Article

Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues [†]

Zixiu Wu [1,2,‡], Simone Balloccu [3,‡], Vivek Kumar [2], Rim Helaoui [1], Diego Reforgiato Recupero [2] and Daniele Riboni [2,*]

[1] Philips Research, High Tech Campus, 5656 AE Eindhoven, The Netherlands
[2] Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
[3] Department of Computing Science, University of Aberdeen, Aberdeen AB24 3FX, UK
[*] Correspondence: riboni@unica.it

[†] This paper is an extended version of our paper published in the Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022. Z.W. was a knowledge contributor during the dataset collection process and was given access to AnnoMI after it was made available in the ICASSP submission.

[‡] These authors contributed equally to this work.

Abstract: Research on the analysis of counselling conversations through natural language processing methods has seen remarkable growth in recent years. However, the potential of this field is still greatly limited by the lack of access to publicly available therapy dialogues, especially those with expert annotations, but it has been alleviated thanks to the recent release of AnnoMI, the first publicly and freely available conversation dataset of 133 faithfully transcribed and expert-annotated demonstrations of high- and low-quality motivational interviewing (MI)—an effective therapy strategy that evokes client motivation for positive change. In this work, we introduce new expert-annotated utterance attributes to AnnoMI and describe the entire data collection process in more detail, including dialogue source selection, transcription, annotation, and post-processing. Based on the expert annotations on key MI aspects, we carry out thorough analyses of AnnoMI with respect to counselling-related properties on the utterance, conversation, and corpus levels. Furthermore, we introduce utterance-level prediction tasks with potential real-world impacts and build baseline models. Finally, we examine the performance of the models on dialogues of different topics and probe the generalisability of the models to unseen topics.

Keywords: dialogue; counselling; motivational interviewing; natural language processing; dataset

check for updates

Citation: Wu, Z.; Balloccu, S.; Kumar, V.; Helaoui, R.; Reforgiato Recupero, D.; Riboni, D. Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues. *Future Internet* **2023**, *15*, 110. https://doi.org/10.3390/fi15030110

Academic Editor: Paulo Quaresma

Received: 15 February 2023
Revised: 6 March 2023
Accepted: 8 March 2023
Published: 14 March 2023

1. Introduction

Patient health can be significantly improved by changes in behaviour, such as reducing alcohol consumption [1]. Counsellors, however, may have difficulty in convincing patients to adopt such changes. Thus, motivational interviewing (MI) [2] has been developed as an effective counselling approach that evokes motivation for change from the client themselves (A client in counselling is not necessarily a patient, so the word “client” is used in this study). Correspondingly, coding systems such as MISC (Motivational Interviewing Skill Code) [3] and MITI (Motivational Interviewing Treatment Integrity) [4] are commonly used to identify MI codes and aspects related to the therapist and client.

Recent years have seen significant interest in the research of linguistic and statistical MI analysis. The first computational model for identifying reflection, a key skill in MI, was introduced by Can et al. [5]. More broadly, the modelling of MI-related aspects such as codes and therapist empathy has been approached with methods based on classical machine learning [6–8] (e.g., support vector machines) and deep learning [9–12] (e.g., recurrent neural networks). In terms of data resources, Pérez-Rosas et al. [13] recently published a corpus of MI conversations automatically transcribed from video-sharing websites, where some dialogues showcase high-quality MI and the others illustrate low-quality MI.

Despite the progress, natural language processing (NLP) for MI is still an extremely low-resource domain, owing to privacy-related restrictions. As research in this field has been conducted primarily on private/undisclosed annotated MI dialogues, it has been challenging to replicate and further develop prior work. Previously, to the best of our knowledge, the only publicly available dataset of MI conversations was created by Pérez-Rosas et al. [13] through the automatic captioning of YouTube/Vimeo videos. However, the transcript quality is compromised by the considerable transcription errors from automatic captioning, which can make the transcripts difficult to understand. Pérez-Rosas et al. [13] also analysed two MI codes—reflection and question—based on the dataset annotations from trained students, but these annotations are unavailable at the time of writing.

To alleviate the scarcity of resources for NLP-for-MI research, we previously introduced AnnoMI [14], a dataset of 133 MI-adherent and non-adherent therapy conversations that are (a) professionally transcribed from YouTube/Vimeo videos demonstrating MI, and (b) annotated for key MI aspects by experienced MI practitioners. We also obtained explicit permission from the video owners for creating, releasing, and analysing the dataset. We note that "MI-adherent" and "MI non-adherent" are synonyms for "high-quality" and "low-quality" respectively, which describe therapy quality rather than video/transcription quality.

In this work, we expand [14] significantly as follows:

1. We release the full version of AnnoMI, which has several fine-grained additional attributes. We also elaborate on the details of data collection and processing, including the results of a post-annotation survey for the annotators, which suggest that the dataset reflects real-world high- and low-quality MI even though its dialogues are from demonstration videos.
2. We present detailed, visualised statistical analyses of the expanded dataset to examine its patterns and properties.
3. We establish two AnnoMI-based utterance-level classification tasks with potential for real-world applications: therapist behaviour prediction and client talk type prediction. We also experiment with various machine learning models as baselines for these tasks to facilitate comparison with future methods.
4. We explore the performance of these models on different topics, as well as their generalisability to new topics.

The motivation for predicting therapist behaviour and client talk type is that accurate models for these tasks can automatically label utterances and therefore facilitate therapy quality monitoring and provide feedback for the therapist (Section 2.1), thus ultimately improving counselling quality. We also note that points 3 and 4 are distinctly different from previous work on automated MISC/MITI coding, since (a) the dataset is annotated following a MISC/MITI-inspired scheme rather than the original MISC/MITI, and (b) we are focused on establishing baselines for the new tasks and probing the impact of dialogue topics on the performance, instead of pursuing the state-of-the-art.

After listing the background and related work in Section 2, we describe our MI video acquisition, transcription, and annotator recruitment in Section 3. The annotation scheme is detailed in Section 4, while we show the results of inter-annotator agreement in Section 5. We present thorough analyses of AnnoMI in Section 6 and then introduce the utterance-level prediction tasks and their baselines in Section 7. Topic-related analyses are shown in Section 8 and discussions over the creation and application of AnnoMI are given in Section 9, before this work is concluded in Section 10.

The data collection process is described in its entirety—as opposed to only the elements related to the newly introduced utterance attributes—to facilitate a better understanding of this work. The same applies to the analyses in Section 6.1, which have appeared in [14], as they provide a basis and context for the subsequent analyses and experiments in this study.

2. Background and Related Work

2.1. MI Coding

The gold standard for examining counsellor adherence to therapy protocols is behavioural observation and coding [15], which provides feedback and evaluation of therapy sessions. During the coding process, trained annotators assign labels to therapist skills/behaviours such as reflection and client behaviours such as change talk. Session-level ratings on qualities such as empathy are often also included.

A variety of coding schemes have been proposed, including the Motivational Interviewing Skill Code (MISC) [3] and the Motivational Interviewing Treatment Integrity Code (MITI) [4]. However, as manual coding is costly and time-consuming, automatic coding of utterance-level behaviour and related tasks such as automatic rating of therapist empathy have garnered significant research interest in recent years.

2.2. Available Resources

MI conversation resources are scarce. As real-world therapy often contains sensitive topics and information, counselling dialogues are mostly privately owned or proprietary (e.g., therapy transcripts from Alexander Street [https://alexanderstreet.com/products/counseling-therapy-video-library], accessed on 14 February 2023). As for resources, annotated MI corpora such as [16] have been built from sources such as wellness coaching phone calls and leveraged for tasks such as utterance-level code prediction [17] and empathy prediction [18], but they mostly remain publicly inaccessible.

To the best of our knowledge, the only freely and publicly available MI corpus to date is [13], created based on automatic transcripts of MI videos on YouTube/Vimeo. Pérez-Rosas et al. [13] also collected annotations with respect to reflections and questions for the corpus and conducted related analyses, but these annotations are not available at the time of writing. Moreover, considerable transcription errors from automatic captioning are present in the corpus (Section 3.2), thus limiting the quality of the dataset.

2.3. Text-Based Approaches to MI Analysis

In terms of text-based approaches to automatic coding, Can et al. [5] used N-grams and similarity features to develop the first model for identifying reflection, while Atkins et al. [7] used a labelled topic model to generate MI codes. More recently, deep-learning-based models have been utilised. For example, Xiao et al. [10] and Tanana et al. [19] studied the use of RNNs for behaviour prediction, followed by Gibson et al. [11], who did so under a multi-label multi-task setting to improve the performance, as well as Cao et al. [12], who also investigated the forecasting of the codes of upcoming utterances.

For therapist empathy modelling, an early approach was from Xiao et al. [6] with an n-gram language model. Gibson et al. [8] leveraged language features inspired by psycholinguistic norms, while Gibson et al. [9] used LSTMs to produce turn-level behavioural acts further processed to predict empathy. Separately, Wu et al. [20,21] explored leveraging links between therapeutic and general conversation empathy to tackle therapist empathy prediction in low-resource scenarios.

2.4. Speech-Based and Multimodal Methods for MI Analysis

For utterance-level code prediction, Singla et al. [22] proposed an LSTM-based [23] end-to-end model that directly predicts codes given speech features without using automatic speech recognition. Most other works leveraging speech features for code prediction exploit multimodal features, such as [24,25], where LSTMs with joint prosodic and lexical features are utilised.

For session-level therapist empathy modelling, more speech-only methods have been proposed, including Xiao et al. [26], who studied prosodic features such as jitter and shimmer from speech signals, as well as Xiao et al. [27], who investigated speech rate entrainment. Flemotomos et al. [28] proposed an automatic rating tool for MI sessions using speech and language features, predicting a range of session-level codes including empathy and MI spirit, in addition to utterance-level codes.

3. Creating AnnoMI

Considering the scarcity/absence of publicly available conversation datasets of real-life MI and their privacy-related legal and ethical restrictions, we rely instead on demonstrations of MI-adherent and non-adherent therapy from online video-sharing platforms, in a similar vein to [13]. With explicit consent from the video owners, we obtain professional transcripts of the demonstrations and recruit MI experts to annotate the transcripts following a scheme covering key MI elements.

3.1. MI Demonstration Videos

To balance privacy restrictions and the faithfulness of therapy, we leverage MI demonstrations on video-sharing websites (YouTube and Vimeo). We identified 346 videos that demonstrate high- and low-quality MI, using key phrases including "good motivational interviewing" and "bad MI counselling". According to the literature [2], high-quality MI is centred on the client and conducted with empathy, whereas low-quality MI is characterised by frequent instructions and suggestions.

We label each video to be high-quality MI or low-quality using its title (e.g., "Motivational Interviewing—Good Example"/"The Ineffective Physician: Non-Motivational Approach"), as well as descriptions and narrator comments (e.g., "Demonstration of the motivational interviewing approach in a brief medical encounter"). We consider such labelling to be verified automatically, as the video uploaders are professional MI practitioners and organisations focused on healthcare and behaviour change. We also note that the definition of high- and low-quality MI is clear in the literature ([1,2], inter alia); therefore, the high/low MI quality divide is consistent across different institutions/therapists and different demonstrations.

We gained explicit permission from the content owners (as well as the individuals appearing in the videos if applicable) for us to use their videos to create, analyse, and publicly release a transcript-based MI dialogue dataset. We eventually obtained permission to use 119 (42 overlapped with [13]) of these videos, which contain 133 complete conversations—a video may contain multiple dialogues. Moreover, 110 of the dialogues showcase high-quality MI, with 8839 utterances in total, and the other 23 dialogues illustrate low-quality MI, with 860 utterances in total. As shown in Figure 1, high-quality MI dialogues are generally longer than low-quality ones, with several surpassing 200 utterances in length, but most dialogues have less than 100 utterances. A pair of high- and low-quality MI session excerpts, both about smoking cessation/reduction, are presented in Table 1.

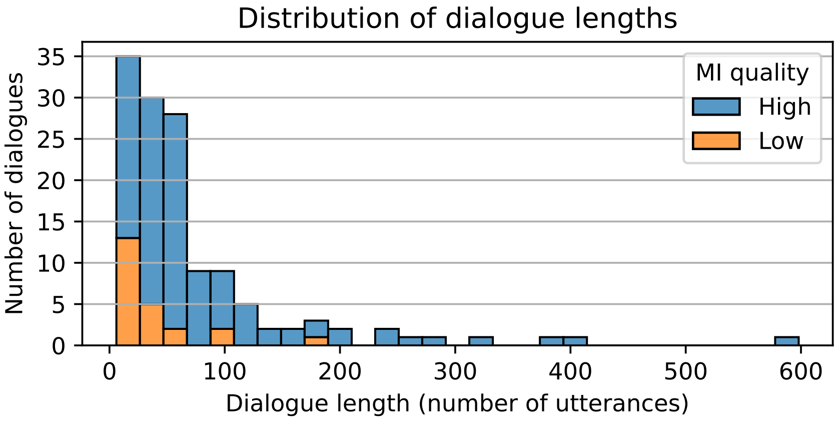

Figure 1. Distribution of dialogue lengths (number of utterances per dialogue).

Table 1. High- and low-quality MI conversation snippets, where the goal is smoking cessation/reduction. ***T***: therapist; ***C***: client.

High-Quality MI
T: Um, I did wanna talk to you though. I'm a little bit concerned looking through his chart of how many ear infections he's had recently. And I-I noticed that you had checked the box that someone's smoking in the home. So I was wondering if you can tell me a little more about that.
C: Well, um, It's just me and him and I do smoke. Um, I try really hard not to smoke around him, but I-I've been smoking for 10 years except when I was pregnant with him. But it– everything is so stressful being a single mom and-and my having a full-time job. And so it's just—that's why I started smoking again.
T: You have a lot of things going on and smoking's kind of a way to relax and destress.
C: Yeah. Some people have a glass of wine. I have a cigarette.
T: Sure. And it sounds like you're trying not to smoke around him. Why did you make that decision?
Low-Quality MI
T: Well, now's the time to quit. It's really gotten to the point where you can't keep smoking. Not only for him, like I said, but also for you. You're putting yourself at risk for lung cancer, for emphysema, for oral cancers, for heart disease, for all kinds of things-
C: I know, I know. I've heard– People have told me before, I've heard all that. I just don't know how to do it. How am I supposed to quit? It's-it's so hard.
T: Well, there's all kinds of things you can use now. It's not as hard as it used to be. You can use nicotine replacement. There's patches, there's lozenges, there's gum, there's the inhaler, there's nasal spray. We can talk about medications. You can try Chantix, you can try Zyban, there's quit smoking groups you can go to, there's hotlines you can call.
C: I just don't have time for any of that.

The imbalance with respect to high- and low-quality MI dialogue volumes can be attributed to the following: (a) fewer low-quality MI video owners responded to our request or consented; (b) low-quality MI videos are relatively scarce on Youtube/Vimeo, possibly because MI adherence demonstrations are deemed more valuable as "good examples" and thus filmed and uploaded more.

3.2. Transcription

Using a professional transcription service (https://gotranscript.com/, accessed on 14 February 2023), we collected fluent and faithfully transcribed MI conversations from the videos, whereas the transcripts of [13] were produced by automatic captioning. While a step of verifying video content–caption matching is reported in [13], in practice, we find a considerable number of incorrectly transcribed words/phrases and mismatched speaker (therapist/client) labels in the corpus of [13], which can significantly hinder text understanding. Table 2 presents transcript snippets from [13] and AnnoMI of the same video to exemplify the marked difference in transcription quality between the two datasets. AnnoMI is also free from other noise, such as narrations, but retains context-relevant details, including "hmm", "right", and speaker sentiment/emotion [29–31] indicators such as "[laugh]".

3.3. Expert Annotators and Workload Assignment

As MI annotation requires specialised knowledge, we rely on experienced MI practitioners to annotate the transcripts. Specifically, 10 therapists found through the Motivational Interviewing Network of Trainers (https://motivationalinterviewing.org/trainer-listing, accessed on 14 February 2023), an international organisation of MI trainers and a widely recognised authority in MI, were recruited for the task. All the annotators had a high proficiency in English and prior experience in practising and coding MI. Informed consent was collected from all the annotators.

Table 2. Transcription quality comparison between AnnoMI and [13]. Red: incorrectly transcribed word; Blue: omitted words/phrases; Orange: words from the other speaker that should have started a new utterance; ~~strikethrough~~: incorrectly transcribed word within such a misplaced utterance; {C}/{T}: missing client/therapist utterance.

AnnoMI
C: Right. Well, it would be good if I knew, you know, that my kids are taken care of too- ***T***: Yeah. ***C***: - so I'm not worried about them while I'm at work. ***T***: Right. Yeah. Because you're- you're the kind of parent that wants to make sure your kids are doing well. ***C***: Right. ***T***: Yeah. Um, so tell me, what would it take to get you to like a five in confidence, to feel a little bit more confident about getting work? ***C***: Well, I mean, being able to make the interviews would be the priority. ***T***: Okay, Yeah. ***C***: Um, so chi- you know, taking care, having some childcare, having- ***T***: Mm-hmm. ***C***: - having someone I trust that I can call when I know I've got an interview. ***T***: Yeah. Because you definitely need to go to an interview in order to get the job. ***C***: Right. Yeah. ***T***: So having taken care of that part, having some reliable childcare would definitely help. ***C***: Yeah.
[13]
C: one it would be good if I knew you know that my kids are taking care of ("too") - yeah so I'm not worried about them law in the work right yeah ***T***: because you're you're the kind of parent that wants to make sure your kids are doing well great ({C}) yeah um so tell me what would it take to get you to like a five in confidence to feel a little bit more confident ("about") getting work ***C***: well I mean being able to make the interviews would be the priority again ({T}) um so try you know taking care having some child care I mean having ({T}) someone I trust that I can call when I you know what that interview because you definitely need to go to an interview ~~of~~ in order to get ~~three~~ ("the job") ***T***: ~~yeah~~ yeah so having taken care of that part having some reliable child care ("would definitely help") ***C***: yeah definitely not

Overall, each expert annotated 19 to 20 transcripts, with total lengths around 144 min in terms of the total duration of the original 19 to 20 videos. To facilitate computation of inter-annotator agreement (IAA), we selected 7 common transcripts to be annotated by all experts, based on 3 criteria: (1) they should add up to approximately 1/3 (45 min) of the workload of each annotator; (2) they should cover diverse topics (6 out of the 7 transcripts have distinct topics); (3) they should cover both high- and low-quality demonstrations (5 showing high-quality MI and 2 showing low-quality). We tried various combinations of transcripts before we found one combination that satisfied the criteria above. During the annotation process, no expert was aware that a part of their workload would be used to compute the IAA. Each of the 126 (133-7) non-IAA transcripts was annotated by one expert due to our budget limits.

We note that the IAA results of AnnoMI are not directly comparable with those of other annotated MI corpora, since the former are calculated based on the annotations of 10 experts, while the latter often come from much fewer (e.g., 2 or 3) annotators, and it is usually less likely to reach the same or a higher level of IAA with more annotators. This also means that the attributes of AnnoMI that do have good IAAs are indeed reliably annotated.

3.4. AnnoMI and "Real-World" MI

For AnnoMI to be useful for real-world applications, it is crucial that its dialogues reflect both high- and low-quality MI in the real world. Therefore, we surveyed the

10 annotators after they completed their tasks, asking them whether they felt that the AnnoMI dialogues resembled real-world MI , and we eventually received responses from 6 annotators. As shown in Figure 2, 83% of the responses "agree" or "somewhat agree" that the therapist utterances and the dialogues overall reflect real-world MI, and the figure is 66% for the client utterances. The clear majority in each case shows that AnnoMI indeed sufficiently captures the characteristics of real-world MI, even though the dialogue sources are demonstrations.

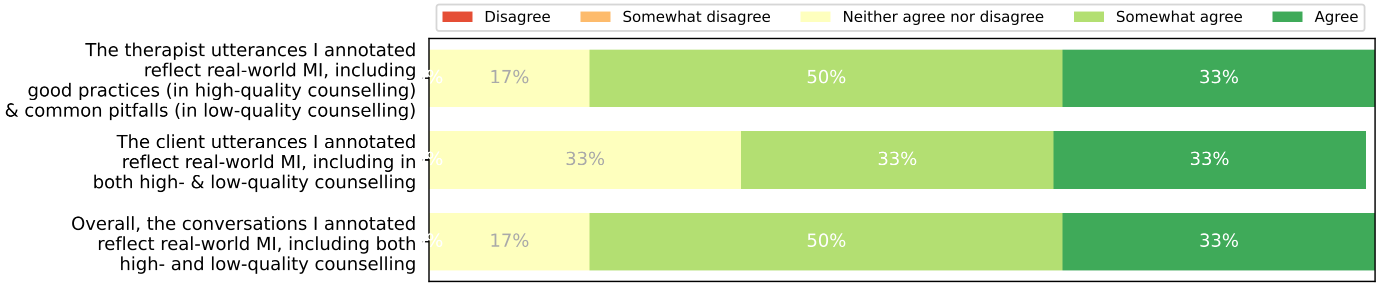

Figure 2. Results of survey of annotators regarding whether AnnoMI reflects real-world high- and low-quality MI.

We note that researchers, especially those in corporate environments, are faced with a very challenging legal and regulatory landscape in the field of NLP for counselling, due to privacy-related concerns and rules in different jurisdictions. Therefore, a dataset such as ours can be used significantly more broadly, since it does not have any privacy implications or legal issues concerning different jurisdictions.

4. Annotation Scheme

We design a detailed annotation scheme to study therapist and client behaviours, based on the MI literature, existing coding protocols (MISC/MITI), and feedback from therapists. At the dialogue level, we asked the annotators to briefly describe the dialogue's goal, e.g., "smoking cessation". Thus, we summarise in Table 3 the top 10 topics in terms of (a) the number of conversations that have these topics, and (b) the total number of utterances in these conversations. At the utterance level, the annotation scheme is as shown in Table 4. When annotating an utterance, an annotator could also see the preceding and subsequent utterances for more context.

4.1. Therapist Utterance Attributes

4.1.1. (Main) Behaviour

In MI, three fundamental yet crucial skills to achieve effective counselling are asking, informing, and listening [1]. In view of this principle and the related components of mainstream coding schemes for MI, we consider Question, Input, and Reflection as major therapist behaviours that correspond to asking, informing, and listening, respectively. In cases where more than one behaviour is present in an utterance, e.g., a question after an input, the expert is asked to further select the Main Behaviour. Other is listed as the fourth and default option, where no Question, Input, or Reflection appears in the utterance.

We also list Question, Input, and Reflection as separate attributes of therapist utterances in order to investigate their sub-types, as laid out in the sections below.

Table 3. Top 10 topics in AnnoMI in terms of (a) number and percentage of dialogues that have those topics, and (b) total number and percentage of utterances in those dialogues.

Topic	Dialogues
Reducing alcohol consumption	28 (21.1%)
Smoking cessation	21 (15.8%)
Weight loss	9 (6.8%)
Taking medicine/following medical procedure	9 (6.8%)
More exercise/increasing activity	9 (6.8%)
Reducing drug use	8 (6.0%)
Reducing recidivism	7 (5.3%)
Compliance with rules	5 (3.8%)
Asthma management	5 (3.8%)
Diabetes management	5 (3.8%)
Other	33 (24.8%)
Topic	**Utterances**
Reducing alcohol consumption	1914 (19.7%)
Reducing recidivism	1303 (13.4%)
Smoking cessation	1106 (11.4%)
Diabetes management	709 (7.3%)
Reducing drug use	578 (6.0%)
Taking medicine/following medical procedure	574 (5.9%)
More exercise/increasing activity	525 (5.4%)
Weight loss	396 (4.1%)
Avoiding DUI	394 (4.1%)
Changing approach to disease	315 (3.2%)
Other	2107 (21.7%)

Table 4. Utterance-level multi-choice annotation scheme. (+) implies presence of utterance attribute (e.g., “Simple reflection“ entails that Reflection exists in utterance), while (-) indicates absence thereof (e.g., “No reflection” label implies that Reflection is not present in utterance).

Therapist Utterance Attributes	Label
(Main) Behaviour	Question Input Reflection Other
Question	Open question (+) Closed question (+) No question (-)
Input	Information (+) Advice (+) Options (+) Negotiation/goal setting (+) No input (-)
Reflection	Simple reflection (+) Complex reflection (+) No reflection (-)
Client Utterance Attribute	**Label**
Talk Type	Change Neutral Sustain

We note that this work is more focused on the use of asking, informing, and listening in the AnnoMI dialogues. For this purpose, the original annotation scheme was more ambitious and had several non-MITI/MISC annotation fields, but they are not included in this paper due to their very low IAAs (Fleiss’ kappa), and thus the annotation scheme presented in this section may resemble a subset/regrouping of MISC to some readers. This work does not seek to compare directly with previous studies that use the complete MISC/MITI for annotation.

4.1.2. Question

Therapists use asking to develop an understanding of the client and their problems. Therefore, we include Question as a therapist behaviour and define any question as open or closed in accordance with mainstream MI coding conventions. An open question allows a wide range of possible answers and may seek information, invite the client's perspective, or encourage self-exploration, while a closed question implies a short answer such as Yes/No, a specific fact, a number, etc. [3]. Some examples are given in Table 5.

Table 5. Example labelling for therapist Question from the dataset.

Utterance	Question Type
Do you have children in your house?	Closed (Yes/No answer)
How much does it actually cost you a week?	Closed (Number)
Okay. What kind of alcohol do you drink at parties?	Closed (Specific fact)
So what is a typical week for you as far as your alcohol use is concerned?	Open (Seek information)
Okay. So how do you feel about being here today?	Open (Invite client's perspective)
So, when you think about what you like and don't like about your drinking, where do you wanna go from here?	Open (Encourage self-exploration)

4.1.3. Input

The primary manner of communicating knowledge to the client is informing. Based on MISC coding [3] and insights from a professional therapist regarding the patterns of informing in the AnnoMI transcripts, we use the term Input to include a wide range of conveyed knowledge and consider 4 subtypes: providing information, giving advice, presenting options, and setting goals (negotiation). Some examples are given in Table 6. When an utterance contains more than one type of Input, the annotators choose the main type of Input to make the labels mutually exclusive and facilitate utterance-level NLP applications.

Table 6. Example labelling for therapist Input from the dataset.

Utterance	Input Type
You're not alone in feeling that way. Binge drinking can feel normal to some people.	Information
So that's a hormone that allows you to utilise sugar in your body.	Information
I want you to be healthy. And I don't want to see you coming back in here for something else. So I'm really gonna recommend that you try to cut down to that amount.	Advice
That's why I recommend that all my adolescent patients not drink at all.	Advice
So, what have you looked into about, um, you know, advocacy in that area or expungement or anything like that?	Options
Okay. So, exploring some yoga classes. Is doing yoga in your living room appealing to you at all?	Options
So for you being in your class, when that bell rings, then you know, this is the goal.	Negotiation/goal setting
Do you think you could go two months without drinking?	Negotiation/goal setting

4.1.4. Reflection

Reflection is an essential means of listening. In using reflections, the therapist shows that they are listening to and understanding the client, which is effective in helping people to change. Following MISC, we consider two reflection types: simple and complex. A simple reflection shows an understanding of the client's words but contains little additional meaning—for example, by repeating the client's statement. In comparison, a complex reflection conveys a deeper level of understanding of the client's point of view and adds substantial meaning to the client's statement, using techniques such as metaphors and exaggeration [1]. Two pairs of contrasting simple and complex reflections to the same client statement are presented in Table 7. Clearly, the simple reflections identify the client's emotion/situation but do not go beyond the overt content of the client's statement, while the complex reflections "continue the paragraph" by interpreting the client's words and anticipating what they might reasonably say next.

Table 7. Example labelling for therapist Reflection based on the dataset.

	Scenario 1 — Smoking Cessation	
Speaker	**Utterance**	**Reflection Type**
Client	Um, I try really hard not to smoke around him, but I-I've been smoking for 10 years except when I was pregnant with him. But it– everything is so stressful being a single mom and-and my having a full-time job. And so it's just– that's why I started smoking again.	
Therapist 1	Things are very stressful for you right now.	Simple
Therapist 2	You have a lot of things going on and smoking's kind of a way to relax and de-stress.	Complex
	Scenario 2 — Reducing Alcohol Consumption	
Speaker	**Utterance**	**Reflection Type**
Client	Um, I've been really trying not to, but, you know, weekends come around, and, um, all my friends are kind of partying and stuff, and it's been hard to, like, break that habit.	
Therapist 1	It's quite a challenge for you.	Simple
Therapist 2	Mm-hmm. So, there's this external pressure coming from the people you care about to sort of stay in the scene.	Complex

4.2. Client Utterance Attribute (Talk Type)

According to the MI literature [1], clients usually feel ambivalent about adopting positive behaviour change, and thus the desirable outcome of MI is for the client to pick up pro-change arguments and talk themselves into changing, provided that it aligns with their aspirations and values. This type of talk that favours change is known as "change talk". Conversely, a "sustain talk" conveys resistance to behaviour change and favours the status quo. On the other hand, a "neutral talk" indicates no preference for or against change. Hence, we name Change Talk, Sustain Talk, and Neutral Talk as three types of the client Talk Type attribute. Table 8 presents some examples of these talk types in different scenarios, such as reducing alcohol consumption.

Table 8. Example labelling for client Talk Type from the dataset.

Utterance	**Talk Type**
Yeah, I just want to do what's right.	Change
Well, that was fine until I came here, um, but now that I know about the health risk, um, I have something I gotta think about.	Change
Um, I mean, the 10 drinks seems like not a lot for me and my tolerance.	Sustain
Yeah, whatever. I Know you got to do your job, but I don't care.	Sustain
Yeah, I would like to play soccer in college.	Neutral
And um, I think she used to look after me because she used to do the cooking and stuff like that.	Neutral

5. Inter-Annotator Agreement (IAA)

5.1. Default Measure: Fleiss' Kappa at Utterance Level

We use Fleiss' kappa [32] as the default measure for calculating utterance-level inter-annotator agreement (IAA) over the annotations on the 7 transcripts. We consider 3 methods of calculation: ALL, ALL(STRICT), and BINARY. ALL applies to all the utterance attributes, while the other two modes apply to Input, Reflection, and Question only.

Specifically, since these three attributes have a default "absence" option (i.e., no input, no reflection, and no question, as shown in Table 4), we compute a two-class presence-vs.-absence (i.e., BINARY) IAA for them, in addition to the fine-grained all-class IAA (i.e., ALL). When computing ALL-IAA for Question, for example, we consider the original label space: {Open question (+), Closed question (+), No question (-)}, where (+) means that there is a question in the utterance and (-) means that there is not. Conversely, we only consider the presence-vs.-absence {(+), (-)} space when calculating BINARY-IAA.

We also calculate ALL(STRICT)-IAA, which computes IAA within the original label space but on a more challenging subset of utterances, motivated by the observation that it is substantially more difficult to distinguish between the presence (+) labels than between presence (+) and absence (-). For example, differentiating between "Simple reflection (+)" and "Complex reflection (+)" is more difficult than between Reflection and non-Reflection. Therefore, we compute ALL(STRICT) on the utterances where at least one annotator chose a presence (+) option. For Reflection, for example, we calculate ALL(STRICT)-IAA on the utterances where at least one annotator selected "Simple reflection (+)" or "Complex reflection (+)".

5.2. Results of Default IAA Measure

All Fleiss'-kappa-based IAAs are listed in Table 9. Following [33], we group the IAAs into slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.80–1.00) agreement. We consider an attribute predictable if its IAA shows moderate or better agreement.

Table 9. Inter-annotator agreements on utterance-level annotations, in Fleiss' kappa. Orange, blue, cyan, and green indicate fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.80–1.00) agreement, respectively.

Therapist Utterance Attribute	IAA Setting	IAA
Input	ALL(STRICT)	0.34
	ALL	0.51
	BINARY	0.64
Reflection	ALL(STRICT)	0.32
	ALL	0.50
	BINARY	0.66
Question	ALL(STRICT)	0.54
	ALL	0.74
	BINARY	0.87
(Main) Behaviour	ALL	0.74
Client Utterance Attribute	**IAA Setting**	**IAA**
Talk Type	ALL	0.47

We notice that for the utterance attributes where BINARY and ALL(STRICT) are applicable, the order of agreement is, without exception, ALL(STRICT)-IAA < ALL-IAA < BINARY-IAA, which proves the challenge of the subset used for computing ALL(STRICT)-IAA, as well as the ease of annotating the absence/presence of a particular utterance attribute.

The annotators show fair agreement on Input and Reflection under ALL(STRICT), which reveals the difficulty of annotating these attributes despite their inclusion in MISC/MITI, particularly when their presence in an utterance cannot be easily ruled out. Nevertheless, the IAA jumps to substantial agreement for Input and Reflection under the BINARY setting, which suggests the presence of distinguishable linguistic features unique to these two attributes.

Encouragingly, Question, (Main) Behaviour, and Talk Type all record moderate or better IAAs under all settings, which shows the text-based predictability and therefore the existence of distinct linguistic features of these attributes.

5.3. Supplementary IAA Measure: Intraclass Correlation

Following MITI, we also use intraclass correlation (ICC) to analyse (Main) Behaviour and Talk Type at the label level to gain more insights and facilitate comparison with other studies. For each label, we count the number of occurrences of utterances annotated with the label in each session by each annotator. Thus, each of the 10 annotators has 7 label counts corresponding to the 7 IAA transcripts. Then, ICC is computed to describe how much of the total variation in the label counts is due to differences among annotators. Moreover, following MITI, the ICC scores are obtained using a two-way mixed model with absolute agreement and average measures.

As Table 10 presents, all the (Main) Behaviour and Talk Type labels have excellent (0.75–1) [34] agreement scores, which shows the reliability of these annotations. Nevertheless, Change Talk and Sustain Talk have slightly lower ICCs—around 0.9—compared to the other ICCs that are almost 1.0, which somewhat echoes the lower Fleiss'-kappa-based IAA of Talk Type compared to that of (Main) Behaviour.

Table 10. Inter-annotator agreement as intraclass correlation.

(Main) Therapist Behaviour	ICC
Input	0.975
Reflection	0.991
Question	0.997
Other	0.996
Client Talk Type	ICC
Change	0.916
Neutral	0.986
Sustain	0.890

5.4. IAA and Full Dataset Release

We release the full version of AnnoMI, which has the following attributes:

- **Question**: {Open question, Closed question, No question}
- **Input**: {Information, Advice, Options, Negotiation/Goal-Setting, No input}
- **Reflection**: {Simple reflection, Complex reflection, No reflection}
- **(Main) Behaviour**: {Question, Input, Reflection, Other}
- **Talk Type**: {Change, Neutral, Sustain}

For the 7 IAA transcripts that are annotated multiple times, we release the annotations from each expert, but we take a majority vote over multiple annotations in order to facilitate dataset analysis (Section 6) and experiments (Sections 7 and 8). Compared to the original version [14], which only contains (Main) Behaviour and Talk Type annotations, the full version can enable the development of fine-grained classifiers operating at the label sub-type level, such as simple and complex reflection. While we do not pursue such development in this study due to the lower ALL(STRICT) IAAs of some attributes, future work may explore the treatment of these labels as probabilistic for model training and other research purposes.

6. Dataset Analysis

We analyse the annotations via visualisations. Unless otherwise specified, (Main) Behaviour represents the behaviour of an utterance. For example, if a therapist utterance consists of a reflection and a question but Reflection is annotated as the Main Behaviour, we consider the utterance to be a reflection instead of a question, in order to facilitate further analysis.

We also note that while there are clear correlations between utterance attribute distribution and MI quality in some cases, they do not necessarily point to causation, especially given the relatively low amount of data and potential sampling bias.

6.1. General (Main) Behaviour and Talk Type Distributions

As Table 11 demonstrates, the most marked contrast between therapist behaviours in MI-adherent and non-adherent therapy is the proportions of Reflection and Input. The average MI-adherent therapist employs Reflection in 28% of their utterances, whereas it is only 7% in non-adherent therapy, echoing the MI requirement of trying to understand the client's perspective and communicating it. On the other hand, Input is given 33% of the time in low-quality MI but only 11% in high-quality MI, which, together with the statistics of Reflection, conforms to the observation [1] that high-quality MI emphasises understanding the client as opposed to speaking from their own point of view. The correlation between MI quality and the share of Question and Other is relatively weak.

Table 11. (Main) Behaviour distributions in high- and low-quality MI.

	High-Quality MI	**Low-Quality MI**
Reflection	28%	7%
Question	28%	32%
Input	11%	33%
Other	33%	28%

As for Talk Type, Change Talk is more frequent in high-quality MI—25% vs. 17%—whereas Sustain Talk has a stronger presence in low-quality MI—11% vs. 15% (Table 12). These contrasts are, nevertheless, less obvious than those found in Reflection and Input. Possible explanations include (a) some clients in low-quality MI could adopt tepid change-talk-like speech such as "Yeah, maybe" only to end the counselling quickly, and (b) some clients in high-quality MI are simply more reluctant to change but the therapist still respects this, as is recommended in MI. On the other hand, most (64–68%) client utterances belong to the neutral talk category regardless of MI quality, for which the prevalence of short utterances such as "Mhmm" and "Uh huh" can be a major contributing factor.

Table 12. Talk Type distributions in high- and low-quality MI.

	High-Quality MI	**Low-Quality MI**
Change	25%	17%
Neutral	64%	68%
Sustain	11%	15%

6.2. Posterior (Main) Behaviour and Talk Type Distributions

MI guidelines have specific recommendations on how a therapist should respond when the client talks in certain ways, and a client may also react to the therapist in particular patterns. We therefore probe the posterior distributions of next-turn therapist behaviours (/client talk types) given the current-turn client talk type (/therapist behaviour). Denoting u_t^T as the therapist utterance at turn (time step) t and u_{t+1}^C as the client reply in the following turn, the posterior distribution of client talk types can be represented as $p(Talk\_Type(u_{t+1}^C) \mid Behaviour(u_t^T))$. Similarly, the posterior distribution of therapist behaviours can be formulated as $p(Behaviour(u_{t+1}^T) \mid Talk\_Type(u_t^C))$

Figure 3 presents the posterior distribution of client talk types (i.e., $p(Talk\_Type(u_{t+1}^C) \mid Behaviour(u_t^T))$). While neutral talk is clearly the majority talk type of the client response, in most cases, $p(Talk\_Type(u_{t+1}^C) = \text{Change} \mid Behaviour(u_t^T))$ is substantially larger in high-quality MI than in low-quality MI regardless of $Behaviour(u_t^T)$, which shows that an MI-adherent counsellor is more likely to evoke Change Talk from the client, irrespective of specific therapist behaviours. On a more granular level, Question is the most likely (31%) therapist behaviour in high-quality MI to evoke Change Talk, which may be because some therapist questions lead to Change Talk more often, such as asking the client what steps they could take towards a behaviour change or how confident they are about adopting a change. Interestingly, Input results in more Change Talk (21%) than any other therapist behaviour in low-quality MI, but it is also the therapist behaviour that prompts the most (23%) sustain talk, which may suggest that the effect of frequent input—characteristic of low-quality MI, as shown in Table 11—is far from certain in terms of evoking Change Talk and reducing sustain talk.

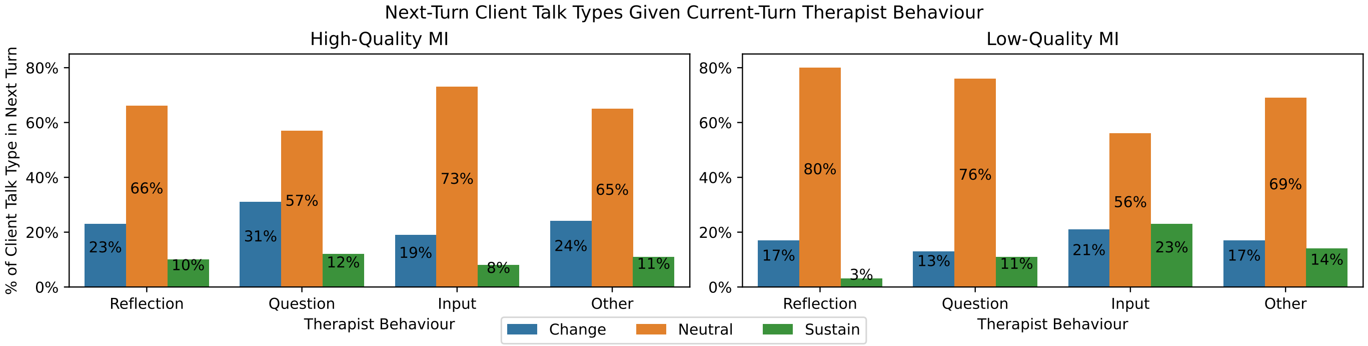

Figure 3. Distribution of next-turn client talk types given different therapist behaviours in the current turn.

Figure 4 shows the posterior distribution of therapist behaviours (i.e., $p(Behaviour(u^T_{t+1}) \mid Talk\_Type(u^C_t)))$. One can observe that MI-adherent therapists in general use considerably more reflections than non-adherent therapists do—30% vs. 12%—in response to Change Talk, which suggests that high-quality MI utilises Reflection to reinforce willingness to change. On the other hand, the most commonly shown therapist behaviour in response to Sustain Talk in high-quality MI is Reflection (37%), while the dominant pattern of reacting to Sustain Talk in low-quality MI is Input (54%). This contrast serves as strong evidence that MI-adherent therapy focuses more on showing empathy and trying to understand the client when faced with resistance, including through Reflection, whereas a non-adherent therapist is more likely to try to challenge, correct, or persuade the client through more Input—a common mistake in MI non-adherent therapy [1].

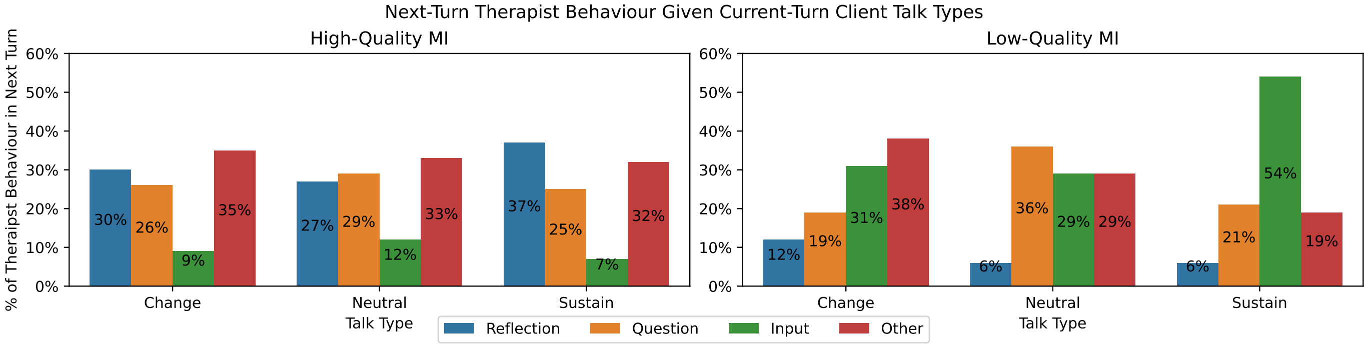

Figure 4. Distribution of next-turn therapist behaviours given different client talk types in the current turn.

6.3. (Main) Behaviour and Talk Type as Conversation Proceeds

Following [13], we divide each conversation into 5 parts: $[0.0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$ and $(0.8, 1.0]$, in order to probe conversational properties at different dialogue stages. Specifically, we examine the distributions of different therapist behaviours and client talk types at these stages.

Among the trends shown in Figure 5, one can observe in both high- and low-quality MI that the proportion of Question gradually decreases as the therapist gathers more information about the client from the progressing conversation. The amount of Reflection, on the other hand, generally fluctuates within a small interval throughout a dialogue in both high- (27–31%) and low-quality MI (2–8%), which means that Reflection is common throughout a high-quality MI session and rare throughout a low-quality one. Finally, the proportion of Input rises during the middle stages ($(0.4, 0.8]$) in both high- and low-quality MI, but the increase is substantially more pronounced in low-quality MI sessions (from ∼30% to ∼60%) than in high-quality ones (from ∼10% to ∼15%), which further indicates that a non-adherent therapist tends to talk from their own perspective more as the conversation develops.

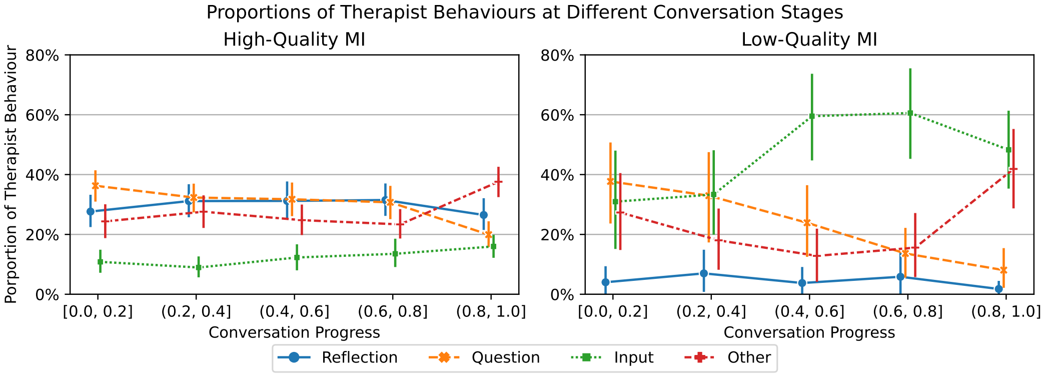

Figure 5. Proportions of therapist behaviours in different conversation stages in high- and low-quality MI. The marked data points are sample means, and the error bars around them are calculated using bootstrapping with a 95% confidence interval.

The trends of different client talk types are displayed in Figure 6. A clear shift is shown in high-quality MI: there are similar amounts of Change Talk and Sustain Talk at the beginning of a conversation, but Change Talk becomes more present steadily and eventually reaches around 40% at the end of a dialogue, while the share of Sustain Talk diminishes gradually at the same time and drops to around 7%. In other words, the desired effects of MI-adherent therapy, namely change talk evocation and sustain talk reduction, become increasingly prominent with the progression of a session. In low-quality MI, however, during the early and middle conversation stages (i.e., $[0.0, 0.6]$), the proportion of Sustain Talk soars from approximately 10% to a little over 40%, while the number for Change Talk remains under 10%. Interestingly, the later stages (i.e., $(0.6, 1.0]$) show the opposite trend, as the growing share of Change Talk surpasses the declining proportion of Sustain Talk, finishing at around 30% and 12%, respectively, at the end. Nevertheless, the absolute %[Change Talk]–%[Sustain Talk] difference is clearly larger at the end of a high-quality MI session in general.

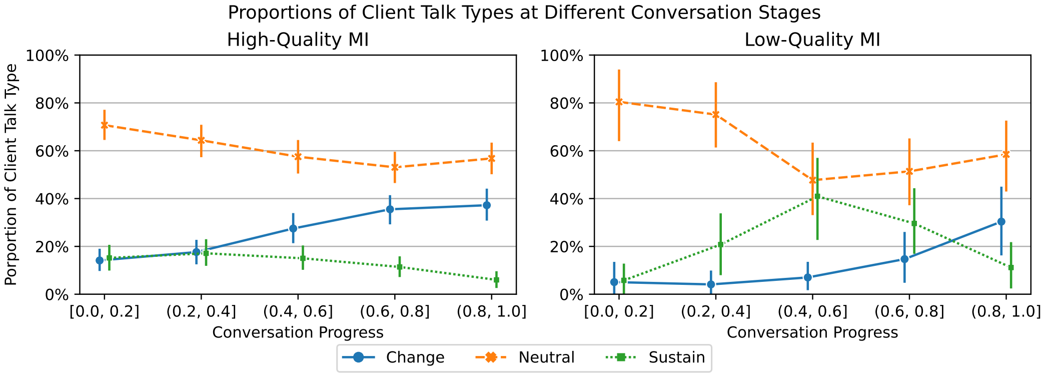

Figure 6. Proportions of client talk types in different conversation stages in high- and low-quality MI. The marked data points are sample means, and the error bars around them are calculated using bootstrapping with a 95% confidence interval.

6.4. Utterance Length Distributions

Following [13], we study the lengths (number of words) of utterances of different types. To better represent the distribution of individual utterance lengths, we opt for violin plots to render a kernel density estimation of each underlying distribution, with the first, second, and third quartiles marked as dashed lines. This applies to Figures 7–9, though Figure 9 shows the distribution of utterance length ratios instead of absolute lengths.

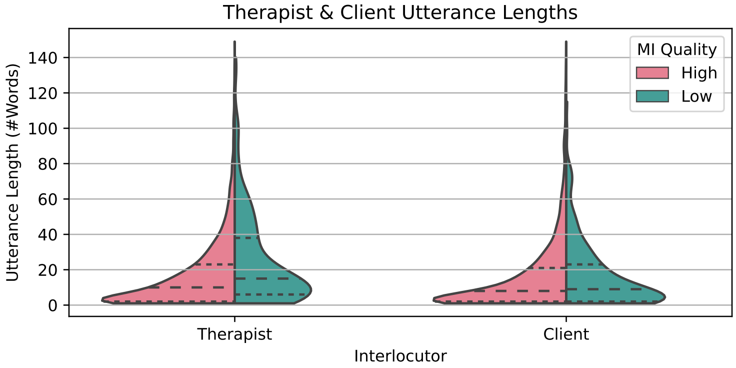

Figure 7. Lengths (number of words) of therapist and client utterances in high- and low-quality MI.

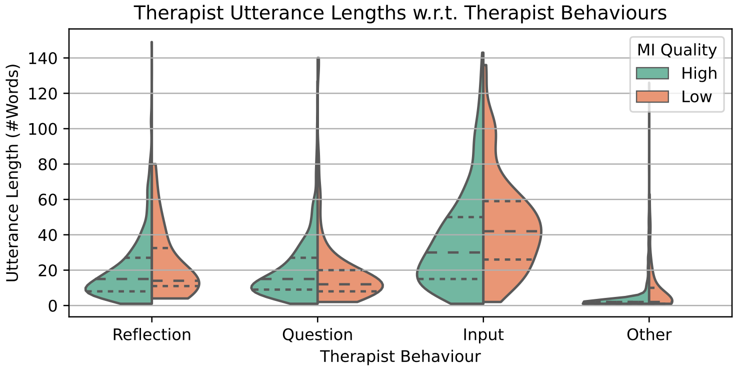

Figure 8. Utterance lengths (number of words) of different therapist behaviours in high- and low-quality MI.

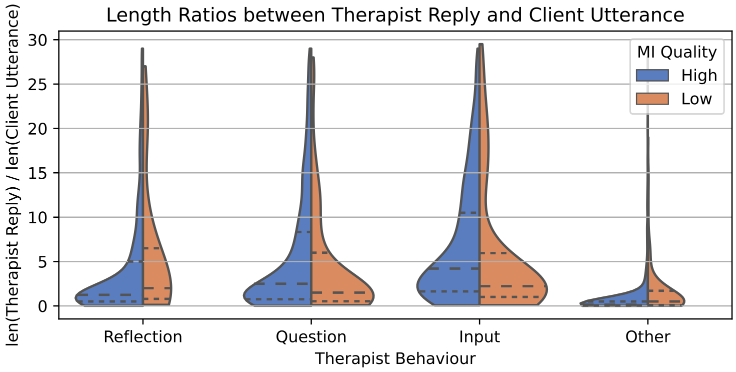

Figure 9. Ratios between the length (number of words) of the next-turn therapist response (broken down into 4 types of therapist behaviours) and that of the current-turn client utterance.

Figure 7 shows the therapist and client utterance length distributions in high- and low-quality MI. It is clear that the client utterance distributions are similar in MI-adherent and non-adherent sessions, whereas therapist utterances are generally shorter in high-quality MI than in low-quality MI, which is another indicator that an MI-adherent therapist takes more time to actively listen to and understand their client.

Figure 8 shows a more fine-grained therapist utterance length distribution with regard to each therapist behaviour. For Reflection, the median utterance length is roughly the same in high- and low-quality MI, but the proportion of shorter utterances is clearly larger in the former. In terms of Question, an MI-adherent therapist tends to pose slightly longer questions than their non-adherent counterpart, which may suggest that an MI-adherent therapist more often asks tailored and nuanced questions. Input is substantially longer in both high- and low-quality MI, but input from an MI-non-adherent therapist is generally 10 words or more longer than that from an adherent counsellor, indicating the relatively larger degree to which an MI-non-adherent therapist talks from their own perspective. Finally, the generally short (no more than a few words) utterances of the Other behaviour show that, similar to neutral talk, these utterances mostly carry little meaning and are often simply used to facilitate the conversation.

We also investigate the length ratio between a therapist reply and its immediately preceding client utterance, which shows how much longer the therapist "talks in return". As illustrated in Figure 9, Reflection has smaller length ratios in high-quality MI than in low-quality MI (e.g., for high-quality MI, smaller length ratios near the bottom of the Y-axis dominate the distribution), while Question shows the opposite, both of which are in line with the previous observation of the absolute lengths of Reflection and Question utterances in Figure 8. However, the Input length ratios are generally larger in high-quality MI sessions than in low-quality ones, which could be attributed to the fact that an MI-adherent therapist usually asks for permission first before providing input. More specifically, the therapist might say "May I explain to you what those numbers in your blood test results mean?", and the client would simply say "Yes" or "Sure, why not", before the therapist replies with a substantially longer Input utterance, thus leading to a larger utterance length ratio.

6.5. Frequent 3-Grams

Table 13 lists the most frequent 3-grams in the therapist utterances of each behaviour and in the client utterances of each talk type. It is clear from the table that an MI-adherent therapist tends to use "it sounds like" to initiate a reflection—as is recommended by MI guidelines—more often than a non-adherent counsellor. Otherwise, however, the frequent 3-grams reveal little about the characteristics of utterances of different types or MI qualities. This suggests that utterance-level semantic differences are more nuanced and contextualised.

Table 13. Most frequent 3-grams of (1) therapist utterances of different (main) behaviours and (2) client utterances of different talk types, in high- and low-quality MI.

	Therapist			**Client**	
	High-Quality MI	**Low-Quality MI**		**High-Quality MI**	**Low-Quality MI**
Reflection	"it sounds like" (78) "sounds like you" (56) "a little bit" (51) "you do n't" (43) "a lot of" (39)	"'re gon na" (4) "you do n't" (3) "you 're here" (3) "you 're gon" (3) "you 've already" (2)	Change Talk	"I do n't" (188) "do n't know" (68) "I 'm not" (42) "do n't want" (30) "I think I" (30)	"I do n't" (8) "I guess I" (6) "I think I" (5) "do n't know" (4) "I-I guess I" (4)
Question	"do you think" (91) "a little bit" (62) "me a little" (35) "little bit about" (33) "I 'm wondering" (30)	"do you think" (11) "you think you" (6) "a lot of" (5) "that you 're" (5) "you 're not" (5)	Neutral Talk	"I do n't" (261) "do n't know" (142) "I 'm not" (53) "do n't really" (47) "I did n't" (27)	"I do n't" (27) "I 'm not" (7) "do n't know" (7) "I 've been" (6) "I have n't" (5)
Input	"a lot of" (32) "a little bit" (27) "one of the" (16) "that you 're" (13) "you 'd be" (13)	"a lot of" (15) "you need to" (11) "that you 're" (9) "'s gon na" (8) "that you 've" (7)	Sustain Talk	"I do n't" (135) "do n't know" (57) "I 'm not" (28) "it 's not" (23) "do n't really" (23)	"I do n't" (14) "I 'm not" (8) "do n't know" (5) "It 's just" (5) "I just need" (4)
Other	"for coming in" (12) "that you 're" (8) "a little bit" (8) "coming in today" (7) "I do n't" (7)	"that 's certainly" (2) "so it 's" (2) "you 're not" (2) "you 're still" (2) "be able to" (2)			

6.6. Utterance Embedding Distribution

To further investigate the semantic-level differences between utterances of different types, we probe the clustering of utterance embeddings. Specifically, we obtain the utterance embeddings using a language model [35] that is lightweight and performs well on sentence embedding tasks (https://www.sbert.net/docs/pretrained_models.html, accessed on 10 March 2023) as a sequence-level encoder. Through t-SNE (maximum 1000 iterations, perplexity = 30) [36]—an unsupervised, non-linear technique for visualising high-dimensional data—Figures 10 and 11 show that there is no obvious clustering of utterances of the same therapist (Main) Behaviour or client Talk Type, which is evidence that more advanced machine-learning-based methods are needed to distinguish between utterances of different types.

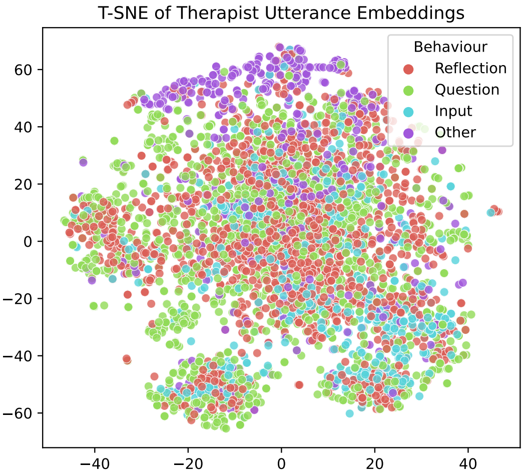

Figure 10. t-SNE of therapist utterance embeddings.

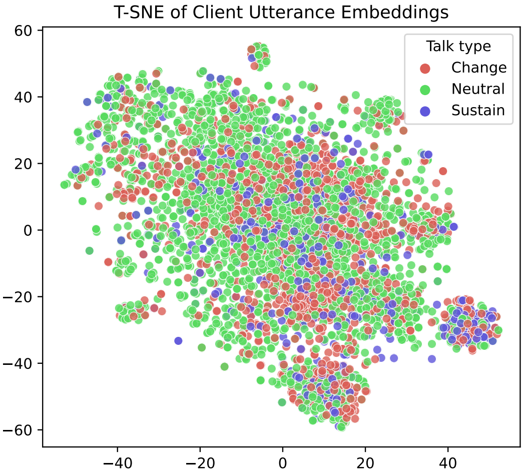

Figure 11. t-SNE of client utterance embeddings.

7. Utterance-Level Prediction Experiments

From the annotation labels, various utterance-level prediction tasks can be readily defined. In this section, we focus on two tasks: therapist behaviour prediction and client talk type prediction. We introduce these tasks as examples of potential real-world applications of AnnoMI, in order to inspire future tasks based on the dataset. From a practical point of view, an accurate prediction model of therapist behaviour and client talk type can automatically label utterances and thus facilitate therapy quality monitoring and provide feedback for the therapist (Section 2.1), ultimately improving counselling quality. In experimenting with these tasks, we also examine the impact of the relatively lower IAAs of some utterance attributes on the prediction performance.

While imbalance exists between the high- and low-quality dialogue volumes, we expect its impact on the tasks to be minor, since they are not related to MI quality directly. For future work exploring session- or utterance-level MI quality classification, however, remedies such as data augmentation will be needed to address the imbalance.

Each task allows a single utterance as the input and requires a class label as the output. We experiment with 4 machine learning models, as listed below. We implement the BERT variants with AdapterHub [37,38] (https://github.com/Adapter-Hub/adapter-transformers, accessed on 14 February 2023), the CNN models with Keras (https://keras.io/, accessed on 14 February 2023), and the other models with Scikit-learn [39].

- **BERT w/o Adapters**: BERT-base-uncased [40] fine-tuned on AnnoMI.
- **BERT w/ Adapters**: BERT-base-uncased with adapters [37,41] fine-tuned on AnnoMI. Adapters are a small set of task-specific parameters that can be easily plugged into transformer [42] models, so that only the lightweight adapters are updated during fine-tuning, while the rest of the model is frozen.
- **CNN**: convolutional neural networks initialised with word2vec embeddings [43] and fine-tuned on AnnoMI.
- **Random Forest**: random forest with tf-idf features.

We also use 2 random baseline classifiers for comparison:

- **Prior**: random prediction based on the class distribution in the training set;
- **Uniform**: random prediction based on the uniform distribution of the classes.

Since duplicate utterances are present in AnnoMI, especially in the categories of Other and Neutral Talk (e.g., “Uh-huh” and “OK”), we perform de-duplication as a preprocessing step. Specifically, if multiple identical utterances have the same label, we randomly select one of them to keep and remove the others. The distribution of (Main) Behaviour and Talk Type after this step is shown in Table 14.

Table 14. Distribution of (Main) Behaviour and Talk Type after de-duplication. Overall, (Main) Behaviour has 3796 unique examples and Talk Type has 3685.

(Main) Behaviour				Talk Type		
Reflection	**Question**	**Input**	**Other**	**Change Talk**	**Neutral Talk**	**Sustain Talk**
34%	36%	16%	14%	29%	57%	14%

Considering the relatively small size of the dataset, we conduct 5-fold cross-validation (CV) at the utterance level with stratification with regard to utterance labels, so that (1) the class distribution in each fold is close to being identical, and (2) each time, 4 folds are used as training and validation data and 1 fold is used as test data. The training to validation data ratio is 9:1, and we select the best-performing checkpoint (for CNN and BERT models) based on the performance on the validation set, so that we can test the checkpoint on the test set. We use Macro F1 as the validation and test metric, since it is commonly used for classification tasks and robust to class imbalance. For Prior and Uniform, whose outputs are random, we run the models 1000 times on the test data and calculate the average performance. Therefore, each of the 6 models listed above eventually has 5 performance values from 5-fold CV, and we take the mean as the final performance value of the model.

To address class imbalance, we introduce two versions for each training set: Original Unbalanced and Augmented Balanced. The former keeps the original data in each CV training set, while the latter leverages a Pegasus [44]-based neural paraphraser (https://huggingface.co/tuner007/pegasus_paraphrase, accessed on 14 February 2023) in order to augment the non-majority classes so that the size of each class in Augmented Balanced reaches that of the majority class in Original Unbalanced.

7.1. Task 1: Therapist Behaviour Prediction

We first investigate therapist behaviour prediction. Given a therapist utterance, the task is to predict its (Main) Behaviour. As shown in Table 15, the BERT variants score the highest, with macro F1s at 0.72, followed by CNN at 0.6 and Random Forest at approximately 0.5. Compared to the random baselines (Prior and Uniform) with macro F1s below 0.25, the trained models, especially the BERT variants, have clearly learned contextualised semantics. No substantial difference exists between the results of BERT w/o Adapters and BERT w/ Adapters. The effects of augmentation are minor and universally negative for the BERT variants and CNN.

Table 15. Macro F1 and per-class F1 scores of (main) therapist behaviour prediction. All results averaged from 5-fold cross-validation. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced.

Result Format	**Original Unbalanced (Augmented Balanced)**				
(Main) Therapist Behaviour Prediction					
Model	**F1-Macro**	**F1-Reflection**	**F1-Question**	**F1-Input**	**F1-Other**
BERT w/ Adapters	0.72 (0.70↓)	0.77 (0.75↓)	0.86 (0.84↓)	0.63 (0.60↓)	0.64 (0.62↓)
BERT w/o Adapters	0.72 (0.70↓)	0.77 (0.75↓)	0.85 (0.85)	0.63 (0.60↓)	0.64 (0.62↓)
CNN	0.60 (0.58↓)	0.64 (0.63↓)	0.70 (0.70)	0.50 (0.48↓)	0.56 (0.52↓)
Random Forest	0.50 (0.50)	0.56 (0.53↓)	0.58 (0.54↓)	0.41 (0.45↑)	0.46 (0.46)
Prior	0.25 (0.24↓)	0.34 (0.29↓)	0.36 (0.30↓)	0.16 (0.20↑)	0.14 (0.18↑)
Uniform	0.24 (0.24)	0.29 (0.29)	0.30 (0.29↓)	0.20 (0.20)	0.18 (0.18)

The order of per-class F1 by the best-performing models (BERT variants) is generally Input $\approx$ Other $<$ Reflection $<$ Question, which largely correlates with the order of proportions of those labels in the task (Table 14). The performance gap between different classes is not reduced in the Augmented Balanced scenario, which shows that this issue cannot be resolved by simple paraphrasing-based class-wise data augmentation. Interestingly, Question shows better (e.g., $\Delta = 0.09$ $F1$ for BERT w/ Adapters) performance than Reflection despite having similar amounts of examples, which may be because (1) Question utterances generally have syntactic cues such as question marks and are therefore easier to classify, and (2) Question has higher IAAs than Reflection (Table 9) and is thus less noisy.

By inspecting the confusion matrix (Figure 12) of BERT w/ Adapters in the Original Unbalanced setting, we observe that Input and Other utterances are most frequently misclassified into Reflection. Since Input and Other are less than half in size compared to Reflection (Table 14), this imbalance may have contributed to the instances of misclassification. On the other hand, Reflection and Question are similarly sized but Question contributes less to the misclassifications of Input and Other, which may again be linked to its syntactic cues and less noisy labels, as mentioned before.

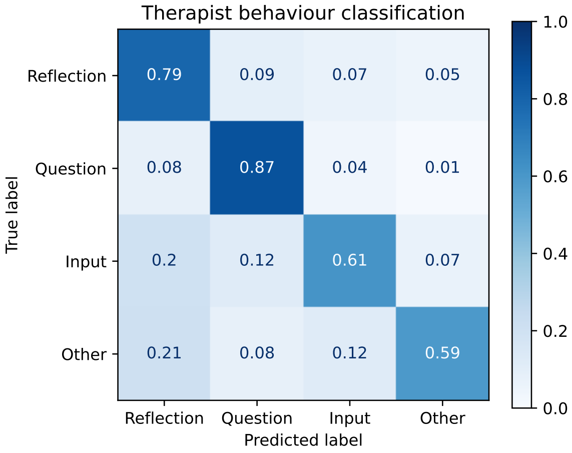

Figure 12. Confusion matrix of BERT w/ Adapters for (main) therapist behaviour prediction in the Original Unbalanced setting. Normalised by row.

7.2. Task 2: Client Talk Type Prediction

Client talk type prediction aims to produce the correct client Talk Type label given a client utterance. As shown in Table 16, this task records universally lower scores than Task 1 for all the trained models—the best BERT-variant performance scores are around 0.55 for macro F1, while CNN and Random Forest score less than 0.47, irrespective of data augmentation. Two factors likely responsible for the performance gap between the two tasks are dialogue context and annotation noise. In some cases, the talk type of a client utterance can only be determined with context grounding. For example, "Yeah" as a reply to "So you work out every day?" is a neutral talk, but it should be Change Talk when it follows "Don't you ever wish things were different?". Moreover, the IAA (Fleiss' kappa) for client talk type is around 0.47, while it is 0.74 for therapist behaviour, which suggests that annotating the talk type is more challenging and therefore more noise is present in the labelling. Inevitably, such noise makes it more difficult to optimise the trainable models.

Table 16. Macro F1 and per-class F1 scores of client talk type prediction. All results averaged from 5-fold cross-validation. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced.

Result Format	Original Unbalanced (Augmented Balanced)			
	Client Talk Type Prediction			
Model	**F1-Macro**	**F1-Change Talk**	**F1-Neutral Talk**	**F1-Sustain Talk**
BERT w/ Adapters	0.55 (0.53↓)	0.51 (0.53↑)	0.74 (0.67↓)	0.39 (0.37↓)
BERT w/o Adapters	0.53 (0.52↓)	0.49 (0.51↑)	0.71 (0.67↓)	0.39 (0.39)
CNN	0.47 (0.46↓)	0.45 (0.44↓)	0.65 (0.63↓)	0.31 (0.31)
Random Forest	0.39 (0.44↑)	0.38 (0.40↑)	0.71 (0.65↓)	0.10 (0.26↑)
Prior	0.33 (0.31↓)	0.29 (0.31↑)	0.57 (0.42↓)	0.14 (0.20↑)
Uniform	0.31 (0.31)	0.31 (0.31)	0.42 (0.42)	0.20 (0.20)

Among the talk types, Neutral Talk has the best performance, followed by Change Talk and Sustain Talk, which matches the class distribution (Table 14), similar to the finding in Task 1 (Section 7.1). Interestingly, in some cases, the inter-class performance gap is reduced thanks to data augmentation. For example, the gap between Change Talk and Neutral Talk is 0.23 F1 in Original Unbalanced but 0.14 F1 in Augmented Balanced, even though most of it is attributed to the performance decrease of 0.07 F1 on Neutral Talk. Unsurprisingly, both Change Talk and Sustain Talk are frequently misclassified as Neutral Talk, even by the best-performing model, BERT w/ Adapters, as can be seen in the confusion matrix (Figure 13). Using dialogue context as an additional input may reduce misclassification to a certain extent, as hypothesised before, but class imbalance may ultimately become the bottleneck for performance improvement [12]. We leave further probing to future work.

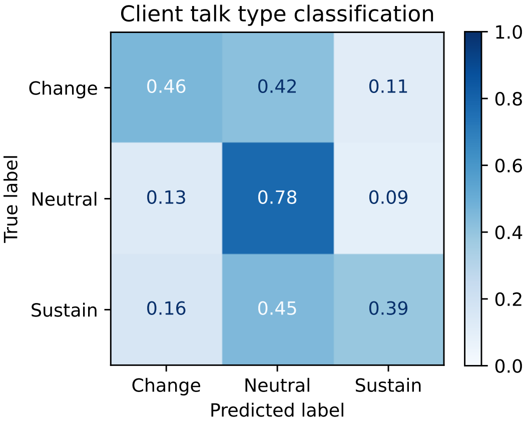

Figure 13. Confusion matrix of BERT w/ Adapters for client talk type prediction in the Original Unbalanced setting. Normalised by row.

Overall Remarks: We establish the single-utterance BERT variants as baselines for both tasks to facilitate comparison with more advanced models in future work. For example, one possibility for improvement would be to formulate Tasks 1 and 2 as BERT-style next-sentence prediction tasks, using the dialogue context as the first sequence and the therapist/client utterance as the second. This setup would enable more contextualised prediction, especially for client talk type prediction.

8. Topic-Specific and Cross-Topic Performance

Apart from performance over the entire AnnoMI, we also explore how the models fare in conversations of different topics, hypothesising that some topics may be more challenging for certain models on particular tasks. Importantly, as the generalisability to topics unseen during training is a major desideratum of reliable models with real-world impact, we probe cross-topic model performance by training on data of all but one topic and testing on examples from that topic.

Based on the topic coverage of AnnoMI (Table 3), we select three topics—reducing alcohol consumption, reducing recidivism, and smoking cessation—for probing the topic-specific and cross-topic performance of all the trained models on the two tasks defined in Section 7, since between 10% and 20% of the utterances in AnnoMI belong to conversations on these topics. We focus on the Original Unbalanced setting, since the results from Augmented Balanced are similar. All the results are measured by macro F1 and are summarised in Table 17.

Table 17. Topic-specific and cross-topic performance (macro F1) in the Original Unbalanced setting for (1) therapist behaviour prediction and (2) client talk type prediction. The 3 topics are reducing alcohol consumption, reducing recidivism, and smoking cessation. ↑/↓: cross-topic performance is higher/lower than topic-specific performance.

Result Format	**Topic-Specific → Cross-Topic**		
Topic	**Reducing Alcohol Consumption**	**Reducing Recidivism**	**Smoking Cessation**
	(Main) Therapist Behaviour Prediction		
BERT w/ Adapters	0.74 → 0.74	0.63 → 0.62↓	0.70 → 0.72↑
BERT w/o Adapters	0.72 → 0.75↑	0.65 → 0.66↑	0.72 → 0.70↓
CNN	0.59 → 0.55↓	0.50 → 0.52↑	0.64 → 0.60↓
Random Forest	0.49 → 0.49	0.40 → 0.36↓	0.53 → 0.48↓
	Client Talk Type Prediction		
BERT w/ Adapters	0.55 → 0.52↓	0.41 → 0.43↑	0.56 → 0.51↓
BERT w/o Adapters	0.54 → 0.52↓	0.41 → 0.42↑	0.55 → 0.50↓
CNN	0.47 → 0.45↓	0.39 → 0.39	0.50 → 0.43↓
Random Forest	0.42 → 0.38↓	0.33 → 0.34↑	0.37 → 0.32↓

8.1. Topic-Specific Performance

To obtain the performance on topic T_i, we re-use the 5-fold CV models for the two tasks (Section 7), but we test each model only on a T_i-specific subset of the corresponding test fold. Specifically, the subset consists entirely of utterances that are originally from conversations of topic T_i. By averaging the performance of the 5 models on their respective T_i-specific test-fold subsets, this method covers all T_i utterances and thus yields a reliable measure of the T_i-specific performance of each model type.

Generally, it is clear that the model performance, especially that of the BERT variants, follows the topic-wise ordering below:

- **Therapist Behaviour Prediction:** reducing alcohol consumption > smoking cessation > reducing recidivism;
- **Client Talk Type Prediction:** reducing alcohol consumption ≈ smoking cessation > reducing recidivism.

One contributing factor to the performance gaps between different topics could be topic coverage, namely the number of utterances from sessions of a particular topic, as better coverage entails more data used for training. For example, reducing alcohol consumption has more utterances than reducing recidivism (Table 3) and correspondingly also better performance.

However, it is also clear that the performance on reducing recidivism conversations is considerably lower than on smoking cessation, despite the slightly larger coverage of reducing recidivism. This is more likely because the utterances of the topic themselves are more semantically challenging for the task, and it also shows the necessity to include a wide range of topics in a counselling dialogue dataset.

8.2. Cross-Topic Performance

It is often important for trained models to generalise to unseen domains. While conversations of different topics are not completely different domains, the results shown in Section 8.1 illustrate that the models indeed have varying levels of performance depending on the topic. Hence, to complement Section 8.1, where models trained on dialogues of all topics are examined for their topic-specific performance, we probe model generalisability by removing a topic T_i from the training set completely and then analysing its performance on a T_i-only test set.

Concretely, we adopt a leave-one-topic-out approach by training on all the AnnoMI utterances from conversations that do not have topic T_i and testing on all the AnnoMI utterances from dialogues that only have topic T_i. Conversations with multiple topics that include T_i are not present during training or testing. We note that the test set in this setup is effectively identical to that of Section 8.1, which allows for cross-topic and topic-specific performance to be compared fairly.

For therapist behaviour prediction, the performance of the BERT models remains stable when moving from topic-specific to cross-topic, which shows that (1) the models are generalisable to new topics for this task, and (2) therapist language is relatively consistent in conversations of different topics.

For client talk type prediction, on the other hand, consistent and more noticeable (as much as 0.05 F1 for the BERT models) performance drops can be seen for reducing alcohol consumption and smoking cessation. While this may indicate that client language varies more across topics, we note that client talk type prediction generally has lower performance than therapist behaviour prediction (Section 7), and thus it may be a more challenging task in general and need more training data irrespective of topic.

9. Discussion

While AnnoMI contains transcripts of MI demonstrations instead of real therapy sessions, we believe that it is the closest approximation possible without privacy violations, while the precise transcription and the accompanying expert annotations further make it more reliable and versatile than similar datasets (e.g., [13]). We note that most of the source videos are from professional therapists and research organisations/institutes dedicated to relevant topics (e.g., reducing substance use), and therefore the authenticity of the demonstrated client–therapist interaction can be considered reliable, as confirmed by the survey responses from the professional annotators. It could also be interesting to explore the domain gap between the corpus and undisclosed real-world therapy datasets. In particular, as the average duration of the source videos is 7 minutes and thus shorter than usual real-world counselling sessions, we will in future work replicate our experiments on other corpora with longer sessions and then compare the results with those obtained based on AnnoMI.

We also note that while client talk type has comparatively lower IAA scores, the performance difference between the trained models and random baselines is substantial, proving the reliability of the annotations on these attributes. As we experimented with attribute prediction based on the utterance of a single turn only, the lack of contextualisation is also likely to have contributed to the relatively lower performance, which we leave to future work to address.

Compared to the original AnnoMI released in [14], the full version introduced in this work has the same dialogues but contains additional fine-grained annotations with respect to label sub-types for Question, Input and Reflection, such as Open Question and Simple Reflection. The quality of the full version is therefore similar to the original version, except that some of the new attributes have relatively lower IAAs, but future work may still consider leveraging these annotations as probabilistic labels for model training and other research purposes.

For applications, AnnoMI can be readily used to develop NLP/ML models for MI fidelity, such as generating feedback to help train and supervise counsellors. Example use cases of this nature include (1) categorising current-turn therapist behaviour and/or client talk type, as explored in Sections 7 and 8, and (2) forecasting next-turn client talk type and/or MI-adherent therapist behaviour. Beyond these natural language understanding settings, AnnoMI can also be used for natural language generation to assist human therapists, such as providing suggestions on what a counsellor could say next, given the past utterances of an ongoing session.

10. Conclusions

We release the full version of AnnoMI [14], a dataset of professionally transcribed and expert-annotated conversations that demonstrate high- and low-quality motivational interviewing. Based on the rich annotations by experienced counsellors, we thoroughly analyse various counselling-related properties at the utterance, dialogue, and corpus levels. We also create relevant utterance-level prediction tasks and establish baseline models. Finally, we examine the topic-specific model performance on these tasks and probe the generalisability of the models to new topics.

AnnoMI represents a powerful resource for research in the important direction of counselling-related natural language processing. For future work, we plan to explore rich dialogue contexts as as additional input for the therapist behaviour and client talk type prediction tasks. We also plan to investigate other applications of AnnoMI with real-world impacts, such as assisting counsellors with real-time session analytics and next-turn suggestions.

Author Contributions: Conceptualization, Z.W., S.B. and V.K.; methodology, Z.W., S.B. and V.K.; software, Z.W. and S.B.; validation, Z.W.; formal analysis, Z.W.; investigation, Z.W., S.B. and V.K.; resources, R.H., D.R.R. and D.R.; data curation, S.B. and V.K.; writing—original draft preparation, Z.W.; writing—review and editing, S.B., V.K., R.H., D.R.R. and D.R.; visualization, Z.W.; supervision, R.H., D.R.R. and D.R.; project administration, Z.W.; funding acquisition, R.H., D.R.R. and D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882, and partially funded by the ASTRID project (Fondazione di Sardegna, annualità 2020), CUP: F75F21001220007.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University of Aberdeen (1 April 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: In this work, we created and analysed the full version of AnnoMI, which is available under the Public Domain License at https://github.com/uccollab/AnnoMI, accessed on 10 Mar 2023.

Acknowledgments: We would like to thank Ehud Reiter and Mark Aloia for their guidance and support. We would also like to thank the anonymous reviewers for their insightful comments and suggestions in improving this paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rollnick, S.; Miller, W.R.; Butler, C. *Motivational Interviewing in Health Care: Helping Patients Change Behavior*; Guilford Press: New York, NY, USA, 2008.
2. Miller, W.R.; Rollnick, S. *Motivational Interviewing: Helping People Change*; Guilford Press: New York, NY, USA, 2012.
3. Miller, W.R. (University of New Mexico, Albuquerque, New Mexico, USA); Moyers, T.B. (University of New Mexico, Albuquerque, New Mexico, USA); Ernst, D. (Denise Ernst Training and Consultation, Portland, Oregon, USA); Amrhein, P. (Montclair State University, Montclair, New Jersey, USA) Manual for the motivational interviewing skill code (MISC). 2003. *Unpublished Manuscript*.
4. Moyers, T.B.; Rowell, L.N.; Manuel, J.K.; Ernst, D.; Houck, J.M. The motivational interviewing treatment integrity code (MITI 4): rationale, preliminary reliability and validity. *J. Subst. Abus. Treat.* **2016**, *65*, 36–42. [CrossRef] [PubMed]
5. Can, D.; Georgiou, P.G.; Atkins, D.C.; Narayanan, S.S. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
6. Xiao, B.; Can, D.; Georgiou, P.G.; Atkins, D.; Narayanan, S.S. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Hollywood, CA, USA, 3–6 December 2012; pp. 1–4.
7. Atkins, D.C.; Steyvers, M.; Imel, Z.E.; Smyth, P. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implement. Sci.* **2014**, *9*, 1–11. [CrossRef] [PubMed]
8. Gibson, J.; Malandrakis, N.; Romero, F.; Atkins, D.C.; Narayanan, S.S. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
9. Gibson, J.; Can, D.; Xiao, B.; Imel, Z.E.; Atkins, D.C.; Georgiou, P.; Narayanan, S.S. A Deep Learning Approach to Modeling Empathy in Addiction Counseling. In Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, USA, 8–12 September 2016; pp. 1447–1451. [CrossRef]
10. Xiao, B.; Can, D.; Gibson, J.; Imel, Z.E.; Atkins, D.C.; Georgiou, P.G.; Narayanan, S.S. Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks. In Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 908–912.
11. Gibson, J.; Atkins, D.; Creed, T.; Imel, Z.; Georgiou, P.; Narayanan, S. Multi-label multi-task deep learning for behavioral coding. *IEEE Trans. Affect. Comput.* **2019**, *13*, 508–518. [CrossRef] [PubMed]
12. Cao, J.; Tanana, M.; Imel, Z.; Poitras, E.; Atkins, D.; Srikumar, V. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5599–5611.
13. Pérez-Rosas, V.; Wu, X.; Resnicow, K.; Mihalcea, R. 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, Florence, Italy, 28 July–2 August 2019; pp. 926–935.
14. Wu, Z.; Balloccu, S.; Kumar, V.; Helaoui, R.; Reiter, E.; Recupero, D.R.; Riboni, D. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6177–6181.
15. Bakeman, R.; Quera, V. Behavioral observation. In *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics*; APA Handbooks in Psychology®; American Psychological Association: Washington, DC, USA, 2012; pp. 207–225. [CrossRef]
16. Pérez-Rosas, V.; Mihalcea, R.; Resnicow, K.; Singh, S.; An, L. Building a motivational interviewing dataset. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, 16 June 2016; pp. 42–51.
17. Pérez-Rosas, V.; Mihalcea, R.; Resnicow, K.; Singh, S.; An, L.; Goggin, K.J.; Catley, D. Predicting counselor behaviors in motivational interviewing encounters. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 1, Long Papers, pp. 1128–1137.
18. Pérez-Rosas, V.; Mihalcea, R.; Resnicow, K.; Singh, S.; An, L. Understanding and predicting empathic behavior in counseling therapy. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1426–1435.
19. Tanana, M.; Hallgren, K.A.; Imel, Z.E.; Atkins, D.C.; Srikumar, V. A comparison of natural language processing methods for automated coding of motivational interviewing. *J. Subst. Abus. Treat.* **2016**, *65*, 43–50. [CrossRef] [PubMed]
20. Wu, Z.; Helaoui, R.; Recupero, D.R.; Riboni, D. Towards Low-Resource Real-Time Assessment of Empathy in Counselling. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, Online, 11 June 2021; pp. 204–216.
21. Wu, Z.; Helaoui, R.; Kumar, V.; Reforgiato Recupero, D.; Riboni, D. Towards Detecting Need for Empathetic Response in Motivational Interviewing. In Proceedings of the Companion Publication of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 26–29 October 2020; pp. 497–502.

22. Singla, K.; Chen, Z.; Atkins, D.; Narayanan, S. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3797–3803.
23. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
24. Chen, Z.; Singla, K.; Gibson, J.; Can, D.; Imel, Z.E.; Atkins, D.C.; Georgiou, P.; Narayanan, S. Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6605–6609.
25. Singla, K.; Chen, Z.; Flemotomos, N.; Gibson, J.; Can, D.; Atkins, D.C.; Narayanan, S.S. Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy. In Proceedings of the 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018 ; pp. 3413–3417.
26. Xiao, B.; Bone, D.; Segbroeck, M.V.; Imel, Z.E.; Atkins, D.C.; Georgiou, P.G.; Narayanan, S.S. Modeling therapist empathy through prosody in drug addiction counseling. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
27. Xiao, B.; Imel, Z.E.; Atkins, D.C.; Georgiou, P.G.; Narayanan, S.S. Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
28. Flemotomos, N.; Martinez, V.R.; Chen, Z.; Singla, K.; Ardulov, V.; Peri, R.; Caperton, D.D.; Gibson, J.; Tanana, M.J.; Georgiou, P.; et al. Automated evaluation of psychotherapy skills using speech and language technologies. *Behav. Res. Methods* **2022**, *54*, 690–711. [CrossRef] [PubMed]
29. Reforgiato Recupero, D.; Cambria, E. ESWC'14 challenge on Concept-Level Sentiment Analysis. *Commun. Comput. Inf. Sci.* **2014**, *475*, 3–20. [CrossRef]
30. Dridi, A.; Reforgiato Recupero, D. Leveraging semantics for sentiment polarity detection in social media. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2045–2055. [CrossRef]
31. Recupero, D.R.; Alam, M.; Buscaldi, D.; Grezka, A.; Tavazoee, F. Frame-based detection of figurative language in tweets [application notes]. *IEEE Comput. Intell. Mag.* **2019**, *14*, 77–88. [CrossRef]
32. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378. [CrossRef]
33. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
34. Cicchetti, D.V.; Sparrow, S.A. Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *Am. J. Ment. Defic.* **1981**, *86*, 127–137. [PubMed]
35. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5776–5788.
36. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
37. Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; Gurevych, I. AdapterHub: A Framework for Adapting Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 46–54.
38. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019 ; pp. 4171–4186.
41. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
43. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
44. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 11328–11339.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.