

AutoAnnoMI: A Framework to Automate the Annotations of MI Conversations using LLMs

Sepehr Ahmadi¹ and Zafarullah Mahmood²

^{1,2}Department of Electrical & Computer Engineering, University of Toronto

Word Count: 1995 **Penalty:** 0%

1 Introduction

Decisions related to behaviour change often involve a complex interplay of psychological, emotional, and situational factors. A significant barrier in this process is *ambivalence*, or the experience of having mixed feelings about making a change. This ambivalence often impedes the adoption of healthier behaviours or the cessation of harmful habits. The role of ambivalence in hindering behaviour change has long been recognized and forms the basis of some of the most successful behavioural therapy techniques, namely motivational interviewing.

Motivational interviewing (MI) is an evidence-based [1] form of counselling with the goal to “evoke and strengthen a person’s motivation for change” [2]. The MI counsellor helps facilitate the client’s journey from ambivalence to *change talk* — a state where the client expresses a desire or commitment to change [3].

The success of MI can be attributed to its foundational principles, which include the use of open-ended questioning, providing personal affirmations, engaging in reflective listening, and offering thoughtful summarization [4]. MI counsellors undergo training to master these techniques, and their proficiency is assessed through their ability to apply these conversational strategies effectively. Typically, MI experts are hired to rate an interview between a counsellor and a client based on specific coding systems, such as the Motivational Interviewing Skills Code (MISC) [5]. These ratings offer valuable feedback to counsellors and pinpoint areas for skill enhancement.

The task of annotating an MI conversation is tedious. It requires carefully analyzing every statement exchanged between the counsellor and the client and labelling them using specific behavioural codes [5]. It therefore requires considerable MI expertise and could be viewed as a significant obstacle in expanding the training and assessment of MI counsellors.

We hypothesize that if MI annotation is either fully or partially automated, it will not only reduce the burden on MI experts but could pave the way for the scalable training and assessment of MI counsellors. Such automation will improve the accessibility and broader adoption of MI in healthcare settings like smoking cessation.

One feasible way to automate MI annotation is by employing Large Language Models (LLMs). Recently, LLMs such as GPT-4 have shown remarkable performance on a wide range of natural language processing tasks, including annotation. Specifically, LLMs have been shown to annotate utterances based on prompts that specify the annotation protocols and, optionally, some example annotations [6]. Prompt engineering [7] of LLMs differs from traditional machine learning as it facilitates the LLMs to generalize across complex tasks without requiring task-specific training data or fine-tuning [8].

In this work, we explore the design of an LLM-in-the-loop annotation application that helps experts in labelling MI transcripts.

2 Background & Related Work

Providing detailed feedback to counsellors has long been promoted as an effective MI training technique [9]. To this end, the developers of MI came up with MISC, a comprehensive behavioural coding system to evaluate gains in training [5]. A subset of MISC is the Motivational Interviewing Treatment Integrity (MITI) code, which is designed to assess and enhance practitioners’ MI proficiency. It offers structured feedback for skill improvement in both research and non-research environments.

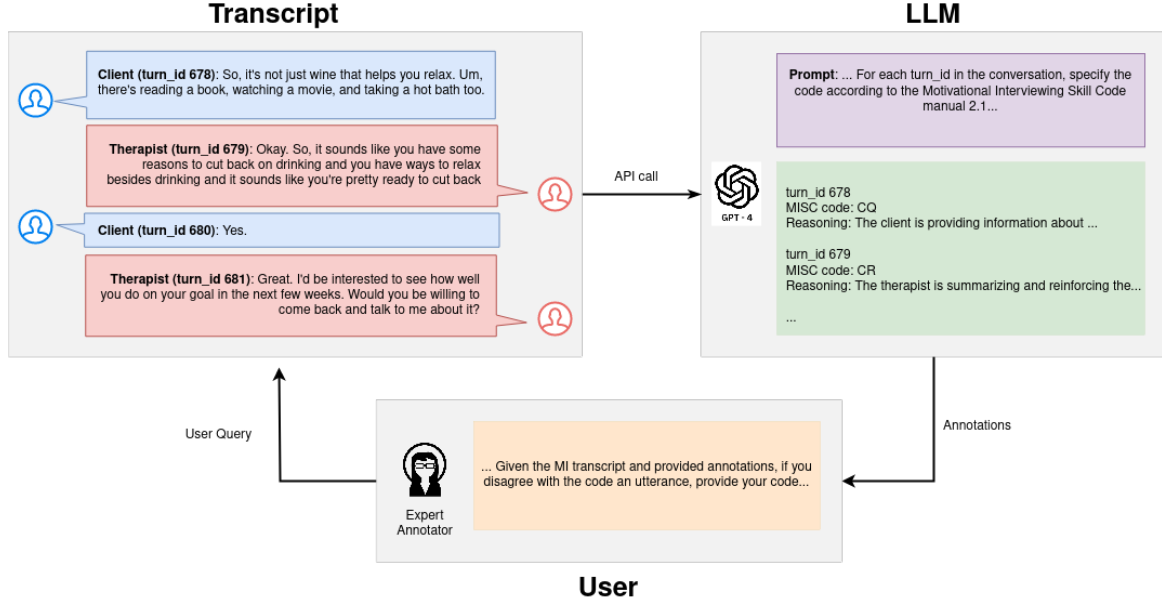


Figure 1: Our annotation framework where the LLM first annotates the transcripts by providing the MISC labels, and the expert annotator reviews them and suggests changes.

Very little research has been conducted on the effectiveness of natural language processing (NLP) systems in labelling MI conversations or generating appropriate therapist behaviour, with a few notable exceptions. For example, Wu et al.(2022) [10] curated a dataset of expert-annotated MI counselling dialogues and fine-tuned small NLP models [11] to predict the therapist’s behaviour. Another example is the work by Brown et al. (2023) [12], who used a 1.5 billion parameter GPT-2 XL model to generate therapist reflections and a fine-tuned BERT model to classify the quality of the generated reflections.

One obstacle to this research is the absence of a sufficiently large dataset of MI conversations. The largest publicly available MI dataset we discovered was AnnoMI [10]. It consists of “133 faithfully transcribed and expert-annotated demonstrations of high- and low-quality” MI conversations [10]. As previously mentioned, the expansion of such a dataset is constrained by the extensive time and effort required from expert annotators. Our research aims to automate MI annotation, potentially easing this bottleneck and facilitating the development of a larger, more diverse dataset.

In the following sections, we describe our work in detail, starting from our choice of the data and its processing.

3 Data and Data Processing

To study the problem of automating MI annotation, we used the AnnoMI dataset. It contains over 13000 annotated utterances from real MI conversations and annotates therapist utterances with two useful codes: reflection (“simple/RES” or “complex/REC”) and question (“open” or “closed”). Figure 2 shows some example rows from the dataset. In our work, we use the dataset to measure the performance of LLMs in accurately predicting the type of reflection a therapist has made. A therapist’s reflection on a client’s utterance can be categorized as either “simple” (merely paraphrasing what the client has) or “complex” (providing a new perspective or highlighting an inconspicuous aspect of the client’s statement). Figure 3 highlights the difference between a simple and a complex reflection by a therapist. Understanding this difference requires considerable MI expertise and thus provides an opportunity to prove the usability of LLMs in annotating MI conversations.

Prior to the use of this dataset, we performed some exploratory analyses and noticed places in the dataset where interlocutor labels were misplaced, likely by accident. To fix these mislabels, we asked GPT-4 to locate potential mislabels in each conversation. We also manually verified the suggestions by GPT-4 before modifying the dataset. In total, we identified three conversations with mislabels.

To test the generalizability of our application to other MISC codes, we labelled a subset of therapist utterances with the additional ADP/ADW (advice with or without permission) codes. Apart from dropping some low-quality and duplicate MI conversations, we did not perform any additional cleaning. To avoid incurring API

transcript_id	topic	utterance_id	interlocutor	timestamp	utterance_text	reflection_exists	reflection_subtype	question_exists	question_subtype
41	0	reducing alcohol consumption	41	client	00:04:38	Maybe having less wine in the house.	NaN	NaN	NaN
42	0	reducing alcohol consumption	42	therapist	00:04:41	Having less wine in the house will help?	True	simple	True

Figure 2: Example rows from the AnnoMI dataset

Client: "I know I should be exercising more, but I just can't find the time. It's been really hard with my work schedule." Therapist: "It sounds like you're feeling really challenged to fit exercise into your busy schedule."	Client: "I know I should be exercising more, but I just can't find the time. It's been really hard with my work schedule." Therapist: "You're recognizing the importance of exercise, yet finding it really challenging to integrate it into your busy life. It sounds to me your job demands are taking precedence over your personal health goals."
Simple reflection	Complex reflection

Figure 3: An example of a simple and complex reflection by a therapist.

costs while testing proprietary LLMs like OpenAI’s GPT-4, we created two smaller test sets from the original dataset:

1. **Reflection classification (RES/REC/None) test set** consisting of 384 utterances. The number of examples for each classification label is:
 - Simple reflection (RES): 96
 - Complex reflection (REC): 96
 - No reflection (None): 192
2. **Advice classification (ADW/ADP/None) test set** consisting of our 45 hand-labelled examples, distributed as follows:
 - Advice with permission (ADP): 11
 - Advice without permission (ADW): 19
 - No advice present (None): 15

4 Architecture and Software

We formulated the meta-task of annotation as a collection of various sub-tasks for detecting and classifying the behaviour codes in the therapist’s utterances. We specifically focused on two sub-tasks: (a) detection and classification of reflection in therapist’s utterances and (b) detection and classification of therapist advice. In the rest of the report, we refer to these as **RES/REC/None** and **ADP/ADW/None** tasks. These tasks can be viewed as a 3-class classification, and LLMs can be prompted to provide the most correct class label.

For all our prompting experiments, we used OpenAI’s GPT-4 Turbo [13]. For each utterance, we sent an API request by providing the task descriptions and examples as system prompts and the actual utterances as user prompts.

We considered different prompt components as hyperparameters and varied them to measure their impact on classification accuracy. Our “prompt hyperparameters” included:

1. **Task description**, which could be *simple* (minimal details about motivational interviewing) or *elaborate* (explicit instruction to adhere to MISC guidelines). We wanted to see if providing a better task description improves the test accuracy.
2. **Conversation history**, varying from 0 (no previous conversation turns) to a maximum of 5 previous turns. We hypothesized that providing more conversation context would help the LLM predict the correct

Experiment No.	Components of a prompt				Accuracy (%)
	Task description	Number of examples provided	Context provided	Chain-of-thought	
1	Simple	0	None	No	74
2	Simple	3	None	No	75
3	Elaborate	3	None	No	76
4	Elaborate	0	Upto 5 conversation turns	No	88
5	Elaborate	3	Upto 5 conversation turns	No	88
6	Elaborate	3	Upto 5 conversation turns	Yes	90

Table 1: Test set accuracy on the RES/REC/None task for different prompting experiments.

Experiment No.	Components of a prompt				Accuracy (%)
	Task description	Number of examples provided	Context provided	Chain-of-thought	
1	Elaborate	0	2 conversation turns	No	84
2	Elaborate	2	2 conversation turns	No	84
3	Elaborate	0	6 conversation turns	No	82
4	Elaborate	2	6 conversation turns	No	71
5	Elaborate	1	4 conversation turns	No	75

Table 2: Test set accuracy on the ADP/ADW/None task for different prompting experiments.

label, as the answer may not be obvious from the last conversation turn. We capped the number of utterances in one example at 5, as using more adversely impacted the model’s performance, which is intuitive given its decreased attentiveness to the key last utterance.

3. **Number of few-shot examples**, which can be zero (also called zero-shot prompting) or go up to 3. It is well-known that giving task demonstrations to LLMs improves response quality [14]. We refrained from using additional examples as the model’s performance plateaued on the training data.
4. **Chain-of-thought reasoning**, which can be either switched “on” or “off”. Recent success in encouraging LLMs to “think aloud” and explicitly mention the steps taken to arrive at a conclusion inspired us to experiment with strategy.

We combined these hyperparameters to create multiple experiments and calculated the final test accuracy on both RES/REC/None and ADP/ADW/None tasks.

5 Measure of success

Using a majority class classifier will result in an accuracy of 50% on RES/REC/None and 42% on ADP/ADW/None. Furthermore, a relatively simple prompt results in an accuracy of 74% and 84%, respectively. We therefore consider our system (which is jointly described by both the LLM and the system prompt) to succeed on the task if it achieves an accuracy higher than that of a simple prompt.

6 Quantitative Results

Table 1 and Table 2 display the test set accuracies of different prompting experiments on the RES/REC/None and ADP/ADW/None tasks, respectively. As our test set is fairly balanced, the accuracy seems to be a reasonable metric.

From the Table 1, it is evident that most of the accuracy gains come from providing more context (88%). This makes intuitive sense: an MI expert may need to analyze more conversation history before concluding if a therapist’s utterance contains reflection. Additionally, few-shot and chain-of-thought prompting help in further improving the accuracy. The best accuracy (90%, experiment 6) on this task comes from a prompt with an elaborate task description along with 3 examples of the task and the use of chain-of-thought reasoning in the system prompt and up to 5 conversation turns in the user prompt.

Our results on the ADP/ADW/None task (Table 2) are mixed: the best-performing results come from employing basic prompting strategies. (84% accuracy). We performed an extensive qualitative analysis of our experiment

results to explain the discrepancy in the performance of prompts on these two tasks.

7 Qualitative Results

Upon closer examination of the cases where LLM’s classification of the advice or reflection label differed from the expert label in the dataset, we conclude that the dataset labels are ambiguous and can be contested. This highlights how individual experts differ in their interpretation of behaviour codes, and further motivates us to create a dataset including inter-annotator agreement on various MISC codes. We show an example below which highlights how the LLM’s reasoning is correct, even on the “wrong” answer:

Client:	Yeah.
Therapist:	But on the other hand, it does sound like there are things that you- that you value that you don’t want the drinking to interfere with.
Dataset label:	Simple reflection
LLM label:	No reflection
LLM reasoning:	The therapist here is rather asking a question that follows up on a client’s previous statement and does not make a reflective statement. Therefore, no reflection can be identified in this therapist’s utterance.

Table 3: Example of incorrect classification by LLM. The dataset label (simple) seems wrong.

Client:	I didn’t realize that drinking had anything to do with blood pressure.
Therapist:	Mm-hmm. Yeah, it can sometimes.
Client:	Oh, well. Well, then I guess that could be a problem.
Therapist:	My perspective as a doctor, it would be good if you could think about cutting back a little bit or doing something different about your drinking. What-what do you think about that?
Dataset label:	Advice with permission
LLM label:	Advice without permission
LLM reasoning:	The therapist’s utterance contains advice, as indicated by the phrases "it would be good if you could think about cutting back a little bit" and "doing something different about your drinking." This is advice because it suggests a course of action for the client to consider regarding their drinking habits. However, the therapist does not explicitly ask for permission before giving this advice. Instead, the therapist frames the advice as their perspective ("My perspective as a doctor"), and then follows it with a question seeking the client’s thoughts ("What-what do you think about that?"). This question does not constitute asking for permission but rather seeks the client’s reaction to the advice already given. Therefore, the therapist’s utterance is an example of advice WITHOUT permission.

Table 4: Example of incorrect classification by LLM, But the reasoning is sound (Correct answer is ADP).

Similar conclusions can be drawn about the examples from our labelled ADP/ADW/None test set. We note that our labelled test set is not large enough to draw any meaningful conclusions about the success of our prompting strategy. In addition, our labels may not align with the MISC manual as we lack expertise in MI. Nonetheless, we find that the LLM reasoning is sound even in some cases where it fails to provide the correct label.

8 Discussion and Learnings

Based on the results from these two sub-tasks, we conclude that LLMs such as GPT-4 are reasonably equipped to provide the correct MISC behaviour codes given enough context and examples in the prompt. We show one such prompt in [Appendix A](#). We further conclude that recent advancements in prompt engineering, such as few-shot and chain-of-thought prompting, help improve the elicitation of correct answers by the LLMs. The inconclusive results from the ADP/ADW/None task highlight the importance of expert labels and domain knowledge in labelling. Looking back, it would have been beneficial to concentrate on labelling more data. We could then confidently conclude whether the prompting strategies could generalize on tasks of similar complexity.

Although we are satisfied with the results of our experiment, we could have formulated the annotation task as a multi-class classification task, detecting the presence of multiple behaviour codes like question, reflection, and advice in the same prompt. This could potentially save us from going over our project budget. We plan to experiment with this idea in the future. As some of the utterances have ambiguous labelling, we further plan to explore the idea of eliciting the confidence of LLMs in their answer [15]. This way, we can build an application that has the option to defer to experts in cases where the LLM’s confidence is low.

9 Individual Contributions

S. No.	Project member name	Contribution
1	Zafar	Created initial GitHub code and experiment skeleton
2		Conducted exploratory data analysis
3		Utilized GPT-4 to identify mislabeled transcripts in the AnnoMI dataset
4		Stratified splitting of the dataset
5		Designed simple prompt for reflection classification
6		Explored Few-shot, Chain-of-thought prompts for reflection classification
7		Experiment to provide more conversation context to the reflection task
8		Experiment to expand the task description for the reflection classification
9		Analyzed failure cases for reflection classification
10	Sepehr	Developed initial script & prompt for reflection classification
11		Labeled ADP/ADW/None test and train set
12		Wrote script to help identify examples containing potential advice labels
13		Replicated prompting techniques from reflection to advice label
14		Improved chain-of-thought reasoning on the advice classification task
15		Conducted error analysis for advice detection & classification

Table 5: Description of the contribution to the project by each team member

References

- [1] G. Bischof, A. Bischof, and H.-J. Rumpf, “Motivational interviewing: An evidence-based approach for use in medical practice,” *Deutsches Aerzteblatt Online*, vol. 118, Feb. 2021. DOI: [10.3238/arztebl.m2021.0014](https://doi.org/10.3238/arztebl.m2021.0014).
- [2] W. R. Miller and S. Rollnick, “Ten things that motivational interviewing is not.,” *Behavioural and cognitive psychotherapy*, vol. 37 2, pp. 129–40, 2009. [Online]. Available: <https://api.semanticscholar.org/CorpusID:35320869>.
- [3] P. Amrhein, “How does motivational interviewing work? what client talk reveals,” *Journal of Cognitive Psychotherapy*, vol. 18, pp. 323–336, Dec. 2004. DOI: [10.1891/jcop.18.4.323.64001](https://doi.org/10.1891/jcop.18.4.323.64001).

- [4] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change, 3rd edition*, ser. Applications of motivational interviewing. New York, NY, US: Guilford Press, 2013, pp. xii, 482–xii, 482.
- [5] P. Amrhein, W. R. Miller, T. Moyers, and D. Ernst, *Manual for the motivational interviewing skill code (misc)*, Jan. 2008. [Online]. Available: <https://digitalcommons.montclair.edu/psychology-facpubs/27/>.
- [6] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, “Can large language models aid in annotating speech emotional data? uncovering new frontiers,” *arXiv preprint arXiv:2307.06090*, 2023.
- [7] A. Ziegler and J. Berryman, *A developer’s guide to prompt engineering and llms - the github blog*, 2023. [Online]. Available: <https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/>.
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [9] T. B. Moyers, L. N. Rowell, J. K. Manuel, D. Ernst, and J. M. Houck, “The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity,” en, *J. Subst. Abuse Treat.*, vol. 65, pp. 36–42, Jun. 2016.
- [10] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, “Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues,” *Future Internet*, vol. 15, no. 3, 2023, ISSN: 1999-5903. DOI: [10.3390/fi15030110](https://doi.org/10.3390/fi15030110). [Online]. Available: <https://www.mdpi.com/1999-5903/15/3/110>.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: [1810.04805](https://arxiv.org/abs/1810.04805) [cs.CL].
- [12] A. Brown, A. T. Kumar, O. Melamed, I. Ahmed, Y. H. Wang, A. Deza, M. Morcos, L. Zhu, M. Maslej, N. Minian, V. Sujaya, J. Wolff, O. Doggett, M. Iantorno, M. Ratto, P. Selby, and J. Rose, “A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study,” en, *JMIR Ment. Health*, vol. 10, e49132, Oct. 2023.
- [13] OpenAI, *New models and developer products announced at devday*, 2023. [Online]. Available: <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>.
- [14] R. L. L. I. au2, I. Balaevi, E. Wallace, F. Petroni, S. Singh, and S. Riedel, *Cutting down on prompts and parameters: Simple few-shot learning with language models*, 2021. arXiv: [2106.13353](https://arxiv.org/abs/2106.13353) [cs.CL].
- [15] V. Shrivastava, P. Liang, and A. Kumar, *Llamas know what gpts don’t show: Surrogate models for confidence estimation*, 2023. arXiv: [2311.08877](https://arxiv.org/abs/2311.08877) [cs.CL].

Appendices

A Example prompt on the ADP/ADW/None task

The following is a snippet from a motivational interviewing conversation. Show your thinking process step by step and then return "ADP", "ADW" or "none" based on whether the therapist's utterance has advice WITH permission, advice WITHOUT permission, or no advice.

Your reasoning should align with the Motivational Interviewing Skills Code (MISC) 2.1 manual. Remember, utterances with advice will usually contain language that indicates that advice is being given: should, why don't you, consider, try, suggest, advise, you could. Advice WITH permission can be in the form of a request from the client or in the therapist asking the client's permission to offer it. Watch for indirect forms of permission asking, such as giving the client the option to disregard the advice.

Some examples of advice without permission from MISC:

- Consider buying more fruits and vegetables when you shop.
- You could ask your friends not to drink at your house.

Now look at the following example and work with me through the reasoning.

Example:

Client: I've heard about those benefits. I guess I'm just worried about finding the time and staying motivated.

Therapist: Those are valid concerns. Would you like some suggestions on how you might overcome these challenges?

Client: Sure, I'd appreciate that.

Therapist: One approach could be starting with small, manageable goals, like a short walk each day or a few minutes of stretching. It's also helpful to find activities you enjoy. And remember, it's okay to start slow and build up gradually. What do you think about trying something like this?

Answer: The therapist uses ADP by first asking permission before giving advice, discussing the benefits of the suggested change, and then seeking the client's thoughts on the advice given. Hence the answer is: ADP

B Permissions

	Team Members	
	Sepehr	Zafar
Permission to Post Video	Yes	Yes
Permission to Post Final Report	Yes	Yes
Permission to Post Source Code	Yes	Yes

Table 6: Permission to post items on the course website.