A Multi-Persona Chatbot for Hotline Counselor Training

Orianna DeMasi, Yu Li, Zhou Yu
University of California, Davis
`[odemasi, yooli, joyu]@ucdavis.edu`

Abstract

Suicide prevention hotline counselors aid individuals during difficult times through millions of calls and chats. A chatbot cannot safely replace a counselor, but we explore whether a chatbot can be developed to help train human counselors. Such a system needs to simulate intimate situations across multiple practice sessions. Open-domain dialogue systems frequently suffer from generic responses that do not characterize personal stories, so we look to infuse conversations with persona information by mimicking prototype conversations. Towards building a "Crisisbot" hotline visitor simulation, we propose a counseling strategy annotation scheme and a multi-task framework that leverages these counselor strategies to retrieve similar examples, generate diverse sub-utterances, and interleave prototype and generated sub-utterances into complex responses. We evaluate this framework with crowdworkers and experienced hotline counselors. The framework considerably increases response diversity and specificity, with limited impact to coherence. Our results also show a considerable discrepancy between crowdworker and counselor judgements, which emphasizes the importance of including target populations in system development and evaluation.

1 Introduction

Crisis hotlines provide the opportunity for individuals to reach out, be heard, and get help during intense times of need. They are invaluable, highly used resources, as evidenced by The National Suicide Prevention Lifeline centers alone handling 2.2 million calls during 2018 (Vibrant Emotional Health, 2019). Further, many helplines have expanded services to chat, in addition to voice calls, to support texting and increased demand. To help visitors who contact them, helplines rely on counselors trained in various active listening tech-

Figure 1: Snippet of a conversation illustrating a counselor practicing counseling with Crisisbot. Realistic visitor messages are often complex and include multiple issues that characterize a personal situation.

niques and de-escalation strategies (Gould et al., 2013; Knox et al., 2012; Labouliere et al., 2020).

As in other high-stakes settings, training hotline counselors suffers from the fundamental challenge that counselors must practice counseling skills without putting any distressed hotline visitors in danger. Crisis counselors usually obtain experience by practicing with experienced counselors who role-play hotline visitors. Role-playing has been shown to improve crisis intervention (Cross et al., 2011), but requires significant resources that can be challenging for hotlines to provide on tight budgets. For example, it requires experienced counselors' time that could instead be used to help hotline visitors in need.

To provide a realistic, low-risk practice environment to augment crisis hotline counselor training, we seek to develop Crisisbot: a dialogue agent that can simulate a hotline visitor and be used to train human counselors. A chatbot clearly cannot safely and thus ethically take on a counseling role, so we instead explore this direction of using a chatbot to support human counselors. In particular, we consider whether a chatbot can be used to simulate a

visitor to better train human counselors. By chatting with Crisisbot, novice counselors could safely practice counseling and build confidence before interacting with hotline visitors.

In addition to returning relevant messages, Crisisbot needs to simulate multiple visitor stories to begin to expose counselors to the astonishing breadth of scenarios that they may encounter. Multiple stories are also needed to support multiple practice sessions, and these visitor stories must be intimate conversations with realistically complex messages (as in Figure 1). However, collecting pre-defined personas in addition to example conversations burdens a limited pool of skilled workers, so we seek to develop stories by mimicking example conversations directly.

Developing personas for chatbots is challenging, as generic responses remain a shortfall of dialogue generation models (Li et al., 2016; Roller et al., 2020). Generic responses are especially problematic for characterizing distinct personas and for carrying engaging conversations. Without some specificity in responses, different scenarios cannot be simulated for multiple training sessions.

To simulate multiple, specific, and distinct personas, we propose a counselor strategy annotation scheme and a novel multi-task training framework that constructs persona-relevant responses from example transcripts rather than pre-defined personas. Our framework leverages counselor strategies, a pre-trained generative model, utterance segmentation, and sub-utterance selection. It allows for interleaving generated and prototype text to construct complex responses, which are more realistic.

We explore the utility of our strategy annotation scheme and multi-task framework in the context of hotline counseling. We use our scheme to annotate a corpus of realistic conversations and then use our framework to leverage these annotated conversations for response construction. We first test the trained model components through interactive chats with untrained crowdworkers to assess response quality and experience. We then test the framework with experienced hotline counselors to additionally assess how realistic and useful the system is perceived to be. The interactive conversations are evaluated with both subjective human judgements of user experience and automatically calculated measures of response diversity.

We find that our approach considerably increases the diversity of and the presence of persona-

relevant sub-utterances in chatbot responses. We also find discrepancy between crowdworker and counselor evaluation. This discrepancy stresses the importance of our two-fold evaluation approach and the general need for testing within a target setting, especially for specialized systems.

While additional improvements are needed for conversational coherence between turns, our annotation scheme and modeling approach contribute to curating chat experiences with richer responses that characterize distinct personas. We study this framework in the context of counseling, but it could be used more broadly in applications that require simulating multiple scenarios. We make anonymized conversations available for research use[1].

2 Related work

Our work builds on a growing interest in using natural language processing to address challenges in mental health care. Unlike previous work that has predominantly focused on studying the quality of care provided or delivering mental health care with chatbots, we focus on training hotline counselors with a Crisisbot. With this approach, we must address the persistent challenge of generic dialogue response generation (Li et al., 2016; Roller et al., 2020) to avoid tedious and unrealistic training experiences. We attempt to construct personas by mimicking personas from prototype conversations with a multi-task framework that enables constructing responses from pre-written and generated text.

2.1 NLP for hotline and therapy insights

Recent interest in hotline conversations has analyzed the structure of real conversation transcripts with the hope of better understanding what makes hotline conversations perceived to be helpful (Althoff et al., 2016; Zhang et al., 2019; Zhang and Danescu-Niculescu-Mizil, 2020). Similar work has considered talk therapy conversations and tried to understand conversations and predict counseling acts (Pérez-Rosas et al., 2017, 2019) or forecast therapists' next actions (Cao et al., 2019).

2.2 Chatbots for therapy care and training

An alternative line of work has followed the seminal Eliza system (Weizenbaum, 1966) and attempted to create dialogue agents to provide therapeutic benefit, i.e., "therapybots" (Fitzpatrick et al.,

---

[1]Please contact first author for conversations

2017; Inkster et al., 2018). Hotlines are too high-risk to responsibly replace human counselors with chatbots, so, for safety, we instead seek to support mental health professionals with improved training and practice. Our approach is most similar to work that explored building a chatbot to train talk therapists (Tanana et al., 2019), who have different methods than hotline counselors. This work did not explore enabling multiple personas, but trained a single general persona and found generic responses to be prevalent and tedious for crowdworkers.

2.3 Personality in chatbots

There has recently been a surge of interest in generating coherent and consistent dialogues grounded on pre-defined persona profile information from the PersonaChat dataset (Zhang et al., 2018; Dinan et al., 2020). Approaches to enforce consistent personas on this dataset have included retrieving relevant profile facts (Zhang et al., 2018), retrieving and refining relevant utterances (Weston et al., 2018), increasing the probability of copying a word from the profile (Yavuz et al., 2019), tuning to discourage inconsistent responses (Li et al., 2019a), re-ranking candidate responses (Welleck et al., 2019), and combining natural language inference with reinforcement learning (Song et al., 2019). Unfortunately, these methods fall short of generating responses that are as grammatical, diverse, engaging, and descriptive as natural human generated conversation (See et al., 2019; Roller et al., 2020). Further, pre-defined profile facts, which these works rely on, are not present or realistic for our application, which only has example conversations. To leverage the rich example transcripts for simulating personas, we propose a multi-task training framework.

2.4 Multi-task fine-tuning for dialogue response generation

Recent advances in large pre-trained models (Devlin et al., 2018; Radford et al., 2018) have led researchers to consider leveraging these models for dialogue response generation. A particular approach has been using multi-task training, such as for generating persona-relevant responses (Wolf et al., 2018) and adapting to domains (Li et al., 2019b). We build on this multi-task approach by using components that enable us to interleave pre-written and generated sub-utterances to construct descriptive and persona-consistent responses.

2.5 Retrieving and generating messages

Previous work has considered combining retrieval and generation methods for dialogue systems. The majority of this work has taken a retrieve and modify approach, e.g., use a retrieved prototype response to guide the generation of a full response for social chat (Cai et al., 2019; Weston et al., 2018; Wu et al., 2019) or tech support systems (Pandey et al., 2018). Additional work retrieved facts to generate knowledge grounded systems (Dinan et al., 2018). Most closely related to our work are studies that use rankers to choose between generated or retrieved responses in social chatbots (Song et al., 2016; Papaioannou et al., 2017a,b; Shalyminov et al., 2018) and in knowledge-grounded systems (Yang et al., 2019). Our work differs from these approaches, which used fully generated or fully retrieved messages, by interleaving both selected text and generated text into complex messages. Further, unlike other work, we do this to curate personas and as a component of a larger multi-task framework that combines counselor strategy detection, diverse generation, and a fine-tuned transformer model.

3 Dataset

To scaffold conversations and help the model identify similar contexts, we develop a counselor strategy annotation scheme. We use this scheme to annotate a dataset of realistic hotline training conversations that was previously collected with experienced hotline counselors (DeMasi et al., 2019). The counselors were asked to role-play hotline conversations, as during training, and to preserve the privacy of all hotline visitors. These conversations are representative of how counselors currently acquire experience during training. However, as they are synthetic, we are able to consider models that include repeating portions of text without violating the privacy of vulnerable individuals.

Conversations were required to be 40 total turns in length and allowed to be up to 60 total turns. Additional dataset statistics can be found in Table 1. The dataset was collected in multiple phases with unique conversations collected first and then additional examples built by paraphrasing this initial set. Here, we consider the conversations collected in all phases to initially fine-tune a pre-trained language model. Then, for the full multi-task framework, we consider an annotated subset of the unique conversations and not the paraphrased conversations.The

Dataset component	Count
Annotated conversations	160
Unannotated conversations	94
Paraphrased conversations	300
Counselor messages	8,319
Visitor messages	9,062

Table 1: Dataset statistics. For paraphrased conversations, counselors were asked to reword messages from other conversations to generate message variability with less burden.

Class	Counselor strategy	Example sub-utterance
Functional	greet	"Hello!"
	self-intro	"I'm a counselor."
Active listening	feelings	"It sounds like you feel overwhelmed."
	open-prompt	"Would you like to tell me more?"
Procedural	identify-resource	"Have you talked to anyone about this?"
	identify-risk	"Are you having thoughts of hurting yourself?"

Table 2: Example strategies from our scheme based on training materials. Each counselor sub-utterance was annotated with one of 25 strategies.

corpus contains extremely diverse scenarios that are usually characterized by relatively long messages with multiple sentences. The considerable diversity of scenarios makes our task challenging, but emphasizes the need for Crisisbot to simulate multiple personas.

3.1 Sub-utterance segmentation

Counselor and visitor messages were automatically segmented into sub-utterances, e.g., separate sentences, with a sentence tokenizer. Leading greetings, positive, and negative responses were also separated as sub-utterances for consistent strategy annotation. Segmentation and spelling errors were manually corrected during annotation. Splitting sub-utterances was necessary for annotation, as counselors are often encouraged to reply with multiple components, e.g., to acknowledge feelings before asking a question. For visitors, segmenting messages enables construction of more complex messages, e.g., generating a sub-utterance to respond to an unseen question and then selecting an informative sub-utterance from a prototype conversation to elaborate on the response.

3.2 Counselor strategy annotation

To understand input counselor messages and enable the model to look up similar contexts, we develop a counselor strategy annotation scheme and annotate the strategies of counselor sub-utterances in 160 transcripts. The counselor annotation scheme was based on conversations with professionals who train hotline counselors and training materials. It was iteratively refined and expanded by the researchers to cover sub-utterances observed. The annotation scheme consists of 25 annotations, including active listening techniques, de-escalation strategies, and social norms (examples in Table 2, additional details in the Appendix). This approach relates to previous work that annotated conversa-

tion strategies for persuasion (Wang et al., 2019), negotiation (He et al., 2018), and motivational interviewing (Cao et al., 2019; Pérez-Rosas et al., 2017), but these annotations are specialized for a hotline counseling context. However, they could be generalized to related applications, and used to extract related conversational exchanges to generate diverse response components. They could also be used to contrast with similar domains and understand the distinction of hotline counseling.

4 Method

To simulate realistic hotline visitors, we seek to construct complex responses that characterize specific, consistent personas across multiple turns. Towards persona consistency, we base each conversation on a prototype conversation from the corpus and try to mimic the visitor's persona in that conversation. As a result, we do not generate personas, but model conversations on individual training example conversations.To do this, we propose an end-to-end multi-task modeling framework that leverages our annotation scheme and a large pre-trained language model. This framework fine-tunes a generative language model while learning two additional model components: a strategy detector to select similar conversational exchanges, i.e., message-response pairs, and a next sub-utterance classifier that can interleave generated sub-utterances and sub-utterances selected from the prototype conversation into a coherent response. An overview of using these components to construct responses is in
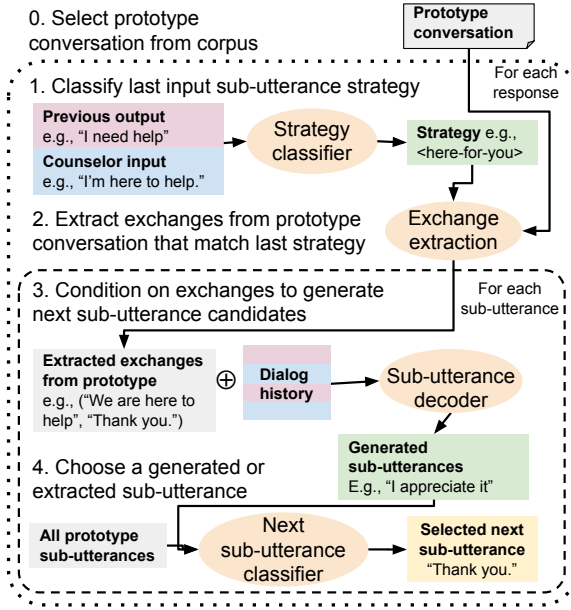
Figure 2: Overview of process to construct responses to messages input by a counselor. The framework bases each conversation on a single prototype conversation and learns three components to construct responses: a counselor strategy classifier to select relevant exchanges to condition the generator on, a generator that decodes diverse sub-utterances, and a classifier that selects to append either a generated sub-utterance or one selected from the prototype conversation.

Figure 2. We combine generated and selected prototype sub-utterances so that the model can respond with diverse, specific responses and appropriately respond to unseen contexts.

4.1 Detecting counselor strategies

To identify user intent, i.e., strategies used in a counselor's input message, we use a counselor strategy classifier. This classifier is trained on the counselor strategies that we annotated for each sub-utterance, as counselor messages typically contain multiple sub-utterances with different strategies (as shown in the introductory example (Figure 1). The counselor strategy classifier predicts the probability of a strategy s_t^i for the i-th sentence at turn t to be

$$p(s_t^i) = softmax\left(W_s \cdot \left[h_{t-1}^l; h_t^i\right]\right)$$

where h_{t-1}^l and h_t^i are hidden states from an underlying fine-tuned language model for the last sentence l of the visitor's message at the previous turn $t-1$ and at the end of the $i-$th sentence of the counselor's input message at turn t, respectively. The weights W_s are learned. At each training step,

this classifier results in a cross-entropy loss across strategies $\mathcal{L}_{strat}$.

4.2 Extracting conversational exchanges based on strategies

We use the last detected counselor strategy to identify similar, relevant conversational exchanges in the prototype conversation. Relevant exchanges are selected as counselor message and visitor response pairs where the last counselor sub-utterance strategy matches the last strategy in the counselor's current input message.

4.3 Generating diverse sub-utterances conditioned on exchanges

To generate sub-utterances characteristic of a visitor's persona, we condition generation from a fine-tuned language model on the selected conversational exchanges. The exchanges, and the speaker roles within them, are distinguished from the current context messages by pre-pending special tokens to each message. The examples and roles are further distinguished from current conversation roles by token embeddings, as in previous work distinguishing speaker roles (Wolf et al., 2018). This approach differs from imbuing conversation with factual knowledge (Dinan et al., 2018) or profile facts (Wolf et al., 2018; Li et al., 2019b) in that we condition on conversational exchanges and in how we select the exchanges.

4.4 Choosing a selected or generated sub-utterance

To add detail to responses, while maintaining persona consistency, we allow sub-utterances to be selected from the prototype conversation. We also allow generated sub-utterances, so that the model can appropriately respond to any input. For example, our system needs to generate a response to an unseen direct question, but then may elaborate with text selected from the prototype conversation.

We consider sub-utterances as individual units and train a next-sub-utterance classifier to select between generated sub-utterance candidates or selecting a sub-utterance from the prototype conversation. This classifier takes the form

$$p(y_t^i) = softmax\left(W_y \cdot h_t^i\right)$$

where y_t^i is the prediction of whether a candidate sub-utterances is the true $i-$th sub-utterance in turn t. Here h_t^i is the last hidden state of extending turn

t with an $i-$th candidate sub-utterance. This model component results in a loss $\mathcal{L}_{next}$.

To train this model component, a set of five examples is constructed for each sub-utterance in the ground truth response: the true next sub-utterance, three randomly selected sub-utterances, and one sub-utterance from the visitor's recent conversation history, i.e., everything the generator considers. All of these candidates are selected from the same prototype conversation and the inclusion of a recently said sub-utterance is used to decrease repetition.

4.5 Multi-task training

The above described model components are trained in a multi-task fashion by minimizing the composite loss function

$$\mathcal{L} = \mathcal{L}_{strat} + \mathcal{L}_{next} + \lambda\mathcal{L}_{LM}$$

where $\mathcal{L}_{LM}$ is the cross-entropy loss from the fine-tuned pre-trained language model and the other losses are as described above. The parameter λ controls the focus on the language model and is empirically set to 2.0 in all experiments. The language model that we use is first tuned for dialogue in a multi-task framework on the PersonaChat dataset (Wolf et al., 2018), then fine-tuned on the full corpus of transcripts collected, and finally tuned on our annotated subset of transcripts in the above framework.

5 Experimental setup

To evaluate the proposed full framework and ablated versions, we consider human judgements of technical conversation quality and both automatic measures and human judgements of engaging persona development. We collect these measures first with untrained crowdworkers. Then, based on crowdworker assessments, we select a subset of models to have counselors assess. Not all of the models are evaluated by counselors, as a limited number of experienced hotline counselors were able to be recruited and participate. Further, each model needs to be considered multiple times by each counselor in order for them to assess distinctness of personas between conversations. All human ratings used 5-point Likert scales. This study was approved by the internal review board.

5.1 Baseline models

We consider three models with subsets of the full model's components. The **Tuned** model is a pre-trained language model (Wolf et al., 2018) fine-tuned to the full 554 counseling conversations collected. The **+rank** model jointly trains the next sub-utterance classifier and fine-tunes the pre-trained language model to the subset of annotated transcripts. The sub-utterance classifier is the mechanism used for increasing response length and diversity by interleaving generated sub-utterances with those selected from the prototype conversation. The third model **+cond** trains the strategy classifier and fine-tunes the pre-trained language model dynamically conditioned on example exchanges from the prototype conversation. It uses no sub-utterance ranking. Finally, we consider the **Full** model that fine-tunes the language model with the strategy and next sub-utterance classifiers.

5.2 Training and interaction details

Our annotated dataset was randomly split with 90% of conversations in the training set and the rest used for validation. During testing, we randomly select a prototype conversation from the validation set at the start of the conversation and only consider selecting sub-utterances and conditioning on examples from that prototype conversation. This approach is intended to focus the conversation on a single visitor for persona consistency. To reduce repetition, we remove examples and sub-utterance from further consideration if a sub-utterance is copied from dynamically conditioning on prototype examples or is selected from the set of visitor's sub-utterances.

To encourage visitor coherence, we use five turns of conversation history for context. To accommodate limited context length, we consider the last sub-utterance of the counselor's messages, as this is typically where counselors will advance the conversation and require a response. We found in preliminary testing that this balance improved the visitor coherence and decreased repetition without significant detriment to responsiveness. Hyper-parameter settings are in the Appendix.

5.3 Automatic measures of response diversity

We calculate various metrics to compare the diversity of responses generated by each model. We consider the specificity of a models with the average negative log likelihood (nll.) of each response generated for its context with respect to the tuned model, so that it is comparable across models. We calculate the breadth of responses with the entropy (ent.) of a unigram language model of all tokens

3628

Model	Crowdworker judgements						Automatic measures of diversity				
	coh.	con.	flu.	det.	hum.	use.	nll.	ent.	d1	d2	#tok
Tuned	3.93	**3.93**	4.03	2.97	**3.87**	3.83	1.88	4.38	0.09	0.30	8.15
+rank	**4.00**	**3.93**	**4.07**	3.40	**3.87**	4.03	2.52	4.86	0.10	0.34	9.55
+cond	**4.00**	3.63	4.00	3.23	3.63	3.73	2.81	4.89	**0.13**	**0.39**	7.94
Full	3.93	**3.93**	3.97	**3.77**	**3.87**	**4.23**	**2.99**	**4.97**	0.12	**0.39**	**12.25**

Table 3: Crowdworker evaluation of models on 1-5 scales for <mark>coherency (coh.), consistency (con.), fluency (flu.), level of detail (det.), human likeness (hum.), and usefulness (use.).</mark> Higher is better for all evaluation metrics.

Model	Hotline counselor judgements								Automatic measures of diversity				
	coh.	con.	flu.	det.	hum.	use.	dis.	rea.	nll.	ent.	d1	d2	#tok
Tuned	**2.80**	**2.73**	**4.27**	**2.53**	**2.67**	**3.00**	**3.00**	**2.67**	1.61	4.19	0.18	0.46	8.31
Full	1.87	2.33	3.60	2.40	2.00	2.20	2.90	1.87	**3.00**	**4.87**	**0.19**	**0.50**	**14.93**

Table 4: Hotline counselor evaluation of models on 1-5 scales for <mark>coherency (coh.), consistency (con.), fluency (flu.), level of detail (det.), human likeness (hum.), usefulness (use.), distinctness (dis.), and realisticness (rea.).</mark> Higher is better for all evaluation metrics. Standard deviation for judgements of each dimension ranged 0.45-1.20.

generated and the <mark>diversity of response elements with d1 and d2</mark>, the ratios of the number of unique to number of total unigrams or bigrams generated, respectively (Li et al., 2016), <mark>Finally we consider the length of responses with the average number of tokens in a response (#tok).</mark>

5.4 Human judgements of quality

For conversation quality, we asked both crowdworker and counselor participants to rate Crisisbot on coherency, fluency and grammatical correctness of responses, and human likeness.

5.5 Human judgements of persona

We also sought to gauge the quality of persona simulation in conversations. To do this, we asked all participants to rate each conversation on persona consistency and level of response detail. We additionally asked counselor participants to rate each conversation on how realistic it was of a hotline visitor. After their second and third conversation with Crisisbot, we asked counselors to rate how distinct the persona was from previous conversations. Finally, at the end of each interaction (one conversation for crowdworkers and three conversations for counselors), we asked participants to rate how useful they thought such conversations would be for training. All ratings used 5-point Likert scales.

5.6 Crowdworker user study

Crowdworkers with US-based location qualification who chatted with Crisisbot were required to write at least 10 turns before rating their experience.

The interface they interacted with included suggestions of things a counselor might say and it was suggested that they explore the problems facing Crisisbot. A persona, i.e., prototype conversation, was randomly selected for each conversation.

5.7 Hotline counselor user study

After crowdworkers' preliminary evaluation, we asked experienced hotline counselors to chat with Crisisbot. Counselors were informed of the purpose of the study and asked to interact with Crisisbot for three conversations in two different settings, i.e., three conversations with each of two models. In each setting a counselor interacted with a single model, but three personas from the model and the personas were fixed between models. As with crowdworkers, counselors were required to write at least 10 messages per conversation. A survey followed each conversation. Unlike in the crowdworker interface, the counselor interface did not include examples of things to say.

6 Evaluation Results

Each of the four models was interactively evaluated by <mark>30 crowdworkers</mark> and we calculated measures of model response diversity on all responses generated (Table 3). Based on the initial crowdworker evaluation, two models were selected for counselor testing. Only two models were tested by counselors because testing each model took significant participant commitment, as counselors had to chat with each model multiple times to consider if the model simulated distinct personas. Due to our ability to

Turn	Hotline counselor	Crisisbot
0	*START*	*hi, i'm michael and i kind of need somebody to talk to today. i'm feeling really down.*
1	*hi michael, sorry to hear you are feeling down, but i am so glad you reached out. what's up?*	*i'm having suicidal thoughts lately. my life sucks.*
2	*sounds like things have been really rough, so much that you are thinking about suicide. do you feel like killing yourself today?*	*sometimes i think about it.*
3	*thank you for sharing with me. have you done anything to harm yourself today?*	*no. i'm feeling a bit better now. but i just want to talk to someone. and i don't know how to help my mom because it hurts me to see that she's in so much distress.*
4	*okay, i am definitely here to talk. i hear that you want to help your mom, and that you are worried about her. what is going on with your mom?*	*she's been really depressed lately. my parents want to divorce and i'm so stressed that i can't focus on school.*

Table 5: Excerpt from a selected interaction of a counselor with Crisisbot with the Full model. The entire transcript continues similarly (see Appendix) and highlights high counselor expectations despite reasonable model performance (judged: coh.=1, rea.=3, con.=1, flu.=4, det.=3, hum.=1).

only access a limited pool of experienced counselors, a total of five counselors interacted with the two selected models, resulting in the evaluation of 15 conversations for each model (Table 4).

6.1 Crowdworker evaluation

Crowdworkers generally positively reviewed all model variants. While the Full model had some slightly lower quality measures (e.g. coherence and fluency), it was rated to have the most level of detail and hypothesized utility for training, which coincided with increased response diversity across all measures of diversity (Table 3). This result indicates that crowdworkers had a preference in their user experience for the utility of diverse responses to train different scenarios and could accept minimal impact to other quality measures.

6.2 Hotline counselor evaluation

Full model responses were more diverse across all measures when responding to counselor messages than the Tuned model (Table 4). However, in ratings, counselors seemed to prefer the coherence of vague model responses. They more favorably scored conversations with the Full model, even though the responses were less diverse. The increased scores were likely because the Full model's short responses were perceived as more coherent and consistent.

In addition to ratings, we asked counselors for open feedback about what they liked and disliked about the conversations, and their responses to these questions revealed different conclusions than their ratings. In open responses, counselors did value the variety of the Full model and reported frustration with dull responses from the Tuned model. An exhaustive list of open responses is in Table 6. In open feedback, all of the counselors cited variability in responses and specificity of issues as what they liked about the Full model. When asked what they disliked, themes emerged about counselors feeling that trying to get more information from the Tuned model was difficult and the simulated problem scenarios did not feel real. In contrast, none of the counselors reported disliking anything related to specificity with the Full model, but some issues arose with responsiveness and consistency across turns, which is an ongoing challenge for dialogue systems (Roller et al., 2020).

6.3 Qualitative review of conversations

Reviewing the interactive conversations reveals considerable difference between counselor and crowdworker conversation approaches, enormous variability in counselor ratings, and inconsistency in how counselor ratings relate to researcher perceptions of individual conversation quality. Examples of conversations can be found in the Appendix. Counselors were generally more engaged than crowdworkers, asked followup questions, and, as they are trained to do, tried to explore issues confronting visitors. Conversely, crowdworker messages would be more focused within individual turns and freely hop between topics. This unsurprising difference emphasizes the need for a specialized chatbot to train counselors how to followup.

As the average quantitative ratings reflect in Tables 3 and 4, even relatively coherent conversations that counselors had with Crisisbot received considerably lower scores than a comparable conversation with a crowdworker. These deflated counselor rat-

Model	Counselor liked
Tuned	depression is simulated well, but the train of thought doesn't last long
Tuned	They are good practice for thinking on your toes and receiving answers to your messaging, so you can work on various approaches depending on the situations. Also, there was a range of suicidality and crisis level.
Tuned	First two were generally more coherent than in my previous session. The chatbot seems to respond well to messages showing validation and empathy.
Tuned	They were a good opportunity to work with ambivalence and practice some text-based OARS
Tuned	Content seemed pretty realistic but still not quite coherent enough to be realistic
Full	More crises situations thrown in. Variety from previous submission.
Full	There was a variety of responses that prompted critical thinking and creativity.
Full	i think the statements about depression and past suicidal ideation and attempts are useful for training counselors. likewise the mention of bullying.
Full	Variety of issues presented

Model	Counselor disliked
Tuned	The third chatbot tried to end the conversation early and kept saying thanks, goodbye, etc. but I had to keep the conversation going for testing - not sure if this was an error.
Tuned	Trying to get more information was very difficult when answers were short and didn't really make sense
Tuned	i did not feel like i was talking to someone with a real life or real problems
Tuned	The responses were not connected to the unfolding narrative of the conversation, so it does not replicate the rapport building process that is needed in this work.
Full	When they would mimic my name choice, took me out of it, although I suppose it does happen occasionally.
Full	There were many inconsistencies in the conversation and the flow was not there. It felt like responding to one individual statement at a time more than a series of statements that went together.
Full	i think the bot has to be more flexible in responding to my input–often the topic got changed altogether
Full	It would help if the personas and storylines were more defined and consistent.

Table 6: All open feedback provided by counselors about what they liked and disliked about their interactions.

ings could relate to their increased experience interacting with humans over chat and their training to ask follow up questions that help individuals dive deeper into conversation, but are more challenging for a bot to respond to. Similarly, it could relate to crowdworkers' increased experience with chatbots and thus lower expectations. Regardless, dialogue systems are notoriously difficult to evaluate and it is also possible that our participants struggled to separate potential improvements on fine-grained dimensions from overall chatbot performance, which resulted in different trends between the evaluations.

7 Conclusion

We explored developing a Crisisbot to imitate hotline visitors' varied personal stories for better training human counselors. Towards this goal, we developed a counselor strategy annotation scheme and proposed a multi-task framework. This framework bases conversations on personas from prototype conversations and interleaves generated text with text selected from the prototype. Through two user studies, we observed increased response diversity overall, but a considerable difference between crowdworker and counselor perceptions and inconsistency with how ratings reflect conversations. These results reinforce the known challenge of dialogue system evaluation and cautions the use of crowdworkers alone for developing system insights. The disparity in ratings between counselors and crowdworkers stresses that the involvement of specialized, intended users in system development and evaluation is crucial for success. While perhaps intuitive, this result is sometimes overlooked during system evaluation. Counselor open feedback indicates that additional development is required for coherence across turns, but that our framework adds detail and variety in personas, which could enable better variety for training counselors.

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.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *Computational linguistics-Association for Computational Linguistics*.

Wendi F Cross, David Seaburn, Danette Gibbs, Karen Schmeelk-Cone, Ann Marie White, and Eric D Caine. 2011. Does practice make perfect? a randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills. *The journal of primary prevention*, 32(3-4):195.

Orianna DeMasi, Marti A Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Madelyn S Gould, Wendi Cross, Anthony R Pisani, Jimmie Lou Munfakh, and Marjorie Kleinman. 2013. Impact of applied suicide intervention skills training on the national suicide prevention lifeline. *Suicide and Life-Threatening Behavior*, 43(6):676–691.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Kerry L Knox, Janet Kemp, Richard McKeon, and Ira R Katz. 2012. Implementation and early utilization of a suicide hotline for veterans. *American journal of public health*, 102(S1):S29–S32.

Christa D Labouliere, Barbara Stanley, Alison M Lake, and Madelyn S Gould. 2020. Safety planning on crisis lines: feasibility, acceptability, and perceived helpfulness of a brief intervention to mitigate future suicide risk. *Suicide and Life-Threatening Behavior*, 50(1):29–41.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019a. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *arXiv preprint arXiv:1911.03860*.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2019b. End-to-end trainable non-collaborative dialog system. *arXiv preprint arXiv:1911.10742*.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar encoder-decoder for neural conversation generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1329–1338.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon. 2017a. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. *Alexa Prize Proceedings*.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. 2017b. An ensemble model with ranking for social dialogue. *arXiv preprint arXiv:1712.07558*.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in

counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723.

Igor Shalyminov, Ondřej Dušek, and Oliver Lemon. 2018. Neural response ranking for social conversation: A data-efficient approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5190–5196. AAAI Press.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research*, 21(7):e12529.

Vibrant Emotional Health. 2019. California callers in 2018 to the national suicide prevention lifeline.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. Transfertransfo: A transfer learning approach for neural network based conversational agents. *NeurIPS CAI Workshop*.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1341–1350.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. *arXiv preprint arXiv:2005.04245*.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

3633

Strategy class	Strategy
Functional	greet
	self-intro
	close-conversation
	ask-name
	thanks
Procedural	identify-coping-plan
	identify-resource
	intent-to-self-harm
	identify-self-harm-plan
	access-to-self-harm-plan
	when-self-harm
Active Listening	affirmation
	caring
	shared-understanding
	feelings
	feelings-and-summary
	summary
	looking-for-info-answer
	open-prompt
	reframing
	validation-of-feelings
	vindication
	here-for-you
Other	suggest/advise
	other

Table 7: Counselor sub-utterance strategy annotations.

A Appendices

We present auxiliary materials for the appreciation of our work, including additional detail on our counselor strategy annotation scheme, hyperparameter settings, information about evaluation scales, example model output from interactive testing, and open counselor feedback.

B Counselor strategies

Each counselor sub-utterance was labeled with a conversation strategy, which can be grouped into classes. A full list of the strategies annotated is in Table 7.

C Training details

The framework was trained with 4 gradient accumulation steps, a batch size of 1, the Adam optimizer with a learning rate decay of .75 for 4-5 epochs, based on minimum validation loss. For decoding, The max decoding length was set to 50 tokens and tokens were decoded with temperature 0.7 with nu-

cleus sampling and with probability mass 0.9. For selecting sub-utterances, 10 candidates were generated to consider. To avoid the model only learning to copy from selected exchanges, we condition the language model on exchanges for a randomly selected counselor strategy for 20% of messages during training.

D Human judgment scales

The scales used by crowdworkers and counselor participants were labeled as follows.

- How coherent was the conversation? 1 (didn't make sense) – 5 (could have been a human)

- How realistic was the conversation as a simulation of a hotline visitor? 1 (it wasn't anything like chatting with a visitor) – 5 (very realistic and representative)

- How consistent was the persona? 1 (many contradictory statements) – 5 (consistent and plausible person)

- How fluent and grammatical were the responses? 1 (constant mistakes, poor english) – 5 (no mistakes)

- How detailed were the chatbot's responses? 1 (only vague answers without any detail) – 5 (lots of details)

- Overall, how human-like was the chatbot? 1 (not at all) – 5 (could have been a human)

- If you were training to be a counselor, how useful would these conversations have been for your training? 1 (not at all useful) – 5 (very useful)

E Example conversations

Example conversations between crowdworkers and counselors can be found in Tables 8 -11.

F Selected successful transcript from counselor

The full conversation of the example of a counselor interacting with the Full model that was excerpted in the main text is in Table 12. This conversation highlights low counselor feedback in spite of relatively coherent responses.

Turn	Crowdworker	Crisisbot
0	*START*	*hi, helpline?*
1	*hi, what can i do for you?*	*i don't really know.*
2	*is something wrong? how are you feeling?*	*a little better. it's really hard to breathe. i'm just so overwhelmed.*
3	*that sounds hard. do you want to tell me more?*	*i just feel overwhelmed.*
4	*ok, so you're feeling overwhelmed. that happens to a lot of people. have you tried talking to someone?*	*no. i don't think i want to talk to anyone.*
5	*i understand, talking about problems can be difficult. no pressure. have you tried doing something to relax?*	*no, not really.*
6	*do you have any hobbies you enjoy?*	*not really.*
7	*what do you think could help you relax a bit?*	*i don't really know.*
8	*ok, would you like me to give you some self-help materials?*	*i'll talk to my friend. ok.*
9	*having a close friend to talk to in times like this is great. that should help you!*	*i will. thank you for your help.*
10	*ok, thanks for chatting with me! i hope everything goes well.*	*i will. thank you so much for your help.*

Table 8: Model: Tuned. Judgements: coh.=3, con.=4, flu.=4, det.=2, hum.=3, use.=2

Turn	Crowdworker	Crisisbot
0	*START*	*hey*
1	*hai*	*we all just started at a new middle school and everything is different*
2	*in what way it is different*	*well i got in a fight with a woman at school*
3	*why the fight was happens?*	*because of the fight, i'm having a really hard time at school*
4	*oh my got, leave the problem,*	*that sounds like it could be fun*
5	*can you identify the reason for the fight happens?*	*well last year i had two best friends*
6	*then*	*they got in a fight and it was really upset*
7	*so*	*exactly, so they started fighting*
8	*how you get caught in that*	*so they started fighting about me?*
9	*ohhh*	*exactly, they're such fascists*
10	*exactly*	*i feel like i heard something about that in the announcements the other day, it was the same guy from school, he was supposed to be my friend*
11	*leave this thing, take a long breath, enjoy a favorite music, take a healthy food, go to sleep, and calm down your self, you will be alrite.*	*ok thanks!*
12	*u welcome*	*bye*
13	*bye tc*	*bye*

Table 9: Model: Full. Judgements: coh.=5, con.=4, flu.=5, det.=5, hum.=5, use.=5

Turn	Counselor	Crisisbot
0	*START*	*hi, i need help*
1	*hi, i'm here to help you. my name is bob.*	*i'm michael.*
2	*hi michael? what do you need help with?*	*i don't know.*
3	*things are a bit overwhelming right now? is that right?*	*yeah, i have to go now.*
4	*michael, i'm here if you need me. here to talk, here to hear you*	*thanks, i appreciate it.*
5	*no problem at all! we're here anytime*	*thanks, i appreciate your help.*
6	*it can be hard to reach out for help, even texting this line can be scary*	*yeah, it is.*
7	*it's scary texting, and yet you're doing it! that's amazig*	*yeah, i just really need to talk to someone about it.*
8	*what is it that you want to talk about, michael?*	*i don't really know. i just don't want to talk to anyone about it.*
9	*i'm hearing how you want to talk about it and at the same time, you don't. it's kind of this massive thing, is that right?*	*yeah, it's a lot of my family stuff. i'm not sure how i'll feel to move on.*
10	*mmmm, yeah, i hear how you're really wrestling with uncertainty*	*yeah, it's like my whole life is so much more comfortable to deal with.*

Table 10: Model: Tuned. Judgements: coh.=4, rea.=3, con.=2, flu.=4, det.=2, hum.=3

Turn	Counselor	Crisisbot
0	*START*	*i don't want to be here anymore*
1	*i'm here to listen to you. my name is bob.*	*i'm christopher.*
2	*hi christopher. you said, you don't want to be here anymore. are you having thoughts of killing yourself or harming yourself?*	*yes*
3	*thanks for letting me know about that christopher. do you have a plan for how you would do that?*	*od*
4	*overdose on what? pills?*	*no its ok*
5	*christopher, i'm here. do you currently have the drugs you would use to overdose with you?*	*yeah. they're in my room*
6	*is there anyone else in the house with you?*	*i don't want to bother anyone. i have to go thanks*
7	*christopher, it's no bother. if you can just put the drugs in another room or give them to someone you trust, i'm here to listen*	*ok, thanks.*
8	*it's not a problem at all– please check in with me*	*ok, bye*
9	*goodnight*	*ok*
10	*again, we are available 24/7 on our call line*	*ok*

Table 11: Model: Full. Judgements: coh.=4, rea.=2, con.=2, flu.=2, det.=3, hum.=3

Turn	Counselor	Crisisbot
0	*START*	*hi, i'm michael and i kind of need somebody to talk to today. i'm feeling really down.*
1	*hi michael, sorry to hear you are feeling down, but i am so glad you reached out. what's up?*	*i'm having suicidal thoughts lately. my life sucks.*
2	*sounds like things have been really rough, so much that you are thinking about suicide. do you feel like killing yourself today?*	*sometimes i think about it.*
3	*thank you for sharing with me. have you done anything to harm yourself today?*	*no. i'm feeling a bit better now. but i just want to talk to someone. and i don't know how to help my mom because it hurts me to see that she's in so much distress.*
4	*okay, i am definitely here to talk. i hear that you want to help your mom, and that you are worried about her. what is going on with your mom?*	*she's been really depressed lately. my parents want to divorce and i'm so stressed that i can't focus on school.*
5	*i'm sorry to hear that. divorce is very stressful for the whole family, and it makes sense that you are stressed and can't focus.*	*yeah, i can't even face my own dad right now. he's always pressuring me.*
6	*wow, on top of that, you are feeling pressured by your dad and it sounds like that is a difficult relationship, too. has anything felt helpful to you when you have felt stressed in the past?*	*sometimes i get really overwhelmed by stuff, and i can't focus on school. i feel like i'm a burden and i don't know how to deal with school.*
7	*i hear that life feels very overwhelming. no wonder you can't focus on school! has there been a time in the past when you were overwhelmed, but things got better?*	*hopefully, i'll be able to set an appointment with a therapist soon and also see my girlfriend in the meantime though.*
8	*yes, i hear that talking to your therapist is something that helps you, and also spending time with your girlfriend. are you able to reach out to either of them today?*	*i'm not ready to think about that yet. but i do miss her and she's the only person who's ever really supported me emotionally.*
9	*i see, maybe it feels like you are not ready for that kind of social contact yet? is there something that you can do on your own that helps you feel more relaxed?*	*okay, i'll keep that in mind next time i check in. thank you. bye*
10	*okay, thank you so much for reaching out today. we are always here to talk. how are you feeling after our chat today?*	*not great, but i'm okay. i've been really stressed lately.*

Table 12: Model: Full. Judgements: sim.=1, coh.=1, rea.=3, con.=1, flu.=4, det.=3, hum.=1, use.=3