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

Sepehr Ahmadi and Zafarullah Mahmood

Department of Electrical & Computer Engineering, University of Toronto

Word Count: 1198

Penalty: 0%

1 Introduction

We aim to build a large language model (LLM) based application that assists experts in annotating motivational interviewing (MI) transcripts. Existing MI datasets [1], [2] rely on experienced MI practitioners to annotate transcripts which is both arduous and time-consuming. Moreover, the high cost of this process prohibits sending a reasonably sized dataset to multiple expert annotators and calculating a reliable inter-annotator agreement (IAA) [1].

We propose to build an LLM-in-the-loop annotation framework where an LLM provides annotation suggestions to an expert throughout an annotation session. The expert is free to incorporate or discard LLM suggestions. As the session progresses, the LLM refines its suggestions based on the actions taken by the expert. We hypothesize that this approach will reduce the annotation time without degrading the quality.

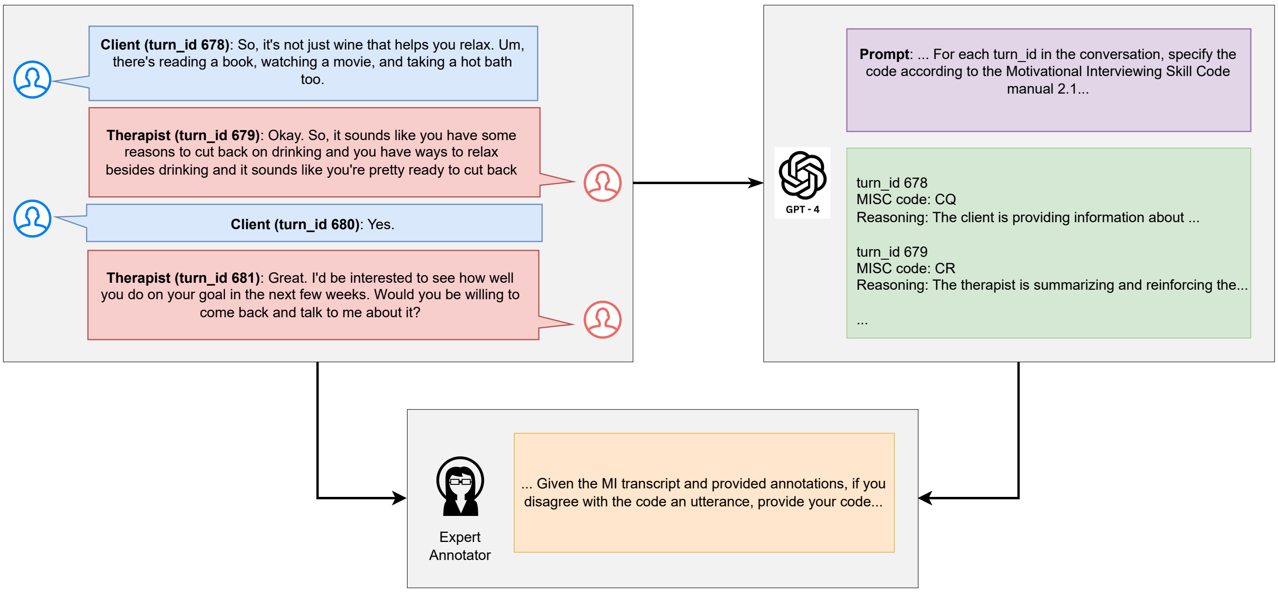
In this work, we will try different prompting strategies on a diverse range of LLMs to generate initial and refined annotation suggestions. To confidently measure the performance of LLMs, we plan to augment existing MI datasets [1] with more MISC labels [3] and report the performance of LLMs on each label category.

2 Background

Motivational Interviewing (MI) is an evidence-based [4] form of counselling with the goal to "evoke and strengthen a person's motivation for change" [5]. It is particularly useful in cases where a client shows ambivalent feelings about changing their behaviour [6] and often lacks clarity on how the behaviour affects their life goals and values [4]. The MI counsellor's job is to facilitate the client's journey from ambivalence to *change talk* — a state where the client expresses a desire or commitment to change [7]. To successfully do this, the counsellor employs, among other techniques, four characteristic conversational strategies: **O**pen-ended questioning, personal **A**ffirmations, **R**eflective listening and thoughtful **S**ummarization [8]. These are collectively known as the OARS skillset for MI.

Although at first glance, MI conversations exhibit high-level patterns, on a granular level, they display the art of thought elicitation, empathetic listening and compassionate counselling deeply embedded in the human experience. Therefore they provide a benchmark to evaluate the NLP systems on their understanding of nuanced conversations. These conversations can also potentially act as training examples for AI-based MI chatbots, where their role is to demonstrate to chatbots the differences between concepts like simple and complex reflections, helpful and unsympathetic affirmations, etc. We, therefore, argue that annotated MI conversations are valuable resources in the evolution of these chatbots.

Motivated by similar arguments, several attempts have been made to create a dataset of annotated MI conversations. The most notable work is by Zixiu et al. who released AnnoMI, a dataset of "133 faithfully transcribed and expert-annotated demonstrations of high- and low-quality" MI conversations [1]. As noted above, a limiting factor in expanding such datasets is the effort and time of expert annotators. In this work, we explore integrating LLMs into the annotation framework to make the annotation process efficient.



Client (turn_id 678): So, it's not just wine that helps you relax. Um, there's reading a book, watching a movie, and taking a hot bath too.

Therapist (turn_id 679): Okay. So, it sounds like you have some reasons to cut back on drinking and you have ways to relax besides drinking and it sounds like you're pretty ready to cut back

Client (turn_id 680): Yes.

Therapist (turn_id 681): Great. I'd be interested to see how well you do on your goal in the next few weeks. Would you be willing to come back and talk to me about it?

Prompt: ... For each turn_id in the conversation, specify the code according to the Motivational Interviewing Skill Code manual 2.1...

turn_id 678
MISC code: CQ
Reasoning: The client is providing information about ...

turn_id 679
MISC code: CR
Reasoning: The therapist is summarizing and reinforcing the...

...

... Given the MI transcript and provided annotations, if you disagree with the code an utterance, provide your code...

Figure 1: The proposed annotation framework where the LLM first annotates the transcripts by providing the transcript and the expert annotator reviews it and suggests changes.

3 Source of Data and Processing

We plan to use the AnnoMI [1] dataset in this work. The dataset contains over 13000 annotated utterances from real MI conversations. It annotates therapist utterances with two useful codes: reflection (“simple” or “complex”) and question (“open” or “closed”). These codes are part of the Motivational Interviewing Skills Code (MISC) 2.1 manual [3], which provides a coding system to assess the MI-consistency of therapist behaviours and responses.

Before using the dataset, we intend to remove several inconsistencies in it. For example, at some locations, the interlocutor (“therapist” or “client”) labels are misplaced. We also note that the dataset deviates in its definition of *utternace* from the MISC manual [3], which defines it as “a complete thought” that can be assigned a code [3]. A single therapist turn or *volley* can have multiple utterances with different codes. The dataset treats each turn as an utterance and only provides one code per turn even if it has multiple thoughts deserving different codes. We propose to fix this inconsistency using GPT-4 followed by manual verification.

As a part of our data augmentation effort, we hope to annotate therapist utterances with an additional Advise (ADP/ADW) code and use the LLM reasoning to provide granular MISC codes for client change/sustain talks (Reason, Other, Taking Steps, Commitment, Follow/Neutral).

3.1 Dataset splitting

For LLM prompt engineering and API setup, we will use a randomly selected subset of 20 AnnoMI conversations stratified by topic and quality and report the metrics on the remaining 113 conversations.

4 LLMs for Automatic MI Annotation

In this section, we discuss our approach to utilizing LLMs for MI annotation, mainly focusing on the open-source Llama 2 model, and later expanding to more advanced models like GPT-3.5 and GPT-4. We outline our strategies for prompting, integration with annotation platforms, and adherence to MISC guidelines, along with efforts to optimize costs and computational efficiency

As most LLM APIs are expensive, we plan to use the open-source 7B parameter Llama 2 model [9] as a baseline. This will allow us to experiment with different prompting strategies, integrate the model with our annotation platform and get the baseline results. Once we complete our baseline setup, we will expand our suite of LLMs to OpenAI GPT-3.5 and GPT-4 models. For these LLMs, we will try zero-shot, few-shot and chain-of-thought [10] prompting and select the best strategy for final testing. We will also record the LLM reasoning and compare them with the MISC guidelines [3].

Week Start	Week End	Estimated Workload (hrs)	Work	Owner
October 30	November 5	6	Setting up LLM APIs	Sepehr
		9	Setting up the annotation platform	Zafar
		9	AnnoMI cleaning	Sepehr
November 6	November 12	12	Evaluation of baseline model	Zafar
		9	Removing dataset inconsistencies	Sepehr
		12	Multicoding each turn	Zafar
November 13	November 19	15	Data labelling for the new Advise (ADP/ADW) code	Sepehr
		18	Initial Results on AnnoMI - Baseline	Zafar
		9	**Progress Report**	Sepehr
November 20	November 26	15	Testing LLMs on AnnoMI and gathering acceptance rates	Zafar
		15	Relevant metrics calculations	Sepehr
		6	Result analysis and Discussion	Zafar
November 27	December 3	18	Code refactoring	Sepehr
		12	**Final Presentation Slides**	Zafar
December 4	December 10	24	**Final Report**	Sepehr

Table 1: Breakdown of weekly workload for our project

Since LLM inference is computationally expensive, we will find ways to reduce our API costs. For example, we will explore how to get annotations for multiple utterances in a single API. Finally, we will report the acceptance rate for each MISC code LLM pair and the average LLM API requests per conversation for each LLM.

5 System Architecture Details

Figure 1 describes our annotation framework which is a user interface that queries the LLM via APIs at various annotation stages. It displays the queried LLM annotations to the expert. Every time the expert accepts or rejects the annotation from LLM, we record this event and in the subsequent LLM query, we present this as a positive or negative example in the prompt. This way, as the annotation progresses, the LLM annotations will be refined and most likely be accepted by the annotator.

6 Plan

Table 1 provides a detailed breakdown of our project plan, including sub-tasks, owners, and estimated time-frames for each task. To ensure effective collaboration, we will conduct regular team meetings and maintain open communication channels throughout the project's lifecycle.

7 Risks

The most prominent risk we face in this project is that LLMs might not be able to provide the right MISC code for utterances and their suggestions will be discarded by the experts. Our preliminary work suggests otherwise. LLMs such as GPT-4 are able to provide the right codes and their reasoning, though debatable, is often sensible. Nevertheless, we will conduct rigorous experimentation to provide the right context to LLMs.

Our team is not an MI expert, and we risk quality issues with our label augmentation. To mitigate this, we will adhere to MISC guidelines and show the LLM results on our labels separately.

8 Conclusion

In this proposal, we discuss the need for an LLM-integrated interface for MI annotation and describe our plan to experiment with harnessing LLMs for this task. We also provide a rough sketch for our annotation framework. Finally, we explain the existing issues with the dataset, and the risks associated with this work and highlight our mitigation plans.

References

[1] 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. [Online]. Available: https://www.mdpi.com/1999-5903/15/3/110.

[2] V. Pérez-Rosas, X. Wu, K. Resnicow, and R. Mihalcea, "What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 926–935.

[3] 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/.

[4] 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.

[5] 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.

[6] S. W. Feldstein Ewing, T. R. Apodaca, and J. Gaume, "Ambivalence: Prerequisite for success in motivational interviewing with adolescents?" *Addiction*, vol. 111, no. 11, pp. 1900–1907, 2016. DOI: https://doi.org/10.1111/add.13286. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/add.13286. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/add.13286.

[7] 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.

[8] 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.

[9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL].

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, *Chain-of-thought prompting elicits reasoning in large language models*, 2023. arXiv: 2201.11903 [cs.CL].