A Human-AI Approach to Few-Shot Learning for Motivational Interviewing-based Counseling

Chengqian Wang[1,†], Arjumand Younus[1,*,†]

[1]School of Information and Communication Studies, University College Dublin, Dublin, Ireland

Abstract

Motivational interviewing (MI) is a counselling approach that supports individuals in resolving ambivalence and progressing towards positive behavioural change. However, training and evaluating MI skills at scale remains challenging due to the need for expert supervision, high-quality annotated dialogue data, and careful adherence to professional and ethical standards. This paper presents our submission to the MIRROR task at IberLEF 2026, which focuses on generating synthetic conversational turns for MI-based counselling across selected behavioural-code pairs, including simple versus complex reflection, open versus closed questions, and persuasion versus giving information. Our approach uses the expert-annotated AnnoMI dataset as a source of high-quality seed conversations and applies few-shot prompting through the ChatGPT OpenAI API to generate labelled synthetic MI dialogue turns. To improve the quality and safety of the generated data, we incorporate human-in-the-loop validation at both the prompt-design and final-review stages, allowing noisy or inappropriate counsellor responses to be removed and ensuring closer alignment with MI principles. The proposed pipeline demonstrates a practical human-AI strategy for producing task-specific synthetic counselling data under limited-data conditions. We also discuss the limitations of the approach, including topic and language constraints, and identify future directions involving fine-tuning, retrieval-augmented generation, and more systematic evaluation of LLM-generated counselling interactions.

Keywords

few-shot learning, publicly available MI data, human-in-the-loop, prompts

1. Introduction

Motivational interviewing (hereon referred to as MI) is a well-documented counselling method that helps people with ambivalence issues towards behavior change [1]. The chief idea behind MI is the encouragement of helpseekers to explore their internal motivation for positive and sustainable behavior change. While a valuable skill, it is also challenging to learn because it requires the ability to adhere to a specific set of rules, which comes through training with experts, and with the global shortage of such a workforce artificial intelligence has been used to provide data-augmented training in MI [2, 3]. MIRROR task [4] at IberLEF 2026 [5] advances this field by introducing a synthetic data challenge where participants are asked to create simulated conversation turns comprising various behavioral codes for an MI setting. More specifically, the organizers of the challenge specifically asked for three pairs of behavioral codes, i.e., simple reflection vs complex reflection, open questions vs closed questions and persuasion vs giving information. Note that behavioural codes represent specific techniques and approaches used by a counselor in a motivational interviewing process strictly adhering to the coding system of Motivational Interviewing Treatment Integrity [6].

This paper presents our submission for this task whereby performance of our submitted datasets on a fine-tuned model for behavioral code classification is the assessment criteria for the task. Our approach relies on few-shot learning using the publicly available collection of 133 transcribed therapy conversations, AnnoMI [7]. We also incorporate human-in-the-loop into our prompt-engineering data-generation pipeline to remove noisy instances of counselor interactions while also ensuring adherence

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

chengqian.wang@ucdconnect.ie (C. Wang); arjumand.younus@ucd.ie (A. Younus)

0009-0009-7921-4741 (C. Wang); 0000-0001-7748-2050 (A. Younus)

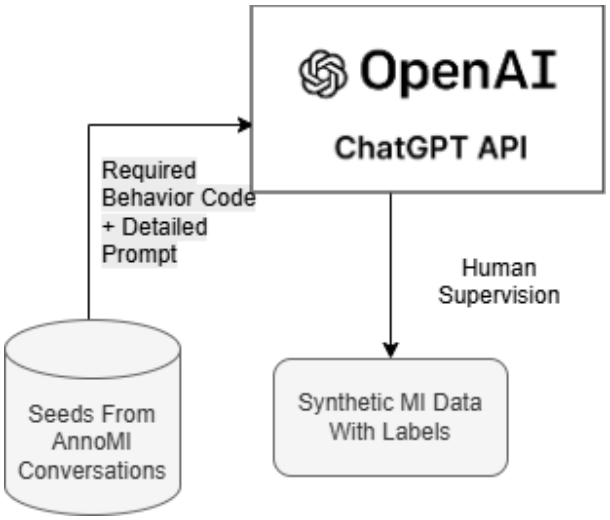

Figure 1: High-level Overview of Our Dataset Generation Process

to safety incentives in the counseling process [8]. We detail various phases of the dataset generation process in subsequent sections of this paper and discuss potential limitations of our approach, as well as current aspects of LLM-based counseling that warrant further investigation.

2. Background

Our work utilises AnnoMI dataset [7] which consists of 133 conversations over a wide range of topics from “smoking cessation" to “reducing recidivism." Expert annotations of high-quality and low-quality motivational interviewing transcripts by therapists is the chief highlight of this dataset; the usefulness of this dataset makes it particularly suited for further research in pyschotherapy on account of utterance-level annotations following a strict adherence to MI literature [9]. Table 1 shows some example rows from the dataset for a single conversation whereby utterances occur between a client and therapist on the topic “smoking cessation"; note that the feature “therapist behavior" reveals important insights, and it is this feature that we incorporate into our few-shot learning prompts for the required behavioral codes. This is explained in further detail in Section 3.

Table 1
Example rows of a single conversation from AnnoMI dataset

Utterance ID	MI Quality	Interlocutor	Utterance	Therapist Behavior	Topic
3	high	client	Yeah, um, normally, I'm smoking, you know, like once in the morning, uh, once during my lunch break and then in the evening maybe like five or six	n/a	smoking cessation
4	high	therapist	Okay. So about eight cigarettes. So over the course of a day, just under half a pack?	n/a	smoking cessation

3. Methodology

Figure 1 gives a high-level overview of our dataset generation process. The chief aspect involves human-in-the-loop validation of the "detailed prompt" together with a final human-in-the-loop validation. This human intervention aspect is incorporated in keeping up with suggestions from the latest research on mental health chatbots [10]; and a vast majority of the harms from these chatbots emanate from a lack of adhering to ethical, professional standards which require careful human input [11].

3.1. Seed Generation and Prompt Specifics

As mentioned in Section 2, AnnoMI conversations serve as the starting point, with each conversation within a specific category read by one of the authors to select the best conversation with a rich set of interactions. Note that categories picked from within AnnoMI dataset are specifically "reflection", "question", and "therapist_input" to match the demands specified in MIRROR task. It is important to note that we always make a selection from a high-quality transcript; and choose only client and therapist utterances within the segment of the transcript where therapist shows the behavior falling within our category of interest with a further criteria that a minimum of 4 utterances by the therapist fall within that category. Figure 2, for example, shows the seed set of therapist and client utterances corresponding to behavioral code pair "persuasion" vs "giving information" which in case of AnnoMI occurs when therapist behavior is equal to "therapist_input."

Using the seed set as examples we then issue a prompt to ChatGPT OpenAI API with the prompt (see Figure 3) detailing in the form of parameters the number of turns or volleys to generate, the seed set in the form of list called *conversations*, and the topic which we generally restrict to "smoking cessation", ""taking medicine / following medical procedure", or "reducing alcoholism."

4. Discussion

The results of the MIRROR task indicate that our human-AI few-shot generation pipeline achieved competitive performance for motivational interviewing (MI) behavioural-code classification. As shown in Table 2, our submission ranked fourth overall, with an average score of 0.7303. This places our system close to the highest-ranked submission, which achieved an average score of 0.7366, and only

["Okay. Well, we do know that an increase in smoking and alcohol is associated with certain kinds of cancer; colon cancer, cervical cancer, breast cancer, lung cancer, and, um, also pancreatic cancer. It's also associated with liver and lung disease as well as heart disease, diabetes, and hypertension. So that's a lot of information, but I'm-", 'Yeah.', "Okay, great. Let's quantify that a little bit. So how many cigarettes would that be?", "Well, right now, if I'm having like five or six, I could try to smoke three in the evening plus, you know, still the couple during the day.", "Okay. Well, I've got a brochure here and this has-- it's a really good brochure. It has some distraction techniques in it in case you get those nicotine cravings. So it's-", 'Mm-hmm.', 'Well, Dr. Kirtley is one of our counselors and he is an expert in-- at discussing the association between alcohol and stress. I want to share something with you. The National Guidelines recommend that women have no more than seven drinks of alcohol per week to-to stay at that low risk level, to avoid the healthcare concerns that we spoke of earlier.', 'Mm-hmm.', "You've indicated that your alcohol consumption is a bit higher than that, putting you at risk. So that's why I asked if you'd be willing to talk to Dr. Kirtley.", "I mean, I would like to talk about, you know, my stress, but I'm still not sure with the whole drinking thing."]

Figure 2: Portion of Conversation From AnnoMI Where Therapist Behavior Is Of Therapist Input

```
prompt_role='You are a therapist using the technique of motivational interviewing for people wanting to change their behavior. \
    Your task is to generate LENGTH conversations, turn by turn in the format CONVERSATIONS but for the topic TOPIC. \
    The therapist in his turn should ask either a close-ended or open-ended question only.  \
    Also output whether the question asked in given turn is close-ended or open-ended.'

def assist_therapist(length_words: int,conversations: List[str],topic: str):
    conversations = ", ".join(conversations)
    prompt = f'{prompt_role}\nLENGTH: {length_words}\nCONVERSATIONS: {conversations}\nTOPIC: {topic}'
    print(prompt)
    return ask_chatgpt([{"role": "user", "content": prompt}])
```

Figure 3: Portion of Conversation From AnnoMI Where Therapist Behavior Is Of Therapist Input

Table 2
Results of iberLEF MIRROR Task

Team	Rank	Information	Questions	Reflections	Avg
UNISON-GOATS	1	0.7728	0.8559	0.5811	0.7366
Codice	2	0.7466	0.8663	0.5942	0.7357
umuteam	3	0.7593	0.8464	0.5875	0.7311
ucdicsnlpteam	4	**0.7603**	**0.8438**	**0.5869**	**0.7303**
steam_uaz	5	0.7749	0.8382	0.5667	0.7266
roman_lab	7	0.7369	0.8311	0.5948	0.7209

marginally behind the second- and third-ranked teams. This result suggests that a relatively lightweight pipeline based on expert-annotated seed data, few-shot prompting, and human-in-the-loop validation can generate useful synthetic MI data under limited-data conditions.

A notable aspect of the result is that the performance gap between the top submissions is small. Our average score differs only slightly from the first-ranked system, indicating that our approach was broadly aligned with the demands of the task despite its methodological simplicity. This is important because our system did not rely on a complex fine-tuning pipeline or a large task-specific data construction process. Instead, it used carefully selected high-quality examples from AnnoMI and prompt-based generation through the ChatGPT OpenAI API, followed by human review. The competitive ranking therefore provides evidence that carefully designed few-shot prompting can be a practical strategy for generating synthetic counselling-style data.

At the level of behavioural-code categories, our system performed particularly well on Information and Questions. For Information, our submission achieved a score of 0.7603, which is close to the strongest score reported in the table. For Questions, the system achieved 0.8438, again placing it within a narrow range of the leading submissions. These results suggest that the generated examples were especially useful for behavioural codes where surface form and communicative function are relatively explicit.

Questions, for instance, often contain clearer syntactic and pragmatic cues, while information-giving may be more directly represented through explanatory or factual counsellor utterances.

In contrast, Reflections remained the weakest category for our system, with a score of 0.5869. However, this was not unique to our submission; all teams reported substantially lower scores for Reflections compared with Information and Questions. This pattern suggests that reflection classification is intrinsically more challenging in the MIRROR task. The distinction between simple and complex reflection often depends on subtle interpretation of the client's prior statement, the counsellor's inferred meaning, and the extent to which the response adds depth, reframing, or emotional interpretation. These distinctions are less likely to be captured by surface-level linguistic cues alone and may require stronger modelling of dialogue context and MI-specific coding principles.

The results also complicate the expected impact of language mismatch. Our generation process relied on English AnnoMI conversations as seed examples, while the MIRROR task required Spanish synthetic counselling data. This mismatch could have been expected to substantially weaken performance. However, the competitive ranking suggests that at least some MI behavioural-code signals may transfer across languages, particularly for categories such as questions and information-giving. This does not remove the limitation, but it indicates that the approach may have captured generalisable conversational patterns relevant to MI coding. At the same time, the relatively weaker performance on Reflections may reflect precisely where cross-lingual and cross-cultural transfer becomes more difficult, because reflective responses rely heavily on nuanced meaning, pragmatic interpretation, and culturally appropriate counselling style.

Another important implication concerns the role of human-in-the-loop validation. In our pipeline, human review was used to refine prompts, remove noisy generations, and support safer counselling-style outputs. The competitive score suggests that this step may have helped maintain sufficient quality for downstream behavioural-code classification. However, the results also show that human validation should be extended beyond general fluency, safety, and MI plausibility. For future behavioural-code generation, reviewers should also explicitly assess label discriminability: whether each generated utterance is a strong and unambiguous example of the intended behavioural code, and whether it is clearly separable from the paired alternative. This is especially important for reflection codes, where plausible counsellor responses may still be difficult to classify reliably.

The use of AnnoMI as a seed dataset also appears to have been valuable. AnnoMI provides expert-annotated counselling dialogues and therefore offers a principled source of MI-relevant examples for few-shot generation. However, mapping AnnoMI therapist-behaviour categories onto MIRROR behavioural-code pairs remains a potential source of noise. Categories such as reflection, question, and therapist input provide a useful starting point, but they may not fully capture the fine-grained boundaries required for the MIRROR labels. The results therefore support the usefulness of AnnoMI-seeded generation while also highlighting the need for more precise code-mapping guidelines between source annotations and target behavioural-code definitions.

Overall, the MIRROR results show that our approach is promising but not complete. The competitive ranking demonstrates that few-shot LLM-based generation, when combined with human supervision and expert-annotated seed conversations, can produce synthetic MI data that is useful for downstream behavioural-code classification. At the same time, the category-level results reveal that not all MI behaviours are equally easy to generate or classify. Information-giving and question-asking appear more amenable to prompt-based synthetic generation, while reflections require deeper contextual and pragmatic modelling.

Future work should therefore focus on improving generation for reflection-related codes, developing Spanish-first or multilingual prompting strategies, and introducing more systematic validation of label quality. Contrastive prompting may be particularly useful, where the model is shown paired examples that distinguish simple from complex reflections or persuasion from information-giving. Retrieval-augmented generation could also help by grounding generated turns in relevant MI examples, while fine-tuning may provide stronger control over behavioural-code consistency. Finally, future evaluation should combine downstream classifier performance with qualitative coding analysis, so that synthetic counselling data is assessed not only by model scores but also by MI fidelity, label clarity, linguistic

diversity, and ethical appropriateness.

Declaration on Generative AI

The authors used ChatGPT (OpenAI) to review and improve the English grammar and readability of this manuscript. After using this tool, the authors reviewed, edited, and approved the suggested changes. The authors take full responsibility for the final content of the paper.

References

[1] W. R. Miller, S. Rollnick, Motivational interviewing: Helping people change, Guilford press, 2012.

[2] M. Mormina, S. Pinder, A conceptual framework for training of trainers (tot) interventions in global health, Globalization and health 14 (2018) 100.

[3] M. Pellemans, S. Salmi, S. Mérelle, W. Janssen, R. van der Mei, Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: Secondary analysis of a clinical trial, Journal of medical internet research 26 (2024) e53562.

[4] L. J. Arellano, C. Olachea, J. Piette, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Irazú Hernández-Farías, Overview of mirror at iberlef 2026: Motivational interviewing response rating via synthetic conversational turns, in: Procesamiento del Lenguaje Natural, Vol. 77, Vienna, Austria, 2026.

[5] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[6] S. P. Lord, D. Can, M. Yi, R. Marin, C. W. Dunn, Z. E. Imel, P. Georgiou, S. Narayanan, M. Steyvers, D. C. Atkins, Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code, Journal of substance abuse treatment 49 (2015) 50–57.

[7] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. R. Recupero, D. Riboni, Anno-mi: A dataset of expert-annotated counselling dialogues, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6177–6181.

[8] C. Wang, A. Younus, K. Shankar, Analysing logs of llm-based counselling-style conversation agents: A perspective from nlp, in: Proceedings of the 18th ACM Web Science Conference 2026, 2026, pp. 332–337.

[9] W. R. Miller, T. B. Moyers, D. Ernst, P. Amrhein, Manual for the motivational interviewing skill code (misc), Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico (2003).

[10] A. Parks, E. Travers, R. Perera-Delcourt, M. Major, M. Economides, P. Mullan, Is this chatbot safe and evidence-based? a call for the critical evaluation of generative ai mental health chatbots, Journal of Participatory Medicine 17 (2025) e69534.

[11] Z. Iftikhar, A. Xiao, S. Ransom, J. Huang, H. Suresh, How llm counselors violate ethical standards in mental health practice: A practitioner-informed framework, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 8, 2025, pp. 1311–1323.