Understanding Client Reactions in Online Mental Health Counseling

Anqi Li[1,2]*, **Lizhi Ma**[2]*, **Yaling Mei**[2]
Hongliang He[1,2], **Shuai Zhang**[1,2], **Huachuan Qiu**[1,2], **Zhenzhong Lan**[2]†
[1] Zhejiang University
[2] School of Engineering, Westlake University
{lianqi, malizhi, lanzhenzhong}@westlake.edu.cn

Abstract

Communication success relies heavily on reading participants' reactions. Such feedback is especially important for mental health counselors, who must carefully consider the client's progress and adjust their approach accordingly. However, previous NLP research on counseling has mainly focused on studying counselors' intervention strategies rather than their clients' reactions to the intervention. This work aims to fill this gap by developing a theoretically grounded annotation framework that encompasses counselors' strategies and client reaction behaviors. The framework has been tested against a large-scale, high-quality text-based counseling dataset we collected over the past two years from an online welfare counseling platform. Our study show how clients react to counselors' strategies, how such reactions affect the final counseling outcomes, and how counselors can adjust their strategies in response to these reactions. We also demonstrate that this study can help counselors automatically predict their clients' states [1].

1 Introduction

There can be no human relations without communication, yet the road to successful communication is paved with obstacles (Luhmann, 1981). Given the individuality and separateness of human consciousness, it is hard to guarantee one can receive the message sent by another. Even if the message is fully understood, there can be no assurance of its acceptance. By getting feedback from their partners, communicators can better understand their communicative states. This allows communicators to adjust their communication strategies to fit better their communication environment, which is crucial for successful communication. However, most work

Figure 1: Examples of how counselors adjust their strategies according to their client's reactions.

on improving the success rates in communication, such as persuasion (Wang et al., 2019) and mental health support (Zhang et al., 2019; Zhang and Danescu-Niculescu-Mizil, 2020; Liu et al., 2021), focuses on speakers' strategies. But little research is on how listeners' reactions shape trajectories and outcomes of conversations. In this work, we address the gap by examining how to use the reactions of clients to predict and improve the quality of psychological counseling, a field that has profound societal and research impact.

Psychological counseling is one of the most challenging and skillful forms of communication (Althoff et al., 2016). Counselors take clients through their mental health concerns while balancing the stress they are experiencing (Zhang and Danescu-Niculescu-Mizil, 2020). To do it well, counselors rely on training and on continuing experience with clients to acquire the consultative skills. However, it is difficult for counselors to get direct feedback on their interventions from clients in practice (Zhang et al., 2019). Besides, due to the lack of accurate assessments of general counseling interventions (Tracey et al., 2014), the prior studies found no noticeable improvement or effectiveness of counselors' interventions after training or counselings (Dawes, 2009; Hill et al., 2015; Goldberg et al., 2016). As a result, some have even argued that psychological counseling is" *a profession with-*

*Equal Contribution.
†Corresponding Author.
[1] You can access our annotation framework, dataset and codes from https://github.com/dll-wu/Client-Reactions.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10358–10376

out any expertise" (Shanteau, 1992). In this regard, one solution to facilitate counselors noticing the effectiveness of interventions is to know clients' feedback during counseling conversations.

However, researchers in the field mainly study counselors' skills and language patterns to provide feedback on interventions (Althoff et al., 2016; Zhang et al., 2019; Pérez-Rosas et al., 2019). They first separate counselings into two groups, high-quality and low-quality counselings. Then, features of counselors' interventions, such as language diversity, ability to handle ambiguity and make progress, are analyzed. In the end, the general patterns of the features of good counseling are reported. Nonetheless, apart from the counselors' interventions, the counseling, as a process of interactive communication, also includes clients' reactions (Avdi and Georgaca, 2007). Importantly, the clients' reactions towards counselors' intervention reflect the feedback on the effectiveness of the interventions (Ribeiro et al., 2013). Thus, to complete the assessment of counselors' interventions from the client's perspective and to provide feedback for counselors, we are motivated to categorize the clients' reactions although identifying their reactions in the psychological counseling is difficult, even more so than categorizing counselors' interventions (Lee et al., 2019; Sharma et al., 2020).

In this paper, we introduce a theoretically grounded annotation framework to map each turn of the conversation into counselors' intentions and their clients' reactions. The framework is applied to label a large-scale text-based Chinese counseling dataset collected from an online welfare counseling platform over the last two years.

Using the annotation, we analyze the associations between clients' reactions and behaviors in the counselling conversation and their assessment of conversation effectiveness. We demonstrate that the counselors' different intentions and strategies elicit different follow-up reactions and behaviors from the clients. Following this analysis, we examine how counselors should adjust their strategies to encourage clients' positive behaviors based on different conversation stages and historical interaction patterns. We also analyze how the counselors address the clients' behaviors that negatively impact the conversation effectiveness. Along with the automatic annotation classifiers we built, the findings of above analyses would help develop user-centered mental health support dialog systems.

2 Related Work

We mostly draw inspiration from conversational analysis in NLP and psychotherapy.

Despite the abundance of NLP research relating to emotional chat (Zhou et al., 2018), emotional support (Liu et al., 2021), and psycho-counseling (Althoff et al., 2016), in most cases, these studies are still in their infancy. Human-human interaction patterns are rarely studied due to the lack of large-scale conversational datasets (Huang et al., 2020). Meanwhile, the main research focus is either on proposing new datasets or studying consultation skills.

Dataset for Mental Health Support. Because of the sensitive nature of mental health data, most of the available mental health support conversation corpora are collected from public general social networking sites or crowdsourcing (Sharma et al., 2020; Harrigian et al., 2021; Sun et al., 2021; Liu et al., 2021). The potential for understanding human-human interaction patterns is limited with these single-turned or crowd-sourced datasets. Althoff et al. (2016) propose a multi-turn mental health counseling conversation corpus collected from a text-based crisis intervention platform, which is the best-related dataset up to now. However, the length of conversation in (Althoff et al., 2016) is shorter than ours (42 vs. 78 utterances), and the analysis mostly focuses on the counselors' utterances. In contrast, we emphasize the understanding and recognition of client reactions, which could facilitate counselors to understand the clients' feedback of their interventions as the psychological counselings proceed.

Understanding Mental Health Support Conversations Using NLP. Many researchers have endeavored to employ machine learning and NLP techniques to analyze mental health support conversations automatically, including modeling social factors in language that are important in the counseling setting (Danescu-Niculescu-Mizil et al., 2013; Pei and Jurgens, 2020; Sharma et al., 2021; Hovy and Yang, 2021), behavioral codes (Tanana et al., 2015; Pérez-Rosas et al., 2017; Park et al., 2019a; Cao et al., 2019), predicting session- or utterance-level quality (Gibson et al., 2016; Goldberg et al., 2020; Wu et al., 2021), and detecting mental health problems (Asad et al., 2019; Xu et al., 2020). However, these studies again mostly focus on studying consultation skills. There are methods (Tanana et al., 2015; Pérez-Rosas et al., 2017)

that try to classify clients' responses but only limit to a particular mental health support genre called motivational interviewing, which has an existing coding scheme with three classes for clients. Our annotation scheme is not genre specific and has more fine-grained analysis, and is more related to research in psychotherapy.

Analysis of Conversation Outcome in Psychotherapy Research. Different from NLP research where most studies focus on the counselor side, in psychotherapy research, the interactions between counselors and clients are widely investigated (Ribeiro et al., 2013; Norcross, 2010; Falkenström et al., 2014). The working alliance between the counselor and clients is a crucial researched element [2] (Norcross, 2010; Falkenström et al., 2014). This is because the formation of working alliance is arguably the most reliable predictor of counseling conversation outcomes (Ribeiro et al., 2013), yet it is difficult for counselors to gauge accurately during counselings. The scores of alliance rated after each counseling from therapists "appear to be independent of . . . alliance data obtained from their patients" (Horvath and Greenberg, 1994). Additionally, limited by the data resource and analysis tools, most alliance analyses in psychotherapy research are either in small sample size (Ribeiro et al., 2013) with only a few sessions or in session level (Hatcher, 1999). We instead conduct a moment-by-moment analysis on a large-scale dataset and pursue an automatic solution.

3 Annotation Framework

To understand interaction patterns between counselors and clients in text-based counseling conversations, we develop a novel framework to categorize the reactions and behaviors of clients as well as the intentions and conversational strategies of counselors (Figure 2). In collaboration with experts in counseling psychology, we adapt and synthesize the existing face-to-face counseling-focused taxonomies, including Client Behavior System (Hill et al., 1992), Therapeutic Collaboration Coding Scheme (Ribeiro et al., 2013), Helping Skills (Hill, 2009), and Client Resistance Coding Scheme (Chamberlain et al., 1984), to the online text-only counseling conversation settings. We have three developers [3] to carefully build the framework, following the consensual qualitative research method (Hill et al., 1997; Ribeiro et al., 2013; Park et al., 2019b). The details of the framework development process are shown in Appendix A.1. We also compare our framework with existing annotation frameworks in Appendix A.2.

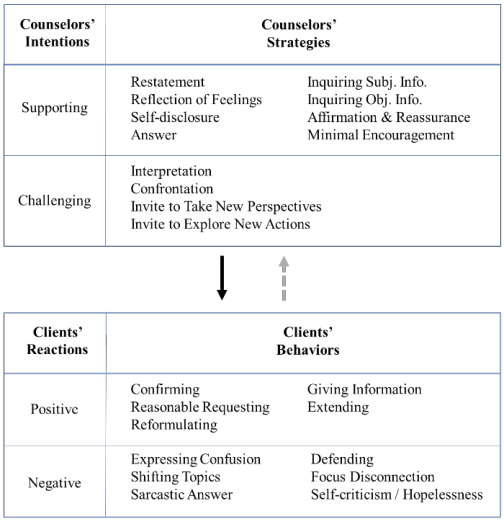

Figure 2: Overview of our proposed framework. It contains the intentions and conversational strategies of counselors, as well as the reactions and behaviors of clients. The **black** arrow indicates the influence of counselors' intervention on clients' reactions and behaviors. The dashed gray arrow indicates that clients' feedback is regarded as signals for counselors to adjust intentions and strategies in conversations.

3.1 Counselor Intentions and Conversational Strategies

Counselor Intentions. Our taxonomy consists of two key counselor intentions, *Supporting* and *Challenging*, providing an outlook of how counselors orient the conversation flow (Ribeiro et al., 2013; Zhang and Danescu-Niculescu-Mizil, 2020).

In a counseling conversation, the counselor must focus on engaging with the client's concerns and providing an empathetic understanding (Rogers, 1957; Hill and Nakayama, 2000). However, overemphasizing the supportive strategies might keep the client from progressing (Zhang and Danescu-Niculescu-Mizil, 2020; Ribeiro et al., 2013). To direct the conversation towards a pos-

[2] **Working Alliance**: "*the alliance represents interactive, collaborative elements of the relationship (i.e., therapist and client abilities to engage in the tasks of therapy and to agree on the targets of therapy) in the context of an affective bond or positive attachment*" (Constantino et al., 2002).

[3] One is a Ph.D. in psychology and a State-Certificated Class 3 Psycho-counselor with 3 years of experience; another is a State-Certificated Class 2 Psycho-counselor with more than 10 years of experience; and the last one is a doctoral student majoring in computer science and the first author of this paper.

itive outcome that benefits clients, the counselor should challenge and prompt the client to make some changes (Mishara et al., 2007; Zhang and Danescu-Niculescu-Mizil, 2020). By analyzing the collected counseling conversations, we do find it common for counselors to employ supportive and challenging strategies alternatively in practice.

Conversational Strategies. Our taxonomy contains eight *Supporting* and four *Challenging* fine-grained conversational strategies. We present detailed definitions and examples in Appendix A.3.

Counselors utilize various conversational strategies to convey their intentions (Hill, 2009). To provide support, the counselors reflect on the contents or feelings the client has shared to make the client feel heard and understood (*Restatement* and *Reflection of Feelings*). The counselor also affirms the client's strengths or normalizes the client's negative emotions by expressing reassurance (*Affirmation and Reassurance*). On the other hand, to prompt the client to make progress, the counselor might point out the client's unreasonable beliefs (*Confrontation*) or encourage him or her to brainstorm solutions (*Invite to Explore New Actions*).

Notably, our annotation framework captures functional details of conversational strategies (Ribeiro et al., 2013). For example, although both *Interpretation* and *Invite to Take New Perspectives* encourage clients to view life from different angles, the way in which the insights are provided differs. *Interpretation* strategy directly provides a new meaning, reason, or explanation to the client's behavior, thought, or emotion from the perspective beyond the client's statement or cognition. For example, "Comparing yourself to others makes you feel unsatisfied with yourself. But everyone's growth has its timeline". While *Invite to Take New Perspectives* strategy usually guides the client to think from a new perspective by asking questions. For example, "If your closest friend heard your appeal, what do you think he would say to you?"

3.2 Client Reactions and Behaviors

Client Reactions. The counselors' interventions elicit the clients' reactions, which is an important criterion for judging the effectiveness of counselors' previous interventions. The clients' reactions towards the counselors' interventions can be categorized as *Positive* or *Negative* as feedback of whether they understand counselors' purposes of using specific intentions and strategies (Leiman and Stiles, 2001; Hill, 2009; Ribeiro et al., 2013). For example, when the counselor utilizes *Affirmation and Reassurance* strategy to show empathy to the client by saying, " You have a great insight into yourself!", the client may experience being understood and respond with confirmation by saying, "Thank you for your accomplishment!"; or the client may find the mere consolation is useless in resolving the dilemma of the moment and then express dissatisfaction with the counselor's intervention by saying "You always comfort me. But is there any concrete advice or suggestions?". The client's negative reactions indicate that the counselor intentions fail to achieve the intentions as expected, indicating the counselor needs to adjust strategies in the ensuing conversations (Thomas, 1983; Zhang and Danescu-Niculescu-Mizil, 2020; Li et al., 2022).

Behaviors. Our taxonomy contains five and six fine-grained behavior types for clients' *Positive* and *Negative* reactions, respectively. Detailed definitions are in Appendix A.4.

Clients react to the counselor's interventions through different behaviors. For example, when the counselor provides a perspective different from a client to help the client understand a distressing experience (*Interpretation*), the client may express approval (*Confirming*) or start introspection (*Extending*); on the contrary, the client may still insist on individual inherent views and directly express disagreement with what the counselor has said (*Defending*) or show disinterest in counselor's words implicitly by changing the topic (*Changing Topics*).

4 Data Collection

To validate the feasibility of our proposed framework in the psychological counseling conversation, we collect a large-scale counseling corpus and carefully annotate a subset of these conversations according to the framework. Our dataset will be made available for researchers who agree to follow ethical guidelines.

4.1 Data Source

We build an online mental health support platform called Xinling to allow professional counselors to provide each client with a free text-based counseling service of about 50 minutes each time, which is a widely recognized basic time setting in psychological counseling. After each conversation,

clients are asked to report their clarity on the approaches to solve existing problems by rating the conversations based on the following aspects: (1) Awareness of the changes that can be made; (2) New perspectives of looking at the problems; (3) Confidence in the ways of coping with the problems; (4) Confidence in the conversations that can lead to desirable outcomes. Clients' self-reported scores on these scales have been recognized as a consistent and major positive indicator of effective counseling (Tracey and Kokotovic, 1989; Hill, 2009). Details of the post-survey are in Table 7 in Appendix B.1. We then collect counseling conversations between actual clients and experienced counselors from this counseling platform.

In the end, we collect 2,382 conversation sessions, 479 of which receive the self-reported scales from the clients. To our knowledge, this is the largest real-world counseling conversation corpus in Mandarin. The statistics of all the collected conversations are presented in Table 1. We observe that, on average, these conversations are much longer than existing conversations collected through crowdsourcing (78.49 utterances compared to 29.8 utterances in ESConv (Liu et al., 2021)), indicating that, in real scenarios, the professional counseling conversations contain more turns of interaction. Meanwhile, clients express longer utterances than counselors (avg. 32.48 characters compared to 24.11 characters) because clients need to give details of their problems and are encouraged to express them in the conversations, while counselors mainly act as listeners.

Category	Total	Counselor	Client
# Dialogues	2,382	-	-
# Dialogues with Scores	479	-	-
# Speakers	848	40	808
# Utterances	186,972	93,851	93,121
Avg. utterances per dialogue	78.49	39.40	39.09
Avg. length per utterance	28.28	24.11	32.48

Table 1: Statistics of the overall conversations.

4.2 Annotation Process

We randomly annotate a subset of sessions (520 sessions) based on the proposed framework[4]. Previous research found it difficult to accurately identify counselors' conversational skills (Lee et al., 2019; Sharma et al., 2020) and challenging to categorize clients' behaviors due to the linguistic diversity (Lee et al., 2019). To ensure high-quality labeling, we carefully select and train 12 annotators offline. To further enhance inter-rater reliability continuously, we design a novel training-in-the-loop annotation process. The overall average inter-rater agreement on labeling counselors' and clients' utterances is 0.67 and 0.59, respectively, validating the reliability of the data. Details about the process of annotators selection and training and the training-in-the-loop policy are displayed in Appendix B. We use a free, open-source text annotation platform called Doccano[5] to annotate.

4.3 Data Characteristics

Table 2 shows the statistics of all the annotations, including counselors' intentions and strategies and clients' reactions and behaviors.

	Categories	Num	Mean Length
Counselors' Intentions and Strategies	*Supporting*	*20608*	*16.80*
	Restatement	4553	24.54
	Reflection of Feelings	729	20.08
	Self-disclosure	122	34.5
	Inquiring Subjective Information	5746	18.06
	Inquiring Objective Information	2424	16.20
	Affirmation and Reassurance	3279	17.99
	Minimal Encouragement	3485	2.53
	Answer	273	17.46
	Challenging	*5198*	*33.95*
	Interpretation	2209	36.30
	Confrontation	141	26.27
	Invite to Explore New Actions	2495	33.57
	Invite to Take New Perspectives	353	25.02
	Others	3593	17.57
	Overall	**29399**	**19.92**
Clients' Reactions and Behaviors	*Positive*	*22136*	*32.72*
	Giving Information	15365	40.91
	Confirming	3789	3.52
	Reasonable Request	908	16.47
	Extending	1904	32.52
	Reformulation	170	33.12
	Negative	*753*	*18.65*
	Expressing Confusion	214	12.31
	Defending	425	20.72
	Self-criticism or Hopelessness	51	17.27
	Changing Topics	20	26.55
	Sarcastic Answer	32	18.53
	Focus Disconnection	11	54.45
	Others	3245	9.19
	Overall	**26134**	**29.40**

Table 2: Statistics of all the annotations, including counselors' intentions and strategies and clients' reactions and behaviors.

Overall, counselors use about four times more

[4]Before annotation, we anonymize all the client's personal information, including name, organization, etc., to protect their privacy.

[5]https://github.com/doccano/doccano

supporting than challenging strategies. The most frequently used strategy is *Inquiring Subjective Information* which helps counselors gain a deeper understanding of clients' cognitive and behavioral patterns by exploring their subjective feelings, thoughts, and reasons behind them. According to challenging strategies, *Confrontation* is used much less than *Interpretation* and *Invite to Explore New Actions*. This phenomenon is in line with the existing theory of helping skills in supportive conversations (Hill, 2009) that *Confrontation* should be used with caution because directly pointing out clients' incorrect beliefs or inconsistencies in conversations is likely to damage the relationship between counselors and clients.

As for clients' reactions and behaviors, clients' *Positive* reactions towards counselors' interventions are significantly more than the *Negative* ones, demonstrating an overall high quality of the collected counseling conversations. The most frequent behavior is *Giving Information*, which corresponds to the amount of counselors' strategy *Inquiring Subjective and Objective Information*, the clients provide the information that the counselors ask for. Besides, *Defending* is the most common negative behavior, reflecting that counselors try to get clients to change their perspectives or behaviors during conversations. Still, clients feel hard to follow and therefore defend and insist on their original cognitive and behavioral patterns. Some more extreme behaviors, such as *Self-criticism or Hopelessness*, rarely occurs, hence post difficulties for us to understand these behaviors and build good classifiers on them.

5 Application to Online Counseling

To illustrate how the proposed framework can be used to monitor and improve the effectiveness of conversations, we conduct the following analyses:

First, we demonstrate clients' positive and negative reactions and behaviors affect the final counseling effectiveness (Section 5.1). We then show how clients react to counselors' intentions and strategies (Section 5.2). Based on these findings, we investigate how counselors can adjust their strategies accordingly to make entire conversations more effective (Section 5.3). Finally, we build a baseline model for automatically labeling each counseling strategy and client behavior (Section 5.4).

5.1 How Client Reactions Indicates Counseling Effectiveness

To derive a simple conversation-level measurement, we calculate the proportion of each reaction or behavior over all the client messages in a conversation. We use the client's perceived total score on the post-conversation survey as an effectiveness indicator.

Reactions The relationship between the distribution of negative reaction types and client-rated conversation effectiveness is analyzed by Pearson Correlation Analysis (Lee Rodgers and Nicewander, 1988). The results show that the proportion of the clients' negative reactions and the conversation effectiveness correlate negatively with correlation coefficient $\rho = -0.2080$ and p-value $p = 1.7591e^{-5}$. Specifically, when clients have more *Negative* reactions to counselors' interventions, they give a lower score of conversation effectiveness (see Figure 3). The findings echo the definition of clients' *Negative* reaction types, which place a negative impact on the effectiveness of counseling conversations.

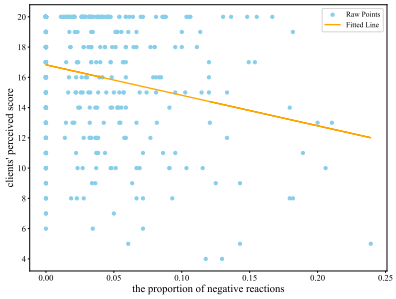

Figure 3: The scatter plot of clients' negative reactions and their perceived conversation-level scores (the blue points) with the best-fit line (the orange line).

Behaviors In order to find out the behaviors that influence conversation effectiveness the most, we fit a lasso model with the proportion of the client's each behavior type as independent variables and the scores of conversation effectiveness as the dependent variable. In the end, we find that the most influential positive and negative behaviors are *Extending* and *Defending* (Detailed results of the importance of each behavior are in Appendix D.2), respectively. which is in line with the fact that counseling conversations are more likely to be effective when clients perceive themselves in a new way or experience changes in their behaviors, thoughts, or

feelings but to be less effective when clients defend their mistaken belief (Hill et al., 1992; Ribeiro et al., 2013).

To further understand the effect of negative behaviors on conversation effectiveness, the average score of the conversations with at least one negative behavior is calculated, which is 15.79, a drop of about 2% from the overall average score (Table 3). The results again indicate that clients' negative behaviors harm conversation effectiveness. Notably, *Defending* happens in most of the sessions that have negative behaviors. The overall low scores with defending behavior indicate that the conversation effectiveness is damaged when the clients start to defend and insist on their original beliefs. Although other negative behaviors such as *Changing Topics* have lower overall scores, they happen in fewer sessions and are less influential in our dataset. Once we have enough data for these categories, we expect their importance to become more apparent.

Categories	Avg. Score	# Sessions
Changing Topics	14.57	14
Sarcastic Answer	14.40	10
Focus Disconnection	13.25	4
Defending	15.46	175
Self-criticism or Hopelessness	14.04	24
Expressing Confusion	16.05	127
All Conversations	16.14	419
Conversations with Negative Behaviors	15.79	239

Table 3: The effect of the occurrence of each negative behavior on the conversation effectiveness.

5.2 Similar Counseling Strategies Leads to Similar Client Reactions

The clients react and behave differently towards counselors' different strategies. We find that counselors' strategies with the same intention lead to similar clients' behaviors. Specifically, strategies belonging to *Challenging* result in a larger proportion of clients' follow-up *Negative* behaviors than those belonging to *Supporting* (4.77% vs. 2.87%). The findings verify the rationality of categorizing the counselors' strategies into *Supporting* and *Challenging*. The detailed analysis is shown in Appendix D.3.

We then explore the influence of the counselors' strategies of *Supporting* and *Challenging* on clients' *Extending* and *Defending* behaviors as these are the most important ones according to the above analysis. As shown in Figure 4, compared with the *Supporting*, the *Challenging* brings a higher proportion of the clients' *Extending* behaviors. Meanwhile, *Challenging* makes the clients *Defending* as well. Therefore, to improve the conversation effectiveness, the appropriate utilization of the counselors' *Challenging* strategies is important, and we will analyze it in the following section.

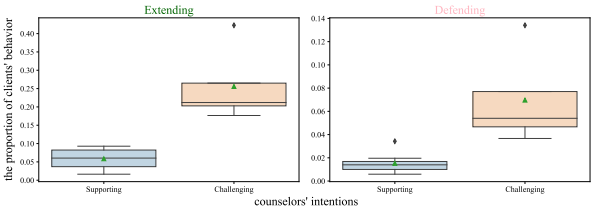

Figure 4: The distribution of clients' *Extending* and *Defending* behaviors after the counselors' *Supporting* and *Challenging* strategies.

5.3 Appropriate Strategy Utilization

To explore how counselors utilize *Challenging* appropriately to make clients behave as *Extending*, rather than *Defending*. We focus on two factors that influence the effectiveness of strategies: conversation stages and interaction patterns in the conversation history between counselors and clients(Althoff et al., 2016).

Conversation Stages. Each conversation is divided uniformly into five stages, and the distribution of clients' certain behaviors after counselors' *Challenging* at each stage is computed. Due to the high proportion of content in the first and last stages (18.70% and 33.86%) being irrelevant to counseling topics (labeled as *Others*), only the content in the middle three stages are analyzed. As shown in Figure 5, the counselors utilize more and more *challenging* as the conversations progress. Meanwhile, both *Extending* and *Defending* increase when clients face counselors' *Challenging*. Since *Extending* is overall more common than *Defending*, this phenomenon suggests that counselors adopt *Challenging* step by step within a counseling session. We will leave the cross-section analysis in future work.

Counselor-Client Preceding Interaction Patterns. The counselor-client preceding interaction is defined as the pair of the counselors' *Supporting* or *Challenging* and the clients' following-up *Positive* or *Negative* reactions. We fit a logistic regression classifier to study how these preceding interaction patterns affect the *Extending* and *Defending* behaviors when facing a *Challenging* strategy. The overall classification accuracy is around

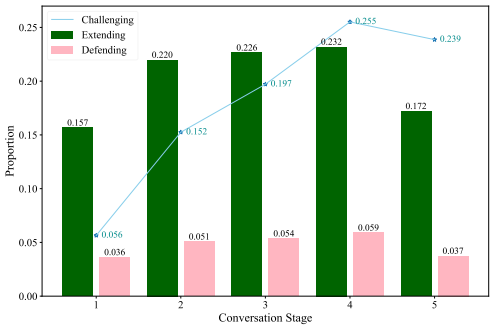

Figure 5: The distribution of clients' *Extending* and *Defending* reactions after counselors' *Challenging* strategy at different stages. The sky-blue line presents the proportion of *Challenging* strategy utilized in each stage.

80%, but we care more about the fitted coefficients, shown in Table 4. As can be seen, if the clients reacted positively to the counselors' *Challenging* before, the probability of the clients' *Extending* reactions increase when the counselors intervene with *Challenging* again, and vice versa. In other words, if counselors detect negative reactions from their clients, especially because of their supporting strategy, they should address those issues before launching into challenging strategies. In the event that they challenge their clients and receive positive reactions, they can continue to use the same strategy.

Interaction Patterns	Coefficients
Supporting - Positive	1.3041***
Supporting - Negative	-9.0643***
Challenging - Positive	3.7189***
Challenging - Negative	-7.3665***

Table 4: Associations between the counselor-client interaction patterns in preceding conversations and clients' current behaviors in response to counselors' *Challenging* interventions. ***p <0.001. The coefficients are from a logistic regression predicting the probability that the clients behave as *Extending* rather than *Defending*.

5.4 Baseline Classifiers for Automatic Label Prediction

To facilitate counselors guessing their clients' states, we train classifiers to categorize counselors' intentions and strategies and identify clients' reactions and behaviors based on a pre-trained Chinese RoBERTa-large model (Cui et al., 2020). Each task assigns a label to each sentence in a long utterance, utilizing conversation history as the context. To improve the domain adaption of pre-trained models (Gururangan et al., 2020; Sharma et al., 2020), we perform the masked language modeling (MLM) task on all the collected conversations and then jointly train each classification task on the annotated data with the MLM as an auxiliary task. More experimental details are shown in Appendix C.1.

As shown in Table 5, the test set of four tasks. The model's performance in categorizing counselors' intentions and strategies is better than identifying clients' reactions and behaviors. The overall performance on identifying clients' reactions is limited by *Negative* reactions (F1-value = 34.78%). The results indicate that clients' reactions are difficult to identify, especially the negative behaviors (Lee et al., 2019; Cao et al., 2019).

The major error in predicting clients' behaviors comes from the confusing *Reformulating* with *Extending*. In both cases, the client is making changes, but the former changes more deeply. Besides, *Defending* is hard to identify due to clients' diverse expressions of resistance. Clients may defend themselves by expressing different opinions from counselors rather than directly denying them, which is difficult for the model to recognize. More detailed classification results are in Appendix C.2.

Task	Acc.	Precision	Recall	Macro-F1
Intentions	$0.9025_{0.0030}$	$0.8821_{0.0046}$	$0.8446_{0.0045}$	$0.8612_{0.0040}$
Strategies	$0.8103_{0.0035}$	$0.7317_{0.0236}$	$0.6533_{0.0082}$	$0.6791_{0.0074}$
Reactions	$0.9490_{0.0016}$	$0.7762_{0.0163}$	$0.6977_{0.0167}$	$0.7214_{0.0138}$
Behaviors	$0.8597_{0.0018}$	$0.5815_{0.0273}$	$0.5190_{0.0140}$	$0.5354_{0.0155}$

Table 5: The overall results of the test set of four tasks: categorizing counselors' intentions and strategies, and clients' reactions and behaviors (Due to the scarce number of *Changing Topics*, *Sarcastic Answer* and *Focus Disconnection*, we filter out these samples when building classifiers). We report averages across five random seeds, with standard deviations as subscripts.

6 Conclusion

We develop a theoretical-grounded annotation framework to understand counselors' strategies and clients' behaviors in counseling conversations. Based on a large-scale and high-quality text-based counseling dataset we collected over the past two years, we validate the plausibility of our framework. With the labeled data, we also find that clients' positive reactions boost their ratings of counseling effectiveness, but negative reactions undermine them. Meanwhile, clients are more likely to *extend* after

counselor *challenge* their beliefs. Moreover, our automatic annotation models indicate that clients' reactions and behaviors are more difficult to predict than counselors' intentions and strategies. Due to the complexity of the data and the lack of labeled data for rare cases, our analysis is relatively shallow. We analyze the weakness of our work in section 7 and will dig deeper into each interaction pattern once we have more data.

7 Limitations

As this is the first large-scale analysis of client reactions in online mental health counseling, there is huge room for future improvement. Here we only list a few problems that we would like to address in the short future. First, although our annotation framework is comprehensive, the data labeled is quite imbalanced. In some rare classes, there are fewer than 50 instances, making it difficult to conduct an in-depth analysis, let alone train an accurate classifier. Therefore, our analysis mostly focuses on the *Extending* and *Defending* behaviors. We will label more data so that rare cases can be better understood and classified more accurately. The accuracy of a classifier is important for real-life applications because it has the potential to mislead counselors. Second, we only have one short post-survey, which limits our coarse-scale analysis. We are adding more and richer post-surveys. Third, while we hope that the lessons learned can be applied to everyday conversations, our analysis has only been limited to psycho-counseling. The lessons learned will be tested against a wider range of use cases. It is important, however, not to overgeneralize our findings as this may harm the naturalness of our daily conversations. After all, the psycho-counseling process is a very special type of conversation.

Acknowledgements

We are grateful to all counselors and clients for agreeing to use their counseling conversations for scientific research, and all annotators for their hard work. We appreciate the engineers who operate and maintain the counseling and annotation platform. Besides, we would like to express our gratitude to Professor Zhou Yu and other teachers, colleagues and anonymous reviewers who provided insightful feedback and suggestions for this project.

Ethics Statement

The study is granted ethics approval from the Institutional Ethics Committee (20211013LZZ001). All the clients and counselors signed a consent form when using our counseling platform, which informed them that the counseling conversations collected on the platform would be used for scientific research purposes, and might be used for scientific research by third parties. During the annotation process, we spared no efforts to manually de-identify and anonymize the data to protect clients' and counselors' privacy. The annotators also signed data confidentiality agreements and acquired ethical guidelines before they got access to the conversation data. Meanwhile, they were paid a reasonable wage for annotation. For the rules of releasing data, the third-party researchers who require access to the raw conversation data must provide us their valid ID, proof of work, the reason they request data (e.g., the research questions), etc. They are required to be affiliated with an non-profit academic or research institution. This includes obtaining the approval of an Institutional Review Board (IRB), having principal investigators working full-time as well as the written approval of institution's office of Research or equivalent office. Additionally, they must sign the Data Nondisclosure Agreement and make promise that they would not share the data with anyone.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Nafiz Al Asad, Md. Appel Mahmud Pranto, Sadia Afreen, and Md. Maynul Islam. 2019. Depression detection by analyzing social media posts of user. In *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 13–17.

Evrinomy Avdi and Eugenie Georgaca. 2007. Discourse analysis and psychotherapy: A critical review. *European Journal of Psychotherapy and Counselling*, 9(2):157–176.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Patricia Chamberlain, Gerald Patterson, John Reid, Kathryn Kavanagh, and Marion Forgatch. 1984. Observation of client resistance. *Behavior Therapy*, 15(2):144–155.

MJ Constantino, LG Castonguay, and AJ Schut. 2002. The working alliance: A flagship for the "scientist-practitioner" model in psychotherapy. *Counseling based on process research: Applying what we know*, pages 81–131.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *CoRR*, abs/1306.6078.

Robyn Dawes. 2009. *House of cards*. Simon and Schuster.

Fredrik Falkenström, Fredrik Granström, and Rolf Holmqvist. 2014. Working alliance predicts psychotherapy outcome even while controlling for prior symptom improvement. *Psychotherapy Research*, 24(2):146–159.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111:21.

Simon B Goldberg, Nikolaos Flemotomos, Victor R Martinez, Michael J Tanana, Patty B Kuo, Brian T Pace, Jennifer L Villatte, Panayiotis G Georgiou, Jake Van Epps, Zac E Imel, et al. 2020. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of counseling psychology*, 67(4):438.

Simon B Goldberg, Tony Rousmaniere, Scott D Miller, Jason Whipple, Stevan Lars Nielsen, William T Hoyt, and Bruce E Wampold. 2016. Do psychotherapists improve with time and experience? a longitudinal analysis of outcomes in a clinical setting. *Journal of counseling psychology*, 63(1):1.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Keith Harrigian, Carlos Alejandro Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. *ArXiv*, abs/2011.05233.

Robert Hatcher. 1999. Therapists' views of treatment alliance and collaboration in therapy. *Psychotherapy Research*, 9(4):405–423.

Hill, Clara, E., Corbett, Maureen, and M. 1992. Client behavior in counseling and therapy sessions: Development of a pantheoretical measure. *Journal of Counseling Psychology*.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

Clara E Hill, Ellen Baumann, Naama Shafran, Shudarshana Gupta, Ashley Morrison, Andrés E Pérez Rojas, Patricia T Spangler, Shauna Griffin, Laura Pappa, and Charles J Gelso. 2015. Is training effective? a study of counseling psychology doctoral trainees in a psychodynamic/interpersonal training clinic. *Journal of Counseling Psychology*, 62(2):184.

Clara E Hill and Emilie Y Nakayama. 2000. Client-centered therapy: Where has it been and where is it going? a comment on hathaway (1948). *Journal of Clinical Psychology*, 56(7):861–875.

Clara E Hill, Barbara J Thompson, and Elizabeth Nutt Williams. 1997. A guide to conducting consensual qualitative research. *The counseling psychologist*, 25(4):517–572.

Adam O Horvath and Leslie S Greenberg. 1994. *The working alliance: Theory, research, and practice*, volume 173. John Wiley & Sons.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 588–602. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Mikael Leiman and William B. Stiles. 2001. Dialogical sequence analysis and the zone of proximal development as conceptual enhancements to the assimilation model: The case of jan revisited. *Psychotherapy Research*, 11(3):311–330.

Anqi Li, Jingsong Ma, Lizhi Ma, Pengfei Fang, Hongliang He, and Zhenzhong Lan. 2022. Towards automated real-time evaluation in text-based counseling.

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.

Niklas Luhmann. 1981. The improbability of communication. *International Social Science Journal*, 33(1):122–132.

Brian L Mishara, Fran̦ois Chagnon, Marc Daigle, Bogdan Balan, Sylvaine Raymond, Isabelle Marcoux, Cécile Bardon, Julie K Campbell, and Alan Berman. 2007. Which helper behaviors and intervention styles are related to better short-term outcomes in telephone crisis intervention? results from a silent monitoring study of calls to the us 1–800-suicide network. *Suicide and Life-Threatening Behavior*, 37(3):308–321.

John C Norcross. 2010. The therapeutic relationship. *The heart and soul of change: Delivering what works in therapy*, pages 113–141.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019a. Conversation model fine-tuning for classifying client utterances in counseling dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1448–1459, Minneapolis, Minnesota. Association for Computational Linguistics.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019b. Conversation model fine-tuning for classifying client utterances in counseling dialogues.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *31st Conference on Neural Information Processing Systems (NIPS)*.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5307–5326. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

Eugénia Ribeiro, António P Ribeiro, Miguel M Gonçalves, Adam O Horvath, and William B Stiles. 2013. How collaboration in therapy becomes therapeutic: The therapeutic collaboration coding system. *Psychology and Psychotherapy: Theory, Research and Practice*, 86(3):294–314.

Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.

James Shanteau. 1992. Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2):252–266.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 194–205. ACM / IW3C2.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

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. *arXiv preprint arXiv:2106.01702*.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79, Denver, Colorado. Association for Computational Linguistics.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.

Terence J Tracey and Anna M Kokotovic. 1989. Factor structure of the working alliance inventory. *Psychological Assessment: A journal of consulting and clinical psychology*, 1(3):207.

Terence JG Tracey, Bruce E Wampold, James W Lichtenberg, and Rodney K Goodyear. 2014. Expertise in psychotherapy: An elusive goal? *American Psychologist*, 69(3):218.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2021. Towards low-resource real-time assessment of empathy in counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online. Association for Computational Linguistics.

Zhentao Xu, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Inferring social media users' mental health status from multimodal information. In *International Conference on Language Resources and Evaluation*.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, Online. Association for Computational Linguistics.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. *arXiv preprint arXiv:1906.07194*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

A Annotation Framework

A.1 Framework Development Process

We have three main taxonomy developers (two are experienced with clinical and emotional support, and one is the first author) to develop the framework, following the consensual qualitative research method (Hill et al., 1997; Ribeiro et al., 2013; Park et al., 2019b). Here, we describe the detailed developing process for the counselor's taxonomy as an example.

Firstly, based on existing taxonomies (Ribeiro et al., 2013; Hill, 2009), we filter those categories that are not appropriate for the text-only conversation settings (e.g., silence, head nods) and create the first version of taxonomy and annotation guideline. Secondly, we randomly select 6-10 conversations and ask all the taxonomy developers to annotate them independently. After the annotation, the developers discuss the differences and confusions among their annotations until reaching a consensus. During this process, they may add, merge or delete certain categories and refine the annotation guideline. We repeat the above step two for five times to obtain the final version of the taxonomy and guideline, including detailed definitions and examples. The Fleiss kappa values (Fleiss, 1971) among the three taxonomy developers in the five iterations are as follows: 0.6255, 0.6721, 0.6819, 0.7085, and 0.7233. The monotonically increasing agreement proves that the iterative process effectively resolves differences among developers. And the substantial agreement ensures the reliability of our taxonomy. During the whole process, we annotate 30 conversations.

A.2 Comparison with Existing Frameworks

We compare our proposed framework with existing annotation frameworks for analyzing dialogue acts of participants in the counseling conversations (see Table 6). Much research has mostly focused on studying counselors' strategies, such as CCU (Lee et al., 2019) and ESC (Liu et al., 2021). Specifically, the ESC framework proposes 7 counselors' support strategies based on three counseling stages. Different from ESC, our framework contains a more comprehensive and finer-grained classification (12 strategies) of counselors' skills based on their intentions. There are methods that attempt to classify clients' responses (Park et al., 2019b; Tanana et al., 2015; Pérez-Rosas et al., 2017). Park et al. (2019b) build a novel Categorization scheme of Client Utterances (CCU) with 5 categories. Such a scheme does not contain clients' immediate feedback on counselors' interventions, especially the negative one, limiting its role in helping counselors adjust their strategies and evaluating counseling effectiveness. In (Tanana et al., 2015; Pérez-Rosas et al., 2017), researchers conduct categorization

on both counselor and client sides based on MISC framework, but they are only limited to a particular mental health support genre called motivational interviewing. Our annotation framework is not genre specific and has more fine-grained analysis.

Framework	Categorization		Not Genre-Specific
	Counselor	Client	
CCU (Park et al., 2019b)		✓	✓
TCA (Lee et al., 2019)	✓		✓
ESC (Liu et al., 2021)	✓		✓
MISC (Tanana et al., 2015) (Pérez-Rosas et al., 2017)	✓	✓	
Our Framework	✓	✓	✓

Table 6: A comparison of our proposed framework with other existing annotation frameworks.

A.3 Definitions of Strategies

Restatement. The counselor reflects the content and meaning expressed in the client's statements in order to obtain explicit or implicit feedback from the client.

Reflection of Feelings. The counselor uses tentative or affirmative sentence patterns to explicitly reflect the client's mood, feelings, or emotional states.

Self-disclosure. The counselor discloses personal information to the client, including but not limited to the counselor's own similar experiences, feelings, behaviors, thoughts, etc.

Inquiring Subjective Information. The counselor explores the client's subjective experience, including thoughts, feelings, states, the purpose of doing something, etc.

Inquiring Objective Information. The counselor asks the client to concretize the imprecise factual information, including details of events, basic information about the client, etc.

Affirmation and Reassurance. The counselor affirms the client's strengths, motivations, and abilities, and normalizes the client's emotions and motivations, and provides comfort, encouragement, and reinforcement.

Minimal Encouragement. The counselor offers minimal encouragement to the client in an affirmative or questioning manner, encouraging the counselor to continue talking and facilitating the conversation.

Answer. The counselor answers the questions that the client asks about the conversation topics.

Interpretation. The counselor gives a new meaning, reason, and explanation to the behaviors, thoughts, or emotions of the client from a perspective beyond the client's statements or cognition, and tries to make the client look at problems from a new perspective.

Confrontation. The counselor points out the client's maladaptive beliefs and ideas, inconsistencies in the statements, or contradictions that the client is unaware of or unwilling to change.

Invite to Take New Perspectives. The counselor invites the client to use an alternative perspective to understand the given experience.

Invite to Explore New Actions. The counselor asks questions to guide the client to think and explore how to take new actions or invites the client to act in different ways during or after the conversation.

A.4 Definitions of Behaviors

Confirming. The client understands or agrees with what the counselor has said.

Giving Information. The client provides information according to the specific request of the counselor.

Reasonable Request. The client attempts to obtain clarification, understanding, information, or advice and opinions from the counselor.

Extending. The client not only agrees to the counselor's intervention, but also provides a more in-depth description of the topic being discussed, including the client's analysis, discussion, or reflection on his or her original cognition, thoughts, or behaviors.

Reformulating. The client responds to and introspects the counselor's intervention while proposing his or her own perspectives, directions of thinking, or new behavioral patterns on current issues.

Expressing Confusion. The client expresses confusion or incomprehension of the counselor's intervention or directly states that he or she has no way to answer or respond to the questions or interventions raised by the counselor.

Defending. The client is stubborn about an experience, glorifies or makes unreasonable justifications for his or her own views, thoughts, feelings, or behaviors, and insists on seeing the experience from the original perspective.

Self-criticism or Hopelessness. The client falls

into self-criticism or self-reproach, is engulfed in a state of desperation and expresses his or her inability to make changes.

Shifting Topics. Faced with the intervention of the counselor, the client's reply does not postpone the previous issue, but shifts to other issues.

Focus Disconnection. The client disengages from what the counselor is discussing, focuses on stating issues of interest, and does not respond to the counselor's intervention.

Sarcastic Answer. The client expresses dissatisfaction with the counselor, and questions or ridicules the counselor's intervention.

B Annotation Process

B.1 Post-survey Scales

To facilitate readers understand clients' self-report results of counseling conversations in our data, we present the questions of the assessment in Table 7. For each question, clients are required to choose only one from the following five options: seldom, sometimes, often, very often, and always, representing 1 to 5 points, respectively.

No.	Questions
1	As a result of this session, I am clearer as to how I might be able to change.
2	What I am doing in the counseling gives me new ways of looking at my problem.
3	I feel that the things I do in the counseling will help me to accomplish the changes that I want.
4	I believe the way we are working with my problem is correct.

Table 7: Questions of assessment after the counseling

B.2 Annotators Selection and Training

Annotators Selection and Training. We select 30 candidates out of more than 100 applicants who are at least undergraduate in psychology or with practical experience in counseling to attend an offline interview. During the interview, all the candidates are asked to learn the annotation guideline and then take three exams. Each exam consists of 50~60 conversation snippets. For each snippet, candidates are required to annotate the last utterance. After each exam, we provide the candidates the annotations to which they assigned incorrect labels in the exam and the corresponding correct labels to help them better understand the guideline. After the interview, the top 12 candidates with the highest average accuracy on the three exams become the final annotators. The highest and lowest accuracies are 72.07% and 64.01%, respectively (refer to Table 8 for more details). We then conduct two-day offline training for these qualified annotators. During training, all the annotators first annotate three conversations simultaneously (305 utterances), which have a ground truth labeled by our psychological experts. Then, the annotators analyze the utterances mislabeled in group meetings.

Annotator ID	Exam1	Exam2	Exam3	Avg.
1	0.6914	0.7582	0.7126	$0.7208_{0.0279}$
2	0.6420	0.7692	0.7356	$0.7156_{0.0538}$
3	0.6790	0.7692	0.6552	$0.7011_{0.0491}$
4	0.5679	0.7802	0.7356	$0.6946_{0.0914}$
5	0.6296	0.7033	0.7356	$0.6895_{0.0444}$
6	0.6173	0.7253	0.7241	$0.6889_{0.0506}$
7	0.6296	0.6923	0.7356	$0.6859_{0.0435}$
8	0.6790	0.7143	0.6552	$0.6828_{0.0243}$
9	0.7161	0.6593	0.6667	$0.6807_{0.0252}$
10	0.5679	0.7142	0.7356	$0.6726_{0.0745}$
11	0.5679	0.7143	0.6782	$0.6535_{0.0623}$
12	0.6296	0.6813	0.6092	$0.6401_{0.0304}$

Table 8: The results of each and average accuracy of the selected top twelve annotators in the three exams, with standard deviations as subscripts in the last column.

Training in the Loop. To further improve the inter-annotator agreement and annotation accuracy, we design the annotation process into six annotation and training stages. In the annotation stage, annotators are asked to record the utterances difficult to label (confusion samples). In the training stage, the psychological experts train each annotator after reviewing the confusing samples (618 samples) in a questions-and-answers document. As shown in Figure 6, the average agreement of the latter stages is higher than the former stages, indicating that the training-in-the-loop policy is effective.

B.3 Data Quality Control

We randomly assign each conversation to three or more annotators and ask them to annotate based on counselors' fine-grained conversational skills and clients' behavior types at the sentence level. Once obtaining the annotated data, we calculate the Fleiss' kappa (Fleiss, 1971) among multiple annotators in each conversation, which measures the proportion of agreement over and above the agreement expected by chance instead of measuring the overall proportion of agreement. For Fleiss' kappa, 0.61∼0.80 is indicated as substantial agreement. Considering the task demand that we have 12 annotators who annotated 13 and 12 categories of

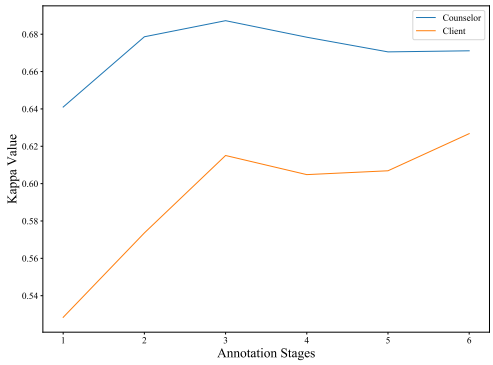

Figure 6: Agreement among annotators on labeling counselors and clients' utterances in each annotation stage.

counselors' strategies and clients' behaviors and reality of time, we take the substantial level of agreement. Finally, the average inter-rater agreement on labeling counselors' and clients' utterances is 0.67 and 0.59 respectively, validating the reliability of the data. And we find that human annotators struggle with some specific categories, such as *Interpretation* versus *Invite to Take New Actions* in counselors' strategies, *Extending* versus *Reformulation* in clients' behaviors, etc. We then utilize a majority vote method to obtain the final labels. For those samples that haven't been labeled by the above method process, we randomly assign them to three or more annotators until we get a majority vote. Overall, we find that compared to annotating counselors' conversational skills, identifying clients' reactions and behaviors is more difficult because they do not act within theoretical frameworks (Lee et al., 2019).

C Automatic Prediction

C.1 Experimental Details

Data Preparation Tasks for both speakers share the same data preparation process. We randomly split the annotated data into a training set (70%), validation set (15%), and test set (15%). Note in the split, all utterances in a conversation are assigned to the same set.

Experimental Settings All the models are implemented with PyTorch deep learning package (Paszke et al., 2017). To make the pretrained model aware of the speaker's information in conversation, we adopt a simple, special tokens strategy by prefixing a special token [SP] or [SK] to each utterance from counselors or clients, respectively. The masking probability in the MLM task is set to 0.15 in the both domain post-training and fine-tuning process. In the fine-turning stage, we initialize weights of feed-forward layers with normal distribution. We set the training epoch as ten and select the model that achieves the best macro-F1 value on the validation set to test on the test set. For both training processes, we adopt cross-entropy loss as the default classification loss. And we use Adam optimizer to train the network with momentum values $[\beta_1, \beta_2] = [0.9, 0.999]$. The learning rate is initialized to $5e-5$ and decayed by using the linear scheduler. The batch size in the training stage is 8. The domain post-training experiment is performed on four NVIDIA A100 GPU, and all the fine-tuning experiments are performed on one NVIDIA A100 GPU. Each fine-tuning experiment costs about 80 minutes.

C.2 Experimental Results

Table 9 shows detailed experimental results about precision, recall and macro-f1 for each category in predicting counselors' intentions and strategies, and clients' reactions and behaviors.

Task	Categories	Precision	Recall	Macro-f1
Intentions	Supporting	0.9194	0.8146	0.9208
	Challenging	0.9578	0.6902	0.8940
	Others	0.9382	0.7473	0.9072
Strategies	Restatement	0.719	0.8891	0.795
	Reflection of Feelings	0.7955	0.5882	0.6763
	Self-disclosure	0.5714	0.5714	0.5714
	Inquiring Subj. Info.	0.8447	0.8671	0.8558
	Inquiring Obj. Info.	0.8248	0.75	0.7856
	Affirmation & Reassurance	0.8055	0.76	0.7821
	Minimal Encouragement	0.9518	0.9478	0.9498
	Answer	0.6522	0.4286	0.5172
	Interpretation	0.664	0.5773	0.6176
	Confrontation	0.6667	0.2222	0.3333
	Invite to Explore New Actions	0.7717	0.7899	0.7807
	Invite to Take New Perspectives	0.3824	0.2766	0.321
	Others	0.9342	0.9148	0.9244
Reactions	Positive	0.9642	0.9757	0.9699
	Negative	0.459	0.28	0.3478
	Others	0.9002	0.9043	0.9023
Behaviors	Giving Information	0.8952	0.9263	0.9105
	Confirming	0.8881	0.9237	0.9055
	Reasonable Request	0.8468	0.8268	0.8367
	Extending	0.4384	0.3546	0.3921
	Reformulating	0.1	0.0345	0.0513
	Expressing Confusion	0.4545	0.4545	0.4545
	Defending	0.2963	0.1569	0.2051
	Self-criticism or Hopelessness	0.75	0.2308	0.3529
	Others	0.9036	0.918	0.9107

Table 9: The RoBERTa classification result for each category in four tasks, including predicting counselors' intentions, strategies and clients' reactions and behaviors.

D Application to Counseling

D.1 Correlation Between Clients' Reactions and Conversation Outcomes

We group all conversations according to the proportion of the clients' *Negative* reactions contained in the conversations, ensuring that the number of conversations in each group is almost the same (except for the first group). We then calculate the mean and standard deviation of the clients' self-reported conversation-level scores in each group. The results are shown in Table 10.

Group	Ratio Span	# Session	Score
1	$0.000 \sim 0.012$	181	$16.62_{3.62}$
2	$0.012 \sim 0.024$	37	$16.76_{3.60}$
3	$0.024 \sim 0.036$	47	$16.81_{3.72}$
4	$0.036 \sim 0.048$	39	$16.33_{3.65}$
5	$0.048 \sim 0.060$	35	$15.74_{3.19}$
6	$0.060 \sim 0.096$	42	$14.74_{4.48}$
7	$0.096 \sim 0.240$	38	$14.16_{5.07}$

Table 10: Grouped conversations according to the ratio of clients' *Negative* reactions included. The last column shows the average scores of conversations in each group, with standard deviations as subscripts.

D.2 Which behavior influences conversation effectiveness the most?

# Variables	Confirming	Giving Information	Reasonable Request	Extending	Reformulating	Expressing Confusion	Defending	Self-criticism or Hopelessness	Shifting Topics	Focus Disconnection	Sarcastic Answer
9	✓	✓	✓	✓	✓	✓	✓	✓			✓
7	✓	✓	✓	✓	✓		✓	✓			
6	✓	✓	✓	✓	✓		✓				
5	✓	✓		✓	✓		✓				
3	✓	✓		✓							
2		✓		✓							
1				✓							

Table 11: Behavior types selected as important independent variables that affect clients' self-reported evaluation of conversation effectiveness by Lasso model, as the coefficient of L1 regularization uniformly increases from 0.001 to 0.1.

D.3 Clients Reactions and Behaviors towards Counselors' Strategies

Figure 7 shows clients' follow-up behavior distribution after the counselor's every strategy in the overall conversations, where the behavior distribution refers to the proportion of the clients' each immediate behavior type. We find that compared with using strategies with *Supporting* intention, counselors' utilization of *Challenging* strategies is more likely to lead to clients' *Negative* behaviors.

We then measure the similarity of the impact of counselors' each strategy on clients' behaviors by calculating the Euclidean distance between clients' follow-up behavior distribution after different strategies (see Table 12).

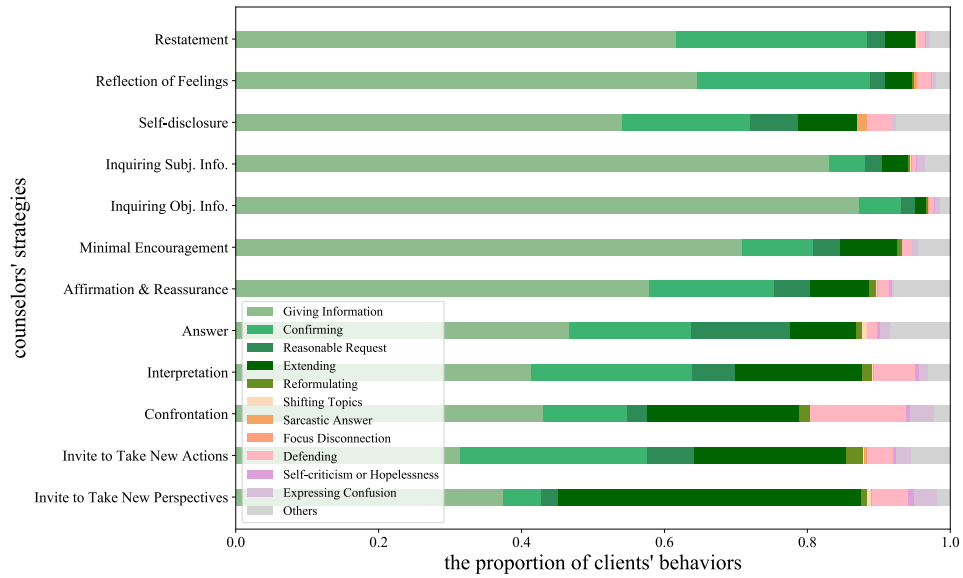

Figure 7: The clients' behaviors distribution after the counselors' each strategy in the overall conversations.

	Restatement	Reflection of Feelings	Self-disclosure	Inquiring Subj. Info.	Inquiring Obj. Info.	Minimal Encouragement	Affirmation & Reassurance	Answer	*Supporting*	Interpretation	Confrontation	Invite to Explore New Actions	Invite to Take New Perspectives	*Challenging*
Restatement	0.0000	0.0452	0.1690	0.3069	0.3337	0.1994	0.1359	0.2381	**0.204**	0.2691	0.3854	0.3626	0.5131	0.3825
Reflection of Feelings	0.0452	0.0000	0.1724	0.2665	0.2928	0.1648	0.1354	0.2518	**0.1898**	0.2887	0.3854	0.3896	0.5184	0.3955
Self-disclosure	0.1690	0.1724	0.0000	0.3369	0.3833	0.2043	0.0630	0.1076	**0.2052**	0.1777	0.2585	0.2767	0.4097	0.2807
Inquiring Subjective Information	0.3069	0.2665	0.3369	0.0000	0.0582	0.1398	0.2932	0.4122	**0.2591**	0.4847	0.5057	0.5930	0.6073	0.5477
Inquiring Objective Information	0.3337	0.2928	0.3833	0.0582	0.0000	0.1877	0.3388	0.4593	**0.2934**	0.5285	0.5532	0.6380	0.6551	0.5937
Minimal Encouragement	0.1994	0.1648	0.2043	0.1398	0.1877	0.0000	0.1585	0.2800	**0.1907**	0.3472	0.3882	0.4540	0.4926	0.4205
Affirmation and Reassurance	0.1359	0.1354	0.0630	0.2932	0.3388	0.1585	0.0000	0.1449	**0.1814**	0.2148	0.3020	0.3130	0.4289	0.3147
Answer	0.2381	0.2518	0.1076	0.4122	0.4593	0.2800	0.1449	0.0000	0.2706	0.1591	0.2688	0.2303	0.3917	**0.2625**
Interpretation	0.2691	0.2887	0.1777	0.4847	0.5285	0.3472	0.2148	0.1591	0.3087	0.0000	0.1921	0.1175	0.3071	**0.2056**
Confrontation	0.3854	0.3854	0.2585	0.5057	0.5532	0.3882	0.3020	0.2688	0.3809	0.1921	0.0000	0.2512	0.2661	**0.2365**
Invite to Explore New Actions	0.3626	0.3896	0.2767	0.5930	0.6380	0.4540	0.3130	0.2303	0.4071	0.1175	0.2512	0.0000	0.3107	**0.2265**
Invite to Take New Perspectives	0.5131	0.5184	0.4097	0.6073	0.6551	0.4926	0.4289	0.3917	0.5021	0.3071	0.2661	0.3107	0.0000	**0.2946**

Table 12: The Euclidean distance between each strategy's follow-up behavior distribution. The columns of *Supporting* and *Challenging* show the average distance of the follow-up behavior distribution between each strategy and all the other strategies belonging to the corresponding intention, where the lower average distance values are bolded.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 4, section 5.4

☑ B1. Did you cite the creators of artifacts you used?
section 5.4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
section 5.4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
ethics statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
section 4, Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
section 4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 4.3, section 5.4

C ☑ Did you run computational experiments?

section 5.4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 5.4, appendix C.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 5.4, appendix C.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 5.4, appendix C.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 5.4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
section 4

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Annotation instructions provided to our annotators are very long and contain many examples of counseling conversations. Therefore, the full text of instructions is not suitable to be put in our paper. But we put the definitions of each category in our annotation framework in Appendix A.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
section 4, appendix B.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
section 4

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
ethics statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
section 4