Towards Automated Real-time Evaluation in Text-based Counseling

Anonymous ACL submission

Abstract

Automated real-time evaluation of counselor-client interaction is important for ensuring quality counseling but the rules are difficult to articulate. Recent advancements in machine learning methods show the possibility of learning such rules automatically. However, these methods often demand large scale and high quality counseling data, which are difficult to collect. To address this issue, we build an online counseling platform, which allows professional psychotherapists to provide free counseling services to those in need. In exchange, we collect the counseling transcripts. Within a year of its operation, we manage to get one of the largest set of (675) transcripts of counseling sessions. To further leverage the valuable data we have, we label our dataset using both coarse- and fine-grained labels and use a set of pretraining techniques. In the end, we are able to achieve practically useful accuracy in both labeling system.

1 Introduction

The last decade has witnessed an increasing demand for psychological counseling in daily life. Although positive effects of psychological counseling are widely recognized (Lambert et al., 1994; Perry et al., 1999), it remains great potentiality to improve the quality of counseling services. For instance, the counselors, especially the novice, may provide inappropriate interventions towards clients’ maladaptive experience, which would undermine the interaction quality between the counselor and client, leading to a negative outcome of the counseling (Stiles et al., 1994). Therefore, developing an evaluation system that can monitor the quality of each interaction between counselors and clients in an ongoing counseling session is necessary. That is, when the evaluation system detects negative effects in the current interaction, it can provide an immediate feedback to the counselors and suggest them to adjust the perspectives of topics or experiences being discussing in the next turn of counseling dialogue. Figure 1 presents the automated evaluation process, in which the evaluation system identifies the client’s status in a timely manner and provides immediate feedback for the counselor in an ongoing session.

Nonetheless, it is not without difficulties to build an evaluation system for psycho-counseling because of some practice issues. 1)-The corpus of psychological counseling is not easy to be accessed due to the confidential nature of psycho-counseling. 2)-It requires the expert-level annotators to label the data. 3)-The advanced machine learning techniques are data-hungry. Adapting machine learning models with few data is also a challenging step. In this paper, we aim to bridge the gap between the machine learning techniques and the evaluation system on psycho-counseling.

To collect the Chinese psycho-counseling corpus in the real-world scenarios, we build an online counseling platform to provide the service of free text-based counseling. From the platform, we collect 675 transcripts of counseling sessions, conducted by professional counselors employing skills.
of various counseling genres, including Cognitive Behavioral Therapy (CBT), Narrative Therapy, Psychodynamic Therapy, etc. Taking 3 consecutive counselor-client dialogue turns as a sample, we obtain 33792 samples in total.

To evaluate the quality of the counseling, we further apply the Clients’ Response-Experience Types Coding Scheme (CRETCS) to annotate our data. The CRETCS is based on the therapeutic collaboration coding system (TCCS) proposed by Ribeiro et al. (2012), and defines the categories of clients’ responses and experiences, reflecting the quality of counselor-client interaction from the perspective of clients. Under the principle of the scheme, we can distinguish whether clients accept counselors’ interventions and the level of risk that clients experience towards preceding interventions. Following the coding scheme as guidance, we manually annotate 4786 samples among all data.

To study the feasibility of leveraging advanced computational methods to automate the real-time evaluation process, we conduct extensive experiments on the proposed data. To improve the performance of model with limited annotated data on a new domain, we conduct additional pre-training phases to adapt our model to the task-specific data, including the unlabeled data and the annotated training data. The empirical results show that a second pre-training phase is an effective way to bring performance gain. In the detailed error analysis, we find that the model might be confused when categorize two labels that share similar linguistic features.

In this paper, we first construct a Chinese benchmark dataset, which was annotated by a CRETCS. To verify the potential that advanced machine learning-based methods can solve the real-time evaluation of interaction quality in ongoing counseling, extensive experiments were studied on various computational approaches. We believe our benchmark and the proposed evaluation system will bring improvements to the assessment of interaction quality in the field of psycho-counseling. The contributions are summarized as follows:

- We first propose a coding scheme for evaluating the quality of the real-time interaction between counselors and clients, and design detailed operational definitions of each category to facilitate our research.
- We build a free online text-based psychological counseling platform and collect 675 transcripts of counseling sessions from the platform. We also construct a dataset, which is annotated following the principle of the proposed coding scheme.
- Thorough experiments are conducted on our new benchmark dataset. Empirical results indicate that computational approaches hold promise to automate the process of real-time evaluation.

Our code and model will be made available freely to the research community after the paper is accepted. Because of the sensitive and private nature, we will also simulate part of this data and make it available by request.

2 Related Work

Our work mainly focuses on the evaluation of psycho-counseling. Here, we briefly give an overview of the assessment methods for the interaction quality in psychological counseling. Thereafter, we also review the computational approaches for the counseling evaluation.

2.1 Assessment of the Interaction Quality in Psycho-counseling

The quality of psychological counseling can be assessed by clients’ self-reports of improvements of maladaptive experiences after attending the counseling (Stiles et al., 1994). Meanwhile, the behavioral coding is another primary method employed to evaluate the quality of counseling sessions, in which trained coders assign specific labels or numeric values to each utterance or evaluate the whole sessions based on coding schemes (Bakeman and Quera, 2012). To evaluate the interaction quality in the counseling, the categorization of each counselors’ and clients’ utterance during counseling is necessary. The works constructed the coding schemes on the utterance level mainly defined and categorized counselors’ strategies, such as support, reflection, question and challenging etc. (Lee et al., 2019; Grandeit et al., 2020; Young and Beck, 1980), which reflect therapeutic approaches instead of the interaction quality between counselor and client. Some coding schemes also have taken into account the categories of clients’ utterances during counseling but are developed specifically for certain types of counseling in which the categories are closely related to the objectives of the corresponding counseling. For instance, Motivational Inter-
viewing Skill Codes (MISC) was designed to provide the utterance-level labels to code counselors’ and clients’ behaviors respectively for Motivational Interviewing (MI) (Miller et al., 2003). While some other studies have designed the coding schemes that can be applied to diverse counseling genres and reflect the interaction between counselor and client. For examples, the categorization of client utterances was developed by Park et al. (2019) to represent clients’ past experience, current situation and changes of problematic thoughts; and the therapeutic collaboration coding scheme (TCCS) was designed by Ribeiro et al. (2012) to analyze the effectiveness of moment-by-moment interaction between counselor and client according to clients’ zone of proximal development (Leiman and Stiles, 2001).

2.2 Computational Approaches for Behavioral Coding of Counseling

A diverse set of works have adopted advanced computational approaches, which automate the behavioral coding, to evaluate the quality of the text-based counseling (Can et al., 2012; Xiao et al., 2016; Atkins et al., 2014; Tanana et al., 2016; Gibson et al., 2019; Flemotomos et al., 2021a). Early works relied mostly on designing hand-crafted feature representations for the text (Can et al., 2012; Atkins et al., 2014; Tanana et al., 2015, 2016; Pérez-Rosas et al., 2017). One of the earliest work (Can et al., 2012) used different sets of hand-crafted linguistic features to automatically identify Reflection. In (Atkins et al., 2014), the authors adopted topic models to predict counselors’ intervention strategies under MISC scheme on the utterance-level. In order to make full use of text information, the work in (Pérez-Rosas et al., 2017) integrated the syntactic, semantic and similarity features to identify counselors’ strategies. In the era of deep learning, a number of solutions employ the advanced neural networks, i.e., convolutional neural network (CNN), recurrent neural network (RNN), Transformer, etc, to extract discriminative features from the text (Xiao et al., 2016; Huang et al., 2018; Cao et al., 2019; Lee et al., 2019; Tavabi et al., 2021; Flemotomos et al., 2021b). For example, Cao et al. (2019) used hierarchical recurrent encoders with attention mechanism to encode dialogues. Some recent works (Grandet et al., 2020; Tavabi et al., 2021) benefit from the pre-trained model, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) etc. To detect the categories of clients’ responses in counseling of diverse genres, Park et al. (2019) develop a pre-trained conversation model using a sequence-to-sequence architecture.

3 Data Collection

In this section, we will detail the proposed dataset and the annotation scheme, which can be served as a benchmark to evaluate the counseling quality from the utterances of clients, for various counseling genres. Our dataset is collected from the free Chinese psychological counseling platform, where the counseling services are provided by professional human counselors. In § 3.1, we first introduce the data selection principles and the statistics of the corpus. Then we will explain the coding scheme CRETCS, which categorizes clients’ utterances according to the types of clients’ responses and experience in § 3.2. Following the coding scheme, we can annotate our data and present a Client Response-Experience Types Dataset (CRETD). The labeling process is discussed in § 3.3. The § 3.4 summarizes the characteristics of our dataset.

3.1 Data Source

Our free online Chinese psychological counseling platform provides dyadic text-based counseling for clients. In our online platform, we hire 30 professional counselors, with each counselor experienced with one or more counseling genres, e.g., CBT, Narrative Therapy, Psychodynamic Therapy. Before the time we build our dataset, our platform has served 1405 clients of different genders, ages, levels of education and regions, and the counseling topics include interpersonal relationships, career, self-development, emotion regulation, etc.

Given the counseling transcripts collected from our platform, we choose 675 counseling sessions following two criteria: (1)-the transcripts include more than 20 counselor-client dialogue turns; (2)-the contents in the transcripts reflect an appropriate counseling process, e.g., discussion of maladapted behaviors, undesirable thoughts, difficulties in life, instead of the chit-chat, such as general inquiries about the counseling platform.

Once we obtain the counseling data, we then pre-process each counseling session by following steps: (1)-in each dialogue turn, we combine multiple consecutive messages from the same speaker, connected by a full stop, such that each spliced
message is an utterance from the speaker. (2)-We anonymize clients’ personal information, including name, organization, etc, for privacy purposes. (3)-In the end, we create our dataset, with each sample consisting of 3 consecutive dialogue turns, started by the counselor. After the pre-processing, we can obtain 33792 unlabeled samples. In Table 1, we show a sample snippet in both Chinese and its English counterpart. Statistically, the average dialogue turns per session are 50 in the counseling data, and in the pre-processed samples, the average length of counselors’ and clients’ utterance is 21 and 32 characters respectively. The label distribution of annotated data is shown in Table 2.

3.2 Annotation Scheme

As mentioned above, coding the counselors’ utterances reflects only the therapeutic approaches rather than the interaction quality between counselor and client during counseling (Zhang and Danescu-Niculescu-Mizil, 2020). As an example shown in Figure 1, the client’s refusal to counselor’s intervention, who tried to change the client’s problematic thoughts, indicating that the counselor’s utterances might not be a suitable intervention as to the client at that moment. However, this utterance of the counselor may be evaluated as an “appropriate” intervention based on the therapeutic techniques solely (Schegloff, 1987; Thomas, 1983; Zhang and Danescu-Niculescu-Mizil, 2020). Thus, it is necessary to consider the clients’ responses when assessing the interaction quality during ongoing counseling.

To address the issue, our idea is to evaluate the effectiveness of counselors’ intervention and the interaction quality by means of analyzing the types of response and experience expressed by the clients.

Referring to Vygosky’s zone of proximal development (Vygotsky et al., 1978), the clients attending the psycho-counseling develop their problem solving skills within the development level and experience different degrees of risk in response to counselors’ interventions (see Figure 2) (Leiman and Stiles, 2001). If counselors’ interventions inspire clients to explore and solve their problems effectively, the clients will follow the interventions within their developmental level (between the actual and potential level). They might feel safe to communicate with the counselor or feel challenging when discussing their maladaptive experience, indicating that the clients are making valid responses. But if the counselors’ interventions fail to capture the clients’ attention (below the actual level) or irritate the clients (above the potential level), the clients might lose interests in the ongoing conversation or start to defend themselves, reflecting that their responses become invalid as to counselors’ interventions. Based on these concepts, in the CRETCS, we categorize clients’ utterances into Validation, Invalidation as to the response types, and Safety, Tolerable Risk, Intolerable Risk and Disinterest as to the experience types, which is adapted from the categorization of clients’ responses in TCCS (Ribeiro et al., 2012). Additionally, we add the label Others to code utterances that counselors and clients are not talking about the clients’ problems, including greetings, scheduling the next appointment and expressing appreciations to counselors etc.

Considering the fact that annotators’ subjective perception on distinguishing "the level of risk" from clients’ experience may affect the accuracy of annotation, we designed a handbook, which contains the detailed and objective operational definitions of all the categories and the corresponding examples, to facilitate the annotation process. The handbook was finalized through repeated procedures of annotation, discussion of resolving differences and reaching consensus, analysis of reasons, and summary of criteria by two research psychologist, one of them holds the certificate of psycho-counseling qualified by Ministry of Human Resource and Social Security of China. In the revision process, the research psychologists annotated three sessions with 158 samples. 42 out of the samples are used as examples in the handbook and the remaining was used to train and test the labeling accuracy of annotators before formally annotation (detailed in § 3.3). Samples of the handbook will...
be shown in the Appendix A.1.

3.3 Data Annotation

Having the utterance samples and coding scheme at our disposal, we are ready to annotate our data. As shown in Table 1, each sample includes three dialogue turns and can be split into dialogue history (the blue text), counselor’s intervention utterance (the orange text), and client’s response utterance (the purple text). Considering the context from dialogue history and counselors’ intervention, we can annotate first the response type and then experience type w.r.t. client’s response utterance according to the coding scheme.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Utterance</th>
<th>Index</th>
<th>Response Type</th>
<th>Experience Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>counselor</td>
<td>听起来像是，本质上不用承担选择的后果和责任。</td>
<td>$t_1$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>client</td>
<td>但是这个被安排的结果，也还是我自己承担嘛，我是觉得差不多。 But I still have to bear the result of this arrangement. I don’t think the difference is too much.</td>
<td>$c_1$</td>
<td>Invalidation</td>
<td>Intolerable Risk</td>
</tr>
<tr>
<td>counselor</td>
<td>免除了自己选择别的可能性的后果 But it could eliminate the consequences of choosing other possibilities.</td>
<td>$t_2$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>client</td>
<td>确实。 Exactly.</td>
<td>$c_2$</td>
<td>Validation</td>
<td>Safety</td>
</tr>
<tr>
<td>counselor</td>
<td>对于没有被安排的其他部分，你怎么看呢？ What do you think about the other parts that are not arranged？</td>
<td>$t_3$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>client</td>
<td>那么，选择太多，结果也不确定。显而易见，我选择了逃避。Because there are too many choices, the result is not certain. Obviously, I chose to escape, which made me worse.</td>
<td>$c_3$</td>
<td>Validation</td>
<td>Tolerable Risk</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 1: A sample snippet of a simulated transcript with its English translated version. Each sample is composed of three consecutive counselor-client dialogue turns. Each client’s utterance is annotated with response and experience type towards the counselor’s utterance with taking dialogue history as context.

3.4 Data Analysis

Our dataset has two characteristics: Hierarchy of labels. Each utterance sample is annotated by two types of labels, i.e., response type and experience type. These two types of labels follow the hierarchical relationship of two scales – the response type is the coarse-level annotation while the experience type is the fine-level annotation. That said, the same response type may result from different experience types. For example, both the experience of Intolerable Risk and Disinterest leads to the Invalidation towards counselors’ interventions.

Imbalanced data distribution. To better understand the dataset, we illustrate the statistics of the proposed dataset in Table 2. As shown in Table 2, the number of Validation is about five times larger than that of Invalidation (i.e., 3553 vs. 559) in the coarse-level annotation. While in the fine-level annotation, the ratio of Safety to Tolerable Risk is about 1.6 : 1 and the number of Intolerable Risk and Disinterest are very similar. Such a data distribution is inline with the fact (Stiles et al., 1994; Ackerman and Hilsenroth, 2003) that counselors spent most of time to intervening gently to support the clients and make them feel safe to form a posi-
Table 2: Data distribution and mean length of each category under response types and experience types, respectively.

<table>
<thead>
<tr>
<th>Response Type</th>
<th>Num</th>
<th>Mean Length</th>
<th>Experience Type</th>
<th>Num</th>
<th>Mean Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td>Safety</td>
<td>2174</td>
<td>24.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tolerable Risk</td>
<td>1379</td>
<td>38.30</td>
</tr>
<tr>
<td>Validation</td>
<td>559</td>
<td>30.68</td>
<td>Intolerable Risk</td>
<td>326</td>
<td>27.86</td>
</tr>
<tr>
<td></td>
<td>233</td>
<td>34.62</td>
<td>Others</td>
<td>674</td>
<td>8.81</td>
</tr>
</tbody>
</table>

4.2 Network Architectures

Baseline We use the pre-trained Chinese BERT base model (Devlin et al., 2018), provided by HuggingFace, as our baseline for both the coarse- and fine-grained classification tasks. Following the common practice (Devlin et al., 2018), we add a feed-forward layer on top of the [CLS] token as a task-specific classifier, where [CLS] is the classification token in BERT.

4.3 Training Methods

Task-adaptive Pre-training. In order to attain a better model with limited annotated data on the domain of psychological counseling, we conduct a second phase of pre-training to adapt our model to the task-specific data, whose effects of gaining performance have been demonstrated in (Chronopoulou et al., 2019; Gururangan et al., 2020). We refer to this method as task-adaptive pre-training or TAPT for short. Given the fact that we have a small size of training data and a larger pool of unlabeled data, we set two kinds of task-adaptive pre-training experiments.

TAPT. In this experiment, we use an auxiliary masked language model (MLM) loss during training in the fine-tuning process. The final loss function is the weighted sum of the task-specific classification loss and the auxiliary MLM loss.

Curated-TAPT. In this pipeline, we first pre-train a BERT model on the unlabeled data with an MLM loss. During training, we select the best checkpoint with the minimal MLM loss value on the validation set. Then we perform fine-tuning on the small set of annotated data on the model of the best checkpoint. Noted in the fine-turning stage, the MLM loss is not used in this experiment.

Curated-TAPT+TAPT. In this setting, we develop a pipeline combining the process of ‘TAPT’ and ‘Curated-TAPT’. That said, the Curated-TAPT is used for pre-training of the model on the task-specific unlabeled data, followed by TAPT for fine-turning on the annotated training data.

4.4 Implementation Details

Model Input. To identify speaker information in counseling, we use [T] and [C] as special tokens to represent the utterance for counselor and client respectively, whose embeddings will also be trained in the pre-training and fine-tuning process. For each sample with the format of (dialogue


1https://github.com/huggingface
history, counselor’s intervention utterance, client’s response utterance) as shown in Table 1, we add a special token ([T] or [C]) at the beginning of each speaker’s utterance and concatenate all utterances into a flat sequence with a separation token ([SEP]) before the client’s response utterance. Taking the sample in Table 1 as an example, the data is presented as a sequence of \( (t_1, c_1, t_2, c_2, t_3, c_3) \), where \( t_i \) and \( c_i \) are the counselor and client’s utterance of the \( i \)-th dialogue turn. In the pre-training stage, the sample is fed in the form of “[CLS][T]t_1[C]c_1[T]t_2[C]c_2[T]t_3[SEP][C]c_3” with some non-special tokens are masked. While in the fine-tuning stage, the input sample is presented as: “[CLS][T]t_1[C]c_1[T]t_2[C]c_2[T]t_3[SEP][C]c_3”.

**Experimental Settings.** All the models are implemented with PyTorch deep learning package (Paszke et al., 2017). In pre-training process, the masking probability in the MLM task is set to 0.15. In the fine-tuning stage, we initialize weights of feed-forward layers with normal distribution. We select the model that achieves best macro-F1 value on the validation set to test on the test set. For all training process, 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. All our experiments are performed on one NVIDIA A100 GPU.

4.5 Experimental Results

Extensive experiments are conducted on the baseline, with each being evaluated with three different pre-training strategies (i.e., TAPT, Curated-TAPT, Curated-TAPT+TAPT). We follow the standard protocol to evaluate the models with ‘accuracy’, ‘precision’, ‘recall’, and ‘macro-F1’ metrics. The results for coarse- and fine-grained classification are illustrated in Table 3 and Table 4, respectively. As compared with the results of coarse-grained classification, we find the fine-grained classification is very challenging. More detailed results about precision and recall for each category are included in the appendix A.2.

We find that: along with the vanilla pre-trained BERT, additional pre-training on task-specific data, like TAPT and Curated-TAPT, can consistently bring performance gain for both-level classification tasks. Another interesting observation is that the TAPT always performs better than the Curated-TAPT. Moreover, applying TAPT after adapting to the extra task-specific unlabeled data (i.e., Curated-TAPT+TAPT) leads to better performance and this pipeline of pre-training brings more improvements on coarse-grained than fine-grained classification compared to each baseline. Specifically, taking the baseline model as a reference in coarse-grain classification, Curated-TAPT brings an improvement of 2.9% on macro-F1, while applying TAPT results in a big boost in performance by nearly 8.2%. As compared to TAPT, we also demonstrate that Curated-TAPT + TAPT can further improve the result by nearly 1% and attain the best performance.

4.6 Error Analysis

In this section, we analyze the performance of the best-performing model for coarse- and fine-grained classification respectively.

**Coarse-grained categorization.** The confusion matrix for the categorization of response and experience types is shown in Appendix A.3. The overall performance of the coarse-grained model is limited by Invalidation with 0.51 F1-score (refer to Appendix A.2 for detailed results). To better understand the confusion between Invalidation and Validation, we manually analyze 46 Invalidation samples that are predicted as Validation and 27 samples vice versa. In this process, we find that the most typical cases are that the model categorizes utterances containing negative words, especially “not” or “no”, into Invalidation because many samples in the Invalidation include these words. For instance, “Counselor: I think you hold a positive attitude towards the future. Client: Yes, you are right. In fact, the overall condition of me is not bad, but I still cannot maintain a close relationship”. Likewise, the majority of samples in Validation contains “yes” and “ok”, the model tends to categorize clients’ utterances into Validation when "yes" or "ok" is detected. But the utterances are manually categorized as Invalidation because the clients fails to respond to counselors’ intervention: “Counselor: You’ve already known that the people you met and what happened in your life does not help you to achieve your goal. Then, what’s your plan? Client: Yes. I’m so tried, coping with all these stuff. It may still end up with nothing.”

Thus, the performance of model might be affected by the shared semantic meaning of common words across the Validation and Invalidation.
### Table 3: The comparative results on the coarse-grained categories of all models. We report averages across ten random seeds, with standard deviations as subscripts. Best task performance is boldfaced.

<table>
<thead>
<tr>
<th>Model Architectures</th>
<th>Additional Pretraining Phases</th>
<th>Acc.</th>
<th>Precision</th>
<th>Recall</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Curated-TAPT</td>
<td>TAPT</td>
<td>Curated-TAPT + TAPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>68.25±0.95</td>
<td>59.54±3.36</td>
<td>54.91±1.82</td>
<td>55.81±2.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>67.73±1.37</td>
<td>59.30±1.50</td>
<td>55.99±1.36</td>
<td>56.67±1.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>68.98±0.65</td>
<td>62.26±2.49</td>
<td>57.53±1.13</td>
<td>58.49±2.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>69.08±0.87</td>
<td>61.98±1.60</td>
<td>59.08±1.32</td>
<td>59.72±0.70</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: The comparative results on the fine-grained categories of all models. We report averages across ten random seeds, with standard deviations as subscripts. Best task performance is boldfaced.

<table>
<thead>
<tr>
<th>Model Architectures</th>
<th>Additional Pretraining Phases</th>
<th>Acc.</th>
<th>Precision</th>
<th>Recall</th>
<th>Macro F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Curated-TAPT</td>
<td>TAPT</td>
<td>Curated-TAPT + TAPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>83.79±1.01</td>
<td>71.59±1.48</td>
<td>69.41±1.89</td>
<td>69.91±1.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85.54±0.73</td>
<td>75.30±1.15</td>
<td>70.93±1.77</td>
<td>71.93±1.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>86.14±1.19</td>
<td>77.54±1.07</td>
<td>74.78±1.95</td>
<td>75.65±1.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>86.42±1.44</td>
<td>78.76±2.09</td>
<td>75.20±2.64</td>
<td>76.27±1.47</td>
<td></td>
</tr>
</tbody>
</table>

### Fine-grained categorization.

The confusion matrix of the fine-grained classification is presented in Appendix A.3. The primary error of the fine-grained model comes from label Disinterest with only 0.26 F1-score and Intolerable Risk with 0.47 F1-score, which is largely caused by data imbalance. In the confusion matrix, some Intolerable Risk samples are predicted as Tolerable Risk, which is reasonable due to the limited dialogue history put into BERT which is useful to distinguish Intolerable Risk from Tolerable Risk. Meanwhile, the model confused the label Safety and Tolerable Risk, which is partly due to the model lacks knowledge of counselors’ intervention strategies. For instance, the counselors ask questions to, first, get demographic or experience-related information from clients where the clients’ answers are labeled as Safety. Second, the counselors’ questions are to further explore or provide new perspective for clients to cope with their existing problems where the proper answers are labeled as Tolerable Risk. Specifically, the case “Counselor: Then, what game is it? Client: [Name of the game], my friends recommended to me, I do like this type of games” is Safety but predicted as Tolerable Risk; the case “Counselor: In this situation, how did you cope with your anxiety? Client: I found some other things to do to ease the anxiety. But after doing so, I became more anxious because I had less time to deal with the stuff I should have done.” is Tolerable Risk but was labeled as Safety. As a result, without knowing the intervention strategies, the model might be confused when categorizing the samples of Tolerable Risk and Safety due to the similar structures of counselor-client interaction.

In addition, Others is the easiest to categorize in both classification tasks as expected. Because samples coded as Others are quite different from other samples in terms of semantics and length.

### 5 Conclusion

In this work, we aim to address the issue of real-time evaluation of psychological counseling with various genres. Considering to avoid inappropriate interventions from the counselors, we adopt a new coding scheme (e.g., TCCS) to annotate the clients’ utterances in our new benchmark dataset. We also study the property of the proposed dataset via various advanced deep learning methods.

In future work, we will develop more advanced machine learning solutions to improve the performance of the evaluation systems.

### Ethical Considerations

In the proposed dataset, we have omitted the personal information for the counselors and clients to protect their privacy and identity. It is also not the case that our model can be used to evaluate all kind of psychological counseling for reasons that every dataset is subject to its intrinsic bias and the trained models will inevitably learn the biased characteristics in the dataset.

### References

Steven J Ackerman and Mark J Hilsenroth. 2003. A review of therapist characteristics and techniques pos-


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


A Appendix

A.1 Definitions of Categories in Clients’ Response-experience Type Coding Scheme

The categories of clients’ utterances, response types (validation and invalidation) and experience types (safety, tolerable risk, disinterest, intolerable risk) are depicted below. The definitions of each category are adapted from the therapeutic collaboration coding scheme (TCCS) (Ribeiro et al., 2012). Meanwhile, examples with English translation of counselor-client interactions, simulated from the original data, are provided to facilitate the understanding of our coding process.

——— Validation – Safety ———–

a. Client agrees with counselor’s intervention but does not expand it when the counselor describes or summarizes the client’s discourse using his/her own or client’s words.

来访者：这是活动给你带来了成就感，但是占用太多时间，对学习和休息都有影响。

来访者：是的。

Counselor: These activities brought you a sense of achievement, but took up too much time and affected your study and rest.

Client: Yes.

b. Client answers counselor’s questions when the counselor explores clients’ feelings, ideas and experience using open questions.

来访者：他们给你介绍男朋友的时候，你是怎末反应呢？

来访者：我决定他们在谈话，要么就说现在还不想找。

Counselor: How did you react when they wish to introduce you to potential boyfriends?

Client: I would tell them that I had already had one or I don’ t want to make a boyfriend now.

c. Client seeks advice, explanations, suggestions from the counselors.

来访者：很希望你可以教教我该怎么办样地去应对。

Counselor: I wish you could tell me how to deal with it in a better way.

——— Validation – Tolerable Risk ———–

a. Client accepts the counselor’s advice/suggestions when the counselor invites the client to act in a different way during or after the counseling.

来访者：你有尝试走出去，去别的地方学习吗？

来访者：没有，我想可以尝试一下。

Counselor: Have you ever tried to go outside and to study elsewhere?

Client: Not yet. But I think I could give it a go.

b. Client reformulates his or her perspective over the experience being explored when the counselor directly or implicitly points out the client’s maladaptive beliefs or behaviors.

来访者：我觉得麻烦到他，其实只是你自己的想法。

来访者：或许真的只是我自己单方面在忧虑这些。

Counselor: In fact, it is just your own thinking that this person was bothered by you.

Counselor: Perhaps it’s true, it is only me who is worrying about these stuff.

c. Client attempts to clarify the sense of or make further explanation to his or her previous statement to the counselor.

来访者：感觉很无助，很委屈。

来访者：嗯嗯，就像是被关在一个笼子里怎么也出不去。

Counselor: I feel really helpless and wronged.

Counselor: You think that you are not understood by others.

Client: Yes, it kinds of like that I was being locked in a cage and could not get out at all.

——— Validation – Intolerable Risk ———–

a. Client feels confused and/or are unable to answer the counselor’s question when the counselor points out the client’s maladaptive beliefs, feelings or behaviors by asking questions or speaking out directly.

来访者：你说那件事对你好像没有影响，但是你依然在意，这是为什么呢？

来访者：想不到原因。

Counselor: You have said that it seems like you were not affected by that thing. But you still care about it. Why?

Client: I have no idea.

b. Client does not accept counselor’s advice or suggestions, but persist on looking at their specific experience or topic from his or her perspective when the counselor implicitly or explicitly invites the client to look at the given experience or topic in an alternative way.

来访者：我看到了“敌人”对你的影响，它对你的帮助和限制是什么？
A.2 Experimental Results for Each Category

Table 5 and Table 6 show the detailed experimental results of each category of the coarse- and fine-grained classification of the best-performed model respectively.
A.3 Confusion Matrix

Figure 3 and Figure 4 show the confusion matrix for the coarse- and fine-grained classification of the best-performed model respectively.

![Confusion Matrix for Coarse-Grained Classification](image)

![Confusion Matrix for Fine-Grained Classification](image)

A.4 The Consent Form of Online-counseling

Below is the consent form used in the current work. Every client gave their consent to attend the online text-based psycho-counseling on our counseling platform and agreed to data usage for the current work.

0: 很高兴能够为您服务。
0: I am very happy we can serve you.
1: 为了保障您的权益，请在开启对话之前阅读并同意使用须知
1: In order to protect your rights, please read and agree to the instructions before starting the conversation
2: 我同意接受此心理咨询平台根据我的困惑提供线上文字咨询服务（以下简称服务）
2: I agree to accept the current counseling platform to provide online text-based counseling services based on my confusion (hereinafter referred to as the service)
3: 我了解现阶段此心理咨询平台提供的服务是AI辅助心理咨询，后台由真人心理咨询师提供服务。
3: I understand that the service provided by the current counseling platform at this stage is AI-assisted psycho-counseling, and the service is provided by a real human counselor.
4: 我理解线上文字咨询服务为
4: I understand that the online text-based counseling service is:
5: 互联网线上文字形式的即时心理困惑解答和心理知识普及服务。该服务为中文服务，且仅在该心理咨询平台上提供。
5: Instant psychological puzzle answering and psychological knowledge popularization services in the form of online text on the Internet. This service is provided in Chinese and is only available on the current counseling platform.
6: 我理解服务内容为:
6: I understand service content is:
7: 围绕心理困惑（包括但不限于：情绪问题、情感问题、家庭关系、人际关系、个人成长、生涯发展等困惑）提供支持与帮助。虽然难以保证彻底改善心理状况和解答困惑。但此心理咨询平台秉持着“有时治愈，常常帮助，总是安慰”的态度竭诚为您服务
7: Provide support and help around psychological confusion (including but not limited to: emotional problems, family relationships, interpersonal...
relationships, personal growth, career development, etc.). Although it is difficult to ensure that the psychological condition is completely improved and the confusion is solved, the current counseling platform upholds the attitude of "sometimes heals, often helps, always comforts" and serves you wholeheartedly.

8: I understand that during the counseling process:

9: The conversation will involve my physical/mental health and emotional state and other related information. I have the right to privacy in the counseling service. In principle, the personal information I disclose can be kept strictly confidential. At the same time, my right to privacy is protected and restricted by the national laws.

10: I understand, based on the national laws, the confidentiality principle has exceptions including but not limited to the following items:

11: When the service seeker or other person is preparing or is performing an act that endangers himself or others’ personal and property safety;

12: When the service seeker is likely to endanger others (such as infectious diseases, etc.);

13: When the information disclosed by the service seeker involves the minor beings or about to be sexually assaulted;

14: When the service seeker or other person is preparing or is performing an act that endangers national security or public security;

15: When the data is desensitized and anonymous, when discussing, counseling, or receiving supervision and training among members of the counseling team;

16: When the service seeker or other person is preparing or is performing an act that endangers national security or public security;

17: When data is desensitized and anonymous, when used in scientific research;

18: When the law requires disclosure.

18: I agree that for the above non-confidentiality, for the purpose of protecting the fundamental rights of me or related personnel, the current counseling platform can only disclose information to a minimum within the scope of necessary personnel. In addition, I understand that since the counseling service is based on Online, although the current counseling platform is committed to protecting users’ privacy to the utmost extent, it is difficult to avoid the possibility of leaking users’ personal information due to network information security loopholes, technical failures, or unauthorized intrusions by others.

19: I understand, when the following situations occur, the current counseling platform is difficult to provide effective counseling services and needs to seek professional offline treatment or counseling services:

20: Have suicidal thoughts or plans;

21: Have thoughts or plans to harm oneself or others;

22: Any mental illness diagnosed by the hospital;

23: Meet any of the diagnostic criteria for mental disorders.

24: I understand, if I have any queries or concerns about the above information, I can consult the counseling team;
提供咨询服务，且可能会建议我寻求专业的线下治疗或咨询服务。

24: 我理解，当我认为造成心理困惑的根源在于以下8种）时，此心理咨询平台无法满足我的咨询需求。

25: 我理解，此心理咨询平台围绕心理困惑（包括但不限于：情绪问题、情感问题、家庭关系、人际关系、个人成长、生涯发展等困惑）提供支持与帮助，但仍然存在一些难以实现的服务。

26: 自杀或其他伤害行为的危机干预；

27: 精神障碍的诊断与治疗；

28: 特殊类药物使用方面的具体建议；

29: 处理重度心理创伤；

30: 为我提供具体的职业、学业等资源或信息；

31: 为我提供对于社会现象的看法和政策解读；

32: 为我解梦（例如告诉我梦的含义、为什么会梦到什么人或事等）；

33: 解答非我本人的心理困惑（例如我的朋友、家人、网络等）。

34: 我理解，当我描述的情况超出此心理咨询平台的服务范围（服务范围不包括以上8种）时，此心理咨询平台无法满足我的咨询需求。

35: 我理解互联网线上文字咨询服务潜在的益处和风险。

36: 其中的益处包括但不限于：

37: 能够更便捷地获取服务，无需前往约定地点。并且，尽管风险很小，但我理解可能存在的风险。

38: 潜在风险包括但不限于：

39: 因为我提供的信息可能是不充分的。我得到的服务无法解答我的困惑或改善我的心理状况；由于可能存在的技术故障或其他不可预见的原因，我无法及时得到我对心理困惑的分析与建议。

40: 我同意，当我提供咨询服务时，此心理咨询平台遵循的是中国大陆地区相关的法律法规，而非我所在地区的相关法律法规。

41: 以上知情同意在我单次或多次使用服务的过程中持续生效。

42: 以上知情同意在服务终止时即失效。