Towards Effective Automatic Evaluation of Generated Reflections for Motivational Interviewing

Zixiu Wu*
z.wu@studenti.unica.it
University of Cagliari
Cagliari, Italy

Rim Helaoui
rim.helaoui@philips.com
Philips Research
Eindhoven, Netherlands

Diego Reforgiato Recupero
Daniele Riboni
diego.reforgiato@unica.it
riboni@unica.it
University of Cagliari
Cagliari, Italy

ABSTRACT

Reflection is an essential counselling skill where the therapist communicates their understanding of the client's words to the client. Recent studies have explored language-model-based reflection generation, but automatic quality evaluation of generated reflections remains under-explored. In this work, we investigate automatic evaluation on one fundamental quality aspect: coherence and context-consistency. We test a range of automatic evaluators/metrics and examine their correlations with expert judgement. We find that large language models (LLMs) as zero-shot evaluators achieve the best performance, while other metrics correlate poorly with expert judgement. We also demonstrate that diverse LLM-as-evaluator configurations need to be explored to find the best setup.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**; **Natural language generation**; • **Applied computing** → *Psychology*.

KEYWORDS

motivational interviewing, reflection, automatic evaluation

ACM Reference Format:
Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Towards Effective Automatic Evaluation of Generated Reflections for Motivational Interviewing. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3610661.3616127

1 INTRODUCTION

Motivational Interviewing (MI, [13]) is a common counselling approach that helps the client find their own motivation for positive behaviour changes, e.g., smoking cessation. In MI, reflection is a key skill, where the therapist shows empathy through a conversational summary of their understanding of the client's words [12, 18]. To assist therapist training in acquiring this skill, recent research [19, 20] has leveraged language models (LMs) to automatically generate candidate reflections given a therapist-client dialogue context.

To assess generated reflections, prior work has used both automatic [19, 20] and human [1, 19, 20, 22, 24] evaluation. Automatic evaluation often relies on reference-based metrics that measure the similarity between a generated text (hypothesis) and the gold-standard text (reference) w.r.t. word overlap (e.g., ROUGE [7]) or embeddings (e.g., BERTScore [25]). Human evaluation involves professional therapists (experts) to rate generated reflections as a good-or-bad binary choice [1] or on a Likert scale w.r.t. fine-grained quality aspects such as reflectiveness [19, 20, 22, 24].

While expert evaluation is of high quality, it is expensive and unscalable. Thus, automatic evaluation that is closely aligned with expert judgement is desirable. However, metric-expert correlation remains under-explored for reflection generation. To the best of our knowledge, the only study covering this topic [20] found weak correlation between expert judgement and reference-based metrics.

For open-domain response generation — a closely related task — reference-based metrics correlate poorly with human judgement [8]. In contrast, reference-free metrics, such as [3, 10, 11] which use language modelling likelihood as a response quality indicator, have better performance, but they are untested for reflection generation.

It is also worth noting that, for some text generation tasks including open-domain response generation, large language models (LLMs) have recently proved to be state-of-the-art reference-free evaluators, where the model is shown the hypothesis and then prompted to directly generate a numerical assessment [6, 9, 21].

In this work, we zoom in on automatic evaluation of one quality aspect: **coherence and context-consistency** (referred to as **coherence** for brevity), which takes both syntactics and semantics into account ([24] & Section 2.1). This focus is motivated by the fact that 1) recent LMs struggle with coherence [5]; 2) coherence is the pre-requisite for being a "good reflection" — a reflection should be a sensible response before being assessed w.r.t. therapy guidelines.

Specifically, we explore various automatic metrics for reflection evaluation, using a recent dataset [24] of generated reflections annotated by MI experts w.r.t. coherence. For reference-based metrics, we test the commonly used BLEU [15], ROUGE-L [7] and METEOR [4]. For reference-free metrics, we experiment with the follow-up likelihood method [3] and diverse LLM-as-evaluator settings.

Our results show that LLM-as-evaluator far outperforms other metrics in terms of correlation with expert judgement, but its performance can vary across different LLM-as-evaluator configurations,

[*] Work done during his industrial PhD at Philips Research and University of Cagliari.

ICMI '23 Companion, October 09–13, 2023, Paris, France
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0321-8/23/10.
https://doi.org/10.1145/3610661.3616127

Table 1: Dialogue context on alcohol use during pregnancy, with 1) its gold-standard therapist reflection, 2) examples of generated reflections and their coherence evaluations. Self-repetitions and mid-sentence changes (e.g., “your-your”) are characteristics of spoken dialogue [23].

Context
· · · (previous turns) **Therapist**: So, you’re feeling like once the baby comes and actually in your arms, then things are going to change, but right now, while you’re pregnant, you’re thinking, “I can kind of hang out in both worlds a little bit”, but part of you may be thinking maybe that’s not the case. It’s time to go to the mom world maybe. **Client**: Yeah. Probably. And I’m also feeling like it’s too late. I already screwed up. **Therapist**: Okay. So, you’ve got a lot of concerns about what’s happened to- up until this point, as far as your drinking. What do you know about, um, drinking alcohol throughout pregnancy? **Client**: Well, I know, like, they say not to, but I thought because my friends didn’t really care, I thought, “Okay. Like maybe-maybe it doesn’t matter. Maybe it’s just what they tell you”. Just like drinking, in general, is bad. **Therapist**: So, you were thinking because your-your friends weren’t pressuring you to stop drinking, that-that it must be okay. **Client**: Yeah.
Gold-Standard Reflection from Therapist
Therapist: And then there’s this other part of you that’s- was thinking maybe it’s not okay.
Examples of Generated Reflections
GPT-2: And you’re-you’re not sure that it is. (Evaluated as Coherent by: 2/3 Laypeople and 3/3 Experts)
GPT-3.5: And so now you’re feeling like it’s too late to change and you’re- you’re worried about what kind of impact your drinking might have had on the baby. (Evaluated as Coherent by: 3/3 Laypeople and 3/3 Experts)

which highlights the need to explore diverse evaluation setups for optimal results. We hope our work will inspire future efforts towards effective automatic evaluation for reflection generation.

2 METHODOLOGY

2.1 Data

We use `EvalReflections` [24], a dataset we collected previously that consists of reflections with coherence evaluations.

`EvalReflections` contains 15 dialogue contexts, each being a therapist-client conversation snippet with 14 turns on average — an example is shown in Table 1. Each context has a gold-standard reflection from the therapist, on average 7.1 reflections from GPT-2 [17] (`gpt2-medium`) and 8.8 from GPT-3.5 [2] (`text-davinci-002`). In total, 107 reflections are from GPT-2 and 133 from GPT-3.5.

Each generated reflection is annotated by 3 laypeople and 3 MI experts as a binary choice w.r.t. coherence, resulting in 6 individual binary labels. Specifically, a reflection is not coherent if it has any of the errors below, as defined in [22]:

- ***Malformed***: poor grammar, unclear references, and/or confusing logic.
- ***Dialogue-contradicting***: contradicts context.
- ***Parroting***: repeats a part of context unnaturally.
- ***Off-topic***: little to no relevance to context.
- ***On-topic but unverifiable***: relevant to context but has content that is unverifiable based only on context.

Accordingly, we define the **coherence score** of a reflection to be the number of laypeople/experts annotating it as coherent. For example, the GPT-2 reflection in Table 1 is considered coherent by 2 out of 3 laypeople and 3 out of 3 experts, which means it has a score of 2 from laypeople and 3 from experts. Overall, laypeople-experts correlation w.r.t. coherence scores is strong on GPT-2 reflections and moderate on GPT-3.5 reflections [16] (Table 2).

Table 2: Laypeople-experts Spearman’s correlations w.r.t. coherence scores. Both values are significant ($p < 0.05$).

	On GPT-2 Reflections	On GPT-3.5 Reflections
Laypeople-Experts Correlation	0.704	0.412

Due to space constraints, we refer the reader to our paper [24] for more details on the annotation procedure and in-depth analysis of laypeople-experts differences in coherence evaluation.

2.2 Reference-Based Metrics

Reference-based metrics measure hypothesis-reference similarity and have two main categories: 1) word-overlap-based metrics that calculate n-gram-level similarity; 2) embedding-based metrics that compute semantic similarity via token/sequence embeddings.

We employ 4 commonly used reference-based metrics, using the gold-standard reflection as the reference:

- **BLEU-4** [15]: precision between hypothesis and reference at the level of up to 4-grams.
- **METEOR** [4]: unigram-level F1-score between hypothesis and reference, considering stemming and synonyms.
- **ROUGE-L** [7]: F1 score based on the longest common subsequence between hypothesis and reference.
- **BERTScore** [25]: token-level F1-score between hypothesis and reference. Precision and recall are computed using greedy matching between hypothesis tokens and reference tokens. Matching is based on token embedding similarity.

Among the metrics above, BLEU-4, METEOR and ROUGE-L are based on word overlap while BERTscore is based on embeddings.

2.3 Follow-Up Likelihood (Reference-Free)

We probe the follow-up likelihood method [3, 10], a reference-free metric which assumes that the next utterance in a dialogue indicates the quality of the current turn. For example, if the next utterance is “You’re really confusing”, the current turn likely lacks clarity.

We adopt the formulation from [3]: Given an LM M, a dialogue history h, a response r and a hypothetical follow-up f like “You’re really confusing”, the conditional log likelihood $L_M(f; h, r)$ of the follow-up quantifies the quality aspect represented by f. For example, if the log likelihood of “You’re really confusing” is high, it means that the model, now playing the role of the listener, is likely to continue the conversation with “You’re really confusing”, which is evidence that the response r likely lacks clarity.

In addition to unfavourable follow-ups like “You’re really confusing”, there are also favourable follow-ups, such as “Wow that is really interesting”. Accordingly, high likelihood of a favourable follow-up means the response likely shows the favourable attribute. We use the favourable/unfavourable follow-ups from [3].

2.4 LLM as Evaluator (Reference-Free)

Following recent LLM-as-evaluator studies [6, 9, 21], we formulate reference-free evaluation as a zero-shot prompting task, where the LLM is prompted to generate a single number as the evaluation result. We divide the prompt into two consecutive segments:

- **Task Body**: contains task description, background (optional), dialogue context, and reflection to be evaluated.
- **Assessment Request**: follows the task body and asks for a single number as the reflection evaluation result.

We consider several types of task body and assessment request as shown below, and we pair each task body type with each assessment request type to explore LLM performance.

2.4.1 Task Body Types. The task body consists of 2 mandatory parts and 2 optional parts:

- Part A (Mandatory): A general description of the task.
- Part B (Optional): A brief explanation of incoherence errors ([22] & Section 2.1) without examples.
- Part C (Optional): A tutorial demonstrating coherent and incoherent reflections, taken directly from Table 4 of [22].
- Part D (Mandatory): Dialogue context and reflection.

The full text for each part is shown in Table 6 (Appendix). Thus, we adopt 3 task body types with increasing background information:

- **Instructions**: A general task description followed by the ⟨context, reflection⟩ pair, i.e. concatenation of Parts A & D.
- **Instructions & Error Explanations**: We insert the brief explanation of incoherence errors after the task description. Thus, the task body is a concatenation of Parts A, B and D.
- **Instructions & Tutorial**: We further add the tutorial after the incoherence error explanation. Thus, the task body is a concatenation of Parts A, B, C and D.

2.4.2 Assessment Request Types. Inspired by [6], we experiment with two assessment request types that ask the model to produce a single number as the assessment:

- **Rating**: a rating on a Likert scale from 1 to 5.
- **Scoring**: a score between 0 and 100.

The full text for each type is shown in Table 7 (Appendix).

3 EXPERIMENTS

We measure the performance of an automatic metric as the Spearman's correlation between its assessments and experts-based coherence scores on GPT-2/GPT-3.5 reflections from `EvalReflections`.

3.1 Reference-Based Metrics

For BERTScore, we use two LMs to acquire embeddings: `roberta-large` and `DeBERTa-XLarge-MNLI`. The former is used in the original paper and the latter is best correlated with human judgement[1].

We show in Table 3 the results of reference-based metrics. There is not a clear trend, as BERTScore outperforms word-overlap-based metrics on GPT-2 reflections but the reverse is true on GPT-3.5 reflections. The correlations are low overall — the highest is 0.159 and many are below 0. This is not surprising, since coherence is not equivalent to similarity to the reference, as coherent reflections can diverge considerably from the gold standard.

Table 3: Spearman's correlations between reference-based metrics and expert judgement. None is significant ($p < 0.05$).

Metric	On GPT-2 Reflections	On GPT-3.5 Reflections
Word-Overlap-Based Metrics		
BLEU-4	-0.113	0.072
ROUGE-L	0.056	-0.023
METEOR	-0.108	0.101
Embedding-Based Metrics		
BERTScore (`roberta-large`)	0.159	-0.046
BERTScore (`DeBERTa-XLarge-MNLI`)	0.099	-0.040

3.2 Follow-Up Likelihood (Reference-Free)

Following [3], we adopt the response-generation model `facebook/blenderbot-400M-distill` as the LM to compute log likelihood.

For GPT-2 and GPT-3.5 reflections separately, we probe the top-5 follow-ups w.r.t. absolute correlation, as strong positive and negative correlations are both desirable: an ideal favourable follow-up should have a strong positive correlation, but an ideal unfavourable follow-up should have a strong negative correlation (Section 2.3).

As Table 4 shows, four of the top-5 follow-ups for GPT-2 reflections are unfavourable, while the favourable one "That's a lot of questions!" can also be interpreted as unfavourable. All 5 follow-ups have negative correlations, some of which are intuitive. For example, the top-1 follow-up "You're not understanding me!" is related to incoherence errors like dialogue-contradicting and off-topic.

Nevertheless, the top-1 follow-up for GPT-2 reflections still only has a weak absolute correlation of 0.236. Therefore, following [3], we also consider grouping the top follow-ups and using their averaged log likelihood. As a result, we notice some improvement, such as the group of the top-3 follow-ups reaching an absolute correlation of 0.261, but overall those are low correlations, especially compared to the correlation of 0.51 with human judgement for open-domain dialogue generation in the original paper [3].

On GPT-3.5 reflections, the top follow-ups are not intuitive. For example, the top-1 follow-up — "Why are you changing the topic?" — is unfavourable but has a positive correlation of 0.279. In fact, all top-5 follow-ups have positive correlations, but only one of them is favourable. Furthermore, the highest top-1 correlation is still weak, although no group of top follow-ups outperforms it.

Based on the results, we postulate that incoherent GPT-2 reflections have more straightforward errors that can be easily captured by follow-ups, while incoherent GPT-3.5 reflections are more subtle, making follow-up-based evaluation unsuitable.

3.3 LLM as Evaluator (Reference-Free)

Following [6], we use 3 advanced OpenAI LLMs[2] as the evaluator:

- **`GPT-3.5`** (`text-davinci-002`): An InstructGPT model [14] of 175B parameters. It was also used in [24] to generate the GPT-3.5 reflections in `EvalReflections`.

[1] Leaderboard (https://github.com/Tiiiger/bert_score) accessed on 9 May 2023.

[2] We used the latest version of those OpenAI models at the time of our experiments.

Table 4: 1) Top 5 follow-ups ranked by absolute Spearman's correlation with expert judgement; 2) Likelihood-averaged follow-up groups and their Spearman's correlations with expert judgement. (+)/(-): favourable/unfavourable follow-up. Significant ($p < 0.05$) correlations are *italicised.*

	On GPT-2 Reflections	
	Individual Follow-Up	Correlation
1	(-) You're not understanding me!	*-0.236*
2	(-) I don't really care. That's pretty boring.	*-0.231*
3	(+) That's a lot of questions!	*-0.229*
4	(-) That makes no sense!	*-0.221*
5	(-) You don't really know much.	*-0.213*
	Likelihood-Averaged Follow-Ups	Correlation
	1 & 2	*-0.254*
	1, 2 & 3	*-0.261*
	1, 2, 3 & 4	*-0.251*
	1, 2, 3, 4 & 5	*-0.258*
	On GPT-3.5 Reflections	
	Individual Follow-Up	Correlation
1	(-) Why are you changing the topic?	*0.279*
2	(-) You don't really know much.	*0.212*
3	(+) Wow that is really interesting	*0.206*
4	(-) Don't change the topic!	*0.198*
5	(-) Let's change the topic.	*0.192*
	Likelihood-Averaged Follow-Ups	Correlation
	1 & 2	*0.257*
	1, 2 & 3	*0.259*
	1, 2, 3 & 4	*0.232*
	1, 2, 3, 4 & 5	*0.246*

Table 5: Spearman's correlations between LLM-as-evaluator metrics and expert judgement. Significant ($p < 0.05$) correlations are *italicised.* Best performance is in Bold.

Task Body	Assessment Request	LLM	On GPT-2 Reflections	On GPT-3.5 Reflections
Instructions	Rating	`ChatGPT`	*0.584*	0.103
		`GPT-3.5`	*0.271*	*0.197*
		`GPT-3.5.1`	*0.489*	*0.239*
	Scoring	`ChatGPT`	***0.586***	0.014
		`GPT-3.5`	*0.276*	*0.280*
		`GPT-3.5.1`	*0.482*	*0.261*
Instructions & Error Explanations	Rating	`ChatGPT`	*0.522*	*0.198*
		`GPT-3.5`	0.032	0.167
		`GPT-3.5.1`	*0.397*	*0.309*
	Scoring	`ChatGPT`	*0.551*	*0.193*
		`GPT-3.5`	*0.350*	*0.268*
		`GPT-3.5.1`	*0.472*	*0.336*
Instructions & Tutorial	Rating	`ChatGPT`	*0.506*	*0.219*
		`GPT-3.5`	0.102	*0.336*
		`GPT-3.5.1`	*0.496*	***0.389***
	Scoring	`ChatGPT`	*0.572*	0.034
		`GPT-3.5`	*0.449*	*0.278*
		`GPT-3.5.1`	*0.486*	*0.343*

- **`GPT-3.5.1`** (`text-davinci-003`): An improved version of `GPT-3.5` that better aligns with human preference.
- **`ChatGPT`** (`gpt-3.5-turbo`): A version of `GPT-3.5.1` that is optimised for chat.

We use a decoding temperature of 0 to ensure deterministic output.

As shown in Table 5, the best-performing LLMs have substantially better correlations with expert judgement than reference-based metrics and follow-up likelihood. Specifically, ⟨Instructions, Scoring, `ChatGPT`⟩ achieves 0.586 on GPT-2 reflections, while ⟨Instructions & Tutorial, Rating, `GPT-3.5.1`⟩ reaches 0.389 on GPT-3.5 reflections. Thus, this finding echoes recent work (e.g., [6, 9]) in affirming the effectiveness of LLMs as zero-shot text evaluators. Interestingly, these two numbers also show that it is more difficult for both LLMs and laypeople (Table 2) to align with experts on GPT-3.5 reflections than on GPT-2 reflections.

We also observe that `ChatGPT` always outperforms `GPT-3.5.1` on GPT-2 reflections, but the reverse is true on GPT-3.5 reflections. Furthermore, `GPT-3.5.1` outperforms `GPT-3.5` in every setting, which shows `GPT-3.5.1` is indeed more capable. As for assessment request types, rating is not consistently better or worse than scoring.

Interestingly, with increasing background information (none → error explanations → tutorial) in the task body, `GPT-3.5.1` consistently achieves better performance on GPT-3.5 reflections. Since GPT-3.5 reflection incoherence is likely more subtle (Section 3.2), this shows that more detailed task background information can help LLMs in some settings to evaluate more challenging examples.

Overall, the superior performance of LLMs makes them promising automatic coherence evaluators. Nevertheless, researchers should explore different evaluation setups, e.g., types of task body and assessment request, in order to find the optimal configuration.

4 CONCLUSION

In this work, we explored various automatic evaluators/metrics for their correlation with expert judgement on the coherence and context-consistency of generated reflections. We found that the best performance was achieved by using LLMs as zero-shot reference-free evaluators, while reference-based metrics and the reference-free follow-up-based evaluator all had poor correlations with expert judgement. Nevertheless, exploration of diverse configurations is needed for the LLM-as-evaluator approach, since different setups can have varying performance levels. For future work, we plan to further investigate LLMs as evaluators, e.g., 1) generating aspect-level explanation (e.g., "low coherence due to sudden topic change at ...") for auto-assigned coherence scores; 2) probing the underlying reasons for the superior performance of LLMs as evaluators.

ACKNOWLEDGMENTS

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

REFERENCES

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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2022. Open-Domain Dialog Evaluation Using Follow-Ups Likelihood. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 496–504. https://aclanthology.org/2022.coling-1.40

[4] Michael J. Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*. The Association for Computer Linguistics, 376–380. https://doi.org/10.3115/v1/w14-3348

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[6] Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. *CoRR* abs/2302.14520 (2023). https://doi.org/10.48550/arXiv.2302.14520 arXiv:2302.14520

[7] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[8] 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, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 2122–2132. https://doi.org/10.18653/v1/d16-1230

[9] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. GPTEval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634* (2023).

[10] Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 225–235. https://aclanthology.org/2020.sigdial-1.28/

[11] Shikib Mehri and Maxine Eskénazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 681–707. https://doi.org/10.18653/v1/2020.acl-main.64

[12] 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* (2003).

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

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[15] 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 Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[16] 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 (2014), 535–536. https://doi.org/10.1016/j.ecns.2014.07.005

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

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

[19] 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3096–3107. https://doi.org/10.18653/v1/2022.acl-long.221

[20] 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, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (Eds.). Association for Computational Linguistics, 10–20. https://aclanthology.org/2020.sigdial-1.2/

[21] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048* (2023).

[22] Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2022. 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)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 116–124. https://aclanthology.org/2022.gem-1.9

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 6177–6181. https://doi.org/10.1109/ICASSP43922.2022.9746035

[24] Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Are Experts Needed? On Human Evaluation of Counselling Reflection Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6906–6930. https://aclanthology.org/2023.acl-long.382

[25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

A PROMPTING DETAILS

Table 6 shows the different parts of the task body, and Table 7 presents the two assessment request types.

Table 6: Parts of task body.

Part A: Task Overview Applicable to: **All task body types**
Task overview: evaluate the quality of a Response Candidate for a multi-turn Dialogue # Definitions 1) "Dialogue": Part of a multi-turn conversation between a therapist and a client 2) "Response Candidate": A response candidate that the therapist could say to the client after the last turn in the Dialogue Note: You may notice some self-repetitions and/or mid-sentence changes within a turn. This is normal, as the text is captured from spoken dialogues.
Part B: Incoherence Error Explanations Applicable to: **Instructions & Error Explanations; Instructions & Tutorial**
Incoherence and Inconsistency A response candidate is incoherent and/or inconsistent with the Dialogue if it has any of the following problems: 1) Malformed: ... 2) Dialogue-contradicting: ... 3) Parroting: ... 4) Off-topic: ... 5) On-topic but unverifiable: ...
Part C: Tutorial Applicable to: **Instructions & Tutorial**
Tutorial Below is an example Dialogue and several illustrative Response Candidates. ## Example Dialogue Therapist: ... ... (intermediate turns) Client: ... ## Illustrative Response Candidates ### Coherent and Consistent Response Candidate Therapist: ... ### Incoherent and/or Inconsistent Response Candidates #### Malformed Therapist: ... #### Dialogue-contradicting Therapist: ... #### Parroting Therapist: ... #### Off-topic Therapist: ... #### On-topic but unverifiable Therapist: ...
Part D: Context + Reflection Applicable to: **All task body types**
Task ## Dialogue Therapist: ... ... (intermediate turns) Client: ... ## Response Candidate Therapist ...

Table 7: Assessment request types.

Type 1: Rating
Rating Rate the Response Candidate on a discrete scale from 1 to 5, where a rating of 1 means “completely incoherent and/or inconsistent with the Dialogue” and a rating of 5 means “perfectly coherent and consistent with the Dialogue”. Rating (1-5):
Type 2: Scoring
Scoring Score the Response Candidate on a continuous scale from 0 to 100, where a score of 0 means “completely incoherent and/or inconsistent with the Dialogue” and a score of 100 means “perfectly coherent and consistent with the Dialogue”. Score (0-100):