re Experts Needed?
On Human Evaluation of Counselling Reflection Generation

Zixiu Wu [2] **Simone Balloccu**[34] **Ehud Reiter**[3] **Rim Helaoui**
Diego Reforgiato Recupero[2] **Daniele Riboni**[2]

Philips Research, the Netherlands [2]University of Cagliari, Italy
[3]University of Aberdeen, UK [4]Charles University, Czechia

z.wu@studenti.unica.it {s.balloccu.19 e.reiter}@abdn.ac.uk
rim.helaoui@philips.com {diego.reforgiato riboni}@unica.it

bstract

Reflection is a crucial counselling skill where the therapist conveys to the client their interpretation of what the client said. Language models have recently been used to generate reflections automatically, but human evaluation is challenging, particularly due to the cost of hiring experts. Laypeople-based evaluation is less expensive and easier to scale, but its quality is unknown for reflections. Therefore, we explore whether laypeople can be an alternative to experts in evaluating a fundamental quality aspect: coherence and context-consistency. We do so by asking a group of laypeople and a group of experts to annotate both synthetic reflections and human reflections from actual therapists. We find that both laypeople and experts are reliable annotators and that they have moderate-to-strong inter-group correlation, which shows that laypeople can be trusted for such evaluations. We also discover that GPT-3 mostly produces coherent and consistent reflections, and we explore changes in evaluation results when the source of synthetic reflections changes to GPT-3 from the less powerful GPT-2.

1 Introduction

Motivational Interviewing (MI, Miller and Rollnick, 2012) is a highly effective counselling practice in healthcare (Moyers et al., 2009), where the therapist focuses on evoking the client's own motivation for behaviour change, such as smoking cessation and alcohol use reduction. In MI, reflective listening is a crucial strategy of showing empathy, where the therapist conveys a brief conversational summary of how they understand what the client said (Miller et al., 2003; Rollnick et al., 2008). An example is shown in Table 1.

Learning effective reflective listening requires considerable training time and expert supervision (Rautalinko and Lisper, 2004; Rautalinko et al., 2007). Therefore, recent studies used language models (LMs) as automatic reflection generators to aid training (Shen et al., 2020, 2022; Ahmed, 2022), where the LM receives a dialogue context as the input and outputs a reflection (Table 1).

Context
Client: Well, I'm here because my mom wants me to be here. **Therapist**: Mm-hmm. **Client**: I don't really wanna be here, but it-it– whatever. **Therapist**: Got it. **Client**: Um, she-she found my stash- **Therapist**: Uh-huh. **Client**: -and she freaked out, and she's going crazy over it. Um, I don't why she was going through stuff in the first place, but whatever, so, now I'm here. (intermediate turns) **Therapist**: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that? **Client**: I am. **Therapist**: Yeah. **Client**: I mean, it's my stuff, I don't know why she's–
Reflection Candidates
Therapist (Human): Right. It's like your private place and, you know, it's– that's– it's your stuff.
GPT-2: It's a very sad thing.
GPT-3: It sounds like you're really upset with her because she invaded your privacy.

Table 1: A dialogue context about reducing substance use, together with its human reflection and two examples of synthetic reflections. Self-repetitions and mid-sentence changes (e.g., "it-it-whatever") are characteristics of the dataset (Wu et al., 2022b).

Human evaluation of reflection generation is crucial, since automatic metrics are often not robust (Liu et al., 2016). For such evaluations, experts (professional therapists) are used due to their deep understanding of the complex and sensitive domain of counselling dialogue. However, expert evaluation is costly and difficult to scale, and previous human evaluations often adopted over-simplified annotation schemes (good vs bad reflection) or worked with short dialogue contexts (5 turns). Evaluation with laypeople (such as crowdworkers) tends to be less expensive (Iskender et al., 2020), but to the best of our knowledge its reliabil-

Proceedings of the 61st nnual Meeting of the ssociation for Computational Linguistics
Volume 1: Long Papers, pages 6906–6930

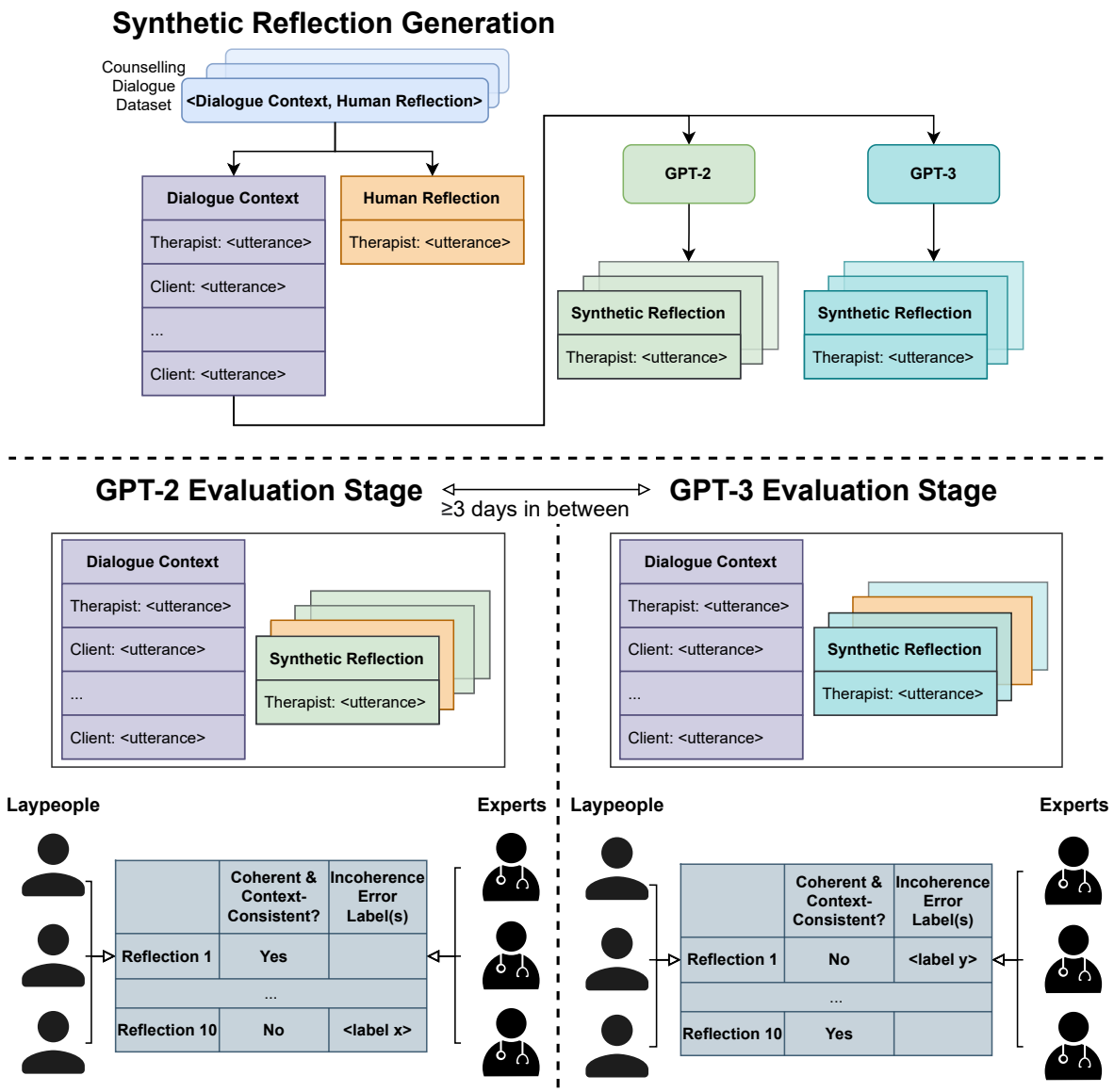

Figure 1: Human evaluation overview. The same human reflections are included in both evaluation stages, mixed with GPT-2 reflections in the GPT-2 stage and with GPT-3 reflections in the GPT-3 stage.

ity for reflections is unknown.

In this work, we investigate if laypeople are a viable alternative to experts for human evaluation of **coherence and context-consistency** (referred to as **coherence** for brevity). This is a weak point of recent generative models (Ji et al., 2022) and also a fundamental quality aspect of reflection generation, since a reflection has to first "make sense" in the context before it can be evaluated against counselling principles.

To this end, we recruit a group of MI experts and a group of laypeople as annotators and analyse their evaluation[1] quality (Figure 1). The workload of each annotator consists of mixed human reflections from actual therapists and synthetic reflections produced by language models (GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020)[2]), and the annotator is not informed of the source of any reflection. For each reflection, the annotator flags whether it is coherent as a Yes/No binary choice. If "No" is chosen, the annotator proceeds to select one or more applicable incoherence error categories. In doing so, our evaluation goes beyond a binary Yes/No scheme and sheds light on the types of inco-

[1]Data available at https://github.com/uccollab/expert_laypeople_reflection_annotation.

[2]We also conducted human evaluation of reflections generated by BART (Lewis et al., 2020), but the results are not included in the main body as the model failed to generate sufficiently diverse reflections (Appendix B)

herence errors made by reflection generators. Notably, we adopt long dialogue contexts — 14 turns on average — to allow for more detailed conversational background to both the reflection generator and the annotator.

Based on the annotations, we conduct in-depth analysis of intra-group agreement among laypeople and among experts, as well as the inter-group correlation between laypeople and experts. We also explore whether more powerful LMs produce more coherent synthetic reflections and how they affect annotations of human reflections. We find that:

I Both laypeople and experts are reliable annotators based on their intra-group agreements on binary coherence evaluation. They also show moderate to strong inter-group correlation.

II Human reflections are more often annotated as coherent than GPT-2 reflections, but it is not the case with the more powerful GPT-3. Interestingly, both laypeople and experts are less likely to annotate a human reflection as coherent when its surrounding synthetic reflections come from GPT-3, though experts are relatively more consistent in this regard.

I represents the first evidence that laypeople are capable of coherence evaluation for reflection generation. II poses an interesting research question on whether synthetic reflections from large LMs can match or outperform human reflections on dimensions deeper than coherence, such as empathy.

2 Related Work

2.1 Human Evaluation for Response Generation

In most studies of response generation, human evaluation is considered the ultimate benchmark, since it can assess quality aspects like interestingness and safety (Deriu et al., 2021; Liu et al., 2016; Thoppilan et al., 2022) that may elude automatic metrics. Typically, the human evaluator rates model-generated responses in an interactive or static setup.

In an interactive setting, the human converses with the dialogue model and evaluates its responses as good/bad (e.g., Shuster et al., 2022) or selects applicable attributes like knowledgeable/engaging/... (e.g., Komeili et al., 2022). In a static setup, the human evaluates responses or entire dialogues on the Likert scale for an attribute (Rashkin et al., 2019; Li et al., 2020, *inter alia*) or compares responses from different models through ranking or A/B testing (Xie and Pu, 2021; Kim et al., 2021, *inter alia*).

Despite their popularity, standard human evaluation protocols suffer from various issues. One such example is subjectivity (Li et al., 2019; Howcroft and Rieser, 2021), in particular in the context of Likert scales. Other issues include the lack of reproducibility across studies and the influence of evaluation instructions (Belz et al., 2023; Huynh et al., 2021; Smith et al., 2022).

2.2 Reflection Generation and Its Human Evaluation

Shen et al. (2020) developed the first LM-based reflection generator. Shen et al. (2022) leveraged commonsense and domain knowledge for reflection generation. Ahmed (2022) adopted a few-shot approach. All those studies used at most 5 turns as the dialogue context, in contrast to the 14 turns on average in our work. Therefore, our generation and evaluation is more context-aware.

For human evaluation, Shen et al. (2020, 2022) asked two experts to evaluate relevance, fluency and "reflection-like-ness" on Likert scales. Ahmed (2022) conducted expert evaluation of GPT-3 generated reflections in a good-vs.-bad setup. Wu et al. (2022a) proposed non-expert evaluation of coherence and context-consistency and developed an error annotation scheme accordingly. We adopt this annotation scheme in our work, but we focus on comparing laypeople- and experts-produced evaluations and investigating if laypeople can be a viable alternative to experts for coherence evaluation.

2.3 Expert and Non-Expert Evaluation for Natural Language Generation

Whether to use experts for NLG evaluation generally depends on the domain. For example, open-domain dialogue generation mostly involves non-experts to assess attributes like engaging-ness and human-ness (e.g., Roller et al., 2021; Komeili et al., 2022), while response generation for specialised domains like mental health (Sharma et al., 2021) and clinical dialogue (Miehle et al., 2018) is largely evaluated by domain experts.

Some human evaluation studies have compared expert and non-expert NLG evaluations, such as for summarisation (Gillick and Liu, 2010; Fabbri et al., 2021), machine translation (Freitag et al., 2021), story generation (Karpinska et al., 2021) and others (e.g., Snow et al., 2008). Many of these works reveal considerable gaps between assessments from

experts and those from crowdworkers. In particular, Freitag et al. (2021) find that automatic metrics outperform crowdworkers in terms of correlation with expert judgement.

3 Methodology

3.1 Synthetic Reflection Generation

We leverage LMs to generate synthetic reflections through fine-tuning and prompting, both of which are based on AnnoMI (Wu et al., 2022b), an expert-annotated dataset of transcribed MI sessions over various topics such as smoking cessation and alcohol use reduction. AnnoMI contains 110 conversations with 4441 therapist turns (utterances), 28% (1256) of which are reflections and we refer to those as "human reflections".

For each human reflection, we concatenate its preceding utterances and keep the rightmost (i.e., temporally most recent) 384 tokens as the dialogue context, which contains 14 previous turns on average. Notably, this is 3 times the context size used in previous work (≤ 5 turns), as we assume richer context enables better reflection generation. Thus, we construct 1256 ⟨context, human reflection⟩ pairs based on AnnoMI.

3.1.1 Fine-Tuning

Following recent work on reflection generation (Shen et al., 2020), we fine-tune GPT-2 (`gpt2-medium`, Radford et al., 2019) on ⟨context, human reflection⟩ pairs. At test time, we use greedy, beam and nucleus (Holtzman et al., 2020) ($p \in \{0.4, 0.6, 0.8, 0.95\}$) decoding to generate diverse synthetic reflections.

3.1.2 Prompting

We also prompt GPT-3 (`text-davinci-002`, Brown et al., 2020) to generate reflections, in light of the impressive generative capabilities of large LMs shown recently (Bhaskar et al., 2022; Goyal et al., 2022, *inter alia*) including for reflection generation (Ahmed, 2022). We use the default temperature (1.0) and $p \in \{0.4, 0.6, 0.8, 0.95\}$ for decoding. We model our prompt as asking GPT-3 to read a series of ⟨context, human reflection⟩ pairs (learning examples) and then to complete a final dialogue context where the reflection is missing (test example).

The test example is always a dialogue context from AnnoMI, but we explore two sources of learning examples — AnnoMI and **textbook** — to diversify the generation. The former (Figure 2a) is simply ⟨context, human reflection⟩ pairs we constructed previously, while **textbook** examples (Figure 2b) are taken from the Motivational Interviewing Treatment Integrity (MITI) coding manual (Moyers et al., 2014). Each textbook example consists of a client statement — which we use as dialogue context — along with a simple reflection and a complex one, where the complex reflection adds more meaning/emphasis to the client statement than the simple one (Miller et al., 2003).

```
Below are a few examples of
how a therapist responds to a
client given the context of their
previous exchanges. Learn
from these examples and write
the therapist response for the
last example.

# Example 1
## Context
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Response
Therapist: $utterance

... (4 other AnnoMI examples)

# Example 6
## Context
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Response
Therapist:
```

(a) Using AnnoMI examples.

```
Below are several examples of
how a therapist responses to a
client using a Simple Reflection
or a Complex Reflection, given
the Conversation History. Learn
from these examples and
complete the last example.

# Example 1
## Conversation History
Client: $utterance
## Reflections
Simple Reflection: $utterance
Complex Reflection: $utterance

... (7 other Textbook examples)

# Example 9
## Conversation History
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Reflections
```

(b) Using textbook examples.

Figure 2: Prompting formats.

3.2 Human Evaluation

We recruit 2 groups of annotators:

- 9 laypeople known to us and with no experience in MI;
- 9 experts found through professional networks, in particular the Motivational Interviewing Network of Trainers[3], an international organisation of MI trainers and a widely recognised MI authority.

3.2.1 Workload

Table 2 presents the annotation workload overview.

To create annotation materials, we randomly sample 15 ⟨context, human reflection⟩ pairs from 15 AnnoMI dialogues. For the context in

[3] https://motivationalinterviewing.org/

Each batch contains	1 dialogue context, 1 human reflection, N synthetic reflections
GPT-2 stage	
Each layperson/expert has	5 batches
Each reflection annotated by	3 laypeople, 3 experts
Synthetic refl. per batch (N)	7.13 on average
Total batches	15
Total human reflections	15
Total synthetic reflections	107
GPT-3 stage	
Each layperson/expert has	5 batches
Each reflection annotated by	3 laypeople, 3 experts
Synthetic refl. per batch (N)	9 (except one batch with 7)
Total batches	15
Total human reflections	15
Total synthetic reflections	133

Table 2: Overview of Annotation Workload.

each pair, we generate 9 semantically diverse synthetic reflections[4] with GPT-3 and 7.13 on average[5] with GPT-2. Thus, for each ⟨context, human reflection⟩ pair, we create 2 annotation batches that each contain the context, the human reflection and synthetic reflections. The two batches differ in that the synthetic reflections in one batch come from GPT-2 while those in the other batch are from GPT-3.

Each annotator is first randomly assigned 5 batches where the synthetic reflections are from GPT-2 (**GPT-2 stage**). After completion of these batches and then a waiting period of at least 3 days (Appendix C), the annotator is randomly assigned 5 more batches where the synthetic reflections are from GPT-3 (**GPT-3 stage**). The task ends when the annotator has finished all 10 batches. Overall, each batch is randomly assigned to 3 laypeople and 3 experts, resulting in each reflection being evaluated 3 times by laypeople and 3 times by experts.

3.2.2 nnotating One Batch

When annotating a batch (Figure 1), the annotator first reads the context and then iteratively annotates all the reflections. The reflections in each batch are shuffled, and the annotator is not informed of the source of any reflection.

For each reflection, the annotator chooses Yes/No regarding whether it is coherent. If the answer is No, the annotator selects one or more applicable error categories. We adopt the error annotation scheme developed by Wu et al. (2022a), since the categories were qualitatively extracted from free-text feedback provided by laypeople w.r.t. model-generated reflections. Therefore, those categories represent a good approximation of what errors our annotators may find in synthetic reflections. Those categories are:

- ***Malformed***: suffers from unclear references, bad grammar, and/or confusing logic.
- ***Dialogue-contradicting***: contradicts context partially or fully.
- ***Parroting***: repeats a part of context unnaturally.
- ***Off-topic***: little to no relevance to context.
- ***On-topic but unverifiable***: relevant to context but including content that cannot be verified based on context alone.

Prior to annotation, the annotator reads a mandatory tutorial about coherence and consistency with examples for each error category, and it remains accessible throughout the annotation process.

3.2.3 Cross-Stage Human Reflection Recurrence

Due to random batch assignment, an annotator may annotate batch b_m in the GPT-2 stage and b_n in the GPT-3 stage where b_m and b_n share the same ⟨context, human reflection⟩. For the annotator in such cases, the shared human reflection is **recurring** across stages, and hence the annotator annotates it twice. To make it less likely that an annotator annotates a recurring human reflection in the GPT-3 stage based on how they recall annotating it in the GPT-2 stage, each annotator waits for at least 3 days[6] between completing their last batch in the GPT-2 stage and starting their first batch in the GPT-3 stage.

4 nnotation Results & nalysis

4.1 Intra-Group greement

We measure intra-group agreement among laypeople and among experts, i.e., how much the annota-

[4]There is one context with 7 instead of 9 GPT-3 reflections due to lack of semantic diversity among generated candidates.

[5]In practice, GPT-2 and BART reflections were evaluated together, and their combined size is the same as GPT-3 reflections'. Thus, there are fewer GPT-2 reflections than GPT-3 reflections. We exclude BART reflections from the GPT-2 stage for fairness considerations. More details in Appendix B.

[6]More details in Appendix C.

	Laypeople		Experts	
	GPT-2	*GPT-3*	*GPT-2*	*GPT-3*
Fleiss'	0.42	0.23	0.44	0.04
Randolph's	0.42	0.30	0.45	0.42

Table 3: Global agreement on *Coherent*/*Incoherent* binary choice.

	Laypeople		Experts	
	GPT-2	*GPT-3*	*GPT-2*	*GPT-3*
Coherent	0.69	0.76	0.66	0.90
Incoherent	0.71	0.51	0.75	0.25

Table 4: Per-label majority agreement ratios on *Coherent* and *Incoherent* separately.

	Laypeople		Experts	
	GPT-2	*GPT-3*	*GPT-2*	*GPT-3*
Parroting	*0.38*	0.45	*0.00*	0.11
Malformed	0.47	*0.00*	0.37	0.00
Off-topic	0.35	0.00	0.55	*0.00*
Dialogue-contradicting	0.34	0.16	0.24	0.30
On-topic but unverifiable	0.20	0.23	0.29	0.12

Table 5: Per-label majority agreement ratios for error categories. *Italic*: less than 10 reflections are given this error category by any annotator.

tors of the same group agree with each other, which gauges the internal consistency of the annotators.

4.1.1 *oherent* and *Incoherent*

We first analyse the global agreement on the binary Yes/No (*Coherent*/*Incoherent*) annotation. We adopt both the classical Fleiss' kappa (Fleiss, 1971) and Randolph's fixed-marginal kappa (Randolph, 2005), because 1) Fleiss' kappa is known to be overly penalising when the marginal label distribution is imbalanced (Feinstein and Cicchetti, 1990) and 2) Randolph's kappa is preferable when the annotators have no prior knowledge of the expected label distribution (Assimakopoulos et al., 2020).

As Table 3 shows, Fleiss' kappa in the GPT-2 stage indicates moderate agreement (Landis and Koch, 1977) for both annotator groups, but in the GPT-3 stage it drops to fair agreement for laypeople and almost zero for experts. The drop may appear to suggest a drastic change in agreement, but deeper analysis reveals a considerable cross-stage change of marginal label distribution that may skew Fleiss' kappa — for example, experts annotate GPT-3 reflections as *Coherent* 82% of the time (§4.3.2) as opposed to 38% for GPT-2 reflections. As an evidence, Randolph's kappa, which is not influenced by marginal label distribution, still shows (Table 3) fair agreement among the laypeople and moderate agreement among the experts in the GPT-3 stage.

Beyond global agreement, we conduct more granular analysis on which one of {*Coherent*, *Incoherent*} is easier to agree upon. Specifically, we follow Tsakalidis et al. (2022) to calculate the **per-label majority agreement ratio** (referred to as "agreement ratio" for brevity) for *Coherent* and *Incoherent* separately. For a label l, its agreement ratio $A\ \ (l)$ is:

$$A\ \ (l) = \frac{\text{\#(reflections assigned } l \text{ by 2 annotators)}}{\text{\#(reflections assigned } l \text{ by} \geq 1 \text{ annotators)}}$$

For example, the agreement ratio of *Coherent* is the number of reflections annotated as *Coherent* by 2 out of 3 annotators (hence majority agreement) divided by the number of reflections annotated as *Coherent* by any annotator.

As Table 4 shows, the agreement ratio of *Incoherent* has a minor lead over that of *Coherent* in the GPT-2 stage. In the GPT-3 stage, however, *Coherent* shows substantially higher agreement ratio than *Incoherent*. Therefore, as the LM grows in power (GPT-2→GPT-3), it becomes easier for annotators to agree on what is *Coherent* than on what is not, and this applies to both groups.

We note that, in Tsakalidis et al. (2022), an example is given label l if the agreement ratio of l is above 0.3 and a majority of annotators assign l to the example. Our results show that both laypeople and experts have agreement ratios that are almost always comfortably higher than the 0.3 threshold, particularly w.r.t. *Coherent* (0.66∼0.90). Thus, also considering the global agreements (Table 3), both laypeople and experts appear to be reliable annotators, and a reflection should be considered *Coherent* if a majority of annotators deem it so.

4.1.2 greement on Error Categories

We also measure agreement ratio for each error category to inspect whether some errors are easier than others for annotators to agree upon.

Based on Tables 4 and 5, one can observe that agreement ratio is generally higher for *Incoherent* than for any error category. While it may be inherently more challenging to annotate an error category than to annotate *Coherent*/*Incoherent* due to the label space size difference (5 vs. 2), this is

	Spearman	Pearson
GPT-2	0.741	0.742
GPT-3	0.444	0.446

Table 6: Correlations between laypeople- and experts-based coherence scores. $p <$ 1e-7 for all 4 values.

still a strong indication that it is easier for annotators to agree that a reflection is *Incoherent* than to agree upon any specific incoherence problem.

Interestingly, *Parroting* has clearly higher agreement ratio among laypeople than among experts in both stages, which means some experts are more tolerant of *Parroting* than others but laypeople are similar to each other in this regard.

4.2 Inter-Group Correlation

We also investigate inter-group correlation, namely the correlation between laypeople and experts w.r.t. their annotations. We measure it based on **coherence scores**: given a reflection and the 3 annotators to whom it was assigned, its coherence score is the number of annotators that flagged it as *Coherent*. Thus, a coherence score has a range of {0, 1, 2, 3}, and each reflection has one score from laypeople and one from experts.

As Table 6 shows, inter-group correlation is strong in the GPT-2 stage and moderate in the GPT-3 stage (Prion and Haerling, 2014). Combined with our previous findings on the intra-group agreement on coherence (§4.1.1), this is further evidence that laypeople can be a viable alternative to experts for scaled-up reflection coherence evaluation. In particular, a binary *Coherent*/*Incoherent* setup may be more suitable, since per-label majority agreement ratios are clearly higher on *Coherent* and *Incoherent* than on the error categories (§4.1.2). Nevertheless, the weaker inter-group correlation in the GPT-3 stage does suggest experts-laypeople differences (we probe them further in §4.3), and it also shows that laypeople-based evaluation is relatively more challenging when the reflections come from powerful LLMs.

4.3 Cross-Stage nnotation Changes

We further investigate how reflections — both human and synthetic ones — are annotated differently in different stages. We focus on the distribution of *Coherent*/*Incoherent* labels and error labels based on the results in Figure 3.

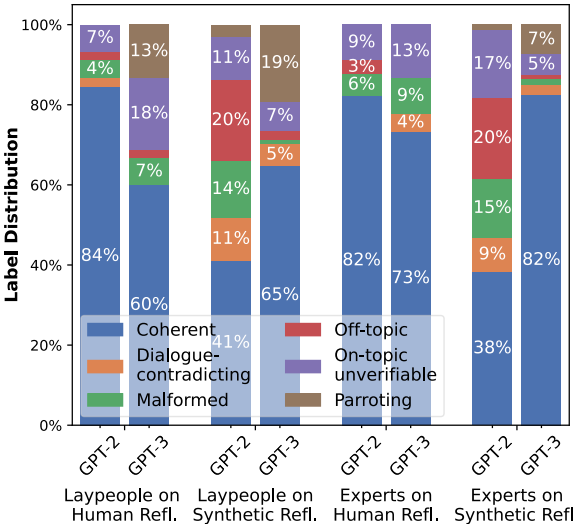

Figure 3: Labels distribution on human and synthetic reflections in the GPT-2 stage and GPT-3 stage. *Incoherent* labels are broken down into fine-grained error categories. Colour-blind-safe and greyscale-safe version is shown in Figure 8.

	ll		Recurrence-Free	
	GPT-2	*GPT-3*	*GPT-2*	*GPT-3*
Laypeople	84%	**60%**	87%	**58%**
Experts	82%	73%	83%	77%

Table 7: *Coherent*/*Incoherent* label distributions for human reflections. We report how often (%) the annotators flag a human reflection as *Coherent*. **Bold**: significant (chi-squared test, p< 0 05) cross-stage shift.

4.3.1 Cross-Stage Shift on Human Reflections

Both laypeople and experts flag human reflections as *Coherent* less often in the GPT-3 stage than in the GPT-2 stage. Therefore, we analyse the distribution of *Coherent* and *Incoherent* labels given to human reflections and examine whether the cross-stage distribution shift is significant. We do so with 2 settings: **ll** and **Recurrence-Free**. **ll** takes into account all the *Coherent* and *Incoherent* labels. **Recurrence-Free** removes the labels from an annotator for a reflection if the reflection is recurring (§3.2.3) for the annotator (i.e., the annotator annotated the reflection in both stages) and therefore removes recurrence-caused annotator bias. As shown in Table 7, under both **ll** and **Recurrence-Free**, both laypeople and experts less often annotate human reflections as *Coherent* in the GPT-3 stage. Notably, the shift of laypeople is significant, while the shift of experts is not.

Beyond the global distribution of *Coherent* and *Incoherent* labels, we also inspect the cross-stage

shift w.r.t. coherence scores (defined in §4.2) of human reflections. With the paired Wilcoxon signed-rank test, we have a similar discovery: laypeople-based coherence scores are significantly ($p < 0\ 05$) lower in the GPT-3 stage than in the GPT-2 stage, while it is not the case for experts.

Also shown in Figure 3, human reflections are clearly more likely ($\geq 11\%$) to be annotated by laypeople as *Parroting* and *On-topic but unverifiable* in the GPT-3 stage. In comparison, error annotations by experts for human reflections are more consistent across stages, with minor ($\leq$ 4%) increases in *On-topic but unverifiable*, *Malformed* and *Dialogue-contradicting*.

Therefore, compared to experts, laypeople are overall more influenced by synthetic reflections when annotating human reflections. This annotation fluidity is a potential concern for laypeople-based scaled-up coherence evaluation.

4.3.2 Cross-Stage Differences on Synthetic Reflections

As Figure 3 shows, GPT-3 reflections are significantly (chi-squared test, $p < 0.05$) more often annotated as *Coherent* than GPT-2 ones by both laypeople and experts, which is not surprising given that GPT-3 is considerably more powerful. Interestingly, while laypeople and experts are similar in *Coherent*/*Incoherent* label distribution for GPT-2, experts are significantly more likely than laypeople to annotate GPT-3 reflections as *Coherent*.

Upon further analysis, we notice that much of the laypeople-experts divide on GPT-3 *Coherent* rate can be attributed to *Parroting*, which is used 19% of the time by laypeople but only 7% by experts. For the other 4 error categories, laypeople are experts behave similarly: the proportion of each category is substantially lower in the GPT-3 stage. This shows that GPT-3 makes most types of incoherence errors less often than GPT-2.

Overall, it is clear that experts are less strict about *Parroting*. This is likely because a reflection summarises what the client said, which may sometimes appear repetitive to a layperson when an expert may consider it good practice. As further evidence, we note that human reflections, which showcase good practice, are not annotated as *Parroting* by experts in either stage, while laypeople do so in the GPT-3 stage (§4.3.1).

4.3.3 Human vs. Synthetic in *oherent* Rate

We compare human and synthetic reflection w.r.t. the proportion of *Coherent* labels[7]. As shown in Figure 3, human reflections are annotated as *Coherent* significantly (chi-squared test, $p < 0.05$) more often than synthetic reflections by both laypeople and experts in the GPT-2 stage. This is not unexpected since human reflections are considered the gold standard. However, the trend is reversed in the GPT-3 stage, even though the lead of GPT-3 over human reflections is not significant. This shows that GPT-3 is capable of producing coherent reflections, and it can even sometimes match or outperform human reflections. It also raises interesting research questions on whether GPT-3 can compete with human reflections on aspects deeper than coherence, such as empathy and adherence to counselling principles.

4.4 Case Study

To gain qualitative insights into the annotations, we show a case study in Table 8 which presents the annotations on the reflections shown in Table 1.

While the human reflection is annotated as *Coherent* by every layperson in the GPT-2 stage, it is flagged by 2 laypeople as *Parroting* in the GPT-3 stage, which may be because those 2 laypeople found the human reflection to be a rephrase of the last client utterance (e.g., "it's your stuff" in the human reflection compared to "it's my stuff" in the client utterance). Notably, this example echoes the overall trend that human reflections are more likely (0%→13%) to be flagged by laypeople as *Parroting* in the GPT-3 stage (§4.3.1).

On the other hand, the human reflection is annotated as *Coherent* by every expert in the GPT-2 stage, but it is flagged by 1 expert as *Malformed* in the GPT-3 stage. We postulate that the fluency of GPT-3 reflections may make the human reflection appear less fluent to some annotators. This may be particularly true when there are faithfully transcribed self-repetitions and mid-sentence changes ("it's-that's-it's your stuff") in the human reflection, even though we explicitly informed the annotators that those are normal.

For comparison, we also analyse the annotations on the examples of GPT-2 and GPT-3 synthetic reflections. The GPT-2 reflection roughly matches the mood of the client but is also generic,

[7]We do not compare at the granular error-category-level due to the different scales of human and synthetic reflections.

Context
Client: Well, I'm here because my mom wants me to be here. **Therapist**: Mm-hmm. **Client**: I don't really wanna be here, but it-it– whatever. **Therapist**: Got it. **Client**: Um, she-she found my stash- **Therapist**: Uh-huh. **Client**: -and she freaked out, and she's going crazy over it. Um, I don't why she was going through stuff in the first place, but whatever, so, now I'm here. **Therapist**: Mm-hmm. **Client**: Um, I've been hanging out with a new cool crowd of people that I really like. **Therapist**: Mm-hmm. **Client**: Uh, a-and-and that's-that's basically it. **Therapist**: Yeah. So-so you've got this new group of friends and-and, um, you-you actually kind of like where you're at with things right now. And your mom was going through your stuff and found your stash, and it's just turned into a, you know, all of this. **Client**: Yeah. **Therapist**: Yeah. **Client**: Yeah, basically. **Therapist**: Yeah. **Client**: Mm-hmm. **Therapist**: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that? **Client**: I am. **Therapist**: Yeah. **Client**: I mean, it's my stuff, I don't know why she's–

Therapist (Human): Right. It's like your private place and, you know, it's– that's– it's your stuff.			
GPT-2 Stage Annotation			
L2	*Coherent*	**E2**	*Coherent*
L3	*Coherent*	**E7**	*Coherent*
L7	*Coherent*	**E8**	*Coherent*
GPT-3 Stage Annotation			
L1	*Coherent*	**E3**	*Coherent*
L4	*Parroting*	**E4**	*Coherent*
L7	*Parroting*	**E5**	*Malformed*

GPT-2: It's a very sad thing.			
L2	*Coherent*	**E2**	*Coherent*
L3	*Coherent*	**E7**	*Off-topic*
L7	*Off-topic*	**E8**	*Off-topic*

GPT-3: It sounds like you're really upset with her because she invaded your privacy.			
L1	*Coherent*	**E3**	*Coherent*
L4	*Coherent*	**E4**	*Coherent*
L7	*Coherent*	**E5**	*Coherent*

Table 8: The complete dialogue context of Table 1 and annotations on reflection examples. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category. ‡: Annotator annotated the human reflection in both stages.

and it is annotated as *Off-topic* by 1 layperson and 2 experts. On the other hand, the GPT-3 reflection is fluent and more specific to the dialogue, and unsurprisingly it is annotated as *Coherent* by all 6 annotators. While those two reflections cannot cover all of the variety of synthetic reflections, their qualitative difference w.r.t. the human reflection is a good example for showing why annotators may be influenced by the surrounding synthetic reflections when they are annotating a human reflection.

5 Conclusion

In this work, we probed whether laypeople can be used as an alternative to experts in evaluating coherence and context-consistency of counselling reflection generation. Accordingly, we asked both laypeople and experts to annotate synthetic reflections generated by LMs and human reflections from actual therapists. We found that both laypeople and experts are reliable annotators and that they also show moderate to strong inter-group correlation, which is the first concrete evidence that laypeople are capable of such annotations, although laypeople are relatively less aligned with experts on GPT-3 reflections. Furthermore, we found that GPT-3 is mostly able to generate coherent and consistent reflections, and we also explored the annotation shift on human reflections when the source of synthetic reflections changes from the smaller GPT-2 to the more powerful GPT-3.

For future work, we plan to mix, in each batch, synthetic reflections from models of different scales, and investigate how the resulting human evaluations might differ. Another direction worth exploring is alternative ways of coherence annotation, such as ranking, for more nuanced human evaluation results. Future work may also re-examine and modify the error categories to increase IAA on error annotations. We also leave potentially IAA-improving annotation procedures to future work, such as using a warm-up exercise task before actual annotation and allowing annotators to discuss with each other to resolve their differences.

Limitations

The main limitation of this work is the quantity of annotated human reflections. Overall, 15 human reflections are annotated, which are outnumbered more than 7:1 by GPT-2 reflections and 9:1 by GPT-3 reflections. If there were more human reflections annotated, we may be able to confirm, among other potential findings, that GPT-3 reflections were indeed significantly more often annotated as *Coherent* compared to

human reflections.

We also note that the laypeople had a longer between-stage waiting period than the experts, because we could not enforce a similarly long waiting period for the experts due to practical reasons (Appendix C). While an ideal setup would keep the same waiting period duration, Appendices C and D show that the duration difference is not critical.

Furthermore, we adopted sequential annotation for reflections within a batch to make the interface easier to navigate for the human annotators, but this also means that the early samples in a batch might indirectly affect the annotation of the later samples. We leave more investigation on this to future work.

Ethics Statement

In this section, we briefly discuss the ethical aspects of our experiments. We do this with regard to our experiment as a whole.

Ethical Review

Prior to our experiment, materials and methodology underwent ethical review by our institution's Ethics Board. The proposal was flagged as ethically compliant and accepted without major revisions.

Risks

Our work inspects the annotation differences between laypeople and experts in the counselling domain (MI and reflections in particular). With these premises, it could be seen as a message that therapy can be fully automated, laypeople can replace therapists in creating such systems and generative models could act as "virtual counsellors". We acknowledge that past work inspected similar options (Fiske et al., 2019; D'Alfonso, 2020; Saha et al., 2022), but we take distance from it. Our work is framed as modelling technological advancements that are solely directed at therapist training. We foresee the use of neural NLG as promising in counselling, but only for supporting trainees. We also point out previous work showing why replacing mental health practices with language models (or AI in general) should not be considered (Le Glaz et al., 2021).

Information and Consent

Prior to starting the annotation, both laypeople and experts received an electronic information sheet containing details on the task, purpose of research, workload and pay. This also included the fact that data would be made available for future research, in accordance with data anonymisation requirements.

Upon starting the annotation, annotators were prompted with a mandatory consent form to confirm their understanding of the terms and conditions and their willingness to take part in the annotation. Annotators were also given an email contact in case of problems during the annotation or any other query. Annotators were automatically prevented from doing the annotation if they did not provide consent.

Demographic Information of nnotators

All annotators were highly proficient in English, which is the language of the dialogues. 5 out of the 9 laypeople were based in the Netherlands while the other 4 resided in Italy. Among the experts, 4 were based in the UK, 1 in the Netherlands, 1 in Hungary, 1 in Italy and 2 in Sweden.

We recruited laypeople who were known to us, as this allowed active monitoring of the annotation task, hence ensuring high quality. While this approach is different from other standard ones (such as using crowdsourcing platforms), we argue that the focus of this work is to understand if fully committed laypeople can be valid annotators, which can be challenging considering the annotation quality issues that crowdsourcing platforms suffer from (Dennis et al., 2020).

We also note that the group of laypeople is diverse in demographics and educational backgrounds. Specifically, the group includes people of 5 nationalities in their 20s, 30s and 40s who range from bachelor's student to professional with a PhD.

To verify the generalisability of our laypeople-based evaluation, future work may replicate our setup on crowdworkers and compare the resulting annotations with ours.

Remuneration

The annotation workload was made explicit in the task (a total of 5 annotation batches in each stage, with a detailed description of what a batch consists of). Annotators were given 30 minutes to complete each annotation batch: laypeople received 19.5 USD/h, while experts received 21.6 USD/hour. This difference is motivated by the generally higher hourly pay of experts. The remuneration is considerably (>50%) higher than the minimum wage levels of the countries of residence of the annotators. It also took most annotators much less than 30 minutes (e.g., 10 to 15 minutes) to complete

a batch, so the effective hourly remuneration was higher than 19.5/21.6 USD.

Data nonymisation

No personal data about the annotators was kept stored at the end of the experiment. During the annotation process, no annotator ever got in touch with anyone involved in the experiments except for the researchers.

cknowledgements

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project (contract no. 812882) and the European Research Council (Grant agreement No. 101039303 NG-NLG). We also thank Craig Thomson and Vivek Kumar for their suggestions.

References

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

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France. European Language Resources Association.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. 2022. Zero-shot opinion summarization with GPT-3. *CoRR*, abs/2211.15914.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *dvances in Neural Information Processing Systems 33: nnual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sean A Dennis, Brian M Goodson, and Christopher A Pearson. 2020. Online worker fraud and evolving threats to the integrity of mturk data: A discussion of virtual private servers and the limitations of ip-based screening procedures. *Behavioral Research in ccounting*, 32(1):119–134.

Jan Deriu, lvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *rtif. Intell. Rev.*, 54(1):755–810.

Simon D'Alfonso. 2020. Ai in mental health. *Current Opinion in Psychology*, 36:112–117. Cyberpsychology.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. ssoc. Comput. Linguistics*, 9:391–409.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res*, 21(5):e13216.

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

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. ssoc. Comput. Linguistics*, 9:1460–1474.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the 2010 Workshop on Creating Speech and Language Data with mazon's Mechanical Turk, Los ngeles, US , June 6, 2010*, pages 148–151. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, ddis baba, Ethiopia, pril 26-30, 2020*. OpenReview.net.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8932–8939. Association for Computational Linguistics.

Jessica Huynh, Jeffrey Bigham, and Maxine Eskénazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *CoRR*, abs/2111.05241.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CM Comput. Surv.* Just Accepted.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1265–1285. Association for Computational Linguistics.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2227–2240. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th nnual Meeting of the ssociation for Computational Linguistics (Volume 1: Long Papers), CL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

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 nnual Meeting of the ssociation for Computational Linguistics, CL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4454–4466. International Committee on Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael D. 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, EMNLP 2016, ustin, Texas, US , November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Juliana Miehle, Nadine Gerstenlauer, Daniel Ostler, Hubertus Feußner, Wolfgang Minker, and Stefan Ultes. 2018. Expert evaluation of a spoken dialogue system in a clinical operating room. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. lbuquerque: Center on lcoholism, Substance buse and ddictions, University of New Mexico*.

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

TB Moyers, JK Manuel, D Ernst, T Moyers, J Manuel, D Ernst, and C Fortini. 2014. Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th nnual Meeting of the ssociation for Computational Linguistics, July 6-12, 2002, Philadelphia, P , US* , pages 311–318. ACL.

Susan Prion and Katie Anne Haerling. 2014. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10(10):535–536.

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

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

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 ssociation for Computational Linguistics, CL 2019, Florence, Italy, July 28- ugust 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. *merican 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, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the ssociation for Computational Linguistics: Main Volume, E CL 2021, Online, pril 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

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

Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies*, pages 2436–2449, Seattle, United States. Association for Computational Linguistics.

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 *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, pril 19-23, 2021*, pages 194–205. ACM / IW3C2.

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 nnual Meeting of the ssociation for Computational Linguistics (Volume 1: Long Papers), CL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3096–3107. Association for Computational Linguistics.

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 nnual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 10–20. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*, abs/2208.03188.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational I, Conv I@ CL 2022, Dublin, Ireland, May 27, 2022*, pages 77–97. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, US , meeting of SIGD T, a Special Interest Group of the CL*, pages 254–263. ACL.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *dvances in Neural Information Processing Systems 33: nnual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze

Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th nnual Meeting of the ssociation for Computational Linguistics (Volume 1: Long Papers), CL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4647–4660. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2022a. Towards in-context non-expert evaluation of reflection generation for counselling conversations. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 116–124, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022b. Anno-mi: A dataset of expert-annotated counselling dialogues. In *IEEE International Conference on coustics, Speech and Signal Processing, IC SSP 2022, Virtual and Singapore, 23-27 May 2022*, 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, CoNLL 2021, Online, November 10-11, 2021*, pages 133–147. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International CM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, nn rbor, MI, US , July 08-12, 2018*, pages 1097–1100. ACM.

Input
⟨client⟩Well, I'm here because my mom wants me to be here.| ⟨therapist⟩Mm-hmm.| . . . | ⟨therapist⟩Yeah.| ⟨client⟩I mean, it's my stuff, I don't know why she's–| ⟨therapist⟩ ⟨listening⟩
Output
Right. It's like your private place and, you know, it's– that's– it's your stuff.

Table 9: Input and output format of fine-tuned models for the ⟨context, human reflection⟩ pair shown in Table 1.

Context
Client: My mother is driving me crazy. She says she wants to remain independent, but she calls me four times a day with trivial questions. Then she gets mad when I give her advice.
Simple Reflection
Therapist: Things are very stressful with your mother.
Complex Reflection
Therapist: You're having a hard time figuring out what your mother really wants.

Table 10: Examples of simple and complex reflections from Moyers et al. (2014).

Modelling & Computation Details

.1 Fine-Tuning

We convert the input dialogue context into a plain-text sequence of utterances with interlocutor labels and utterance separators in between, while the output reflection is simply plain text without special preprocessing. An example is shown in Table 9, which formats the ⟨context, human reflection⟩ pair of Table 1 accordingly. In particular, the " ⟨*listening*⟩" is the cue for the LM to start generating a reflection.

For training, we first divide the 1265 ⟨context, human reflection⟩ pairs into 10 folds, and we then fine-tune the same pre-trained model 10 times independently to generate synthetic reflections for the pairs in each test fold. Each time when fine-tuning a model, we use 8 folds as the training data, 1 as validation data and 1 as test data. We allot pairs from the same dialogue to the same fold in order to avoid overlap between training/validation/test data.

Our experiments are based on the HuggingFace package[8]. We adopt the pre-trained `gpt2-medium` (345M parameters). We use 2e-5 as the learning rate for training, based on a hyperparameter search over different learning rates where the metric is perplexity. The other hyperparameters are fixed, including 8 as the batch size and 42 as the random seed. The fine-tuning stops when perplexity has not improved on the validation data for 3 epochs. We ran the fine-tuning on an NVIDIA V100 GPU (16GB). In total, the fine-tuning and inference took under 50 GPU hours.

.2 Prompting

We used `text-davinci-002`, the largest GPT-3 variant (175B parameters) at the time of experiment. The total cost of generation during the GPT-3 stage was 23.68 US Dollars.

.2.1 Prompting with Textbook Examples

As learning examples, textbook examples are different from `AnnoMI` examples in that 1) textbook examples are written texts instead of transcripts like `AnnoMI`, and 2) the context in a textbook example is considerably shorter than the average `AnnoMI` context which contains 14 utterances.

A simple reflection typically repeats or rephrases what the client has said, while a complex one adds substantial meaning or emphasis and communicates a deeper or richer picture of the client's statement (Miller et al., 2003). An example is shown in Table 10.

A prompt (Figure 2b) begins with an instruction, followed by 8 textbook examples and the test example placed at the end. Thus, the model is prompted to generate 2 synthetic reflections, one simple and the other complex. Considering recent studies (e.g., Zhao et al., 2021) about the impact of few-shot example ordering on the output, we create 3 prompts to generate 3 different sets of {simple reflection, complex reflection}, where the textbook examples in each prompt are identical but with different random orders.

.2.2 Prompting with `nnoMI` Examples

In this prompting method, we do not take simple/complex reflection into account, because human reflections in `AnnoMI` do not have such labels. Similar to prompting with textbook examples, we construct 3 prompts for each test example in order to obtain diverse GPT-3-generated reflections. The difference from prompting with textbook is that we create those 3 prompts by sampling 3 different sets of learning examples instead of shuffling.

[8] https://huggingface.co/

Therefore, the learning example set in each of the 3 prompts is unique, and to ensure fairness the learning examples are not from the same dialogues as the test example.

B Reflection Sampling for nnotation & Inadequacy of B RT

As mentioned briefly in the main body (Footnotes 2 and 5), human evaluation in the GPT-2 stage included both GPT-2 reflections and BART reflections in practice, since we wanted to diversify synthetic reflections from smaller LMs in the GPT-2 stage. For BART, we fine-tuned the pretrained `bart-large` (406M parameters, similar in scale to `gpt2-medium`) in the exact same way we fine-tuned GPT-2, and we also used the same decoding methods for test-time generation.

Overall, for the context in each of the 15 sampled ⟨context, human reflection⟩ pair, we generated 26 synthetic reflections in total with GPT-2, 26 with BART and 36 with GPT-3. In order to ensure smaller LMs and large LMs were equally present in the human annotation of synthetic reflections, we randomly sampled (Appendix B.1) 9 semantically distinct reflections from the 52 GPT-2/BART reflections and also 9 from the 36 GPT-3 reflections for human annotation.

Thus, for each ⟨context, human reflection⟩ pair, we created 2 annotation batches that each contained the context, the human reflection and 9 synthetic reflections. The two batches differed in that the synthetic reflections in one batch came from GPT-2 and BART while those in the other batch were from GPT-3. Both batches were later annotated (§3.2). In other words, GPT-2 and BART reflections were annotated together in the GPT-2 stage. However, BART reflections were vastly outnumbered by GPT-2 and GPT-3 reflections because they were sampled less frequently due to a lack of diversity (Appendix B.2), so we reported only GPT-2 and GPT-3 in the main body for fairness.

Nevertheless, we analyse the annotations on BART-generated synthetic reflections in Appendix B.3, but we note that it is limited by the small quantity of BART reflections and therefore in particular should not be used to compare with the findings w.r.t. GPT-2 and GPT-3 reflections.

B.1 Reflection Sampling Procedure

We grouped reflections through semantic clustering based on their embeddings[9], such that the reflections in each cluster were semantically almost identical. For example, if two reflections were identical except that one had a "Hmm." at the beginning while the other did not, they were grouped into the same cluster. Afterwards, we randomly sampled 9 clusters from all the GPT-2 and BART reflection clusters, and we similarly sampled 9 GPT-3 reflection clusters. Finally, we drew from each cluster the reflection with the most tokens, deeming it as the most semantically rich.

B.2 Lack of Diversity mong B RT Reflections

While we generated the same number (26) of GPT-2 and BART reflections for sampling, in practice there was a considerable lack of diversity among BART reflections that led to them being grouped into fewer clusters and therefore less frequently sampled. Specifically, GPT-2 reflections outnumbered BART reflections 4:1, which means the overall BART:GPT-2:GPT-3 reflection quantity ratio was 1:4:5. Therefore, to ensure fairness, we only reported GPT-2 and GPT-3 reflections in the main body, considering their similar quantities.

To illustrate the lack of diversity among BART reflections, we measure the lexical and semantic diversity of synthetic reflections from GPT-2/BART/GPT-3 using Self-BLEU (Zhu et al., 2018) and average pairwise semantic similarity, respectively.

Self-BLEU is based on BLEU (Papineni et al., 2002) which measures the lexical similarity between two sentences at the n-gram level ($n \in \{1, 2, \quad\}$). Self-BLEU takes all pairs of generated texts (in our case, reflections for the same context), calculates the BLEU score for each pair, and averages the pairwise BLEU scores. Thus, lower Self-BLEU indicates higher diversity among the generated texts. We follow (Zhu et al., 2018) in reporting 2-, 3-, 4-, and 5-gram-level Self-BLEU[10] for BART, GPT-2 and GPT-3 reflections in Table 11. Clearly, BART reflections are substantially more homogeneous than those from GPT-2 and GPT-3. For example, Self-BLEU-4 of BART is at 40.70,

[9] We used the SentenceTransformers package (`https://www.sbert.net/`) and `all-mpnet-base-v2` (Song et al., 2020) as the embedding model.

[10] We calculate Self-BLEU based on the NLTK (`https://www.nltk.org/`) implementation of BLEU.

	B RT	GPT-2	GPT-3
Lexical Diversity			
Self-BLEU-2	48.63	8.44	17.74
Self-BLEU-3	44.36	5.77	14.10
Self-BLEU-4	40.70	4.49	12.02
Self-BLEU-5	37.38	3.75	10.55
Semantic Diversity			
vg. Pairwise Cos. Sim.	0.6952	0.3034	0.4666

Table 11: Overview of lexical (Self-BLEU) & semantic (averaged pairwise cosine similarity) diversity among reflections generated by different models. Lower values indicate more diversity.

	Laypeople	Experts
Coherent	38.1%	77.4%
Dialogue-contradicting	1.8%	3.6%
Malformed	1.8%	0.6%
Off-topic	3.0%	2.4%
On-topic but unverifiable	13.7%	3.6%
Parroting	41.7%	12.5%

Table 12: Label distribution for BART-generated reflections.

compared to the drastically lower 4.49 of GPT-2 and 12.02 of GPT-3.

To compute average pairwise cosine similarity, we 1) compute the cosine similarity between the embeddings (from the same embedding model used for clustering) of the two sequences in each pair of generated reflections for the same context, and then 2) average the similarities of all pairs. As shown in Table 11, the semantic similarity between BART reflections is also considerably higher compared to GPT-2 and GPT-3.

B.3 Label Distribution for B RT Reflections

We show in Table 12 the distribution of labels given to BART reflections. Notably, laypeople and experts show considerable difference ($= 39\%$) in the proportion of *Coherent* labels, which is substantially higher compared to GPT-2 ($= 3\%$) and GPT-3 ($= 17\%$) shown in Figure 3.

Upon further analysis, it is clear that most of the laypeople-experts divide in coherence annotation can be attributed to *Parroting*, which is used considerably more ($= 29\%$) by laypeople than experts. This again echoes the observation in §4.3.2 that laypeople are more strict about *Parroting* than experts.

Qualitatively, Table 13 shows the BART reflection for the case study dialogue (Table 8), which

Context
(intermediate turns)
Therapist: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that? **Client**: I am. **Therapist**: Yeah. **Client**: I mean, it's my stuff, I don't know why she's–

B RT: Okay. So, it's your stuff.			
L1	*Parroting*	**E3**	*Coherent*
L4	*Parroting*	**E4**	*Coherent*
L7	*Parroting*	**E5**	*Coherent*

Table 13: BART-generated reflection for the case study dialogue (Table 8) and its annotations. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category.

clearly mirrors the last client utterance. Matching the trend discussed above, the reflection is annotated by every layperson as *Parroting* but by every expert as *Coherent*.

This finding, together with the low diversity among BART reflections (Appendix B.2), shows that BART has a higher tendency to repeat or rephrase a part of the dialogue context and does not show considerable deviation from this pattern under different decoding parameters. Empirically, this is also our observation of BART reflections in general.

C Waiting Period Between Stages

Initially, we conducted the ⟨Laypeople, GPT-2 stage⟩. We then collected GPT-3-generated reflections and invited the same laypeople for the GPT-3 stage annotation. As those two stages were not planned together, there was about a one-month period in between.

Upon discovering the shifting human reflection annotations (§4.3.1) from the laypeople's results, we recruited the experts to investigate whether the phenomenon was limited to laypeople. Due to time constraint, we were only able to enforce a minimum waiting period of 3 days between the two stages for the experts.

The mean and standard deviation of the waiting period lengths of each annotator group are shown in Table 14. Overall, laypeople had a 39-day gap between the two stages while experts had 7 days.

To probe whether the waiting period difference had an effect, we requested the annotators to fill out a post-annotation questionnaire, where we asked the question "While you were annotating in Phase

	Mean	Standard Deviation
Laypeople	39.1	7.8
Experts	6.9	3.1

Table 14: Waiting period lengths (number of days) between the two stages.

	Yes	No	Maybe
Laypeople	3	1	3
Experts	3	3	1

Table 15: Answers given to the post-annotation question "While you were annotating in Phase 2 (i.e., GPT-3 stage), did you remember seeing any response candidate that you had seen in Phase 1 (i.e., GPT-2 stage)?".

How Often Each nnotator Flags a *Recurring* Human Reflection Identically in Both Stages			
L1	100%	**E1**	100%
L2	100%	**E2**	100%
L3	100%	**E3**	50%†
L4	N/A	**E4**	0%†
L5	100%	**E5**	50%†
L6	50%†	**E6**	N/A
L7	33%†	**E7**	100%
L8	100%	**E8**	100%
L9	50%†	**E9**	67%†
ll	71%	**ll**	73%

Table 16: Overview of how often each layperson (L1∼L9) and each expert (E1∼E9) flags a **recurring** human reflection identically in both the GPT-2 stage and the GPT-3 stage. **N/** : annotator has no recurring human reflections in workload. : annotator does NOT always flag recurring human reflections identically in both stages.

2 (i.e., GPT-3 stage), did you remember seeing any response candidate that you had seen in Phase 1 (i.e., GPT-2 stage)?". We received 7 valid responses from the 8 laypeople who had annotated recurring human reflections, and similarly 7 from the 8 experts that had had recurring human reflections in their workload. Their answers are shown in Table 15.

Clearly, the same number (3) of experts and laypeople remembered seeing recurring human reflections in the GPT-3 stage, but 3 experts answered "No" while 3 laypeople answered "Maybe", which is not surprising since the longer waiting period may have caused more laypeople not to be able to recall exactly. Nevertheless, the fact that the same number of experts and laypeople are positive about seeing recurring human reflections shows that the waiting period for experts was not overly short and may have in fact been sufficient. This is further evidenced by the finding (Appendix D) that laypeople and experts are similarly consistent in annotating recurring human reflections.

D Shifts of Individual nnotators

In §4.3.1, we showed that laypeople and experts as annotator groups are less likely to annotate human reflections as coherent in the GPT-3 stage. In this section, we further inspect whether each layperson/expert annotates human reflections consistently across stages. Since the workload of each annotator consists of **non-recurring** human reflections (appearing in only one stage) and sometimes also **recurring** human reflections (appearing in both stages), we probe the shift of each annotator on these two types of human reflections separately.

We first examine how often each annotator flags recurring human reflections identically (namely choosing "Yes" in both stages or "No" in both) across stages. As shown in Table 16, 8 laypeople and 8 experts have recurring human reflections in their workload. Among those annotators, 3 laypeople and 4 experts fail to annotate all (100%) recurring human reflections identically across stages. Overall, laypeople and experts annotate recurring human reflections identically 71% and 73% of the time, respectively. Those similar numbers are evidence that the laypeople-experts difference in the between-phase waiting period duration (Appendix C) is not critical.

Then, we investigate whether each annotator flags non-recurring human reflections more, equally or less often as *Coherent* in the GPT-3 stage than in the GPT-2 stage. As table 17 shows, 5 laypeople less often annotate non-recurring human reflections as *Coherent* in the GPT-3 stage, 1 does so more often, while the other 3 stay at the same level across stages. Among the experts, 4 give *Coherent* annotations less often, 2 do so more often, while the remaining 3 do not show cross-stage frequency change, which is a similar distribution compared to laypeople. Considering that laypeople and experts have different levels of overall cross-stage shift on non-recurring reflections — 29% for laypeople and 6% for ex-

How Often Each nnotator Flags a *Non-Recurring* Human Reflection as *oherent*					
	GPT-2	*GPT-3*		*GPT-2*	*GPT-3*
L1	100%	50%↓	**E1**	100%	100%
L2	100%	100%	**E2**	100%	75%↓
L3	100%	50%↓	**E3**	100%	67%↓
L4	100%	20%↓	**E4**	75%	100%↑
L5	100%	100%	**E5**	67%	67%
L6	67%	67%	**E6**	80%	60%↓
L7	100%	50%↓	**E7**	75%	100%↑
L8	60%	25%↓	**E8**	67%	67%
L9	67%	100%↑	**E9**	100%	50%↓
ll	87%	58%↓	**ll**	83%	77%↓

Table 17: Overview of how often each layperson (L1 - L9) and each expert (E1 - E9) annotates a **non-recurring** human reflection as *Coherent* in the GPT-2 stage and GPT-3 stage. ↑/↓: increase/decrease in the GPT-3 stage compared to the GPT-2 stage.

perts — we posit that laypeople and experts differ less in the proportion of "shifting" annotators but more in the magnitude of shifts displayed by individual annotators.

E Label Distribution for Differently Generated Synthetic Reflections

Table 18 shows the distribution of *Coherent* and error labels for synthetic reflections from GPT-2 and GPT-3 under different generation settings.

For GPT-2 reflections, larger p values in nucleus decoding cause less coherent reflections, especially when $p \in \{0\ 8, 0\ 95\}$. This is unsurprising, since larger p's give the model more freedom in generation and thus also make it more prone to errors.

For GPT-3, reflections generated through textbook-based in-context learning are overall less coherent than reflections generated through AnnoMI-based in-context learning. This is not surprising, since test examples themselves are from AnnoMI, which means examples from AnnoMI are more useful in helping the model learn to produce coherent reflections for long dialogue contexts.

Among reflections from GPT-3 (textbook), simple reflections are overall more often annotated as *Parroting* than complex ones, especially by laypeople. This is likely because simple reflections mostly repeat/rephrase what the client said, which may appear repetitive to a layperson when an expert would more likely consider it good practice (§4.3.2).

Finally, we note that the *Coherent* rates of GPT-3 reflections can vary considerably under different nucleus decoding p's but without a clear trend, which we leave to future work to probe.

F Data Use & Creation

We leveraged AnnoMI, a dataset available under the Public Domain license. We used it for research purposes, which is consistent with its intended use. While AnnoMI contains therapy dialogues, the data does not reveal personal information since the dialogues are transcripts of professionally produced MI demonstrations. The dataset does not reveal demographic information, but we observe that the dialogues seem to be set in English-speaking countries.

Based on AnnoMI, we created a dataset of human annotations w.r.t. coherence of reflections, and we release it[11] under the CC BY-NC license, which is also compatible with the access conditions of AnnoMI. The human annotations do not reveal any information of the laypeople or experts, and we use L1∼9 to represent the 9 laypeople and E1∼9 to represent the 9 experts. We discussed the demographic information of the annotators in the Ethics Statement.

G nnotation Flow

In practice, each annotation batch contained some parts that are not investigated in this study, which are therefore not shown in the main body. The complete annotation flow is detailed below.

As shown in Figure 4, a batch starts with the annotator reading the context. Then, the annotator reads one reflection and chooses Yes/No regarding whether it is coherent and context-consistent. If the answer is Yes, the annotator assesses the level of empathy displayed in the reflection. If the answer is No, the annotator selects one or more error categories that apply, and in the case of multiple selected errors the annotator further pinpoints the most evident one. Afterwards, the annotator proceeds to annotate the next reflection in the same steps, and the batch ends when all its reflections have been annotated.

H nnotation Interface

The annotation process takes place in the Mechanical Turk Sandbox[12]. Details of the annotation interface are shown in Figures 5, 6 and 7. We note that there is a purposely off-topic reflection in each

[11] Available at https://github.com/uccollab/expert_laypeople_reflection_annotation.

[12] https://workersandbox.mturk.com/

GPT-2 Using Greedy and Beam Decoding

	Greedy		eam Search	
	Laypeople	**Experts**	**Laypeople**	**Experts**
Coherent	50.0%	66.7%	50.0%	44.4%
Dialogue-contradicting	16.7%	0.0%	27.8%	27.8%
Malformed	8.3%	0.0%	0.0%	11.1%
Off-topic	25.0%	0.0%	2.8%	0.0%
On-topic but unverifiable	0.0%	33.3%	8.3%	0.0%
Parroting	0.0%	0.0%	11.1%	16.7%

GPT-2 Using Nucleus Decoding

	Nucleus ($p = .4$)		Nucleus ($p = .6$)		Nucleus ($p = .8$)		Nucleus ($p = .95$)	
	Laypeople	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**
Coherent	56.1%	54.5%	52.4%	54.8%	31.8%	21.2%	22.2%	18.5%
Dialogue-contradicting	12.1%	7.6%	6.5%	11.3%	8.1%	6.1%	11.8%	4.9%
Malformed	4.5%	5.3%	6.5%	3.6%	21.0%	25.8%	27.9%	26.5%
Off-topic	12.9%	16.7%	13.7%	8.3%	28.0%	25.0%	30.3%	37.7%
On-topic but unverifiable	10.6%	14.4%	16.1%	21.4%	9.6%	22.0%	7.7%	12.3%
Parroting	3.8%	1.5%	4.8%	0.6%	1.5%	0.0%	0.0%	0.0%

Simple Reflections From GPT-3, Using Textbook Examples for In-Context Learning

	Nucleus ($p = .4$)		Nucleus ($p = .6$)		Nucleus ($p = .8$)		Nucleus ($p = .95$)	
	Laypeople	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**
Coherent	38.5%	74.4%	40.5%	66.7%	37.5%	72.9%	59.5%	83.3%
Dialogue-contradicting	5.1%	3.8%	7.1%	0.0%	4.2%	0.0%	2.4%	0.0%
Malformed	5.1%	2.6%	2.4%	0.0%	0.0%	1.0%	1.2%	0.0%
Off-topic	10.3%	0.0%	2.4%	2.4%	4.2%	1.0%	4.8%	0.0%
On-topic but unverifiable	0.0%	1.3%	4.8%	7.1%	2.1%	4.2%	13.1%	7.1%
Parroting	41.0%	17.9%	42.9%	23.8%	52.1%	20.8%	19.0%	9.5%

Complex Reflections From GPT-3, Using Textbook Examples for In-Context Learning

	Nucleus ($p = .4$)		Nucleus ($p = .6$)		Nucleus ($p = .8$)		Nucleus ($p = .95$)	
	Laypeople	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**
Coherent	73.3%	75.6%	66.7%	82.2%	57.8%	80.0%	47.6%	90.5%
Dialogue-contradicting	12.2%	11.1%	2.2%	0.0%	2.2%	3.3%	0.0%	0.0%
Malformed	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.0%
Off-topic	2.2%	0.0%	0.0%	0.0%	2.2%	0.0%	0.0%	0.0%
On-topic but unverifiable	3.3%	6.7%	15.6%	11.1%	13.3%	4.4%	9.5%	4.8%
Parroting	8.9%	6.7%	15.6%	6.7%	24.4%	12.2%	38.1%	4.8%

Reflections From GPT-3, Using `nnoMI` Examples for In-Context Learning

	Nucleus ($p = .4$)		Nucleus ($p = .6$)		Nucleus ($p = .8$)		Nucleus ($p = .95$)	
	Laypeople	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**	**Laypeople**	**Experts**
Coherent	75.6%	86.7%	85.7%	92.9%	95.2%	95.2%	82.1%	89.7%
Dialogue-contradicting	6.7%	4.4%	2.4%	0.0%	0.0%	0.0%	3.8%	0.0%
Malformed	0.0%	4.4%	0.0%	4.8%	0.0%	4.8%	0.0%	2.6%
Off-topic	0.0%	2.2%	0.0%	0.0%	0.0%	0.0%	2.6%	2.6%
On-topic but unverifiable	8.9%	2.2%	7.1%	2.4%	0.0%	0.0%	2.6%	2.6%
Parroting	8.9%	0.0%	4.8%	0.0%	4.8%	0.0%	9.0%	2.6%

Table 18: Label distribution on synthetic reflections from GPT-2 and GPT-3 under different generation settings.

batch as an anti-scam mechanism, which is why there appear to be 11 reflections instead of 10 to annotate in those figures.

I Colour-Blind-Safe and Greyscale-Safe Version of Figure 3

Figure 8 shows the colour-blind-safe and greyscale-safe version of Figure 3.

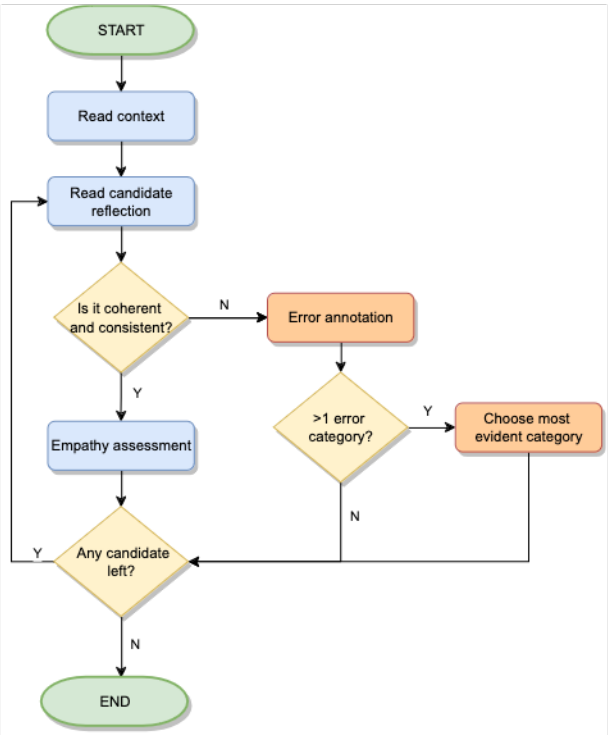

Figure 4: Annotation flow for one batch. Note that in this work we do not investigate annotations w.r.t. empathy assessment or the most evident error category.

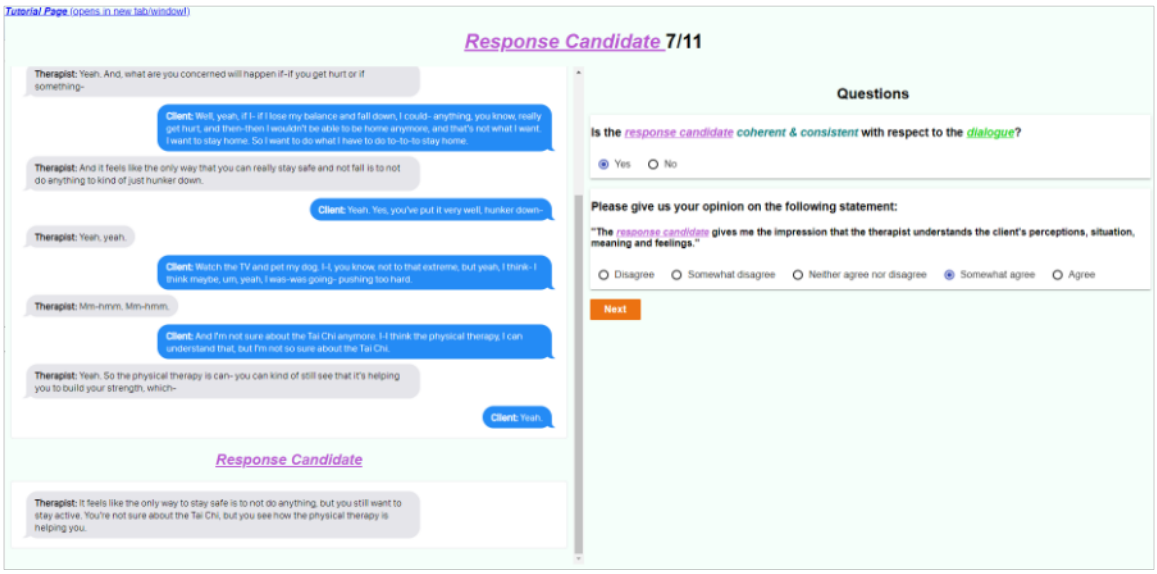

Figure 5: Annotation interface when the annotator annotates a reflection as coherent & consistent.

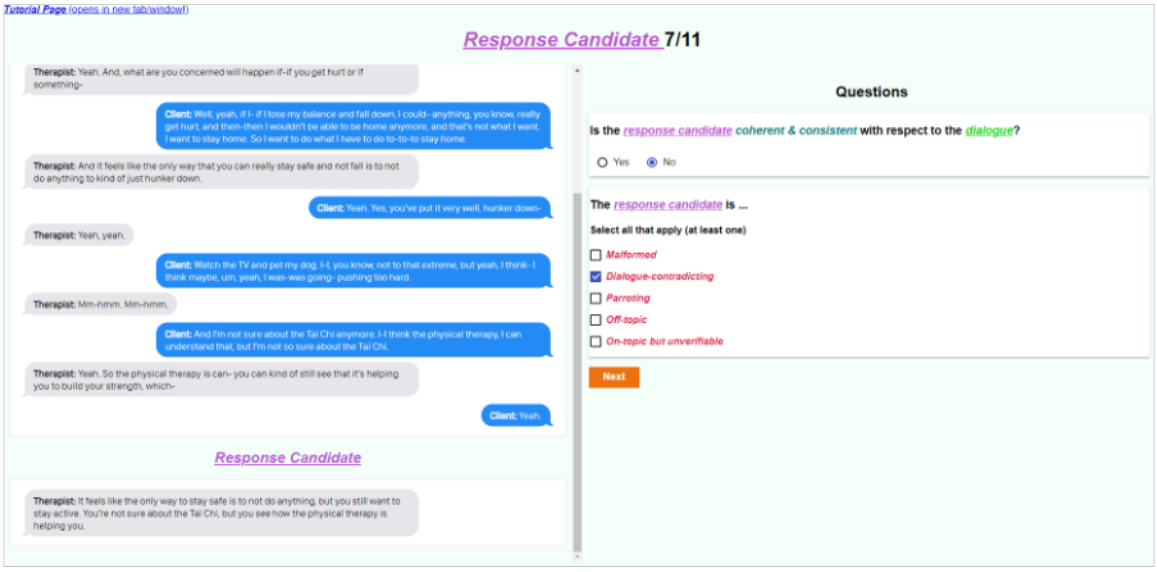

Figure 6: Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses one error category.

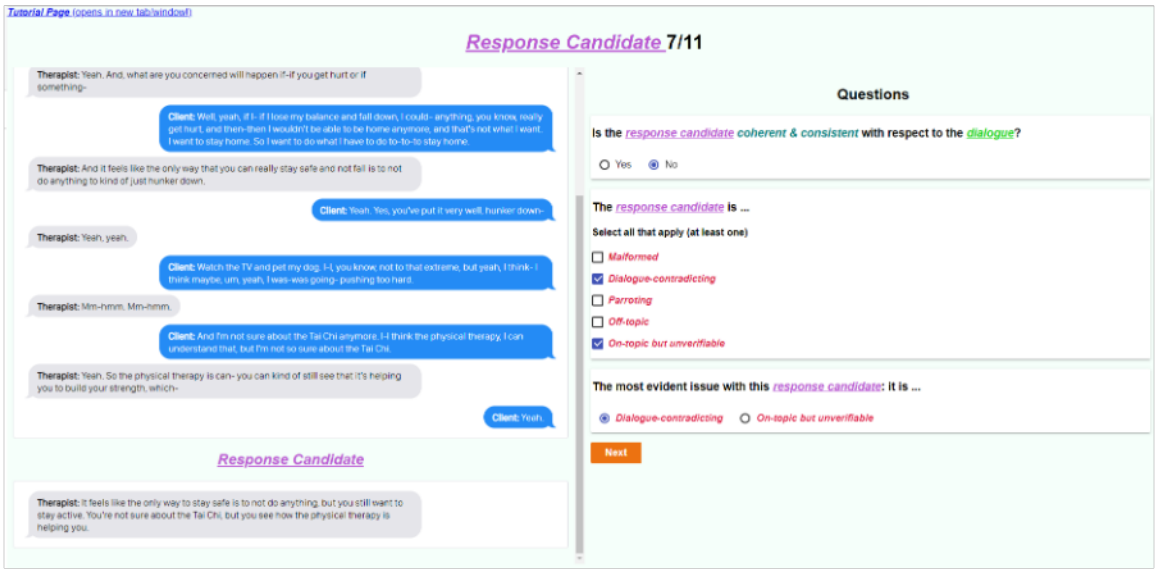

Figure 7: Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses multiple error categories.

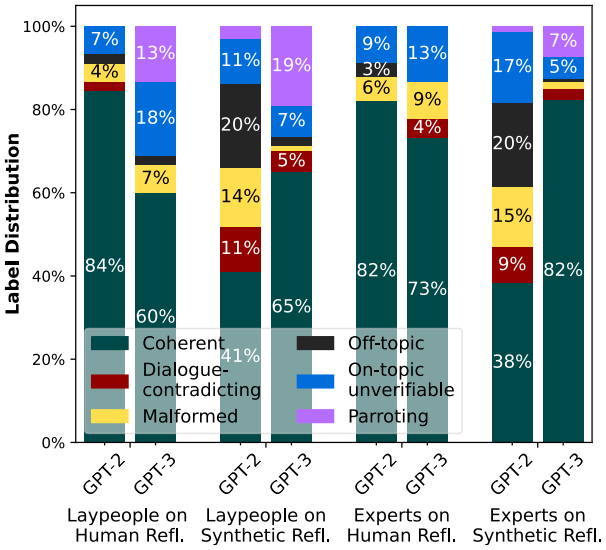

Figure 8: Labels distribution on human and synthetic reflections in the GPT-2 stage and GPT-3 stage. *Incoherent* labels are broken down into fine-grained error categories.

CL 2023 Responsible NLP Checklist

For every submission:

A1. Did you describe the limitations of your work?
Limitations section (unnumbered) after the conclusion

A2. Did you discuss any potential risks of your work?
Ethics Statement -> Risks

A3. Do the abstract and introduction summarize the paper's main claims?
bstract is at the beginning; Introduction is Section 1.

✗ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

1) We used a dataset (nnoMI), and it is mentioned multiple times in the paper. The first mention is in the Introduction. 2) We also created a dataset of human annotations (evaluations), and it is mentioned multiple times in the paper. The first mention is in the Introduction. Section 3 describes our methodology.

B1. Did you cite the creators of artifacts you used?
nnoMI is first cited in the Introduction (Table 1), and formally introduced in Section 3.1

B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
ppendix F

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
ppendix F

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
ppendix F

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.2; Ethics Statement; ppendix F

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1 and ppendix

C Did you run computational experiments?

Section 3.1 and ppendix

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.1 and ppendix

The Responsible NLP Checklist used at CL 2023 is adopted from N CL 2022, with the addition of a question on I writing assistance.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 and ppendix

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 3 & 4; ppendices B, D and E. We note that this is a human evaluation study, so some conventional descriptive statistics are not applicable, e.g., max/mean/single run. Nevertheless, we did use a variety of decoding parameters to generate diverse texts for the annotators to evaluate.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.1, ppendices & B

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3.2 and Ethics Statement

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
ppendix H and Supplementary Material (.zip file uploaded)

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.2 and Ethics Statement

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Statement

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics Statement

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Ethics Statement