In-Context Non-Expert Evaluation of Reflection Generation for Counselling Conversations: A First Study

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

Simone Balloccu
University of Aberdeen
simone.balloccu@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 non-expert human annotations on a) the coherence/consistency of generated reflections and b) incoherence/inconsistency error types. We show that our approach is valid, cost-saving and scalable through inter-annotator agreement on both tasks. Finally, we present detailed analyses of coherence/consistency rates and error types across all the tested models and decoding strategies. This work paves the way for a mass non-expert error annotation phase for generated reflections, which will be followed by an expert-based validation phase.

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.

		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.

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 brittle (Liu et al., 2016) and human evaluation is thus necessitated. Moreover, reflection requires specialised knowledge and counselling is a complex and delicate domain. Ideally, therefore, generated reflections need evaluation by experienced therapists. However, expert annotation is time-consuming and

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

costly (Moyers et al., 2005). Thus, human evaluation in previous work suffers from issues including 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.

In this work, we take the first step towards addressing these issues. We propose 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. 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.

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 descriptions of reflection errors from non-experts and 2) identifying common patterns 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, of which most 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 the scheme, we collect `ReflError Mini`, an initial dataset[3] of 150 annotated reflections distributed across 15 MI dialogues and coming from both PLMs and the ground-truth therapist response. Each dialogue is a real-world-like MI demonstration, and each of the 150 reflections is annotated by 3 non-experts. Overall, we find moderate agreement on whether a reflection is coherent and consistent w.r.t. its context and substantial agreement on the error of incoherent/inconsistent reflections. We also present analyses of the coherence/consistency rates and error type distributions for different models and decoding strategies.

Based on the results, we consider our initial approach to non-expert annotation valid, and we plan 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 Empathetic Dialogue Generation for Peer Support and Counselling

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 or counselling-like settings, such as mental health peer support.

Peer-support EDG follows a ⟨support seeker, supporter⟩ setup, where the model plays the supporter and decides on a supporting strategy — often informed by psychological principles (e.g., Hill, 2009) — before generating a support message. For example, Liu et al. (2021) create a supportive bot that chooses among question, self-disclosure, suggestion etc. Similarly, Sun et al.'s (2021) digital supporter utilises strategies such as interpretation, direct guidance, restatement, etc. On the other hand, Sharma et al. (2021) propose rewriting low-empathy support messages into high-empathy ones using empathetic mechanisms like emotional reactions, interpretations, and explorations.

In comparison, counselling EDG has been less explored due to privacy-related data constraints, but some first works on reflection generation have been done recently. 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 utter-

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

[3]Availble at `https://reflerror.s3.eu-west-1.amazonaws.com/reflerror_mini.zip`

ances 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[4] (Rashkin et al., 2019; Li et al., 2020) 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). Peer-support EDG largely follows this setup, although some studies (e.g., Sun et al., 2021; Liu et al., 2021) also include "helpfulness", where a psychology expert or a dialogue participant assesses if a generated support message is helpful.

For evaluating reflection generation, Shen et al. (2020, 2022) replace "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 by experts in a good-vs-bad binary setup. In comparison, our human evaluation is novel in its explicit focus on context-informed 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 and "extrinsic hallucination" if it cannot be verified based on the input. 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.

[4]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.

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 Figure 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[5]:

> "⟨*client*⟩*The baby was up all night and I'm exhausted.*|⟨*therapist*⟩*So, what you're saying is you've had a rough night?*|⟨*client*⟩*Yes. She was up every three hours to eat, I don't understand it.*|⟨*therapist*⟩~⟨*listening*⟩"

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[6], 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. 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 each model's perplexity (the lower the better), 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) our 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.

Model	GPT-2	BART
Perplexity	17.36	13.29%

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

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. 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, $p \in \{0.4, 0.6, 0.8, 0.95\}$, 5 sequences sampled for each p

Thus, each context leads to 1 response from greedy decoding, 5 from beam and $4 \times 5 = 20$ from nucleus using both GPT-2 and BART, hence 52 responses in total. Nevertheless, repetitions and semantic duplicates can occur among the responses, which is taken into consideration during the human annotation, as will be detailed in §4.

4 Human Annotation

Our non-expert human annotation of reflections consists of two stages. In Stage 1, we gather free-text descriptions of reflection errors from laypeople to formally define error categories and incoherence/inconsistency accordingly. In Stage 2, we conduct a larger scale coherence/consistency and error annotation based on those clear definitions.

4.1 Stage 1: Free-Text Error Description

Since the underlying assumption of our human annotation is that incoherence/inconsistency errors can be spotted by non-experts, we first survey

[5] In practice, we use "⟨asking⟩", "⟨informing⟩", "⟨listening⟩", "⟨other⟩" as the plain-text control codes for Question, Input, Reflection and Other, respectively.

[6] 384 tokens make up N turns where N vares depending on the individual utterance lengths, but on average $N = 14$.

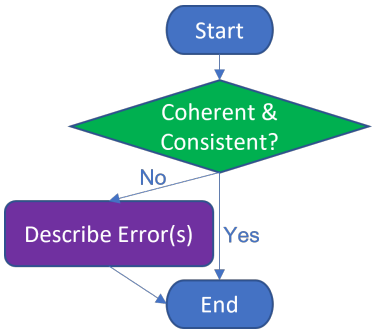

Figure 1: Stage-1 Annotation

laypeople for their own descriptions of reflection errors and then summarise those free-text descriptions into categories.

4.1.1 Annotation Materials

We sample 3 contexts from 3 different dialogues and use their respective ground-truth and model-generated reflections for annotation. Among the 5 returned sequences in beam decoding, we keep the top-beam result and randomly sample another 2 from the remaining 4. Similarly, we randomly sample 3 responses from the 5 sequences from nucleus decoding under each p. For each context, therefore, there are at most 16 responses (1 from greedy decoding, 3 from beam, and $4 \times 3 = 12$ from nucleus) both from GPT-2 and from BART. After removing responses that are complete duplicates or differ in casing/punctuation only, the 3 contexts have 19, 19 and 22 responses each to be annotated.

4.1.2 Annotation Setup

Annotators We found 6 annotators with high proficiency in English and no prior experience in psychology/psychotherapy. Each annotator worked on the same batch of 60 (19 + 19 + 22) reflections for the aforementioned 3 contexts in total.

Annotation Procedure The procedure is illustrated in Figure 1. 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.

4.1.3 Established Error Categories

We use thematic analysis (Braun and Clarke, 2012) to identify common patterns in the annotators' feedback and summarise them into the following error categories. For concrete examples of the categories, see Appendix A.

- **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.

We note that 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**).

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.

4.2 Stage 2: Categorical Error Annotation

Using the error categories established in Stage 1, we conduct Stage-2 annotation on a larger scale.

4.2.1 Annotation Materials

Similar to Stage 1 (§4.1.1), we sample 15 contexts from 15 different dialogues and use their respective ground-truth and model-generated reflections for human annotation. In this stage, however, we remove semantically similar reflections to keep only 9 generated reflections per context in order to 1) maximise the semantic diversity of the reflections to annotate and 2) reduce the burden of annotation.

More specifically, we embed each of the 52 generated reflections of each context with MPNet (`all-mpnet-base-v2`, Song et al., 2020) before

conducting agglomerative clustering to group the 52 reflections into C semantic clusters with a linkage distance threshold σ. Then, the response with the most tokens is chosen as the semantic representative of the cluster, as it is likely to be more contentful. Finally, 9 clusters are randomly sampled out of the C in total, and their respective semantic representatives are later used for human annotation.

Empirically, we found $\sigma = 0.7$ to be a good semantic similarity threshold. For example, a particular cluster of size 3 is

I *"Okay. So, it sounds like you have a really good plan for your future."*

II *"Okay. So, it sounds like you have a pretty good plan in place for what you want to do with your life."*

III *"Okay. So, it sounds like you have a pretty good plan for your future."*

and II is chosen as the semantic representative, since it contains the most tokens. On average, a randomly sampled cluster contains 1.32 reflections.

4.2.2 Annotators

For Stage 2, we followed the same annotator requirements of Stage 1 to recruit 9 non-experts, none of whom participated in Stage 1. The workload of each annotator consists of 5 contexts with 10 responses each — 1 ground-truth reflection and 9 generated ones selected as described in §4.2.1. Thus, each response is annotated by 3 people. Prior to annotation, each participant read a compulsory tutorial that clearly defines the error categories (§4.1.3) and, on that basis, what makes a response incoherent/inconsistent or coherence and consistent.

4.2.3 Annotation Procedure

The procedure is illustrated in Figure 2. Like in Stage 1, the annotators need to choose whether the reflection in a $\langle context, reflection \rangle$ pair is coherent and consistent w.r.t. the context. If their answer is "No", they are asked to choose the applicable error category(ies). In the case of multiple categories being selected, the annotators are required to choose the most evident one in their view.

Additionally, an empathy assessment is presented to the annotators if they consider a reflection coherent and consistent. Specifically, they are shown the statement "The response candidate gives me the impression that the therapist understands the client's perceptions, situation, meaning, and feelings" — the definition of high empathy in MI (Miller et al., 2003) — and they will then choose their level of agreement with the statement on a 5-point Likert scale: {disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree}. "Empathy" is not mentioned, as it can be a complex and vague concept to non-experts. This question aims to enable more insights on how laypeople perceive empathy in MI reflections.

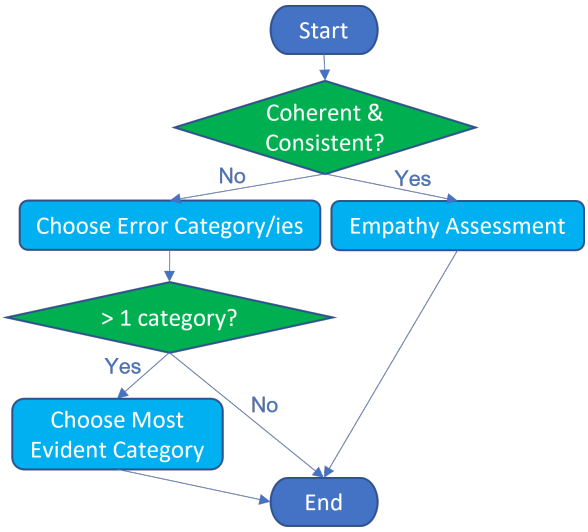

Figure 2: Stage-2 Annotation

5 Annotation Analysis

Since the purpose of Stage 1 is to obtain a system of response-error categories to enable Stage 2, we present only the results of Stage 2 in this section.

5.1 Inter-Annotator Agreement

Overall, there are 150 reflections — 15 ground-truths and 135 generated ones — included in Stage 2, each annotated by 3 people. Therefore, we use Fleiss' kappa (Fleiss, 1971) to compute inter-annotator agreement (IAA), since it does not require the set of raters for each subject[7] to be identical. All IAA results are presented in Table 4.

We first "fill" the most evident error (MEE) annotation where there is only one spotted error category, thus enabling comparison between different MEE annotations. For example, if **Malformed** is chosen as the only error of a particular response, it is automatically deemed the MEE, too. When calculating the IAA for empathy assessment, only the subjects that are considered coherent and consistent by all of the 3 annotators are included. Similarly,

[7]We use "subject", "example", "response" and "reflection" interchangeably when discussing IAA.

Original		
Annotation Item	**IAA**	**#Sub**
Coherent & consistent?	0.43	150
Empathetic?	0.09	24
Most evident error	0.45	51
Derived		
Consensus-adjusted most evident error	0.54	51
Empathetic?$^{(M)}$	-0.04	24
Most evident error$^{(M)}$	0.65	51
Consensus-adjusted most evident error$^{(M)}$	0.75	51

Table 4: Inter-annotator agreements (IAAs). #Sub: number of subjects for IAA. $^{(M)}$: involves category merging. IAA interpretation: poor (< 0), slight (0.01-0.20), moderate (0.41-0.60), substantial (0.61-0.80).

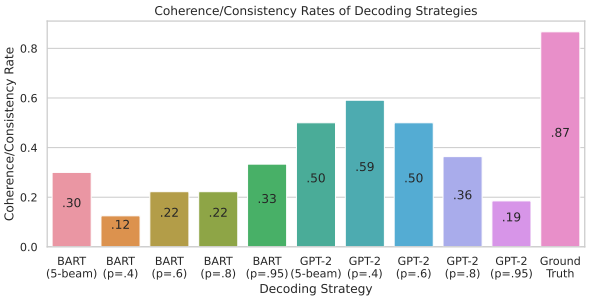

Figure 3: Decoding strategies to coherence rates

when computing the IAA of MEE, only the examples deemed incoherent/inconsistent by all the 3 annotators are taken into consideration.

We observe a moderate agreement of 0.43 over whether a reflection is coherent and consistent, and the IAA is similarly moderate (0.45) for MEE, both indicating the broad consensus on what constitutes a coherent & consistent reflection and what is the most important reason that makes a response incoherent/inconsistent. For the empathy assessment, however, there is only slight (0.09) agreement, which shows the challenge and subjectivity of empathy assessment for non-experts.

We also introduce consensus-adjusted MEE, which replaces the MEE of an annotation with a common error found by all 3 annotators, where applicable. For example, if a generated reflection is {**Malformed**}, {**Off-topic**, **Malformed**} and {**Parroting**, **Malformed**} according to the 3 annotators respectively, the consensus-adjusted MEE is assigned **Malformed** for all of the annotators. If such a consensus error does not exist, consensus-adjusted MEE for each annotator simply takes the original MEE as the default. Unsurprisingly, this approach boosts the IAA to 0.54, as it is designed to increase consensus among the annotations.

Finally, we explore category merging: 1) {**Dialogue-contradicting**, **Off-topic**, **On-topic but unverifiable**} into **Hallucinatory**, 2) {somewhat disagree, disagree} into *disagree*, and 3) {somewhat agree, agree} into *agree* . The effect of 1) on IAA is clear, as it boosts the IAAs of MEE and consensus-adjusted MEE to 0.65 and 0.75, respectively, both in the range of substantial agreement, which is a further evidence of the high level of agreement on response errors. On the other hand, 2) and 3) reduce the IAA of empathy assessment to -0.04, which is relatively surprising.

Overall, we argue that the relatively high IAAs on coherent/consistency and MEE show that larger-scale annotations will likely produce similarly stable results and therefore merit consideration.

5.2 Coherence/Consistency Rates

We inspect the overall coherence/consistency rate of each decoding strategy, which is defined as the percentage of responses generated through a strategy that are annotated as coherent and consistent. To acquire such a label for each annotated reflection, majority voting is used. Also, if a cluster representative is assigned a majority-vote label, all the other reflections in the cluster are given the same label, since they are highly similar semantically.

The coherence/consistency rates of all the strategies with at least 5 majority-vote-labelled reflections are shown in Figure 3. Interestingly, one can observe that even the ground-truth reflections are not always considered coherence/consistent by the annotators. Upon inspection, 2 out of the 15 (13%) of the ground-truths are annotated as incoherence/inconsistent, of which one is considered by 2 annotators as **Malformed** and the other as **On-topic but unverifiable**. The first is due to artefacts from the transcription process, while the second involves content that is not captured by the 384-token context. On the other hand, this also means that in the vast majority of cases, a good reflection like the ground-truths is well-formed and does not refer to information from a relatively distant past, hence it is a reasonable upper-bound for PLM-based dialogue models that use contexts as input.

The gap between the coherence/consistency rates of the models and that of the ground-truth is clear, as the best decoding setup — GPT-2 with nucleus decoding and $p = 0.4$ — only achieves a rate of 0.59. One can also observe that while differ-

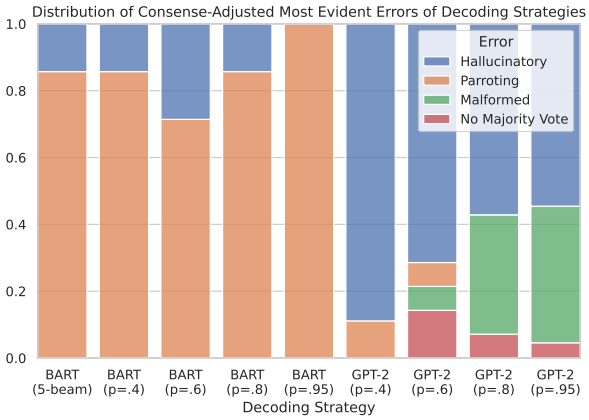

Figure 4: Decoding strategies to their distributions of consensus-adjusted most evident errors

ent p values contribute to wildly different coherence/consistency rates, beam decoding has comparatively higher rates for both GPT-2 and BART.

Overall, there is a negative correlation between the p for nucleus-decoded GPT-2 reflections and the coherence/consistency rate, which is relatively intuitive: higher p gives the model more flexibility during generation, which leads to more diversity but can compromise quality — an observation in other studies (e.g., Dziri et al., 2021) as well. However, the trend is reversed for BART, as increased p is correlated with higher coherence/consistency rates. We leave to future work to examine this phenomenon, but we postulate that both the seq2seq nature of BART — GPT-2 is autoregressive — and its pre-training objectives are important factors.

5.3 Overview of Errors

We also investigate the distribution of most-evident errors of the responses generated under each decoding strategy, namely "if a reflection from a particular decoding strategy is incoherent/inconsistent, how often is it **Malformed/Parroting/Hallucinatory**?".

Similar to §5.2, we take the majority vote on consensus-adjusted MEE for each annotated reflection, after merging {**Dialogue-contradicting**, **Off-topic** and **On-topic but unverifiable**} into **Hallucinatory**. We note that a majority vote is not guaranteed to exist, unlike for coherence/consistency, for example when 2 annotators consider a response incoherent/inconsistent but assign **Malformed** and **Hallucinatory**, respectively, as the only error. For those cases, we consider the consensus-adjusted MEE to be **No Majority Vote**.

Figure 4 presents the distribution of majority-voted consensus-adjusted MEEs for each decoding strategy that has at least 5 responses that are majority-voted as incoherent/inconsistent.

We notice that the predominant error of BART is **Parroting**, regardless of the decoding strategy, while the GPT-2 counterpart is **Hallucinatory**. Considering that larger p leads to slightly higher consistency/coherence rates for BART (§5.2), we can deduce that BART tends to simply repeat a certain part of the context instead of producing relatively novel content, and that it can be alleviated by increasing p but only to a limited extent[8].

For GPT-2, **Malformed** becomes more and more prominent as p rises, while the share of **Hallucinatory** declines steadily. Clearly, while unfaithfulness/ungroundedness is a common theme for GPT-2 reflections, large p's can further cause considerably more text degeneration.

6 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 a whether reflection is coherent and consistent and 2) identification of the errors in an incoherent/inconsistent reflection. Accordingly, we collected an initial dataset of reflections generated by long-context dialogue models, where each reflection is annotated by non-experts following an error scheme that emphasises context-based quality assessment. We showed sufficient inter-annotator agreement on both non-expert tasks and presented analyses of the coherence/consistency rates and error type distributions for different models and decoding strategies. As our assumption is justified by the results, we plan a mass non-expert error annotation phase for generated reflections, which will be followed by an expert-based validation phase.

Limitations

The most obvious limitations of this exploratory study are to do with the models used and the scale of annotation, both to be addressed in future work.

More specifically, our trained response generators, namely `gpt2-medium` and `bart-large`, are orders of magnitude smaller than the latest PLM-

[8]We also experimented with increasing the temperature for BART during decoding to boost output diversity, but it quickly led to predominantly ungrammatical responses, therefore we kept it at 1.0, which is also used for GPT-2 for fair comparison.

based dialogue models, such as LaMDA (Thoppilan et al., 2022) and BlenderBot 3 (Shuster et al., 2022), which often also have dedicated modules to improve response safety and knowledgeability, etc.

As for the annotation scale, our dataset contains 150 annotated reflections, which is large enough to prove the viability of our non-expert annotation design but not sufficient for purposes such as training classifiers to identify hallucinatory reflections.

In our next effort, therefore, we will leverage significantly larger-scale PLMs such as GPT-3 to obtain additional generated reflections, and we will incorporate those responses in a considerably scaled-up non-expert annotation setup.

Ethical and societal implications

The work is driven by the vision of making professional counselling services more readily accessible by lowering the costs of therapist training. Our goal is not to replace human therapists, but rather to assist them and help them develop essential counselling skills like reflective listening, as is approached in this work. While PLM-based reflection generation is a nascent field, it is promising and can be advanced greatly by leveraging high-quality error annotations, which are precisely what we will be gathering on a larger scale as the next step, moving closer to our vision. More broadly, we believe that NLP as a field would have more positive impact on people's lives by playing the AI-in-the-loop role like in our work, instead of trying to replace humans.

Our experiments were reviewed and approved by the Ethics Board of our institution before they were conducted. In particular, we note that counselling dialogues can be extremely sensitive, hence we utilised `AnnoMI`, which consists of demonstrated but real-world-like MI conversations, removing the need to process personal information but without compromising experiment quality.

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, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214.

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

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

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.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. 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.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism,*

Substance Abuse and Addictions, University of New Mexico.

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 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*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

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*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.

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*.

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 Examples for Error Categories

In Table 5, we show the ground-truth reflection for the context, which is coherent and consistent, as well as one example alternative reflection for each error category.

Utt.	Role	Text
		Context
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 5: Example for each error category, as established in §4.1.3