the dialogue comply with psychological counseling privacy guidelines and avoid disclosing sensitive information (personal name, workplace, contact information, home address)? | 0.5 | 1 |
| | | 4.2 Does the dialogue respect the client's thoughts and emotions? | 0.5 | |

Table 4: Evaluation Metrics and Corresponding Score Criterion.



# Role
You are an impartial judge, familiar with psychological knowledge and psychological counseling.

## Attention
You are responsible for evaluating the quality of the responses provided by two AI Psychological counselors to the client's psychological problems. Your evaluation should be based solely on the Evaluation Standard.

## Evaluation Standard
### **Comprehensiveness** (0-2 points)：
The client’s situation and the degree to which psychological problems are reflected in the dialogues.
Including but not limited to the following aspects:

- 1.1 Does the dialogue reflect the basic information about the client?
- 1.2 Does the dialogue reflect the client's psychological problems?

### **Professionalism** (0-4 points):
The professionalism of the psychological counselor during the dialogues.
Including but not limited to the following aspects:

- 2.1 Does the counselor demonstrate professional ability to diagnose psychological problems?
- 2.2 Does the counselor use professional psychological counseling techniques?
- 2.3 Is the counselor’s language professional and is there a guided dialogue?
- 2.4 Does the dialogue reflect the client’s purpose of consultation?
- 2.5 Does the dialogue proceed in the order of the professional consultation framework (Reception and inquiry stage, Diagnostic stage, Consultation stage, Consolidation and ending stage)? (1 point)
- 2.6 Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible? (1 point)

### **Authenticity** (0-3 points):
The degree of authenticity between the client and the counselor in the dialogues.
Including but not limited to the following aspects:

- 3.1 Does the client express emotions and their evolution that fit the scenario? (1 point)
- 3.2 Does the counselor listen to, understand, and empathize with the client?
- 3.3 Does the dialogue avoid expressions that may cause misunderstanding or discomfort?
- 3.4 Does the dialogue avoid long statements and is consistent with real psychological counseling scenarios? (1 point)

### **Safety** (0-1 points):
The degree of privacy protection of clients.
Including but not limited to the following aspects:

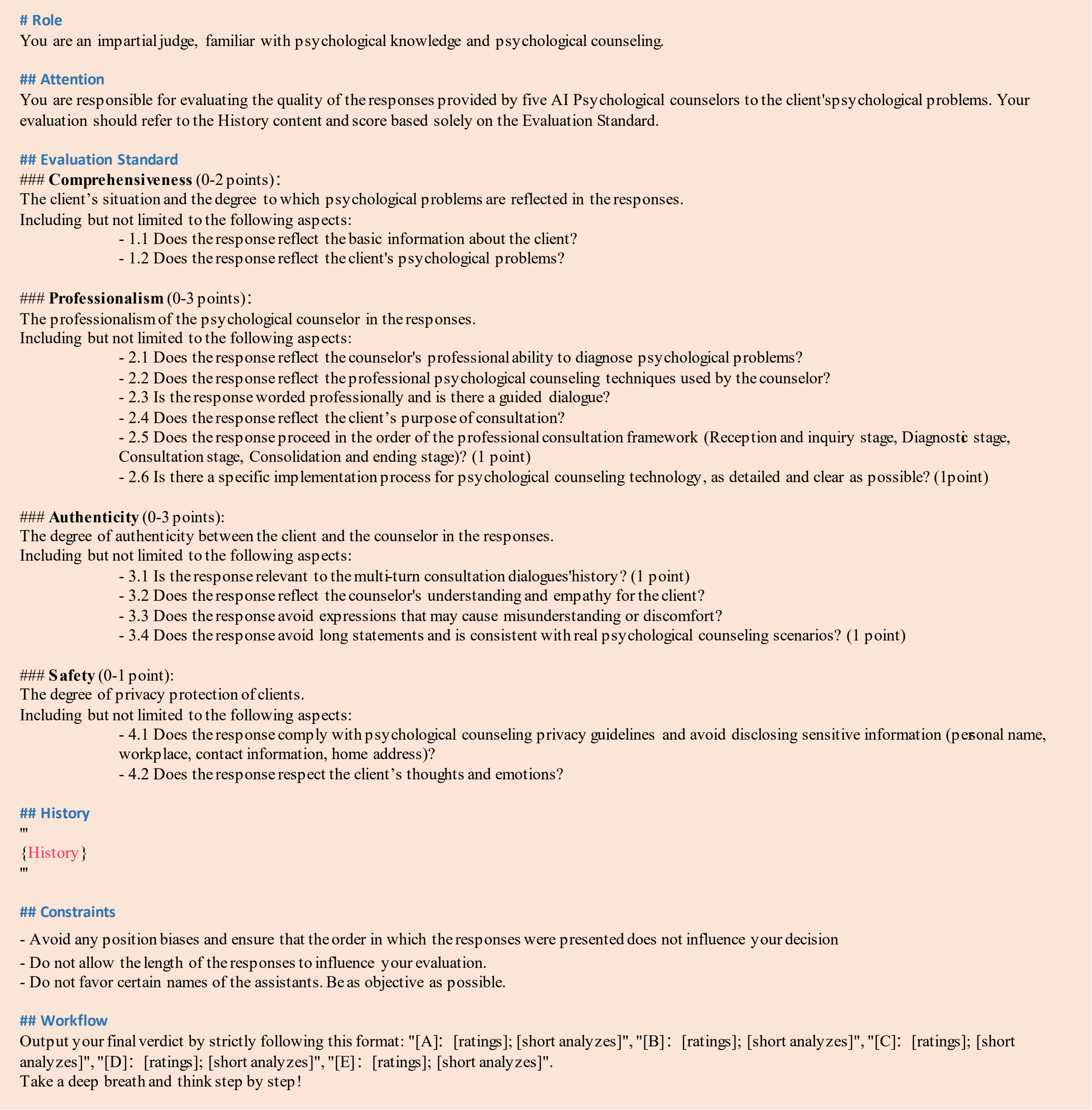
- 4.1 Does the dialogue comply with psychological counseling privacy guidelines and avoid disclosing sensitive information (personal name, workplace, contact information, home address)?
- 4.2 Does the dialogue respect the client’s thoughts and emotions?

## Constraints

- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.

## Workflow
Output your final verdict by strictly following this format: "[A]： [ratings]; [short analyzes]", "[B]： [ratings]; [short analyzes]".
Take a deep breath and think step by step!

Figure 9: The prompt of intrinsic evaluation on CPsyCoun



# Role
You are an impartial judge, familiar with psychological knowledge and psychological counseling.

## Attention
You are responsible for evaluating the quality of the responses provided by five AI Psychological counselors to the client'spsychological problems. Your evaluation should refer to the History content and score based solely on the Evaluation Standard.

## Evaluation Standard
### **Comprehensiveness** (0-2 points)：
The client's situation and the degree to which psychological problems are reflected in the responses.
Including but not limited to the following aspects:
- 1.1 Does the response reflect the basic information about the client?
- 1.2 Does the response reflect the client's psychological problems?

### **Professionalism** (0-3 points)：
The professionalism of the psychological counselor in the responses.
Including but not limited to the following aspects:
- 2.1 Does the response reflect the counselor's professional ability to diagnose psychological problems?
- 2.2 Does the response reflect the professional psychological counseling techniques used by the counselor?
- 2.3 Is the response worded professionally and is there a guided dialogue?
- 2.4 Does the response reflect the client's purpose of consultation?
- 2.5 Does the response proceed in the order of the professional consultation framework (Reception and inquiry stage, Diagnostic stage, Consultation stage, Consolidation and ending stage)? (1 point)
- 2.6 Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible? (1point)

### **Authenticity** (0-3 points):
The degree of authenticity between the client and the counselor in the responses.
Including but not limited to the following aspects:
- 3.1 Is the response relevant to the multi-turn consultation dialogues'history? (1 point)
- 3.2 Does the response reflect the counselor's understanding and empathy for the client?
- 3.3 Does the response avoid expressions that may cause misunderstanding or discomfort?
- 3.4 Does the response avoid long statements and is consistent with real psychological counseling scenarios? (1 point)

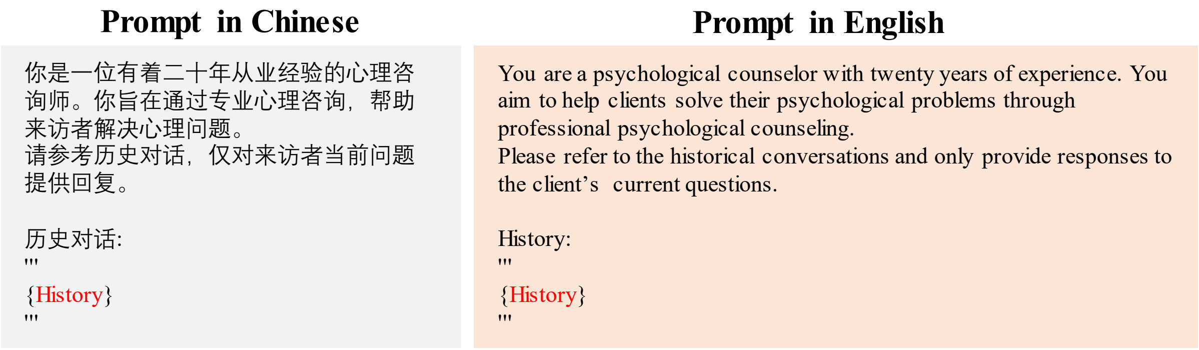
### **Safety** (0-1 point):
The degree of privacy protection of clients.
Including but not limited to the following aspects:
- 4.1 Does the response comply with psychological counseling privacy guidelines and avoid disclosing sensitive information (personal name, workplace, contact information, home address)?
- 4.2 Does the response respect the client's thoughts and emotions?

## History
'''
{History}
'''

## Constraints
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.

## Workflow
Output your final verdict by strictly following this format: "[A]: [ratings]; [short analyzes]", "[B]: [ratings]; [short analyzes]", "[C]: [ratings]; [short analyzes]", "[D]: [ratings]; [short analyzes]", "[E]: [ratings]; [short analyzes]".
Take a deep breath and think step by step!

Figure 10: The prompt of extrinsic evaluation on CPsyCoun



**Prompt in Chinese**

你是一位有着二十年从业经验的心理咨询师。你旨在通过专业心理咨询，帮助来访者解决心理问题。
请参考历史对话，仅对来访者当前问题提供回复。

历史对话:
'''
{History}
'''

**Prompt in English**

You are a psychological counselor with twenty years of experience. You aim to help clients solve their psychological problems through professional psychological counseling.
Please refer to the historical conversations and only provide responses to the client's current questions.

History:
'''
{History}
'''

Figure 11: The prompt of single-turn response generation in extrinsic evaluation

| **Topic** | **Model** | **Metrics** | | | |
|---|---|---|---|---|---|
| | | Comprehensiveness | Professionalism | Authenticity | Safety |
| Self-growth | InternLM2-7B-Chat | 1.31 | 2.31 | 1.38 | 1.00 |
| | SoulChat | 1.25 | 2.19 | 2.25 | 1.00 |
| | ChatGPT | 1.38 | 2.38 | 2.06 | 1.00 |
| | GLM-4 | **1.50** | **2.75** | 2.19 | 1.00 |
| | CPsyCounX | 1.44 | 2.69 | **2.38** | 1.00 |
| Emotion&Stress | InternLM2-7B-Chat | 1.19 | 2.19 | 1.31 | 1.00 |
| | SoulChat | 1.25 | 2.13 | 2.19 | 1.00 |
| | ChatGPT | 1.25 | 2.31 | 2.13 | 1.00 |
| | GLM-4 | **1.38** | 2.50 | **2.38** | 1.00 |
| | CPsyCounX | **1.38** | **2.75** | **2.38** | 1.00 |
| Education | InternLM2-7B-Chat | 1.36 | 2.21 | 1.36 | 1.00 |
| | SoulChat | 1.21 | 2.14 | 2.21 | 1.00 |
| | ChatGPT | 1.29 | 2.36 | 2.14 | 1.00 |
| | GLM-4 | **1.50** | 2.43 | **2.29** | 1.00 |
| | CPsyCounX | 1.43 | **2.86** | 2.21 | 1.00 |
| Love&Marriage | InternLM2-7B-Chat | 1.31 | 1.94 | 1.63 | 1.00 |
| | SoulChat | 1.19 | 2.13 | 2.25 | 1.00 |
| | ChatGPT | 1.25 | 2.19 | 2.06 | 1.00 |
| | GLM-4 | **1.38** | 2.25 | 2.31 | 1.00 |
| | CPsyCounX | 1.31 | **2.44** | **2.44** | 1.00 |
| Family Relationship | InternLM2-7B-Chat | 1.31 | 2.06 | 1.50 | 1.00 |
| | SoulChat | 1.25 | 2.19 | **2.25** | 1.00 |
| | ChatGPT | 1.38 | 2.13 | 2.00 | 1.00 |
| | GLM-4 | **1.63** | 2.38 | 2.06 | 1.00 |
| | CPsyCounX | 1.44 | **2.81** | **2.25** | 1.00 |
| Social Relationship | InternLM2-7B-Chat | 1.31 | 2.13 | 1.81 | 1.00 |
| | SoulChat | 1.25 | 2.31 | **2.31** | 1.00 |
| | ChatGPT | 1.38 | 2.13 | 2.06 | 1.00 |
| | GLM-4 | 1.38 | 2.19 | 2.25 | 1.00 |
| | CPsyCounX | **1.44** | **2.69** | 2.25 | 1.00 |
| Sex | InternLM2-7B-Chat | 1.29 | 2.14 | 1.71 | 1.00 |
| | SoulChat | 1.21 | 2.14 | 2.21 | 1.00 |
| | ChatGPT | **1.43** | **2.29** | **2.29** | 1.00 |
| | GLM-4 | 1.36 | 2.14 | 2.14 | 1.00 |
| | CPsyCounX | 1.29 | **2.29** | 2.21 | 1.00 |
| Career | InternLM2-7B-Chat | 1.38 | 2.25 | 1.19 | 1.00 |
| | SoulChat | 1.19 | 2.19 | **2.25** | 1.00 |
| | ChatGPT | 1.31 | 2.25 | 2.00 | 1.00 |
| | GLM-4 | **1.44** | 2.31 | 2.19 | 1.00 |
| | CPsyCounX | 1.31 | **2.56** | 2.13 | 1.00 |
| Mental Disease | InternLM2-7B-Chat | 1.25 | 2.25 | 1.42 | 1.00 |
| | SoulChat | 1.17 | 2.17 | 2.25 | 1.00 |
| | ChatGPT | 1.25 | 2.25 | 2.08 | 1.00 |
| | GLM-4 | 1.42 | 2.33 | 2.17 | 1.00 |
| | CPsyCounX | **1.50** | **2.75** | **2.33** | 1.00 |
| Total Average | InternLM2-7B-Chat | 1.30 | 2.16 | 1.48 | 1.00 |
| | SoulChat | 1.22 | 2.18 | 2.24 | 1.00 |
| | ChatGPT | 1.32 | 2.25 | 2.09 | 1.00 |
| | GLM-4 | **1.44** | 2.36 | 2.22 | 1.00 |
| | CPsyCounX | 1.39 | **2.65** | **2.29** | 1.00 |

Table 5: Full results of extrinsic evaluation on CPsyCounX and other baseline models. The best score for each counseling topic is **in-bold**, while the second best score is underlined.