Towards In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations

Zixiu Wu
Philips Research & University of Cagliari
zixiu.wu@philips.com

Simone Balloccu
University of Aberdeen
s.balloccu.19@abdn.ac.uk

Rim Helaoui
Philips Research
rim.helaoui@philips.com

Diego Reforgiato Recupero and **Daniele Riboni**
University of Cagliari
{diego.reforgiato, riboni}@unica.it

Abstract

Reflection is an essential counselling strategy, where the therapist listens actively and responds with their own interpretation of the client's words. Recent work leveraged pretrained language models (PLMs) to approach reflection generation as a promising tool to aid counsellor training. However, those studies used limited dialogue context for modelling and simplistic error analysis for human evaluation. In this work, we take the first step towards addressing those limitations. First, we fine-tune PLMs on longer dialogue contexts for reflection generation. Then, we collect free-text error descriptions from non-experts about generated reflections, identify common patterns among them, and accordingly establish discrete error categories using thematic analysis. Based on this scheme, we plan for future work a mass non-expert error annotation phase for generated reflections followed by an expert-based validation phase, namely "whether a coherent and consistent response is a good reflection".

Context		
Utt.	**Role**	**Text**
u_{t-3}	**Client**	The baby was up all night and I'm exhausted.
u_{t-2}	**Therapist**	So, what you're saying is you've had a rough night?
u_{t-1}	**Client**	Yes. She was up every three hours to eat, I don't understand it.
Response (Reflection)		
u_t	**Therapist**	So, she needed to eat every three hours last night and that was really frustrating for you?

Table 1: A 3-turn context and the ground-truth reflection from an MI dialogue.

1 Introduction

Patient health can be greatly improved by changing behaviours such as smoking and alcohol consumption. As patients rarely ask for help with it, healthcare practitioners often need to encourage, counsel and advise them to make changes (Rollnick et al., 2008). An effective counselling approach for this purpose is motivational interviewing (MI, Miller and Rollnick, 2012), which aims to elicit the motivation for change from the client[1] themselves.

In particular, *reflection* — also known as *reflective listening* — is an essential conversational strategy in MI that has been shown to be related to positive counselling outcomes (Moyers et al., 2009). A good reflection conveys to the client that the therapist is listening, hearing and understanding them by reflecting back a short summary of how the therapist understands what the client has said (Rollnick et al., 2008), as shown in Table 1.

Reflection is a crucial skill for counsellors (Braillon and Taiebi, 2020), but its training is time-consuming and reliant on human supervision (Rautalinko and Lisper, 2004; Rautalinko et al., 2007). Therefore, an automatic assistant that offers reflection examples given a particular dialogue context can speed up the process while relieving the burden of supervision. Indeed, recent years have seen studies (Shen et al., 2020, 2022) on reflection generation that fine-tune pretrained language models (PLMs) to produce a reflection given some preceding utterances as the context.

Despite the progress in reflection generation, its evaluation remains a challenge. Automated metrics in language generation tasks are often not robust (Liu et al., 2016) and human evaluation is thus necessitated. Moreover, reflection requires specialised knowledge and counselling is complex and delicate. Ideally, therefore, generated reflections need evaluation by experienced therapists. However, expert annotation is time-consuming and costly (Moyers et al., 2005). Thus, human evaluation in previous work suffers from issues including

[1] A person receiving MI is not necessarily a patient, therefore we use "client" instead of "patient" in this work.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 116 - 124
December 7, 2022 ©2022 Association for Computational Linguistics

simplistic evaluation scheme (e.g., good vs. bad) and small (≤ 50) number of annotated reflections.

Another significant but underexplored weakness is the lack of context. In prior work, dialogue models are given as the input context only a few (≤ 5) preceding utterances. This can be inadequate for models to produce context-aware responses and for human evaluators to provide context-informed assessment, considering that 1) therapy dialogues are relatively long — often between 10 and 120 minutes (Rubak et al., 2005) — and 2) spoken-dialogue utterances are typically short, unlike in written conversations. In particular, sufficient context is important for assessing if a generated text contains hallucination (Ishii et al., 2022), a well-known issue of neural natural language generation where the output is unfaithful/ungrounded w.r.t. the input, for example when a chatbot contradicts what it said previously during a chat with the user (Vinyals and Le, 2015).

To alleviate the time and resource requirement for human evaluation, we advocate for disentangling the human evaluation into two phases: 1) **by non-experts[2]:** whether a generated reflection is coherent and consistent w.r.t. its context and what the issue of an incoherent/inconsistent reflection is; 2) **by experts:** whether a coherent and consistent reflection is a good reflection that conforms to therapy guidelines. We argue that a non-expert is perfectly capable as an evaluator for the first phase, and that this setup saves time and resources as a whole, especially in the second phase. **In this work, we conduct an initial study for the non-expert annotation phase.**

We use longer contexts — 14 turns on average — to better ground reflection generation and human evaluation. We devise a non-expert annotation scheme by 1) collecting free-text error descriptions w.r.t. generated reflections from non-experts and 2) identifying common patterns in the error descriptions and summarising them into discrete categories using thematic analysis (Braun and Clarke, 2012), similar to recent work (e.g., Thomson and Reiter, 2020) adopting bottom-up designs of text error annotation schemes. Thus, we establish {**Malformed**, **Off-topic**, **Dialogue-contradicting**, **Parroting**, **On-topic but unverifiable**} as the error categories, and a generated reflection may suffer from one or more categories of error. Most of these categories require a deeper understanding of the dialogue context but the latter three have not been explicitly included in previous studies on reflection generation.

Based on these error categories, we plan for future work a mass non-expert error annotation phase for generated reflections followed by an expert-based validation phase, namely "whether a coherent and consistent response is a good reflection".

2 Related Work

2.1 Reflection Generation

PLM-based empathetic dialogue generation (EDG, e.g., Rashkin et al., 2019) has seen considerable development in recent years. Of particular interest to us is EDG in counselling, which has taken the form of reflection generation so far. In particular, Shen et al. (2020) build a reflection generator that leverages responses from similar conversations as auxiliary input, while Shen et al. (2022) utilise domain and commonsense knowledge, both studies using only 5 preceding utterances as the context. Ahmed (2022) probes few-shot reflection generation for individual patient statements instead of multi-turn dialogues. Compared to those works, ours differs in its use of long dialogue contexts (14 turns on average) for the generator to enable more context-aware reflections.

2.2 Human Evaluation of Empathetic Dialogue Generation

The standard EDG human evaluation assesses the dialogue-relevance, fluency and empathy[3] (Rashkin et al., 2019; Li et al., 2020b) of a response on a Likert scale. A/B testing has also been used to compare responses from different models (e.g., Xie and Pu, 2021; Kim et al., 2021).

For evaluating reflection generation, Shen et al. (2020, 2022) tweak the standard EDG human evaluation slightly by replacing "empathy" with "reflection-likeness" in {dialogue-relevance, fluency, empathy} to gauge if the response interprets what the client means. Those human evaluation setups are small-scale, with less than 50 sampled reflections per model. On the other hand, 369 responses generated by the patient-statement-based reflection models in Ahmed (2022) are evaluated

[2] "experts" refers to people well-versed in psychology/psychotherapy and "non-experts" refers to the opposite.

[3] Usually rephrased, e.g., "did the responses show understanding of the feelings of the person talking about their experience?" (Rashkin et al., 2019)

Label	Reflection	Question	Input	Other
Prop.	28%	28%	11%	33%

Table 2: Proportion of therapist utterances of each label in high-quality AnnoMI dialogues.

by experts in a good-vs-bad binary setup. In comparison, our human evaluation is novel in its explicit focus on long-context-based error analysis of generated reflections.

One issue not explicitly addressed in EDG human evaluation so far is hallucination, where the output is unfaithful/ungrounded w.r.t. the input. While “off-topic-ness” is roughly equivalent to “dialogue (ir)relevance”, it is only one type of hallucination. Ishii et al. (2022) define a response to be “intrinsic hallucination” if it contradicts the input(e.g., Dziri et al., 2019; Li et al., 2020a) and “extrinsic hallucination” if it cannot be verified based on the input (e.g., Mielke et al., 2022; Roller et al., 2021). Therefore, a hallucinating reflection can be on-topic but contradict the context (intrinsic) or be unverifiable based on the context (extrinsic). Since reflective listening is based entirely on the context, we argue that a hallucinating reflection can cause quick client disengagement, since it is very likely unnatural in the conversation context. Therefore, we take hallucination into consideration explicitly, in contrast to prior work.

3 Modelling of Reflection Generator

3.1 Counselling Dialogue Data: AnnoMI

We utilise AnnoMI (Wu et al., 2022), a corpus of expert-annotated MI counselling sessions. AnnoMI contains both “good” (high-quality) and “bad” (low-quality) examples of MI. Aiming at generating **good** reflections, we leverage the 110 conversations (8839 utterances) of high-quality MI.

Each therapist utterance in AnnoMI is annotated by MI experts as **Reflection**, **Question**, **Input**, or **Other**. Specifically, **Reflection** is reflective listening, **Question** means an open/closed question, **Input** encompasses providing information and suggestions, etc., while **Other** is the default and mostly covers conversation facilitators like “Uh-huh”. The utterances label distribution is shown in Table 2.

3.2 Model Input Format

We train similarly sized `gpt2-medium` (Radford et al., 2019, 355M parameters) and `bart-large` (Lewis et al., 2020, 406M parameters) as reflection generators. Like most open-domain dialogue models, our models generate a response (therapist reflection) based on an N-turn dialogue history (namely the context), where the last turn comes from the client. An illustrative 3-turn context and its ground-truth reflection are shown in Table 1. Pre-trained dialogue models like DialoGPT (Zhang et al., 2020) are not used because they are mostly pre-trained on written conversations with only a few turns as the context, whereas therapy dialogues are spoken and long, causing a large domain gap.

As the volume of AnnoMI reflections is relatively small, we also train the models to generate other types of therapist responses using ground-truth utterance labels as plain-text conditioning codes, inspired by recent work (e.g., Rashkin et al., 2021) of similar approaches. Specifically, we construct the input as a sequence of context utterances with interlocutor labels and utterance separators, appended by the ground-truth therapist response label. For example, the context in Table 1 would become[4]:

> “$\langle client \rangle$*The baby was up all night and I’m exhausted.*$|\langle therapist \rangle$*So, what you’re saying is you’ve had a rough night?*$|\langle client \rangle$*Yes. She was up every three hours to eat, I don’t understand it.*$|\langle therapist \rangle$~$\langle listening \rangle$”

while the ground-truth response is simply

> *“So, she needed to eat every three hours last night and that was really frustrating for you?”*

The underlying assumption is that this will enable more training data for the language modelling of therapy dialogue while better shaping the boundaries of reflections in the latent semantic space.

Thus, a training/validation/test example is simply a $\langle context, response \rangle$ pair representing the $\langle input, output \rangle$. Each context is left-truncated to the most recent 384 tokens to preserve the most recent dialogue turns[5], while each ground-truth response is right-truncated to 128 tokens.

3.3 Training Response Generator

For both GPT-2 and BART we adopt 10-fold cross validation (CV) for training, in order to obtain

[4] In practice, we use “$\langle$asking$\rangle$”, “$\langle$informing$\rangle$”, “$\langle$listening$\rangle$”, “$\langle$other$\rangle$” as the plain-text control codes for Question, Input, Reflection and Other, respectively.

[5] 384 tokens make up N turns where N vares depending on the individual utterance lengths, but on average $N = 14$.

Model	GPT-2	BART
Perplexity	17.36	13.29

Table 3: Perplexity of each reflection generator under cross validation.

a test-time generated response for each example in the dataset (See §3.4). As noted in §3.2, we train a **generic** response generator that can produce any type (namely **Reflection**, **Question**, **Input** or **Other**) of therapist response. The examples of each fold are ensured to be from different dialogues, thus maximising mutual exclusivity between the training (8 folds), validation (1 fold) and test (1 fold) data for each of the 10 CV models. Also, the CV is stratified so that the distribution of ground-truth response types in each fold is the same.

To gauge the performance of the response generators on generating **reflections**, we evaluate them only on the test-fold examples where the ground-truth response is a reflection. Following most recent studies (Thoppilan et al., 2022, Shuster et al., 2022, *inter alia*) on response generation, we report in Table 3 the perplexity (the lower the better) of each model, which quantifies how uncertain a model is about generating the ground-truth reflections in the test data. We do not compare these numbers with other studies because 1) achieving state-of-the-art is not our focus, 2) the dataset and task are unique and have no comparable state-of-the-art, and 3) to the best of our knowledge, there is no study on the utility of perplexity as a metric for reflection generation or counselling dialogue modelling. We also experimented with paraphrasing-based data augmentation, with no significant improvement gained.

3.4 Test-Time Reflection Generation

Once the models are trained, we use them to generate alternative reflections for the context of each ground-truth reflection in AnnoMI, by conditioning the output using the $\langle listening \rangle$ code as before.

Following recent work (e.g., Santhanam et al., 2021) on hallucination in dialogue generation, we experiment with a range of decoding strategies, in order to capture a broad spectrum of potential errors in model-generated reflections. For both GPT-2 and BART, we explore

- Greedy decoding
- 5-Beam decoding, using all of the 5 decoded sequences at the final time step
- Nucleus decoding (Holtzman et al., 2019), $p \in \{0.4, 0.6, 0.8, 0.95\}$, 5 sequences sampled for each p

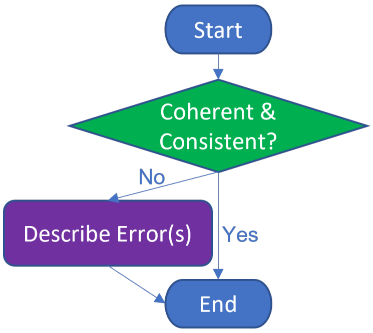


Figure 1: Annotation Flow

4 Human Annotation

As the underlying assumption of our human annotation is that incoherence/inconsistency errors can be spotted by non-experts, we survey laypeople for their own descriptions of reflection errors and then summarise those free-text descriptions into categories.

Annotation Materials We sample 3 contexts from 3 different dialogues and use their respective ground-truth and model-generated reflections for annotation. Based on the responses generated for the 3 contexts, we randomly sample a subset of 60 for human annotation.

Annotators We recruited 6 volunteers with high proficiency in English and no prior experience in NLP or psychology/psychotherapy. Each annotator worked on the same batch of 60 reflections for the aforementioned 3 contexts in total.

Annotation Procedure The procedure is illustrated in Figure 1, and the annotation interface is presented in Appendix A. The annotators are shown each $\langle context, reflection \rangle$ pair and first need to answer whether the reflection feels coherent and consistent given the context. If they choose "No", they are asked to describe the incoherence/inconsistency-causing error(s) of the reflection, otherwise they will proceed to the next example. We note that we do not define "incoherent" or "inconsistent" and instead leave it to the discretion of the annotators, in order to gather more natural insights on response errors. For the same reason, we use the word “response candidate” instead of the more complex term “reflection”, and we do not mention that some response candidates came from models instead of humans.

Inter-Annotator Agreement The inter-annotator agreement (IAA) on the “Coherent & Consistent?” question is 0.37 in terms of Fleiss' kappa (Fleiss, 1971), which is in the “fair agreement” range (0.2-0.4) but close to the “moderate agreement” threshold of 0.4. We attribute the relatively low IAA to two factors: 1) We purposely did not provide a strict definition of “coherence” or “consistency” to the annotators, which led some of them to consider issues like “intimidating tone” as causes for incoherence/inconsistency, but those are actually reserved for the expert-phase, since therapy experts should be the ones to judge whether a response is appropriate in a counselling setting. 2) 6 annotators are involved in the annotation process rather than just 2 to 3 as is commonly done for human evaluation of generated reflections (Shen et al., 2020, 2022), and it is usually less likely to get higher agreement with more raters.

Established Error Categories We use thematic analysis (Braun and Clarke, 2012) to manually and systematically identify common patterns in the annotators' feedback and summarise them into the following error categories:

- **Malformed**: a response that “feels broken” because 1) it has unclear references, 2) it is incomprehensibly ungrammatical, and/or 3) its sentences are issue-free on their own but confusing when combined.
- **Dialogue-contradicting**: a response that contradicts the context, either partially or fully.
- **Parroting**: a response that repeats a certain part of the context in an unnatural way.
- **Off-topic**: a reply that has little to no relevance to the dialogue.
- **On-topic but unverifiable**: an on-topic reply that cannot be verified based on the context.

For concrete examples of the categories, see Table 4.

Other Considerations Good reflections sometimes repeat something that the client has said, for example to affirm it, but those are natural and good practices rather than unnatural repetition (**Parroting**). Also, broadly speaking, **Dialogue-contradicting**, **Off-topic** and **On-topic but unverifiable** reflections are all unfaithful and ungrounded w.r.t. the context, making them all manifestations of hallucination. Finally, we note that a small percentage ($\approx 8\%$) of error descriptions do not contain sufficient information (e.g., “Doesn't feel like a natural response”) and are therefore excluded from the thematic analysis. To account for such generic feedback and also to capture potential errors that do not fit neatly into the categories above, future users of this scheme may optionally create an “Other” category.

5 Conclusion

In this work, we explored non-expert annotation of machine-generated reflections for counselling dialogues, based on the assumption that non-experts are capable of context-informed 1) judgement of whether a reflection is coherent and consistent and 2) identification of the errors in an incoherent/inconsistent reflection. We identified common patterns among the free-text error descriptions from non-experts about generated reflections and accordingly used thematic analysis to establish discrete error categories that emphasised context understanding. Based on these categories, we plan for future work A) a mass non-expert error annotation phase for generated reflections, followed by B) an expert-based validation phase, and the results from both phases will be released to the public.

Limitations

In this preliminary study, our goal is to establish error categories for annotating machine-generated reflections. While we believe the human annotation conducted in this work is sufficient for achieving the goal, its limited annotation scale precludes drawing reliable conclusions from more advanced analysis, such as 1) coherence/consistency rates of different models and decoding strategies, and 2) correlation between human judgement and existing automatic metrics that are commonly used for dialogue generation tasks. Therefore, in our next step, we plan to carry out significantly scaled-up human annotation, in order to facilitate further analysis.

Ethical Aspects

Before starting the experiment, the procedure and materials were carefully reviewed by the University of Aberdeen's Ethics Board. Our experiment proposal was accepted without major revisions. All

		Context
Utt.	**Role**	**Text**
u_{t-16}	**Therapist**	Mm. So, embraces this life of the picket fence and the, you know, the little house and having the baby in this-
u_{t-15}	**Client**	Yeah, suburbs. Yeah.
u_{t-14}	**Therapist**	So, you don't wanna be that- that- that woman that's just sort of living in this house with the baby carriage, and at the same time, that world's kind of saying it's behind you.
u_{t-13}	**Client**	Yeah, I-I guess. Like, I don't wanna- I don't wanna lose my friends because of this. And I don't wanna put pressure on them to change because they're, I mean, they're good people. They're just–
u_{t-12}	**Therapist**	You care about them.
u_{t-11}	**Client**	Yeah.
u_{t-10}	**Therapist**	It's just the scene that they're-they're still in because their life maybe hasn't taken that change that your life has made. And so now, you're here and you're not sure. You want the baby, but you're not sure about the suburbs and the picket fence thing. You're not sure–
u_{t-9}	**Client**	I don't want that.
u_{t-8}	**Therapist**	You don't want that.
u_{t-7}	**Client**	No.
u_{t-6}	**Therapist**	You want this baby.
u_{t-5}	**Client**	Yeah.
u_{t-4}	**Therapist**	And you wanna be, um, take care of your baby as the baby's growing.
u_{t-3}	**Client**	Like, I wanna- I wanna be responsible for this thing.
u_{t-2}	**Therapist**	You don't wanna let your baby down.
u_{t-1}	**Client**	No.
		Ground-Truth Response (Reflection), Coherent & Consistent
u_t	**Therapist**	And at the same time, it's hard to let go of the past, but on the other hand, you're thinking, "I don't have a choice. I've got to look this way." Is that kind of what it feels like?
		Malformed
	Therapist	You-you wanna be the one that, um–you're the one who's gonna pick up the phone, uh, uh–you wanna pick up and, uh –or call them, um –to-to-unquote-the-mother and-to tell them that your-your daughter is going to come home and she's, uh...
		Dialogue-contradicting
	Therapist	OK, so what I'm hearing you say is that taking care of a baby would be too much responsibility for you right now, so you are not ready to have a baby.
		Parroting
	Therapist	Mm-hmm. So, you don't want to let your baby down.
		Off-topic
	Therapist	So you do not wanna allow your kids to go across the street without you.
		On-topic but unverifiable
	Therapist	You've really wanted to be a mom since you were a little girl.

Table 4: Example for each error category, as established in §4.

of the recruited annotators were volunteers. Upon starting the experiment, each annotator was shown a consent form containing all the information regarding the experiment procedure. All workers had to confirm their acceptance of these conditions before proceeding. Workers were given an email contact in case of problems during the experiment. No personal data about the annotators was kept stored at the end of the experiment.

Acknowledgements

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882.

References

Imtihan Ahmed. 2022. *Automatic Generation and Detection of Motivational-Interviewing-Style Reflections for Smoking Cessation Therapeutic Conversations Using Transformer-based Language Models*. Ph.D. thesis, University of Toronto.

Alain Braillon and Françoise Taiebi. 2020. Practicing "reflective listening" is a mandatory prerequisite for empathy. *Patient Education and Counseling*, 103(9):1866–1867.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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 3806–3812.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Y Ishii, ANDREA Madotto, and PASCALE Fung. 2022. Survey of hallucination in natural language generation. *ACM Comput. Surv*, 1(1).

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020a. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020b. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, Stacey ML Hendrickson, and William R Miller. 2005. Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment*, 28(1):19–26.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Erik Rautalinko and Hans-Olof Lisper. 2004. Effects of training reflective listening in a corporate setting. *Journal of Business and Psychology*, 18(3):281–299.

Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. 2007. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy*, 61(2):191–209.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Stephen Rollnick, William R Miller, and Christopher Butler. 2008. *Motivational interviewing in health care: helping patients change behavior*. Guilford Press.

Sune Rubak, Annelli Sandbæk, Torsten Lauritzen, and Bo Christensen. 2005. Motivational interviewing: a systematic review and meta-analysis. *British journal of general practice*, 55(513):305–312.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

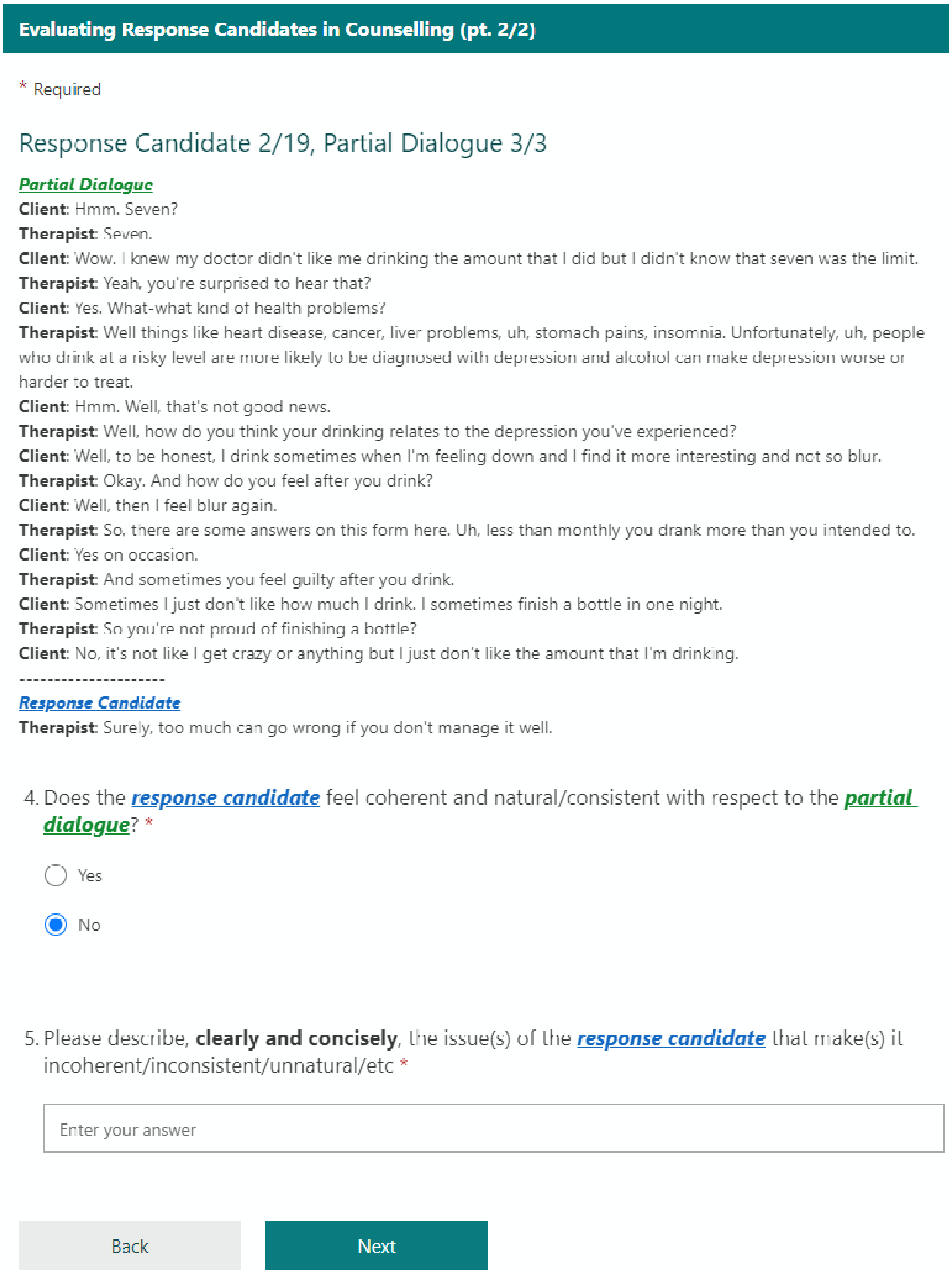
Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

A Annotation Interface

In Figure 2, we show the annotation interface for collecting free-text error descriptions from non-experts about generated reflections.



Evaluating Response Candidates in Counselling (pt. 2/2)

* Required

Response Candidate 2/19, Partial Dialogue 3/3

Partial Dialogue

Client: Hmm. Seven?
Therapist: Seven.
Client: Wow. I knew my doctor didn't like me drinking the amount that I did but I didn't know that seven was the limit.
Therapist: Yeah, you're surprised to hear that?
Client: Yes. What-what kind of health problems?
Therapist: Well things like heart disease, cancer, liver problems, uh, stomach pains, insomnia. Unfortunately, uh, people who drink at a risky level are more likely to be diagnosed with depression and alcohol can make depression worse or harder to treat.
Client: Hmm. Well, that's not good news.
Therapist: Well, how do you think your drinking relates to the depression you've experienced?
Client: Well, to be honest, I drink sometimes when I'm feeling down and I find it more interesting and not so blur.
Therapist: Okay. And how do you feel after you drink?
Client: Well, then I feel blur again.
Therapist: So, there are some answers on this form here. Uh, less than monthly you drank more than you intended to.
Client: Yes on occasion.
Therapist: And sometimes you feel guilty after you drink.
Client: Sometimes I just don't like how much I drink. I sometimes finish a bottle in one night.
Therapist: So you're not proud of finishing a bottle?
Client: No, it's not like I get crazy or anything but I just don't like the amount that I'm drinking.

Response Candidate
Therapist: Surely, too much can go wrong if you don't manage it well.

4. Does the ***response candidate*** feel coherent and natural/consistent with respect to the ***partial dialogue***? *

○ Yes

● No

5. Please describe, **clearly and concisely**, the issue(s) of the ***response candidate*** that make(s) it incoherent/inconsistent/unnatural/etc *

Enter your answer

Figure 2: Annotation Interface