Not All Labels Are Equal: On Predicting Utterance Labels in Mental Health Conversation Data

Anonymous submission

Abstract

When analyzing a mental health conversation between a counselor and his/her client, one examine the semantics underlying the utterances of conversation to understand if the counselor has practiced the appropriate psychotherapy techniques at different points of the conversation. Despite the many breakthroughs in solving NLP tasks, state-of-the-art large language models (LLMs) still perform poorly on this utterance label prediction task. While a simple supervised learning architecture combining an utterance encoder with a linear softmax layer can yield better accuracy, the trained classifiers still suffer from poor quality ground truth labels assigned by human annotators. Motivated by this observation, we propose a *quality-aware framework* that derives quality weights of ground truth utterance labels, trains a target classifier in two stages, and evaluates the target classifier with quality weights. Our experiments on three mental health conversation datasets show that target classifiers trained using our framework yield significantly improved accuracy over classifiers trained not using quality weights, even outperforming the strong LLMs using direct prompting.

Introduction

In mental health conversations, it is important to determine the behavior labels of utterances shared between the counselors and clients. These utterance labels can be used to assess the competency of counselors in applying psychotherapy techniques (Moyers et al. 2016; Yang et al. 2025), or the clients' willingness toward behavior change (Tavabi et al. 2020). While there has been several works on predicting labels of utterances in mental health conversations, the existing automated labeling methods suffer from mediocre prediction accuracies. The label prediction task is not easy due to several reasons. One of the reasons is possibly the lack of high-quality ground truth labels assigned by annotators who are ideally the experts in mental health.

Human annotators may assign labels based on their subjective interpretation of the utterance (in the context of conversation) and the annotators' idiosyncrasies in labeling. There may also be multiple ground truth labels that can possibly be assigned to an utterance. As relabeling is a costly option, we therefore seek to automatically assess the quality of ground truth labels assigned by label prediction models. Our idea is based on two principles:

- A ground truth label supported by majority of models is high quality.
- A non-ground truth label supported by majority of models implies that the ground truth label is not high quality.

Based on the above principles, we introduce a quality-aware framework to train target classifiers that can automatically predict labels of mental health-related utterances, and to evaluate the target classifiers. Instead of the prohibitive costs to use human judges, we train a set of base classifiers to act as judges.

Specifically, we make the following contributions:

- We propose a quality-aware framework to improve the accuracy of utterance label prediction for mental health conversation data. Instead of relying on human experts to verify or re-annotate the ground truth labels, the framework deploys a set base classifiers to evaluate the label qualities.
- We define the quality weight of utterance labels, which can be used for training quality-aware target classifiers and for evaluating them.
- We conduct extensive experiments on different target classifiers under the above framework and evaluate them on three mental health conversation datasets. The target classifiers show improved accuracy when trained using labels with quality weights. These supervised classifiers also significantly outperform the LLM classifiers.

Datasets

We include three datasets with utterance level ground truth labels. PeerMI (Welivita and Pu 2022) and AnnoMI (Wu et al. 2022) are counseling datasets containing sessions of counselors interacting with clients using motivational interviewing techniques (Moyers et al. 2016). In motivational interviewing sessions, the counselor seeks to help a client explore both sides of ambivalence and identify the motivation to change his/her behavior. ESConv is a dataset of emotional support sessions between supporters (or counselors) and seekers (or clients) (Liu et al. 2021). In emotional support conversations, the counselor seeks to understand the cause of emotion distress of the client and offer him/her the appropriate support. Table 1 shows the example instances of counselor behavior or strategy assignment in the three

datasets. The full set of utterance labels and their descriptions can be found in Appendix A. The statistics of the three datasets are shown in Table 2.

Dataset	Context/Utterance	Label
PeerMI	Context: "Client: I've been struggling with drinking lately." Utterance: "Counselor: I can see this is causing you distress. **What specific situations trigger your drinking?**"	Open Question
AnnoMI	Context: "Client: Yeah, it's bad. My friends and I, we just– I mean, it's fun and uh– so I don't know what to do, but I- but I gotta change. I want it" Utterance: "Counselor: **So, it sounds like you're ready. You're ready to give this up.**"	Simple Reflection
ESConv	Context: "Seeker: Yes. There were a argument between two of my friend while trying to resolve the issue they all started targeting me." Utterance: "Supporter: **I have a hard time with my friends too sometimes**"	Self-Disclosure

Table 1: Utterance Label Prediction Task

Dataset	Training	Finetuning	Testing	Total
PeerMI	8374	6717	1720	16,811
AnnoMI	2111	2033	736	4880
ESConv	9133	7433	1810	18,376

Table 2: Dataset Statistics

Related Work

Learning from Noisy Labels. Noises in labels have been well studied in text classification research and assigning label to utterances in mental health conversations is a type of text classification task. In the past, researchers explore the filtering of label noises to improve the accuracy of classification methods (Brodley and Friedl 1999). There are also other researchers who found out that BERT and other large language models demonstrate robust accuracy in text classification (Zhu et al. 2022).

To train classifiers to be noise-tolerant, Natarajan and others proposed weighted 0-1 loss to train binary classification model (Natarajan et al. 2013). This model is however restricted to binary labels and it assumes noise is label specific, not instance specific. In our work, we measure noises at the utterance level (or instance level) as each utterance appears in a distinct conversation context with noises that affect the correctness of label assignment.

Yuan et. al proposed an innovative collaborative learning framework NoiseAL based on active learning to combine two small models with a LLM and to determine and correct noisy labels respectively for improving the classification accuracy (Yuan et al. 2025). This method however adopts incremental learning instead of a simple two-stage learning approach in our work.

Due to the success of LLMs in many well studied NLP tasks, researchers investigate into using instruction-tuned LLMs to preform label prediction. Nevertheless, it have been shown that LLMs yield poorer performance than label supervised BERT and RoBERTa in specialized prediction tasks (Li et al. 2023). Utterance label prediction in mental health conversation is unfortunately one of such specialized tasks.

Utterance Label Assignment in Mental Health Conversation Utterance label assignment have increasingly studied in recent years. Cao et al. studied two such tasks in therapy dialogue analysis: *behavior categorization* and *behavior forecasting*. They proposed a hierarchical neural models using contextualized word embeddings (GloVe and ELMo), along with attention mechanisms to address the tasks. (Han et al. 2024) proposed the CoI framework to enhance LLMs' ability to predict utterance labels by decomposing the reasoning process into three stages: interaction definition, involvement assessment, and valence analysis. When evaluated on real-world MI datasets, CoI outperforms zero-shot, few-shot, and CoT prompting baselines. (Chiu et al. 2024; Sun et al. 2024) developed prompting-based LLM classifiers to predict labels of counselor and client utterances in psychotherapy sessions. Even with strong LLMs GPT-4, these works show that the accuracies of these classifiers are not highly accurate and they overlook the quality issues in ground truth labels.

Quality-Aware Training and Evaluation Framework

Figure 1 depicts our proposed framework to train a target classifier to predict the behavior label of an utterance in a mental health related conversation. The framework consists of three stages, namely: (1) **Quality-based Weighting** to determine the label quality of utterances; (2) **Quality-aware Supervised Learning** to train the target classifier using utterances with labels of varying quality; and (3) **Evaluation** to determine the target classifier's accuracy.

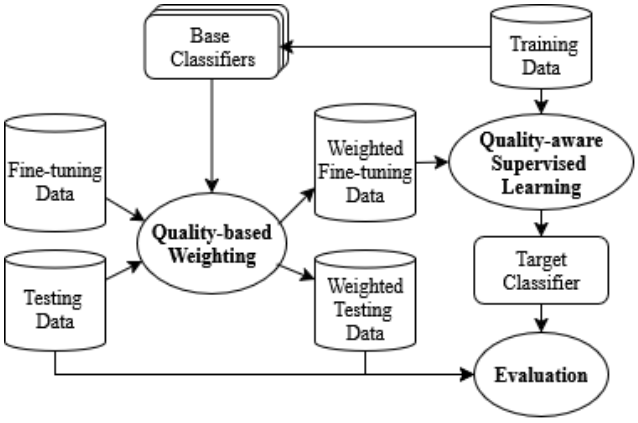


Figure 1: Quality-aware framework for training target classifiers.

The framework divides our labeled utterance data into

training data, finetuning data, and testing data. A set of *base classifiers* based on different underlying models (e.g., machine learning, deep learning and LLM-based models) are trained in the standard way using the same training data for the purpose of determining the quality of labels in the finetuning and testing data in the Quality-based Weighting stage. These base classifiers thus mimic different human experts assigning labels to unseen data (in finetuning and testing data). In this paper, our base classifiers are constructed by adding a linear layer to different language models, including BERT, Llama-3-8B, QWen3-Embedding-8B and Text-Embedding-3-large.

The key idea of Quality-based Weighting is to determine the quality weights of utterance labels in the finetuning and testing data so that one can train the target classifier to focus more on the high quality labels than the low quality ones. Here, we assume that the labels contributed by human annotators in the training data are generally good and hence the base classifiers trained are reasonably accurate. The quality weights of finetuning and testing data are then derived by these base classifiers. In Section , we will introduce the quality weight of an utterance's label considering both the *support of ground truth label* and the *support of another alternative label* by the set of base classifiers.

In the Quality-aware Supervised Learning stage, we train a target classifier using training data followed by finetuning data with quality weight. The former allows the target classifier to achieve reasonable accuracy (similar to base classifiers) while the latter further tunes the target classifier to focus on high quality labels. Finally, we evaluate the target classifier on testing data in which label quality are already determined.

Quality Weights of Utterance Labels

Consensus on Non-Ground Truth Label

As the set of base classifiers predict labels to a given utterance, we measure the consensus among the base classifiers on predicting non-ground truth labels using **non-ground truth entropy** of the predicted labels. Low entropy indicate most base classifiers have largely agreed on the same non-ground truth label, and high entropy suggests a lack of consensus among them.

Let $\mathcal{C}$ denote the set of all labels and $\mathcal{C}'$ denote the set of predicted labels except the ground truth label. We define the non-ground truth entropy of labels assigned to an utterance u_j by all the basic classifier models $M_1, \ldots, M_N$ as follows[1]

$$H_j = -\sum_{c \in \mathcal{C}'} \frac{n_j(c)}{N} \log \frac{n_j(c)}{N} - \frac{n_j(GT_j)}{N} \log \frac{1}{N} \quad (1)$$

where $n_j(c)$ is the number of models predicting label c for utterance u_j. The second component of Eq (1) treats the models predicting the ground truth label GT_j as ones which predict distinctive labels that are not GT_j. Here, $|\mathcal{C}|$ is assumed to be much larger than N, the number of base classifiers. A small H_j indicates strong consensus among models, even if the consensus is on a label different from the ground truth. $H_j = 0$ when all models choose the same label, and $H_j = log|C|$ when every label is chosen by equal number of models As small H_j weakens the validity of ground truth label, we will use it to define the quality weight of each ground truth label.

[1]In our implementation, we incorporate Laplace Smoothing into H_j, i.e., $H_j = \left(-\sum_{c \in \mathcal{C}'} \frac{n_j(c)+\epsilon}{N+|\mathcal{C}|\epsilon} \log \frac{n_j(c)+\epsilon}{N+|\mathcal{C}|\epsilon}\right) - \left(\frac{(|C-C'|-n_j(GT_j))\epsilon}{N+|\mathcal{C}|\epsilon} \cdot \log \frac{\epsilon}{N+|\mathcal{C}|\epsilon}\right) - \left(\frac{n_j(GT_j)(1+\epsilon)}{N+|\mathcal{C}|\epsilon} \log \frac{1+\epsilon}{N+|\mathcal{C}|\epsilon}\right)$.

Ground Truth Label Quality

We assign a quality weight W_j to the ground truth label GT_j of utterance u_j. To compute this weight, we first introduce the following quantities.

Ground Truth Label Support. The ground truth label support is defined as

$$CR_j = \frac{n_j(GT_j) + \epsilon_1}{N + |\mathcal{C}|\epsilon_1},$$

where ϵ_1 is a Laplace smoothing constant. CR_j essentially measures the proportion of base classifier models predict GT_j. The more models predict GT_j, the higher the ground truth label support. In this work, we empirically set ϵ_1 to be 0.1. To explore the effect of ϵ_1 on the final performance, we also conducted experiments with $\epsilon_1 = 0.01$ and $\epsilon_1 = 0.001$, and found that varying this parameter led to only minor changes in performance.

Alternative Label Consensus. We measure the extent to which models agree on an alternative label (non-ground truth label) by

$$CS_j = 1 - \frac{H_j}{\log |\mathcal{C}|},$$

where H_j is the non-ground truth entropy. CS_j approaches 1 when models unanimously agree on a non-ground truth label, and 0 when they predict completely diverse labels. Note that the alternative non-ground truth labels are not included in the finetuning dataset despite full consensus among the base classifier as they have not yet been verified by human annotators.

Quality Weight. We define the quality weight as

$$W_j = \frac{CR_j}{CR_j + (1 - CR_j) \times CS_j}.$$

This formulation ensures that W_j will be close to 1 whenever models cannot reach consensus on an alternative label, even if none of them predict the ground truth label.

Quality Weight and Accuracy Distributions

Base Classifiers Quality work requires base classifiers trained to predict the labels of utterances. In our work, we adopt four base classifiers. Each base classifier consists of a frozen encoder to turn an utterance into an embedding vector followed by a linear layer to predict the label. The four frozen encoder options include: (a) BERT, (b) Llama-3-8B, (c) QWen3-Embedding-8B, and (d) text-embedding-3-large.

Only the linear layer is trained using with cross entropy loss using the training data. The final base classifiers are termed (a) BERT-Linear, (b) Llama-Linear, (c) QWen-Linear, and (d) textEmbeddingLarge-Linear. Cross-entropy loss penalizes the discrepancy between the predicted label distribution $\hat{p}(\cdot|j)$ and the one-hot ground truth label vector y_j. For an utterance u_j with ground truth label GT_j, the cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{j}\sum_{c\in\mathcal{C}} p(c|u_j)\,\log\hat{p}(c|u_j)$$

where $p(c|u_j) = 1$ if $y_j = c$, and 0 otherwise.

Quality Weight Distribution. Figure 2 shows the proportion of utterances in the testing data having ground truth utterance labels of different quality weights. The figure reveals that as much as 23.1%, 17.9% and 28.4% of utterances from PeerMI, AnnoMI and ESConv respectively have very poor quality weights from 0.0 to 0.1 assigned to their ground truth labels. The testing data of PeerMI, AnnoMI and EsConv have more than 50% ground truth labels assigned with high quality weights in [0.8,1.0]. Similar quality weight distributions are also found in the training and finetuning datasets. With many utterances having low quality weights, it is thus a concern that they may introduce noises when training or evaluating the target classifiers.

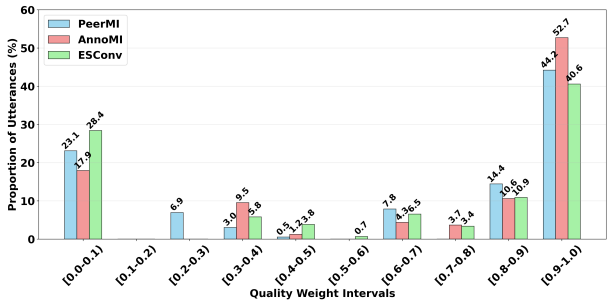


Figure 2: Distribution of quality weights in the testing data.

Case Examples. Example 1 shows an utterance in AnnoMI is assigned the ground truth label "Simple Reflection". All the four base classifiers unanimously predict a different but reasonably correct label "Closed Question". In this case, the ground truth label support (CR) and alternative label consensus (CS) are 0.0227 and 0.9877 respectively. A low quality weight of 0.023 is thus assigned to the ground truth label.

Example 1

Context:
Counselor: Okay. And how long has this been going on?
Client: Well, I, uh, I've traced it back to a holiday I had in Egypt about four years ago. I had, um, a horrible bout of sickness and diarrhea, and, uh, and I don't think my stomach's ever been really right since then.

Utterance:
Counselor: Mm-hmm. Okay. Okay. So you think it began about four years ago after this holiday?

Example 2 illustrates another utterance with ground truth label "closed question". Among the four base classifiers, two predict "closed question", one predicts "open question", and another predicts "complex reflection". As only two base classifiers support the ground truth label, the ground truth label support is only moderate with $CR = 0.4773$. On the other hand, the alternative label consensus $CS = 0.0010$ is very low due to the lack of consensus among the remaining base classifiers. Hence, we assign a very high quality weight ($W = 0.9989$) to the ground truth label.

Example 2

Context:
Counselor: –wasn't that– Is that right? Yeah? So, "Do I live a party lifestyle– party-girl kind of lifestyle, or am I got some other dreams to follow?" So, they're hard calls to make. And-and it all– And there are no three ways. We will have to make those calls at some point.
Client: Mm.

Utterance:
Counselor: And, my sense, is that some of the challenge for you, at the moment? Yeah?

Ground Truth Labels and Their Alternatives

For each dataset, we identify utterances whose ground truth labels are assigned low quality weights ($W^q < 0.1$). These utterances have at least two base classifiers predicting alternative labels (different from their ground truth ones). We summarize these utterances of different (ground truth label, alternative label) combinations in a confusion matrix shown in Figure 3. Each cell in the matrix represents the count of utterances that share a given ground truth label but also an alternative label. We use the matrix to visualize the confusion in different (ground truth label, alternative label) pairs in the dataset.

In **PeerMI**, we notice the following confusions. Utterances with `Complex Reflection` as ground truth label are frequently predicted with `Give Information` alternative label. This occurs when the utterance offers a reflective statement that is factual or educational, or when the counselor paraphrases the client's concern at the same time adding new information or interpretation. Another major confusion arises between `Advise with Permission` and Advise without Permission. These two labels are similar semantically: both provide recommendations or guidance, but differ in whether the counselor explicitly seeks the client's consent before offering advice. In conversations, the presence or absence of a permission phrase (e.g., "Would you like me to share some thoughts?") can be subtle or implied. Hence, even human annotators and by extension, machine classifiers, may find it difficult to reliably distinguish between them. Thirdly, utterances with `Advise without Permission` ground truth labels are often associated with `Give Information` alternative label. Both behaviors involve delivering content or direction to the client. The distinction depends on whether the statement carries a prescriptive tone (advice) or remains purely factual (information).

In **AnnoMI**, the most common label confusions arise between `Open Question` and `Closed Question`, as well as between `Simple Reflection` and `Complex Reflection`. Both open and closed question types serve to guide exploration, yet open questions are designed to evoke elaboration, while closed questions aim to clarify or confirm information. The fine-grained linguistic difference between the two can easily lead to inconsistent labeling across annotators or models.

In **ESConv**, the confusion mostly occur for utterances assigned with the `Others` ground truth label. They are are often given alternative labels: `Affirmation` and `Question`. This may be caused by the flexible and context-dependent nature of emotional support conversations. Human annotators may assign `Others` as the ground truth label when the supportive intent is subtle, while the base classifiers capture these subtle cues and align them with `Affirmation`. Similarly, when the counselor's utterance implicitly invites the seeker to elaborate, it can be perceived as interrogative, leading to the confusion between `Others` and `Question`. Additionally, utterances labeled with `Reflection` ground truth label are frequently associated with the `Affirmation` alternative label. This confusion arises because reflections often serve a dual function, both mirroring the client's feelings and offering emotional validation. The distinction between reflection and affirmation is therefore weak.

Quality-aware Supervised Learning

In quality-aware supervised learning, we have both training data (unweighted) and weighted finetuning data to train a target classifier. While standard cross entropy loss applies to the training data, we introduce the quality weight W_j into the loss function for this round of supervised learning using the fine-tuning data. We adopt an instance-weighted cross-entropy loss as defined below:

$$\mathcal{L}_{\text{WCE}} = -\frac{\sum_j W_j \sum_{c \in \mathcal{C}} p(c|u_j) \log \hat{p}(c|u_j)}{\sum_j W_j}$$

Here, utterances with high-quality ground truth labels (larger W_j) contribute more significantly to the optimization, while utterances of smaller W_j contribute less.

In our work, instead of training a target classifier with both training data and weighted fine-tuning data using a completely new classification model, we further train each base classifier with weighted fine-tuning data to derive the target classifier to study the effect of weighted cross-entropy loss finetuning.

Experiment and Results

Experiment Setup

We conduct experiments to evaluate the effectiveness of weight-aware supervised learning on multiple mental health conversation datasets and encoding model backbones. We randomly assign 50%, 40%, and 10% of the sessions in each dataset into training, finetuning, and testing data respectively (see Table 2). In the training and finetuning data, we further reserve 10% of the sessions for validation, so as to select the best model during training.

We first conduct the stage 1 training of the target classifier using training data, followed by stage 2 training using the finetuning dataset. In stage 1, we optimize the classifier model using the cross-entropy loss $\mathcal{L}_{\text{CE}}$. In stage 2, we apply the weighted cross-entropy loss $\mathcal{L}_{\text{WCE}}$ instead. We report the accuracy of the target classifier after stage 1 and stage 2 denoted by S1 and S1+S2 respectively. To determine effectiveness of weighted cross-entropy loss, we also report the S1+S2 accuracy when $\mathcal{L}_{\text{CE}}$ is used in stage 2 training (denoted by S1+S2 w/o weights).

Supervised target classifiers. We evaluate the target classifiers using the following models for encoding utterances in the target classifier, namely: (a) bert-base-uncased; (b) Qwen3-Embedding-8B; (c) Llama-3-8B-Instruct; and (d) text-embedding-3-large. For BERT, Llama, and Qwen models, we extract the final hidden state corresponding to the last token from the final transformer layer to represent the entire utterance embedding. For the text-embedding-3 models, each utterance is directly fed into the OpenAI API to obtain its sentence-level vector representation. This results in the following supervised target classifiers: (1) Bert-Linear; (2) Qwen-Linear; (3) Llama-Linear; and (4) textEmbeddingLarge-Linear respectively.

LLM target classifiers. For comparison, we also include several pretrained generative models that directly performs label classification through in-context prompting (see Appendix B). These LLM classifiers adopt several state-of-the-art LLMs: (a) Qwen3-8B; (b) Llama-3-8B-Instruct; and (c) GPT-4o.

Jointly trained encoding model and linear layer. For comparison, we also include a BERT-Linear-Joint model, in which both the encoder and linear layers are fine-tuned jointly. This allows us to evaluate the effectiveness of quality-aware supervised learning in joint training.

Metrics

Accuracy For a model M_i, its prediction on utterance u_j is denoted as $\hat{L}_{ij}$. The accuracy is defined as:

$$Acc(M_i) = \frac{\sum_j \mathbf{I}(\hat{L}_{ij} = GT_j)}{\sum_j 1}$$

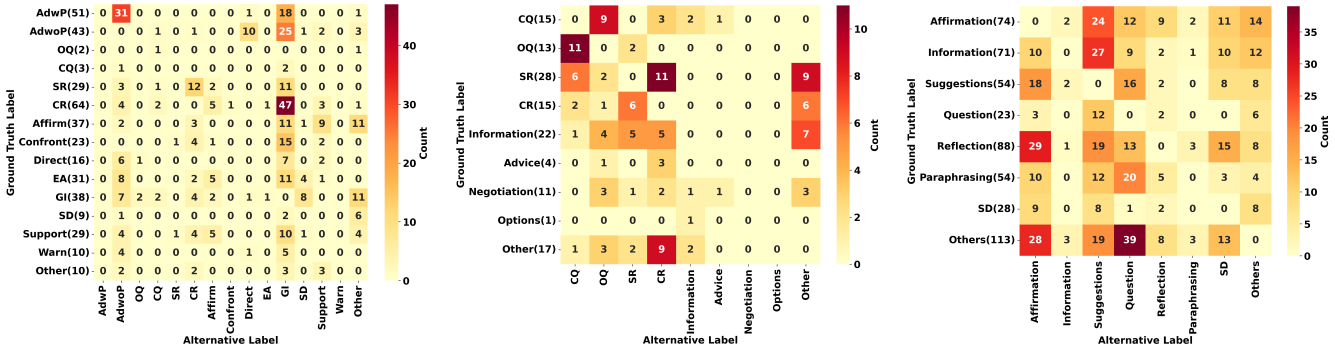
where $\mathbf{I}()$ is a function that returns 1 when the input expression is true, and 0 otherwise.

Weighted Accuracy Given the quality weight W_j of each utterance u_j, we extend the conventional definition of accuracy by weighting the contribution of each ground truth instance. This weighted accuracy is defined as:

$$Acc^w(M_i) = \frac{\sum_j W_j \times \mathbf{I}(\hat{L}_{ij} = GT_j)}{\sum_j W_j}$$

Results and Analysis

Table 3 summarizes the accuracy results of target classifiers for different utterance encoding models and datasets. We derive several interesting findings from the results.



(a) PeerMI (b) AnnoMI (c) ESConv

Figure 3: Ground truth labels with quality weights $W^q < 0.1$ and their alternative labels. `AdwP/AdwoP` refers to *Advise with/without Permission*. `OQ/CQ` stands for *Open/Closed Question*. `SR/CR` corresponds to *Simple/Complex Reflection*. `GI` represents *Give Information*, `EA` denotes *Emphasize Autonomy*, and `SD` indicates *Self-Disclose*. The total count for each label is shown after the label.

Effect of quality-aware supervised learning. Firstly, target classifiers with two-stage training (S1+S2) show noticeably improvement in accuracy (Acc and Acc^w) compared with only stage 1 (S1) training only. Unlike the other target classifiers, textEmbeddingLarge-Linear does not appear to benefit much from two-stage training. This may be attributed to textEmbeddingLarge-Linear's high accuracy relative to other classifiers in stage 1 training. As quality weights are derived from predicted labels of base classifiers after stage 1 training, they may not offer textEmbeddingLarge-Linear much more useful information to improve its accuracy in stage 2 training.

Secondly, quality-aware supervised learning improves target classifier's accuracy. Target classifiers finetuned with weighted cross-entropy loss (i.e., S1+S2) consistently outperform their counterparts using non-weighted cross-entropy loss (i.e., S1+S2 w/o weight). This suggests that the proposed quality-aware supervised learning successfully guides the models to focus on more reliable samples during optimization.

Choice of Utterance Encoder. Thirdly, model capacity correlates positively with accuracy results. Among the embedding-based models, textEmbeddingLarge-Linear achieves the highest accuracy, outperforming both Qwen-Linear and Llama-Linear. This suggests that text embedding large model may have been pretrained with very large data allowing it to generate more discriminative and transferable embedding representations for label classification than general-purpose language models or other smaller large language models. In addition, both Llama-Linear and Qwen-Linear consistently outperform BERT-Linear, reflecting that large-scale pretraining yields stronger representational power and better generalization across mental health conversation datasets.

Joint supervised learning option. Fourthly, Bert-Linear-Joint outperforms Bert-Linear, since fine-tuning the Bert encoder allows the model to better adapt to domain-specific features. However, directly fine-tuning larger LLMs remains a challenge. Further investigation is needed to determine whether full fine-tuning of LLMs can outperform frozen-embedding classifiers.

Weighted accuracy evaluation. The weighted accuracy results of target classifiers are usually better than their unweighted accuracy results. This is interesting because the classifiers are more likely to predict the high quality ground truth labels correctly. This findings is observed across all target classifiers.

Relationship with quality weight distribution. Among the three datasets, the results show that ESConv is the most challenging dataset, followed by PeerMI and AnnoMI across all the target classifiers. Interestingly, this findings is consistent with the distribution of quality weights in the three datasets as shown in Figure 2. Specifically, ESConv has the largest proportion of ground truth labels with low quality weights, PeerMI has the next largest proportion, and AnnoMI has the least proportion.

Comparison with LLM target classifiers. Table 3 shows that the LLM target classifiers including GPT-4o performs poorly on this label prediction task. Their accuracy results (weighted and non-weighted) are significantly worse than all our supervised target classifiers. This suggests that direct prompting on pre-trained LLMs does not yield good accuracy.

Conclusion

In this paper, we address the problem of predicting utterance labels in mental health conversation data when not all the ground truth labels are perfect. As a target classifier may suffer from these labels with uneven quality, we propose a quality-aware framework that distinguishes the label quality and fine-tunes the target classifiers using ground truth labels assigned with quality weights in the fine-tuning dataset. The quality weight of a ground truth label is derived from its sup-

Model	PeerMI		AnnoMI		ESConv	
	Acc	**Acc^w**	**Acc**	**Acc^w**	**Acc**	**Acc^w**
Llama-3-8B-Instruct	0.1732	0.1861	0.1617	0.1659	0.1613	0.1763
Qwen3-8B	0.4140	0.4999	0.5299	0.6514	0.5359	0.7404
GPT-4o	0.5128	0.6561	0.4443	0.4817	0.5381	0.7359
Bert-Linear-Joint						
- S1	0.6203	0.8587	0.6834	**0.8895**	0.5779	0.8666
- S1+S2	**0.6406**	**0.8751**	**0.6989**	0.8840	**0.5939**	**0.8736**
- S1+S2 w/o weight	0.6226	0.8313	0.6902	0.8769	0.5856	0.8476
Bert-Linear						
- S1	0.5854	0.7884	0.6059	0.7625	0.5348	0.7737
- S1+S2	**0.5936**	**0.8077**	**0.6372**	**0.7888**	**0.5519**	**0.7993**
- S1+S2 w/o weight	0.5912	0.7984	0.6209	0.7720	0.5469	0.7870
Qwen-Linear						
- S1	0.6104	0.8603	0.6467	0.8542	0.5436	0.8242
- S1+S2	**0.6261**	**0.8695**	**0.6888**	**0.8809**	**0.5607**	**0.8436**
- S1+S2 w/o weight	0.6145	0.8457	0.6671	0.8435	0.5392	0.8143
Llama-Linear						
- S1	0.5976	0.8431	0.6739	0.8747	0.5447	0.8081
- S1+S2	**0.6244**	**0.8661**	0.6888	**0.8768**	**0.5574**	**0.8248**
- S1+S2 w/o weight	0.6168	0.8416	**0.6929**	0.8584	0.5419	0.7975
textEmbeddingLarge-Linear						
- S1	0.6331	**0.8981**	0.6902	**0.8994**	0.5828	**0.8827**
- S1+S2	**0.6383**	0.8909	**0.6929**	0.8856	**0.5977**	0.8782
- S1+S2 w/o weight	0.6360	0.8957	0.6834	0.8679	0.5801	0.8757

Table 3: Accuracy Results of Target Classifiers (**S1**: Model is trained for Stage 1 only; **S1+S2**: Model is trained for both Stages 1 and 2; **S1+S2 w/o weight**: Model is trained for both Stages 1 and 2 using cross-entropy loss only.)

port by a set of base classifiers as well as the existence of an alternative label agreed among the base classifiers. Our experiments show that this framework allows the target classifiers to achieve more accurate label prediction results. We also show that target classifiers using small utterance encoders appear to benefit more from the quality weighted ground truth labels than those using large encoders. For more comprehensive results, we plan to cover more types of target classifiers in the future. In particular, jointly fine-tuning the encoder and linear layer (e.g., Bert-Linear-Joint) will likely improve the accuracy results although this will incur more training costs. The framework can also be easily applied to other conversation datasets outside the mental health domains.

References

Brodley, C. E.; and Friedl, M. A. 1999. Identifying Mislabeled Training Data. 11: 131–167.

Cao, J.; Tanana, M.; Imel, Z.; Poitras, E.; Atkins, D.; and Srikumar, V. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5599–5611. Florence, Italy: Association for Computational Linguistics.

Chiu, Y. Y.; Sharma, A.; Lin, I. W.; and Althoff, T. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. arXiv:2401.00820.

Han, G.; Liu, W.; Huang, X.; and Borsari, B. 2024. Chain-of-Interaction: Enhancing Large Language Models for Psychiatric Behavior Understanding by Dyadic Contexts . In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 392–401. Los Alamitos, CA, USA: IEEE Computer Society.

Li, Z.; Li, X.; Liu, Y.; Xie, H.; Li, J.; Wang, F.-l.; Li, Q.; and Zhong, X. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Moyers, T. B.; Rowell, L. N.; Manuel, J. K.; Ernst, D.; and Houck, J. M. 2016. The motivational interviewing treatment integrity code (MITI 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65: 36–42.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Sun, X.; Pei, J.; Wit, J. d.; Aliannejadi, M.; Krahmer, E.; Dobber, J. T.; and Bosch, J. A. 2024. Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds.,

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 5609–5621. Torino, Italia: ELRA and ICCL.

Tavabi, L.; Stefanov, K.; Zhang, L.; Borsari, B.; Woolley, J. D.; Scherer, S.; and Soleymani, M. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 406–413.

Welivita, A.; and Pu, P. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, 3315–3330.

Wu, Z.; Balloccu, S.; Kumar, V.; Helaoui, R.; Reiter, E.; Recupero, D. R.; and Riboni, D. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6177–6181. IEEE.

Yang, Y.; Achananuparp, P.; Huang, H.; Jiang, J.; Kit, P. L.; Lim, N. G.; Ern, C. T. S.; and Lim, E.-P. 2025. CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration. In *62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuan, B.; Chen, Y.; Zhang, Y.; and Jiang, W. 2025. Hide and Seek in Noise Labels: Noise-Robust Collaborative Active Learning with LLM-Powered Assistance. *arXiv preprint arXiv:2504.02901*.

Zhu, D.; Hedderich, M. A.; Zhai, F.; Adelani, D. I.; and Klakow, D. 2022. Is BERT robust to label noise? A study on learning with noisy labels in text classification. *arXiv preprint arXiv:2204.09371*.

A: Labels of Mental Health Conversation Datasets

The set of labels used in our three datasets are shown in Tables 4, 5, and 6.

B: In-Context Prompt for LLM Target Classifiers

To predict utterance labels using large language models, we design three prompts(Shown in Figure 4, 5 and 6) that guide the models to infer labels directly from conversational context, without additional supervision or fine-tuning. Each prompt corresponds to the utterance classification task of a dataset.

Label	Description
Advise with Permission	Making suggestions after asking for client's permission
Advise without Permission	Making suggestions without asking for client's permission
Affirm	Providing positive reinforcement or support
Closed Question	Asking questions that can be answered with yes/no
Complex Reflection	Reflecting client's meaning with added interpretation
Confront	Challenging or disagreeing with the client
Direct	Giving direct instructions or commands
Emphasize Autonomy	Highlighting the client's freedom of choice
Give Information	Providing factual information, education, or advice
Open Question	Asking questions that cannot be answered with yes/no
Self-Disclose	Sharing personal experiences or feelings
Simple Reflection	Restating or slightly rephrasing what the client said
Support	Expressing empathy, understanding, or encouragement
Warn	Warning about negative consequences
Other	Any behavior not covered by the above codes

Table 4: PeerMI's Utterance Labels

Label	Description
Closed Question	Asking questions that can be answered with yes/no
Open Question	Asking questions that cannot be answered with yes/no
Simple Reflection	Restating or slightly rephrasing what the client said
Complex Reflection	Reflecting client's meaning with added interpretation
Information	Providing factual information, education, or advice
Advice	Offering recommendations or guidance
Options	Presenting alternative choices or possible courses of action
Negotiation	Engaging the client in collaborative discussion
Other	Any behavior not covered by the above codes

Table 5: AnnoMI's Utterance Labels

Label	Description
Question	Asking questions to understand, explore, or guide the conversation
Affirmation	Providing positive reinforcement, validation, or comfort
Providing Suggestions	Offering advice, recommendations, or solutions
Reflection of feelings	Mirroring or acknowledging the seeker's emotions
Information	Sharing factual information, resources, or knowledge
Restatement	Repeating or rephrasing what the seeker said
Self-disclosure	Sharing personal experiences or feelings
Others	Any strategy not covered by the above categories

Table 6: EsConv's Utterance Labels

Motivational Interviewing Treatment Integrity (MITI) Code Classification

Session Context:
{conversation_history}

Target Utterance
{utterance}

Available MITI Codes:

- Give Information: Providing factual information, education, or advice
- Advise without Permission: Making suggestions without asking for client's permission
- Advise with Permission: Making suggestions after asking for client's permission
- Affirm: Providing positive reinforcement or support
- Complex Reflection: Reflecting client's meaning with added interpretation or emphasis
- Simple Reflection: Restating or slightly rephrasing what the client said
- Open Question: Asking questions that cannot be answered with yes/no
- Closed Question: Asking questions that can be answered with yes/no
- Confront: Challenging or disagreeing with the client
- Direct: Giving direct instructions or commands
- Emphasize Autonomy: Highlighting the client's freedom of choice
- Self-Disclose: Sharing personal experiences or feelings
- Support: Expressing empathy, understanding, or encouragement
- Warn: Warning about negative consequences
- Other: Any behavior not covered by the above codes

Instructions:

1. Consider the target utterance in the above context of a motivational interviewing counselling session.
2. Determine the most appropriate MITI code for this target utterance.
3. Output the MITI code only.

Figure 4: The prompt for behavioral code prediction of PeerMI.

Therapist Behavior Classification

Session Context:
{conversation_history}

Target Utterance:
{utterance}

Available Behavior Codes:

- closed question: Asking questions that can be answered with yes/no or brief responses
- open question: Asking questions that invite detailed, open-ended responses
- simple reflection: Restating or paraphrasing what the client has said
- complex reflection: Reflecting client's meaning with added interpretation, inference, or emphasis
- information: Providing factual information, education, or sharing knowledge
- advice: Giving direct advice, recommendations, or suggestions to the client
- negotiation: Discussing treatment plans, goals, or collaborative decision-making
- options: Presenting multiple choices or alternatives for the client to consider
- other: Other therapist behaviors not fitting the above categories

Instructions:

1. Consider the target utterance in the above context of a motivational interviewing counseling session.
2. Determine the most appropriate behavior code for this target utterance.
3. Output the behavior code only.

Figure 5: The prompt for behavior code prediction of AnnoMI.

Emotional Support Strategy Classification

Session Context:
{conversation_history}

Target Utterance:
{utterance}

Available Strategies:

- Question: Asking questions to understand, explore, or guide the conversation
- Affirmation and Reassurance: Providing positive reinforcement, validation, or comfort
- Providing Suggestions: Offering advice, recommendations, or solutions
- Reflection of Feelings: Mirroring or acknowledging the seeker's emotions
- Information: Sharing factual information, resources, or knowledge
- Restatement or Paraphrasing: Repeating or rephrasing what the seeker said
- Self-disclosure: Sharing personal experiences or feelings
- Others: Any strategy not covered by the above categories

Instructions:

1. Consider the target utterance in the above context of a emotional support session.
2. Determine the most appropriate strategy code for this target utterance.
3. Output the strategy code only.

Figure 6: The prompt for emotional support strategy prediction of ESConv.