Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts: Validation and Application

Soliman Ali[*] **Jiading Zhu**[*] **Alex Guo**[*] **Xiao Nan Ye**[*] **Qilin Gu**[*] **Jodi Wolff**[†]
Carolynne Cooper[*†] **Osnat C. Melamed**[*†] **Peter Selby**[*†] **Jonathan Rose**[*† §]
[*]University of Toronto
[†]Centre for Addiction and Mental Health, Toronto, ON, Canada

Abstract

Motivational Interviewing (MI) is a widely-used talk therapy approach employed by clinicians to guide clients toward healthy behaviour change. Both the automation of MI itself and the evaluation of human counsellors can benefit from high-quality automated classification of counsellor and client utterances. We show how to perform this "coding" of utterances using LLMs, by first performing utterance-level parsing and then hierarchical classification of counsellor and client language. Our system achieves an overall accuracy of 82% for the upper (coarse-grained) hierarchy of the counsellor codes and 88% for client codes. The lower (fine-grained) hierarchy scores at 68% and 76% respectively. We also show that these codes can be used to predict the session-level quality of a widely-used MI transcript dataset at 87% accuracy. As a demonstration of practical utility, we show that the slope of the amount of change/sustain talk in client speech across 106 MI transcripts from a human study has significant correlation with an independently surveyed week-later treatment outcome ($r = 0.28$, $p < 0.005$). Finally, we show how the codes can be used to visualize the trajectory of client motivation over a session alongside counsellor codes. The source code and several datasets of annotated MI transcripts are released.

1 Introduction

There is significant activity using Large Language Models (LLMs) to assist with and directly perform mental health talk therapy (Heinz et al., 2025; Tingley, 2025). These efforts require LLMs not only to engage in the therapeutic dialogue, but also monitor the conversation for problems and measure/-classify its elements to assess whether it meets high quality standards (Bakeman and Quera, 2012). In the past, for human-based counselling, manual classification has been used to train and judge humans. Pre-trained LLMs have become proficient at performing this classification, and so can be leveraged for the tasks of assessing counsellor fidelity to treatment standards and the analysis of the relationship between client language and clinical outcomes (Amrhein et al., 2003).

In this paper, we present a transcript classification approach for a specific kind of talk therapy known as *Motivational Interviewing* (MI) (Miller and Rollnick, 2023), a widely-used counselling approach for facilitating healthy behaviour change. The classification system is based on the Motivational Interviewing Skills Code (MISC) (Houck et al., 2010), the original annotation scheme for MI. It provides comprehensive, mutually exclusive, utterance-level labels for language from the counsellor (typically a clinician) and client (the patient/subject).

The *AutoMISC* system uses pretrained LLMs to perform utterance-level behavioural code annotation of MI transcripts under the MISC 2.5 taxonomy. We validate *AutoMISC* in a number of ways: first by comparing its annotations (on both closed-source and open-source LLMs) to expert-aligned human annotators. Then, we show that its fine-grained annotations align with annotations given in the AnnoMI dataset (Wu et al., 2023). The annotations can also be used to predict the binary counselling quality ratings at the session level of the High/Low Quality Counselling dataset (Pérez-Rosas et al., 2019). To demonstrate its broader utility, we show that the annotations of transcripts from a smoking cessation study correlate with the study outcome metric: the change in client-reported confidence to quit smoking (a validated proxy of actual behaviour change (Gwaltney et al., 2009; Abar et al., 2013)). The key contributions of this paper are:

1. An automated system for utterance-level

[§]Corresponding author: jonathan.rose@utoronto.ca

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 25–54
December 23, 2025

MISC 2.5 (Houck et al., 2010) behavioural coding of MI transcripts.
2. Validation of *AutoMISC* across open and closed-source LLMs by measuring (1) performance against expert-aligned human annotations, and (2) performance on public annotated datasets.
3. An empirical comparison of flat versus hierarchical prompting strategies for behavioural coding.
4. A novel application of this automated annotation where we show a statistically significant correlation between client language and the change in their confidence that they could succeed in a behavior change.
5. Three datasets totalling 506 transcripts annotated automatically, two of which include manually annotated subsets, to support future work in automated evaluation of MI transcripts.
6. Release of an open-source software package.

The following section describes prior work in the area of automated evaluation of therapy transcripts. Section 3 gives a brief background on Motivational Interviewing itself and the MISC coding framework. Section 4 describes the design of the *AutoMISC* system, its parameters and how we determine ground-truth labels. Section 5 describes validation methods and results for the system. Section 6 shows how to visualize the codes and describes a transcript-based metric and its correlation with the therapy outcome.

2 Related Work

2.1 Automated Behavioural Coding in Psychotherapy

Early approaches to automated behavioural coding in psychotherapy relied on linguistic features selected and engineered by experts (Can et al., 2012; Pérez-Rosas et al., 2017) or topic modeling (Atkins et al., 2012, 2014) to detect specific behaviours such as asking questions and providing reflections, occasionally combined with another modality such as accoustic features (Aswamenakul et al., 2018). Later, neural network-based approaches emerged (Tanana et al., 2015; Gibson et al., 2016; Xiao et al., 2016; Huang et al., 2018; Cao et al., 2019; Ewbank et al., 2021), improving classification accuracies in behavioural coding tasks by offering a more expressive and implicit model of the dialogues. More recent work has used BERT-based transformer models (Devlin et al., 2018; Liu et al., 2019) to extract contextual embeddings from counsellor and client utterances (Tavabi et al., 2021; Brown et al., 2023; Pellemans et al., 2024; Xie et al., 2024; Cohen et al., 2024), sometimes complemented by other features such as voice (Tavabi et al., 2020) and facial information (Nakano et al., 2022), which are then passed to downstream neural network-based classifiers. These approaches performed well when the behavioural task is sufficiently constrained, although extensive training is required on datasets annotated with high-quality labels. Among the strongest results is by Cohen et al. (2024), which achieved a macro F1 score of 0.42 with 70% accuracy on 10 counsellor codes under the MITI coding framework (Moyers et al., 2016), and macro F1 of 0.72 with 72% accuracy on three client codes.

The adaptation of LLMs in this space initially explored fine-tuning approaches (Hoang et al., 2024), however, these approaches are limited by the scarcity of publicly available MI datasets, and labelled datasets are even rarer (see following section). More recent efforts have demonstrated that LLMs can be effectively prompted for behavioural coding without fine-tuning, through either zero-shot prompting (Brown et al., 2024; Mahmood et al., 2025a), few-shot prompting (Sun et al., 2024), or in-context learning (Chiu et al., 2024), achieving high accuracy when compared with human labels. Notably, with few-shot prompting, Sun et al. (2024) achieved Macro F1 scores of 0.31 on 16 counsellor codes and 0.32 on 10 client codes under MISC 2.1 (Miller et al., 2003), an earlier version of the MISC framework.

Despite these advances, prior work still has limitations in their behaviour coding capabilities. Many approaches focus exclusively on either counsellor or client speech, and often target only a small subset of behaviours. For MI in particular, no existing work has attempted fully automated coding of both speakers under the complete MISC 2.5 framework (Houck et al., 2010). Moreover, prior work rarely connects automated behaviour coding to treatment outcomes, and few projects release code or software to support reproducibility or real-world use.

2.2 MI Datasets

There are several public, anonymized datasets supporting the task of MI behavioural coding. These include the High/Low Quality Counseling dataset (Pérez-Rosas et al., 2019), Counsel-Chat (Welivita and Pu, 2022), AnnoMI (Wu et al., 2023), MI-

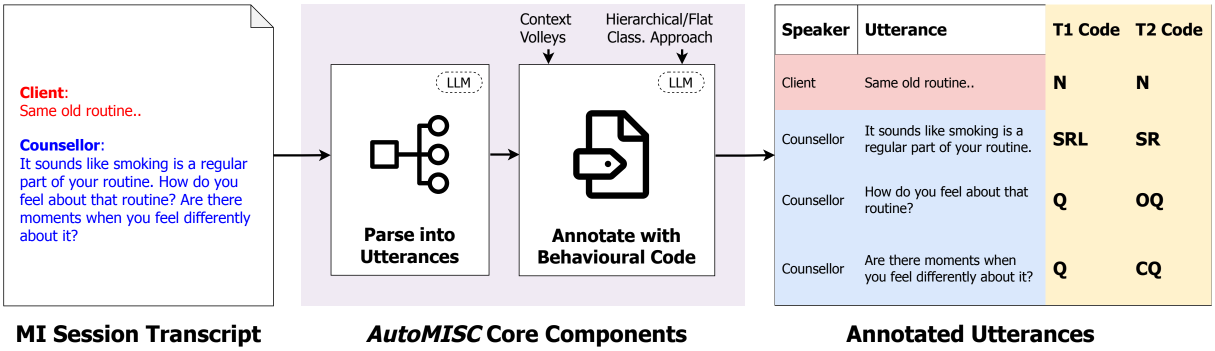


Figure 1: Overview of the *AutoMISC* system. The input to the system is an MI transcript. The system first segments the transcript into utterances, and then annotates them with behavioural codes. The output is the resulting sequence of annotated utterances, which can then be used to compute summary scores or visualize session trajectories.

TAGS (Cohen et al., 2024), and BiMISC (Sun et al., 2024). The datasets vary in their sources, as well as the levels of granularity in the labels they provide. While these datasets have supported progress in behavioural coding, most lack full MISC 2.5 coverage, are not publicly accessible, or offer only coarse labeling. There remains a need for high-quality, fully annotated MI datasets aligned with an existing behavioural coding framework such as MISC 2.5, to support more complex tasks such as fine-grained modelling of MI transcripts and prediction of client behaviours.

3 Motivational Interviewing

Motivational Interviewing is a talk therapy approach that a counsellor (often a medical provider) applies to help a client (a patient or subject) move towards and achieve a target behaviour change, typically related to health. The conversation is meant to be collaborative, rather than directive, and focuses on *guiding* the client in exploring their motivations for change and connecting them to their underlying values. A counsellor uses specific kinds of utterances, such as open-ended *questions* to evoke motivation and *reflections* (which are restatements of client's words, possibly to connected to relevant ideas and facts) to encourage further contemplation around the target behaviour.

As clients express themselves, counsellors listen carefully for two categories of motivational language: *change talk* (Miller and Rollnick, 2023), which indicates motivation, commitment or action towards change, and *sustain talk*, which reflects reasons to maintain the status quo. Most clients exhibit both, indicating an internal state of *ambivalence* in which they wish to change but also identify reasons preventing them from changing. A key goal in MI is to help resolve this ambivalence by inviting and strengthening change talk, while acknowledging but not reinforcing sustain talk.

In successful MI, as the therapeutic alliance develops, there is a progression in client change talk from a *preparatory* stage (expressions of desire, ability, reasons, or need for change) to a *mobilizing* stage (expressions of commitment, activation, or taking steps towards change). This progression reflects increasing client readiness for change and is predictive of actual behavioural outcomes (Miller and Rollnick, 2023; Amrhein et al., 2003).

3.1 The MISC 2.5 Coding Framework

Behavioural coding schemes are a key method by which the quality of the counsellor is judged, and also whether the client language is progressing towards or away from the behaviour. These schemes assign labels to conversational content at the *utterance* level – a single unit of thought. Within a given speaker turn, which we will refer to as a *volley*, a counsellor or client may express multiple utterances in sequence. Thus it is important to first parse volleys into a set of utterances prior to assigning behavioural codes.

We use the MISC framework (Houck et al., 2010) because it was intended for research and provides a comprehensive, mutually exclusive, fine-grained taxonomy for both counsellor and client codes. This contrasts with other frameworks such as the MITI (Moyers et al., 2016) which was developed to assess only the integrity of MI counselling by providers, and does not assess client language.

The MISC 2.5 framework defines 19 counsellor codes and 17 client codes[1]. The basic counsellor

[1]Although not listed in the MISC 2.5, we include "Activation+/-" in the client code set based on definitions

strategies (questions and reflections), as well as client codes (change and sustain talk) described in Section 3 have several sub-types in MISC 2.5. For example, counsellor reflections are further subdivided into *Simple Reflection* (SR) which simply mirrors a client's statement, and *Complex Reflection* (CR) in which the counsellor both mirrors and adds meaning or insight. The full classification taxonomy is provided in Figure A.1 in Appendix A.1.

MISC also provides session-level *summary scores* computed from frequency counts and ratios of behavioural codes across the session, intended as heuristic indicators of session quality in research and training contexts. These include:

- **Percentage MI-Consistent Responses (%MIC)**: the proportion of counsellor behaviours classified as MI-Consistent i.e. directly prescribed in Miller and Rollnick (2023). Higher values indicate greater adherence to MI standards.
- **Reflection-to-Question Ratio (R:Q)**: the ratio of reflective statements to questions posed by the counsellor. Values between 1 and 2 are considered good (Moyers et al., 2016).
- **Percentage Change Talk (%CT)**: the proportion of client utterances coded as Change Talk, with higher values associated with improved behavioural outcomes (Apodaca and Longabaugh, 2009).

4 *AutoMISC* System Design

Figure 1 illustrates the pipeline of the *AutoMISC* system. First, volley is parsed into utterances, then each utterance is annotated with a behavioural code. The input to *AutoMISC* is a single volley-separated file of a transcript which identifies the speaker as either counsellor or client. The outputs from the system are (1) the parsed and annotated corpus, and (2) MISC session-level summary scores. The following sections describe the core components of *AutoMISC* in further detail.

4.1 Separation of Volleys into Utterances

The parser module separates each volley in a conversation into one or more utterances. This is not simply separation into sentences as an utterance can be expressed in multiple sentences or portions of a single sentence. This makes the task semantically complex, and so we use a prompted pre-trained Large Language Model model to perform this task. The prompt begins with definitions of *volley* and *utterance* from the MISC manual and then the general task of separation of utterances. It includes four few-shot example input-output pairs sourced from the MISC manual. The full parser module system prompt is provided in Appendix A.2.

in Miller and Rollnick (2023).

4.2 Automated Coding

The classification of each utterance into a behavioural code is handled by the annotator module, which is also a prompted large language model.

A key decision is whether to use a hierarchical classification approach, or a flat one. This was motivated by our manual coding work (described below in section 4.3) where we found it very helpful to decompose the task into two steps, first classifying into a higher-level grouping of similar MISC codes that we call *Tier 1* codes, then to the fine-grained MISC code (the *Tier 2* codes). We hypothesized that a language model might see performance gains from this decomposition (at the cost of doubling the number of inference calls). For client utterances, the three Tier 1 categories are intuitively Change Talk (C), Sustain Talk (S), and Neutral Talk (N). For counsellor utterances, we grouped the 19 fine-grained codes into six groupings based on (human-perceived) semantic similarity and ease of disambiguation. The full set of Tier 1 and Tier 2 codes is shown in Figure A.1 in Appendix A.1. We compare this to a *flat* approach in which the model selects directly from the full set of Tier 2 codes in Section 5.2.2.

A second key parameter for the annotator module is to decide how much prior conversation context is needed for high classification accuracy. The module takes in a parameter called *number of context volleys* which sets how many volleys prior to the one under consideration to include in the prompt. We hypothesized that performance would improve with additional context up to a point of diminishing returns, discussed further in Section 5.2.1.

Each prompt to the annotator module includes a task description, the available label set, the context window, and finally the target utterance for classification. In the hierarchical mode, the Tier 2 prompt is templated to include only the candidate codes associated with the selected Tier 1 label. Prompt templates are provided in Appendix A.3. Once annotation is complete, the summary scores described in Section 3.1 are computed.

4.3 Consensus Labels & Annotator Alignment with Experts

To evaluate and refine *AutoMISC*, we created a reference dataset of known-good human annotations, which we will refer to as the *consensus labels*. To do so we used a combination of members of our research team which includes both computer engineers and experienced MI clinicians specializing in smoking cessation. To produce reliable annotations, we first trained a team of three undergraduate research interns and one graduate student to annotate transcripts from a public dataset (Mahmood et al., 2025b) using the MISC 2.5 schema. We used an iterative process in which the goal was to achieve substantial inter-rater reliability, commonly quantified as Fleiss' Kappa $\kappa \geq 0.6$ (Cicchetti et al., 1992). The iterative process was as follows:

1. The four annotators independently label five transcripts.
2. The inter-rater reliability (IRR) is computed using Fleiss' κ across all codes, counsellor and client.
3. If $\kappa < 0.6$ for any category, an alignment meeting is held, together with expert MI clinicians to resolve discrepancies.

We completed two iterations: In the first round, annotators labelled the first five transcripts from the dataset (a total of $n = 367$ utterances) but did not meet the IRR threshold for all codes. A two-hour alignment meeting was held, during which consensus labels were produced for that sample. In the second round, annotators labelled a new set of five transcripts ($n = 454$ utterances), after which the IRR target was reached. Training was deemed complete, and consensus labels were consolidated across both sets, yielding a reference set of $n = 821$ utterances (580 from the counsellor, 241 from clients). Figure C.2 in Appendix C gives the pairwise Cohen's Kappa matrices between raters before and after training.

4.4 Classification Prompt Evolution

The initial classification prompts for the annotator module were derived directly from the definitions of behavioural codes in the MISC 2.5 manual and Miller and Rollnick (2023). These were evolved based on classification performance against the consensus labels of the reference dataset, using OpenAI's GPT-4o [2]. There were two key issues found with the prompts: The first concerned Open versus Closed Questions (OQ vs CQ): *AutoMISC* initially overused the OQ label. This was resolved by improving the prompt so that questions answerable with a "yes", "no", or short factual response should be coded as CQ in the Tier 2 counsellor classification prompt, as shown in Appendix A.3.

The second issue concerned Imperative-MI-Inconsistent vs Imperative-MI-Consistent (IMI vs IMC). Here the issue is that an imperative/directive statement is only MI-Consistent if permission was granted to do so, and that permission may be one or *more* volleys prior to the utterance being coded. It was observed that these permissions could be delivered in subtle ways, which were hard to detect. This was addressed by adding a Chain of Thought reasoning process around permission to the end of the T1 counsellor classification prompt, as shown in Appendix A.3.

5 Validation of Automatic Coding

The system is validated primarily via macro F1 score and accuracy, measured on the first 10 conversations (a total of $n = 821$ utterances) from the MI transcript dataset (Mahmood et al., 2025b), using the consensus labels described in Section 4.3 as ground truth . We also validate against the labels of the AnnoMI dataset (Wu et al., 2023), and we show that the annotations can predict counselling quality in the HLQC dataset (Pérez-Rosas et al., 2019).

5.1 Experimental setup

AutoMISC is configured with three input parameters: (1) the language model used for annotation, (2) the classification structure (hierarchical vs. flat), and (3) the number of prior volleys provided as context to the model (the latter two introduced in Section 4.2). The models chosen were selected for diversity both in model provider, using both open- and closed-source, and a range of model sizes, as follows: OpenAI's GPT-4o [2] and GPT-4.1[3], Alibaba's Qwen3-30b-a3b[4], and Google's Gemma-3-12b[4]. The OpenAI models were accessed through the company's for-pay APIs, and the other models were run on an M3 Macbook Pro with 32GB of RAM and makes use of the native GPU acceleration. Wall-clock inference times per utterance

[2] `gpt-4o-2024-08-06`
[3] `gpt-4.1-2025-04-14`
[4] Quantized to 4-bit parameters

were approximately 2 seconds for the OpenAI models, 7 seconds on the Qwen model and 16 seconds on the Gemma model. The utterance parsing step was done by GPT-4o in all cases, to enable direct comparison of classification/coding/annotation accuracy between the different models.

5.2 Parameter tuning

Figure 2 gives the classification performance (macro F1 score and accuracy) versus the number of context volleys for GPT-4.1, separated into different plots by speaker (counsellor/client) and classification approach (hierarchical vs. flat). Results for the other three models are given in Appendix D. The accuracy is greater than F1 because the most common behavioural codes achieve good accuracy across the 19 counsellor codes and the 17 client codes.

5.2.1 Number of Context Volleys

For counsellor codes, Figure 2 (top row) shows that performances improves with additional context up until 2-3 volleys, after which it plateaus or declines. The initial increase is likely due to the fact that all the "IMC" codes require permission to be granted in a preceding volley. The degraded performance with longer contexts might be attributed to the model attending to less relevant context in the earlier volleys.

The client coding performance appears to simply plateau or degrade with added context. This is likely because change and sustain talk is self-evident within an utterance and may even shift rapidly between change talk and sustain talk within the same volley (Miller and Rollnick, 2023), making additional context less informative.

5.2.2 Hierarchical vs. Flat Classification Approach

Figure 2 shows that the hierarchical classification approach is almost uniformly better across all tested models and context window sizes, but the flat approach achieves similar or even higher macro F1 scores in a few configurations, mostly on the client codes.

5.3 Validation Results

Table 1 gives the F1 and accuracy scores for the model and parameter settings that achieved the highest macro F1 score. Complete numerical results across all configurations are in Appendix D.

The highest-performing model and configuration overall was GPT-4.1 using 3 prior volleys as context and the hierarchical classification structure. It achieves a macro F1 score of 0.42 and 68% accuracy on the full set of 19 MISC counsellor codes. On the 17 client codes it achieves an F1 score of 0.41 and 76% accuracy. The smaller open-source models achieved competitive results on both counsellor and client coding. For instance, Gemma-3-12b reached 0.40 Macro F1 on client codes, outperforming the larger Qwen3-30b-a3b model.

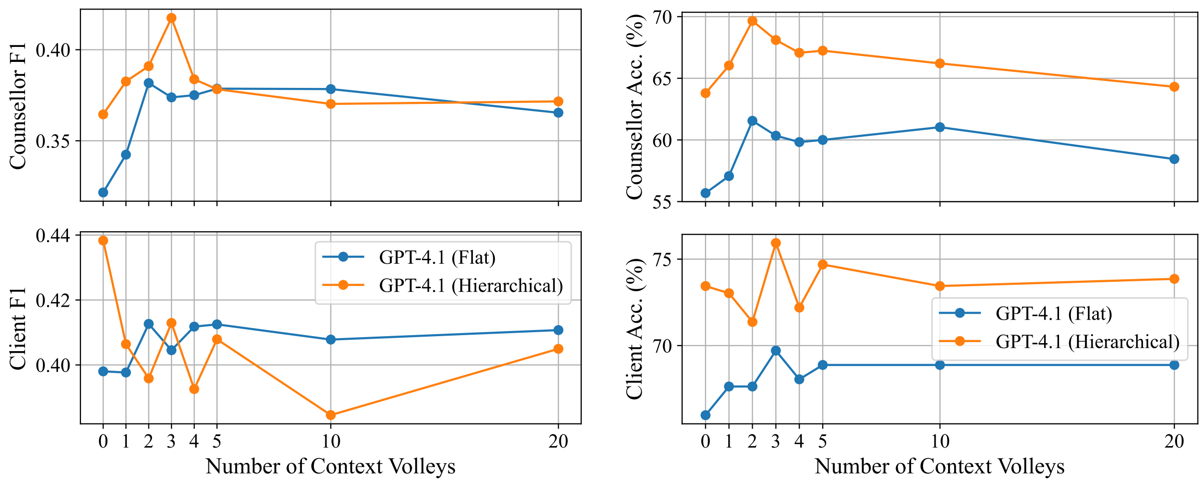
Table 2 compares *AutoMISC*'s classification performance to prior work reported in the original publications introducing the (Sun et al., 2024) and MI-TAGS (Cohen et al., 2024) datasets. In spite of the larger label spaces covered, our results meet or exceed these results across both speaker roles.

Confusion matrices for the best performing models/configurations are included in Appendix C.

5.4 Supplementary Validation Experiments

As supplementary measures of validation, we compare *AutoMISC*'s output to existing datasets. In Appendix D.1 we compare directly to AnnoMI's annotations (Wu et al., 2023) by mapping to their custom volley-level scheme, achieving 65% accuracy ($n = 4882$) on counsellor codes and 77% accuracy ($n = 4817$) on client codes. In Appendix D.2 we show that *AutoMISC*'s outputs can predict the binary session quality rating in the HLQC dataset (Pérez-Rosas et al., 2019) at 87% accuracy.

Since the consensus set from our experiment was small ($n = 821$ utterances) and imbalanced (Appendix C), we manually annotated a larger, more balanced subset of the HLQC dataset ($n = 1924$ utterances) to use as ground truth for evaluating *AutoMISC*. Sweeping across the same parameters described in Section 5.1, the best-performing configuration was GPT-4.1 with 5 prior volleys of context and the hierarchical classification structure. This achieved a macro F1 score of 0.35 and 46% accuracy on counsellor codes and a macro F1 score of 0.32 and 80% accuracy on client codes. Further discussion and experimental results are provided in Appendix E. We release this manually annotated HLQC subset along with the one from the chatbot study as its larger sizer and more realistic conversations may be valuable for future research in automated MI behavioural coding.



(a) Macro F1 Score vs. Context Volleys

(b) Accuracy vs. Context Volleys

Figure 2: Effect of context size and classification approach (hierarchical/flat) on counsellor and client classification performance (GPT-4.1, $n = 821$ utterances)

Model	Class. Structure	Context Volleys	T1 Couns.		T1 Client		T2 Couns.		T2 Client		T2 Overall	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hierarchical	3	**0.80**	**82**	**0.87**	**88**	**0.42**	**68**	**0.41**	**76**	**0.42**	**70**
GPT-4o	flat	2	–	–	–	–	0.41	61	0.41	65	0.41	62
Qwen3-30b-a3b	hierarchical	0	0.61	69	0.77	78	0.28	55	0.35	63	0.30	57
Gemma-3-12b	hierarchical	1	0.60	70	0.80	81	0.30	54	0.40	59	0.33	56

Table 1: Best accuracy (%) and macro F1 scores with consensus labels across models, classification approach and context window sizes for each speaker and code tier ($n = 821$ utterances).

Work	T2 Couns.	T1 Client	T2 Client
BiMISC	0.31 (16)	0.68 (3)	0.32 (10)
MI-TAGS	0.42 (10)	0.72 (3)	–
AutoMISC	**0.42** (19)	**0.88** (3)	**0.41** (17)

Table 2: Reported macro F1 scores from prior work compared to *AutoMISC*. Values in parentheses indicate the number of classes.

6 Applications: Visualization of Client Trajectories and Correlation with Post-Therapy Outcome

A core assumption in MI is that client language influences and shapes downstream behavioural outcomes. This MISC 2.5 summary scores such as percent change talk offer a coarse measure of client motivation but they obscure the progression of motivation through a session. Amrhein et al. (2003) showed that the change in strength of client commitment language (a subset of change talk) over a session is a good predictor of drug use outcomes at follow-up. This motivates the idea to visualize MI transcripts by plotting utterance behavioural codes over time, an idea common in talk therapy research (Horton et al., 2021).

6.1 Visualization of Client Motivation Trajectories

Figure 3 shows an example *conversational trajectory* which is derived from *AutoMISC* codes of counselor and client speech in a session from the dataset used above in validation. The x-axis shows progression along the session in two ways: the thin vertical lines delineate an utterance, while the solid blue or pink colour delineates a complete volley composed of one or more utterances. The left Y-axis shows the Tier 1 categories of the counsellor speech that were determined by *AutoMISC*. The right Y-axis gives the Tier 2 categories of client speech ordered from the bottom as the strongest sustain talk, and at the top to be the strongest change talk, with neutral talk in the middle. Figure 3 shows a trajectory for a session in which the client's talk shows a somewhat upward trend from sustain talk to strong change talk. It also gives a

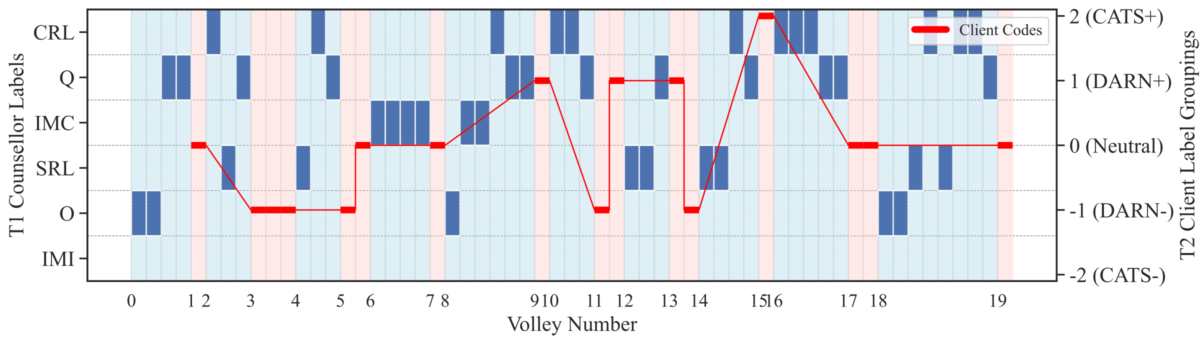


Figure 3: Example Visualization of MI session. Red: Client Speech codes. Blue bars: Counsellor Speech T1 codes

sense of the kinds of MI skills that the counsellor was employing. We feel that this level of detail could play a useful role in the evaluation of the skills of the counsellor and the impact of the session on the client. In the next section we illustrate the latter with a metric computed from the client speech (red line) trajectory.

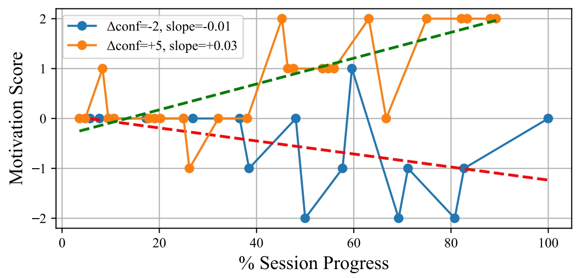


Figure 4: Two sample client motivation trajectories from the smoking cessation study.

6.2 Correlation of Client Code Sequence to Therapy Outcome

In this section we show how the sequence of client codes can be used to create a metric which correlates with a therapy outcome. The metric, which is called the *motivation slope* is computed as the slope of a linear regression on the red line in Figure 3.

The dataset used for validation labels in Section 4.3 also contained a client-reported confidence to quit smoking, reported on a scale of 0-10 prior to the session and one week later. We use the change in confidence (prior to week later) as the outcome measure (Gwaltney et al., 2009; Abar et al., 2013), and compute the Spearman's correlation between several session-level features including the motivation slope, and the change in confidence, for all 106 transcript/outcomes in the dataset. The GPT-4.1 model was used for these codes, with three context volleys and the hierarchical classification approach.

Feature	**Spearman r**	**p-value**
Pre-confidence	-0.11	0.26
Motivation Slope	**0.28**	**< 0.005**
% MIC	0.01	0.07
R:Q	0.10	0.32
% CT	0.17	0.08

Table 3: Spearman correlations between session features and the week-later change in client self-reported confidence to quit smoking ($n = 106$).

Table 3 shows that the motivation slope is significantly correlated with client change in confidence ($r = 0.28$, $p < 0.005$) and is superior to all the other MISC summary scores (and the pre-conversation confidence) none of which have statistically significant correlation. This result shows that significant information is contained in the codes produced by *AutoMISC*, and in so doing gives a form of validation of the quality of the codes produced.

Figure 4 shows two sample client motivation trajectories: one in which the client confidence change was +5 a week later and trajectory is rising (orange), and one with a change of -2 and a falling trajectory (blue). Finally, Figure F.1 in Appendix F gives a scatterplot of motivation slope values vs. change in confidence for all 106 clients.

7 Software & Dataset Release

The source code and three annotated datasets are released publicly along with this paper totalling 506 transcripts. These include the first MISC-labelled releases of the AnnoMI ($n = 133$) and HLQC ($n = 258$) corpora, as well as the smoking cessation transcript dataset ($n = 115$) (Mahmood et al., 2025b), all parsed and annotated at the utterance level. We also release the manual annotations for the subsets of the smoking cessation study ($n = 821$ utterances) and the HLQC dataset ($n = 1924$ utterances). The source code and data are available at: `https://github.com/cimhasgithub/AutoMISC`.

8 Conclusions

We introduce an LLM-based system for fully automated utterance-level annotation of counsellor and client speech in Motivational Interviewing (MI) transcripts under the MISC 2.5 framework. *AutoMISC* achieves classification performance equal to or exceeding prior approaches on expert-aligned annotations, and aligns with annotations in existing datasets like AnnoMI.

We also demonstrate how to use the annotations to predict MI quality in the HLQC dataset. We introduce a novel metric, the *motivation slope*, that correlates significantly with client-reported confidence to quit smoking, a short-term proxy for actual behaviour change. Future work should explore the direct predictive capability when more data is available.

We have shown that *AutoMISC* works both with state-of-the-art APIs and locally hosted models, making it suitable for use in privacy-sensitive settings such talk therapy. In the future, we plan to use these classification tools within fully automated MI systems to track client state change and counsellor adherence to MI. We also plan to employ the tools on evaluation and training of human MI counsellors.

9 Limitations

While *AutoMISC* delivers promising results in automating MI behavior coding, several limitations should be noted. First, the consensus labels we used as ground truths were not directly labeled by MI experts, but instead by annotators aligned by experts. Despite our effort in iteratively refining the labels to meet the IRR threshold, one could argue that such indirect supervision may introduce discrepancies and limit the fidelity of our consensus labels. Second, while our system is grounded in the MISC 2.5 framework (Houck et al., 2010), it does not strictly follow all recommended coding procedures, such as doing a first pass and providing global scores before parsing and assigning behavior codes, nor does it rely on modalities beyond text, such as vocal and visual cues that are essential for accurate interpretation and coding. Our proposed two-tiered coding flow was also designed heuristically and not grounded in MISC 2.5 or any other prior MI literature, whose validity and utility need to be confirmed by future research. Third, our validation experiments are imperfect due to limitations and constraints from the datasets used. For the AnnoMI (Wu et al., 2023) dataset, there might be inconsistencies in the mapping between the MISC labels and their custom volley-level coding scheme; for the HLQC (Pérez-Rosas et al., 2019) dataset, the high and low quality labels for transcripts are provided using their own custom criteria, thus may not always reflect the true quality of the transcripts; for the MI transcript dataset (Mahmood et al., 2025b), the participants were paid and therefore might have had an incentive to report inflated post-therapy outcomes. Finally, while we demonstrate *AutoMISC* 's ability to run on local models to address privacy concerns, our best results are still achieved using proprietary models such as GPT-4.1, leaving room for future work to improve open-source models further and provide better guarantees in regards to privacy.

References

Beau Abar, Brigitte M. Baumann, Cynthia Rosenbaum, Edward Boyer, Douglas Ziedonis, and Edwin D. Boudreaux. 2013. Profiles of importance, readiness and confidence in quitting tobacco use. *Journal of Substance Use*, 18(2):75–81.

Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology*, 71(5):862–878.

Timothy R. Apodaca and Richard Longabaugh. 2009. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. *Addiction*, 104(5):705–715.

Chanuwas Aswamenakul, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Multimodal analysis of client behavioral change coding in motivational interviewing. pages 356–360.

David Atkins, Timothy Rubin, Mark Steyvers, Michelle Doeden, Brian Baucom, and Andrew Christensen. 2012. Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, 26:816–827.

David Atkins, Mark Steyvers, Zac Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation science : IS*, 9:49.

Roger Bakeman and Vicenç Quera. 2012. Behavioral observation. In Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, editors, *APA Handbook of Research Methods in Psychology, Vol. 1: Foundations, Planning, Measures, and Psychometrics*, pages 207–225. American Psychological Association.

Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose. 2023. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Ment Health*, 10:e49132.

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. Association for Computational Linguistics.

Dogan Can, Panayiotis Georgiou, David Atkins, and Shrikanth Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. volume 3.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Domenic V. Cicchetti, Fred Volkmar, Sara S. Sparrow, Donald Cohen, Jacques Fermanian, and Byron P. Rourke. 1992. Assessing the reliability of clinical scales when the data have both nominal and ordinal features: Proposed guidelines for neuropsychological assessments. *Journal of Clinical and Experimental Neuropsychology*, 14(5):673–686. PMID: 1474138.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

M. P. Ewbank, R. Cummins, V. Tablan, A. Catarino, S. Buchholz, and A. D. Blackwell. 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312. PMID: 32619163.

James Gibson, Dogan Can, Bo Xiao, Zac Imel, David Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. pages 1447–1451.

Chad J Gwaltney, Jane Metrik, Christopher W Kahler, and Saul Shiffman. 2009. Self-efficacy and smoking cessation: a meta-analysis. *Psychol Addict Behav*, 23(1):56–66.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

Ayana Horton, Gail Hebson, and David Holman. 2021. A longitudinal study of the turning points and trajectories of therapeutic relationship development in occupational and physical therapy. *BMC Health Services Research*, 21(1):97.

Jonathon Houck, Theresa Moyers, William R Miller, Laura Glynn, and C Hallgreen. 2010. *Manual for the Motivational Interviewing Skill Code (MISC) version 2.5*.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025a. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

Zafarullah Mahmood, Soliman Ali, Jiading Zhu, Mohamed Abdelwahab, Michelle Yu Collins, Sihan Chen, Yi Cheng Zhao, Jodi Wolff, Osnat C. Melamed, Nadia Minian, Marta Maslej, Carolynne Cooper, Matt Ratto, Peter Selby, and Jonathan Rose. 2025b. A fully generative motivational interviewing counsellor chatbot for moving smokers towards the decision to quit. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25008–25043, Vienna, Austria. Association for Computational Linguistics.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

William R. Miller and Stephen Rollnick. 2023. *Motivational Interviewing: Helping People Change*, 4 edition. The Guilford Press, New York, NY.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Yukiko I. Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting change talk in motivational interviewing using verbal and facial information. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 5–14, New York, NY, USA. Association for Computing Machinery.

Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, and Rob van der Mei. 2024. Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: Secondary analysis of a clinical trial. *J Med Internet Res*, 26:e53562.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. 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*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos T.P. Dobber, and Jos A. Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621, Torino, Italia. ELRA and ICCL.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79, Denver, Colorado. Association for Computational Linguistics.

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D. Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 406–413, New York, NY, USA. Association for Computing Machinery.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Kim Tingley. 2025. Kids are in crisis. could chatbot therapy help? `https://www.nytimes.com/2025/06/20/magazine/ai-chatbot-therapy.html`. The New York Times Magazine.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Bo Xiao, Dogan Can, James Gibson, Zac Imel, David Atkins, Panayiotis Georgiou, and Shrikanth

Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. pages 908–912.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13207–13219, Bangkok, Thailand. Association for Computational Linguistics.

A *AutoMISC* System Design Supplementary Material

Figure A.1 shows the full classification taxonomy of *AutoMISC*. Appendices A.2 and A.3 show the prompts for each of the core components of the *AutoMISC* system.

A.1 Classification Taxonomy

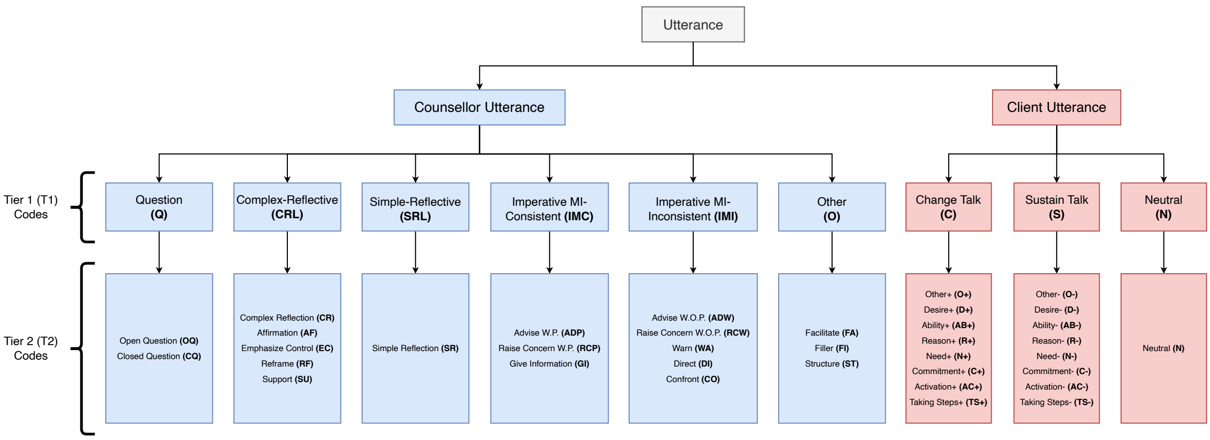


Figure A.1: *AutoMISC* utterance classification taxonomy.

A.2 Parser Module Prompt

The Parser module is fed a system prompt, followed by several input-output pairs from the MISC manual ("few-shots"), and finally the target volley for parsing. It is constrained to return a list of strings using a structured output schema (defined using Pydantic). The prompt and few-shot examples are as follows:

A.2.1 Parser Prompt

You are a highly accurate Motivational Interviewing (MI) counselling session annotator. Your task is to segment the given volley into utterances.

Definitions:

- Volley: An uninterrupted utterance or sequence of utterances spoken by one party before the other party responds.
- Utterance: A complete thought or thought unit expressed by a speaker. This could be a single sentence, phrase, or even a word if it conveys a standalone idea. Multiple utterances often run together without interruption in a volley.

Output Format:

- Return the segmented utterances as a Python list of strings.

```
Input: "Why haven't you quit smoking - are you ever gonna quit?"
Output: ["Why haven't you quit smoking - are you ever gonna quit?"]

Input: "How long since your last drink? Do you feel ok?"
Output: ["How long since your last drink?", "Do you feel ok?"]

Input: "I can't quit. I just can't do it. I don't have what it takes. I just cannot stop."
Output: ["I can't quit.", "I just can't do it.", "I don't have what it takes.", "I just cannot stop
    ."]

Input: "I don't want to go to the bars every day. I don't want my kids to see that. I want my kids
    to have a better life than that."
Output: ["I don't want to go to the bars every day.", "I don't want my kids to see that.", "I want my
     kids to have a better life than that."]
```

A.3 Annotator Module Classification Prompts

The annotator module uses either a hierarchical or flat classification approach. In the hierarchical approach, the model first chooses a Tier 1 code, then selects a Tier 2 code from the subset associated with that Tier 1 category. Following the classification prompt, the annotator module is given a configurable number of volleys prior to the target utterances as context for classification, then the target utterance itself, templated in another prompt we call the "User Prompt". The model output is constrained using a structured output schema (Pydantic) to return only an explanation string and one code abbreviation from either the T1 or T2 grouping. Below we list out the Tier 1, Tier 2 and Flat classification prompts for both counsellor and client, as well as the user prompt.

A.3.1 Tier 1 Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:

1. **C-Reflective (CRL)** - Deeply engages with or affirms the client's perspective.
   - *Behavioural Codes*: Affirm (AF), Support (SU), Complex Reflection (CR), Reframe (RF), Emphasize Control (EC)
   - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
   - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
   - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
   - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually changing the emotional valence of meaning but not the depth.
   - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
2. **S-Reflective (SRL)** - Mirrors or paraphrases the client's statement without adding extra insight (includes summarizing statements).
   - *Behavioural Codes*: Simple Reflection (SR)
   - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or emphasis.
3. **Imperative-MICO (IMC)** - **With client permission**, provides advice, raises a concern, or gives information.
   - *Behavioural Codes*: Advise with Permission (ADP), Raise Concern with Permission (RCP), Give Information (GI)
   - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
   - **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
   - **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
4. **Imperative-MIIN (IMI)** - **Without client permission**, provides advice, raises a concern, warns, directs, or confronts the client.
   - *Behavioural Codes*: Advise Without Permission (ADW), Raise Concern Without Permission (RCW), Warn (WA), Direct (DI), Confront (CO)
   - **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
   - **Raise Concern Without Permission (RCW)**: Without getting permission, points out a possible problem with a client's goal, plan, or intention.
   - **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client takes a certain action.
   - **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
   - **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
5. **Question (Q)** - Asks a question in order to gather information, understand, or elicit the client's story.
   - *Behavioural Codes*: Open Question (OQ), Closed Question (CQ)
   - **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.

- **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with “yes” or “no” under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*

6. **Other/Neutral (O)** - Structural or facilitative utterances that do not engage in MI techniques.
   - *Behavioural Codes*: Filler (FI), Facilitate (FA), Structure (ST)
   - **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
   - **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
   - **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor's final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing between **IMC** and **IMI**.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except IMI and IMC.
2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** (when deciding between **IMC** or **IMI**), you **must first apply this step-by-step reasoning** to determine if permission is given:
   (a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.
   (b) Check for Counsellor's Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.
   (c) If Yes (to 1 or 2): Classify the utterance as IMC (permission has been granted).
   (d) If No: Classify the utterance as IMI (no permission has been granted).
   (e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor's next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client's intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client's earlier request, concern, or area of engagement.
   Apply this permission reasoning chain ONLY when the utterance's function is to provide advice, suggestions, guidance, or information.
   For all other categories (**CRL, SRL, Q, O**), permission is **not relevant**. Assign these categories based on their definitions without using this permission reasoning.

Output Format

- **explanation**: Use brief reasoning for all category assignments expect IMC and IMC. When the category is IMC or IMI, use the full chain of thought for determining permission as the justification.
- **label**: Provide only "CRL", "SRL", "IMC", "IMI", "Q", or "O".

A.3.2 Tier 2 Counsellor Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the counsellor's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

```
{{spec}}
```

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The {{spec}} parameter is replaced by one of the following depending on what the Tier 1 code was:

```
CRL: |
  - **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or
      emphasis to what the client said, conveying a deeper or richer picture of the client's
      statement.
  - **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths
      or efforts.
  - **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side
      with the client.
  - **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client,
      usually changing the emotional valence of meaning but not the depth.
  - **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and
      freedom of choice.
SRL: |
  - **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases
       the client's words or meaning, often with a slight change in wording or emphasis.
IMC: |
  - **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion,
       or offers a solution or possible action.
  - **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem
       with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
  - **Giving Information (GI)**: Provides information to the client, explains something, educates or
       provides feedback, or discloses personal information.
IMI: |
  - **Advise Without Permission (ADWP)**: Offers suggestions or guidance WITHOUT asking or receiving
       permission.
  - **Confront (CON)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade,
      criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
  - **Direct (DIR)**: Gives an order, command, or direction. The language is imperative.
  - **Raise Concern Without Permission (RCWP)**: Without getting permission, points out a possible
      problem with a client's goal, plan, or intention.
  - **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client
      takes a certain action
Q: |
  - **Closed Question (CQ)**: A question is closed if it can be answered with ``yes'' or ``no''
      under any grammatically valid interpretation, even if that answer is awkward, contextually
      unhelpful, or unlikely.
       To determine is a question is CQ, always check for its grammatical structure first. If the
           utterance can be interpreted in a way that permits a yes/no response, you must classify
           it as CQ.
       This includes any form of:
          - any utterance containing or beginning with grammatical constructions that use auxiliary
              or modal verbs, existence/presence checks, or binary/framed prompts must be labeled
              as CQ. These include, but are not limited to, questions that:
             - Begin with or contain modal/auxiliary verbs such as:
                Can, Could, Do, Does, Did, Are, Is, Was, Were, Will, Would, Have, Has, Had, Might,
                    May, Should, Shall, Must followed by a subject and verb/complement.
             - Ask about existence, availability, or presence using forms like:
                Is there, Are there, Do you have, Have you got, Would it be, Could it be, Might it
                    be, Is it possible that...
             - Implicitly or explicitly present binary choices or confirmatory framing, including
                  structures like:
                Do you ever, Would you say, Are you thinking about, Would you like, Is this
                    something you, Have you thought about, Do you feel like, Do you think, Does it
                    feel like, Do you notice...
             If the utterance contains any clause that permits a grammatically valid yes/no or short
                  factual response, even if additional elaboration is possible, it must be labeled
                  CQ.
```

```
          - even if it appears to invite elaboration.
          - confirmation or factual information-seeking
          - curiosity about presence/absence of something
          - request for specific information or choices
        If there is any ambiguity between CQ and OQ, always label it as CQ.
  - **Open Question (OQ)**:
        A question is open only if it cannot be answered with ``yes'' or ``no'' in any grammatically
             valid or logically plausible way.
        The question must structurally require an elaboration, explanation, or descriptive narrative
            that goes beyond a binary or fixed-option response.
        Questions that seem to encourage elaboration but could be reduced to a yes/no response are
            still CQ, not OQ.
        Use this label only when there is no grammatical path to yes/no answers -- no exceptions.
O: |
  - **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. ``Mm
      -hmm'', ``I see'', ``Go on''
  - **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
  - **Structure (ST)**: Gives information about will happen directly to the client throughout the
      course of treatment or within a study format, in this or subsequent sessions.
```

A.3.3 Tier 1 Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

1. **Change Talk (C)** - The client expresses a stance toward **changing** the target behavior.
    - **Commitment** to change (e.g., stating/implying an intention to change, considering alternatives, making plans to change).
    - **Reasons** for change (including personal, health, or emotional factors).
    - **Desire** to change (e.g., "I really want to quit.").
    - **Optimism** about their ability to change (e.g., "I think I can do it.").
    - **Need** to change (e.g., "I have to stop before it gets worse.").
    - **Recent steps** toward change (e.g., "I cut back this week.").
2. **Sustain Talk (S)** - The client expresses a stance toward **maintaining** the target behavior.
    - **Commitment** to maintaining the target behaviour (e.g., stating/implying an intention to continue, dismissing alternatives, making plans to continue).
    - **Reasons** for maintaining the target behaviour (e.g., stress relief, social reasons).
    - **Desire** to continue the target behaviour (e.g., "I enjoy it too much to quit.").
    - **Pessimism** about their ability to change (e.g., "I don't think I can quit.").
    - **Need** to maintain the target behaviour (e.g., "I need cigarettes to cope.").
    - **Recent steps** reinforcing the target behaviour (e.g., "I bought another pack today.").
3. **Neutral (N)** - The utterance does not clearly support or oppose change.
    - Following along with the counsellor without expressing a stance.
    - Asking questions (e.g., "What are the benefits of quitting?").
    - Providing factual or general statements about the behaviour.

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only "C", "S", or "N".

A.3.4 Tier 2 Client Prompt Template

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to assign a category label to the client's final utterance in a given session excerpt.
Classification Categories
The utterance must be assigned one of the following labels:

`{{spec}}`

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label.

Final instructions

1. Analyze the client's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

The `{{spec}}` parameter is replaced by one of the following depending on what the Tier 1 code was:

```
C: |
  - **Desire (D+)**: The client expresses a desire to change the target behaviour, e.g. "I want to
     quit smoking".
  - **Ability (AB+)**: The client expresses optimism about their ability to change, e.g. "I think it
     's possible for me to quit".
  - **Reasons (R+)**: The client provides reasons for changing the target behaviour, e.g. "My
     children are begging me to quit".
  - **Need (N+)**: The client expresses a need to change the target behaviour, e.g. "I've got to
     quit before it gets worse".
  - **Commitment (C+)**: The client expresses a commitment to change, e.g. "I'm going to quit
     smoking".
  - **Activation (AC+)**: The client leans towards action, e.g. "I'm willing to give it another try".
      This includes suggestions of alternatives to the target behaviour.
  - **Taking Steps (TS+)**: The client mentions recent steps towards change, e.g. "I cut back on
     smoking this week".
  - **Other (O+)**: The client makes a statement that supports change but does not fit into the
     other categories. This usually includes problem recognition or hypotheticals.
S: |
  - **Desire (D-)**: The client expresses a desire to maintain the target behaviour, e.g. "I enjoy
     smoking too much to quit".
  - **Ability (AB-)**: The client expresses pessimism about their ability to change, e.g. "I don't
     think I can quit".
  - **Reasons (R-)**: The client provides reasons for maintaining the target behaviour, e.g. "
     Smoking is the only way I can relax".
  - **Need (N-)**: The client expresses a need to maintain the target behaviour, e.g. "I need to
     have my morning cigarettes".
  - **Commitment (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. "I
     'm not going to quit smoking".
  - **Activation (AC-)**: The client leans towards inaction, e.g. "I'm not ready to quit yet". This
     includes suggestions of maintaining the target behaviour.
  - **Taking Steps (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g.
     "I bought two packs today".
  - **Other (O-)**: The client makes a statement that supports maintaining the target behaviour but
     does not fit into the other categories. This usually includes problem recognition or
     hypotheticals.
N: |
  - The utterance does not clearly support or oppose change. There is no further categorization, so
     just use "N".
```

A.3.5 Flat Counsellor Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the counsellor's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Affirm (AF)**: Communicates something positive or complimentary about the client's strengths or efforts.
- **Support (SU)**: Sympathetic, compassionate, or understanding comments, which agree or side with the client.
- **Complex Reflection (CR)**: A reflective listening statement that adds significant meaning or emphasis to what the client said, conveying a deeper or richer picture of the client's statement.
- **Reframe (RF)**: Suggests a different meaning for an experience expressed by the client, usually changing the emotional valence of meaning but not the depth.
- **Emphasize Control (EC)**: Acknowledges, honours, or emphasizes the client's autonomy and freedom of choice.
- **Simple Reflection (SR)**: A reflective listening statement which simply repeats or paraphrases the client's words or meaning, often with a slight change in wording or emphasis.

- **Advise With Permission (ADP)**: After receiving permission, gives advice, makes a suggestion, or offers a solution or possible action.
- **Raise Concern With Permission (RCP)**: After getting permission, points out a possible problem with a client's goal, plan, or intention. Always phrased as the counsellor's concern.
- **Giving Information (GI)**: Provides information to the client, explains something, educates or provides feedback, or discloses personal information.
- **Advise Without Permission (ADW)**: Offers suggestions or guidance WITHOUT asking or receiving permission.
- **Raise Concern Without Permission (RCW)**: Without getting permission, points out a possible problem with a client's goal, plan, or intention.
- **Warn (WA)**: Provides a warning or threat, implying negative consequences unless the client takes a certain action.
- **Direct (DI)**: Gives an order, command, or direction. The language is imperative.
- **Confront (CO)**: Directly disagrees, argues, corrects, shames, blames, seeks to persuade, criticizes, judges, labels, moralizes, ridicules, or questions the client's honesty.
- **Open Question (OQ)**: A question is open only if it cannot be answered with "yes" or "no" in any grammatically valid or logically plausible way. The question must structurally require an elaboration, explanation, or descriptive narrative that goes beyond a binary or fixed-option response.
- **Closed Question (CQ)**: A question is closed if it is about confirmation, factual information-seeking, curiosity about presence/absence of something, request for specific information or choices, or *it can be answered with "yes" or "no" under any grammatically valid interpretation, even if that answer is awkward, contextually unhelpful, or unlikely.*
- **Filler (FI)**: Pleasantries such as "good morning", "nice weather we're having", etc.
- **Facilitate (FA)**: Simple utterance that functions as a "keep-going" acknowledgement e.g. "Mm-hmm", "I see", "Go on"
- **Structure (ST)**: Used to make a transition from one topic or part of a session to another. Also used to give information about will happen directly to the client throughout the course of treatment or within a study format, in this or subsequent sessions.

Category assignment instructions

1. **General instructions**
   (a) Analyze the given context and counsellor's final utterance.
   (b) Identify its primary function.
   (c) If the utterance involves **advice, suggestions, or information**, follow the **Permission Chain of Thought Guide** below before choosing ADP, ADW, RCP, or RCW.
   (d) For other types of utterances, assign the category directly.
   (e) Justify your choice in 1-2 sentences for category assignment except ADP, ADW, RCP, or RCW.

2. **Permission Chain of Thought (Only when assigning IMC or IMI)**
   When the utterance involves **giving advice, suggestions, guidance, or information** , you **must first apply this step-by-step reasoning** to determine if permission is given:
   (a) Check for Client Permission or Interest Expression: Examine the recent client utterances in the provided context. Has the client given permission—either explicitly by directly asking for advice, suggestions, or ideas or implicitly by showing openness, curiosity, or requesting information in a way that reasonably invites guidance. Both explicit and implicit permission are equally valid—there is no difference in weight between them. If either is present, permission is considered granted.
   (b) Check for Counsellor's Prior Permission-Seeking: Has the counsellor previously asked for permission to give advice, suggestions, or information and received agreement? If so, permission is also considered granted.
   (c) If Yes (to 1 or 2): You may classify the utterance as ADP/RCP (permission has been granted).
   (d) If No: Classify the utterance as ADW/RCW (no permission has been granted).
   (e) **Carry Permission Forward:** Once permission—explicit or implicit—is granted, it remains **active** for all **topically related** suggestions, guidance, or information, **even if the counsellor's next utterance introduces a shift in topic or phrasing**. **Do NOT revoke permission just because the surface topic evolves naturally**, as long as the advice remains part of the **same overarching discussion or client goal**. **Permission only expires** if there is a **clear and substantive topic shift**, or if the client **disengages** or **withdraws interest**. In most cases, permission is granted in **recent client utterances**, but **prior permissions—especially implicit ones—can remain valid across multiple counsellor turns** if the conversation stays aligned with the client's intent or focus. You should **assume permission is still valid** unless there is strong evidence that the advice no longer relates to the client's earlier request, concern, or area of engagement.
   Apply this permission reasoning chain ONLY when the utterance's function is to provide advice, suggestions, guidance, or information.

Output Format

- **explanation**: Use brief reasoning for all category assignments except ADP/ADW/RCP/RCW. When the category is one of these, use the full chain of thought for determining permission as the justification.
- **label**: Provide only the appropriate label abbreviation.

A.3.6 Flat Client Prompt

You are an expert annotator of Motivational Interviewing (MI) counselling sessions. Your task is to classify the client's final utterance in a given session excerpt into one of the following groupings of MISC 2.5 behavioural codes:

Classification Categories:
The utterance must be assigned one of the following labels:

- **Desire+ (D+)**: The client expresses a desire to change the target behaviour, e.g. "I want to quit smoking".
- **Ability+ (AB+)**: The client expresses optimism about their ability to change, e.g. "I think it's possible for me to quit".
- **Reasons+ (R+)**: The client provides reasons for changing the target behaviour, e.g. "My children are begging me to quit".
- **Need+ (N+)**: The client expresses a need to change the target behaviour, e.g. "I've got to quit before it gets worse".
- **Commitment+ (C+)**: The client expresses a commitment to change, e.g. "I'm going to quit smoking".
- **Activation+ (AC+)**: The client leans towards action, e.g. "I'm willing to give it another try". This includes suggestions of alternatives to the target behaviour.
- **Taking Steps+ (TS+)**: The client mentions recent steps towards change, e.g. "I cut back on smoking this week".
- **Other+ (O+)**: The client makes a statement that supports change but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Desire- (D-)**: The client expresses a desire to maintain the target behaviour, e.g. "I enjoy smoking too much to quit".
- **Ability- (AB-)**: The client expresses pessimism about their ability to change, e.g. "I don't think I can quit".
- **Reasons- (R-)**: The client provides reasons for maintaining the target behaviour, e.g. "Smoking is the only way I can relax".
- **Need- (N-)**: The client expresses a need to maintain the target behaviour, e.g. "I need to have my morning cigarettes".
- **Commitment- (C-)**: The client expresses a commitment to maintain the target behaviour, e.g. "I'm not going to quit smoking".
- **Activation- (AC-)**: The client leans towards inaction, e.g. "I'm not ready to quit yet". This includes suggestions of maintaining the target behaviour.
- **Taking Steps- (TS-)**: The client mentions recent steps reinforcing the target behaviour, e.g. "I bought two packs today".
- **Other- (O-)**: The client makes a statement that supports maintaining the target behaviour but does not fit into the other categories. This usually includes problem recognition or hypotheticals.
- **Neutral (N)**: The utterance does not clearly support or oppose change. This can include following along with the counsellor without expressing a stance, asking questions (e.g., "What are the benefits of quitting?"), or providing factual or general statements about the behaviour.

Output Format

- **explanation**: Briefly justify your choice in 1-2 sentences.
- **label**: Provide only the appropriate label abbreviation.

Final Instructions

1. Analyze the counsellor's final utterance.
2. Identify its primary function and intent.
3. Provide a brief explanation for your choice.
4. Assign the appropriate label based on the categories provided above.

A.3.7 User Prompt

Session Transcript
The following is an excerpt of a MI counselling session transcript:

{{ transcript }}

Target Utterance for Classification
Below is the target {{ speaker }} utterance in the session excerpt:

{{ utterance }}

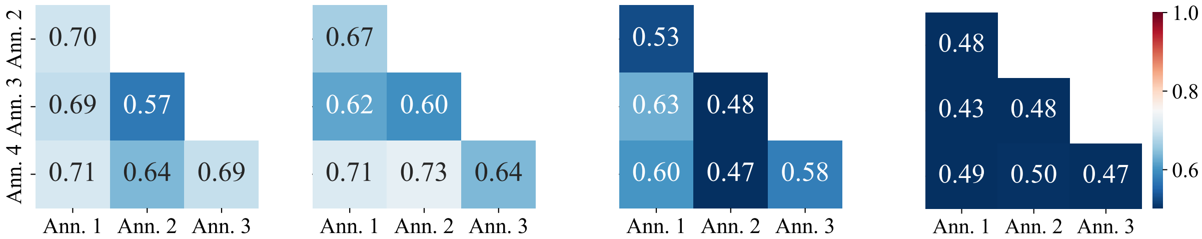
B Expert Alignment of Annotations

B.1 Inter-rater reliability before vs. after alignment

Figures B.1 and B.2 show the Cohen's Kappa between each pair of manual annotators before and after alignment, respectively. The process is described in full in Section 4.3.

B.2 Annotator and MI Expert demographics

Table B.1 lists the demographic information of both the manual annotators and the expert MI clinicians who participated in the transcript labelling alignment meeting described in Section 4.3.



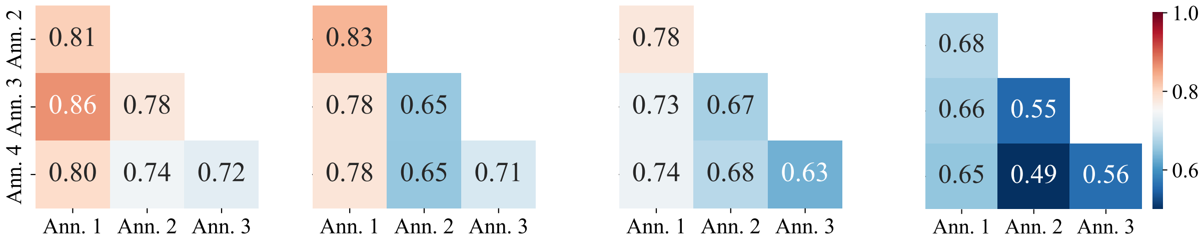
T1 Counsellor
(a) Fleiss' κ = **0.66**

T1 Client
(b) Fleiss' κ = **0.66**

T2 Counsellor
(c) Fleiss' κ = **0.54**

T2 Client
(d) Fleiss' κ = **0.47**

Figure B.1: Pairwise Cohen's Kappa (and Fleiss' Kappa between all annotators) **before** alignment ($n = 367$).



T1 Counsellor
(a) Fleiss' κ = **0.79**

T1 Client
(b) Fleiss' κ = **0.74**

T2 Counsellor
(c) Fleiss' κ = **0.70**

T2 Client
(d) Fleiss' κ = **0.60**

Figure B.2: Pairwise Cohen's (and Fleiss' Kappa between all annotators) **after** alignment ($n = 454$).

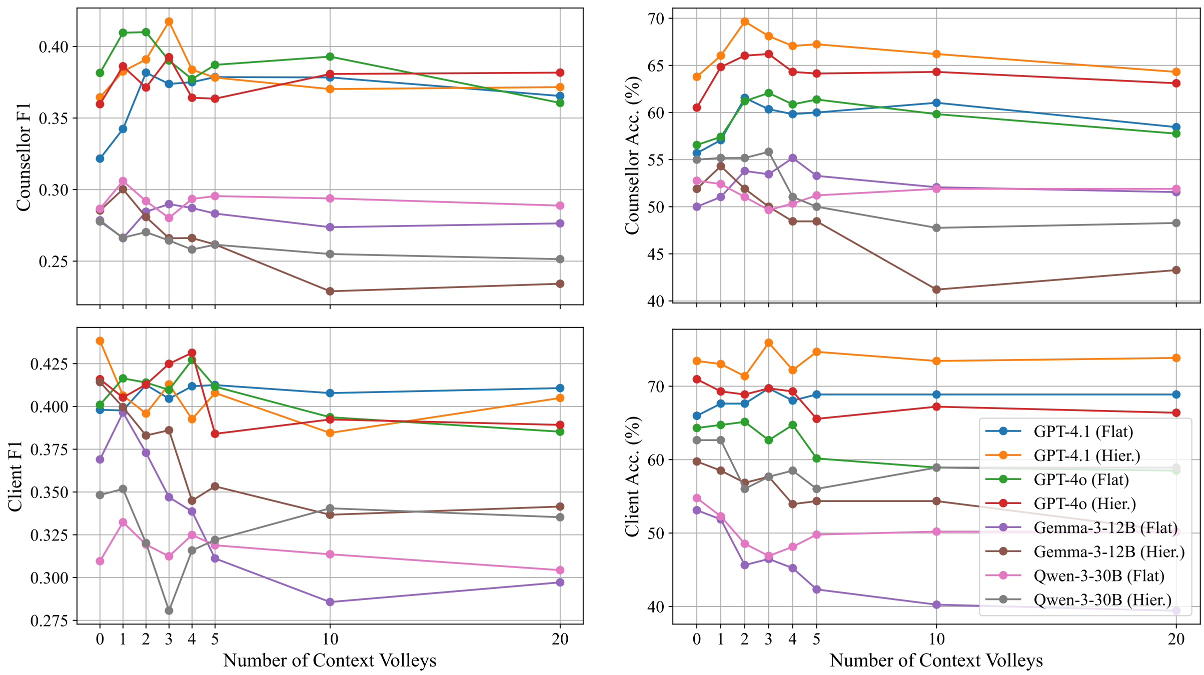
	Anno. 1[1]	**Anno. 2**[2]	**Anno. 3**[2]	**Anno. 4**[2]	**Expert 1**[3]	**Expert 2**[4]	**Expert 3**[5]
Sex	Male	Female	Male	Male	Female	Female	Male
Age Group (years)	20-29	20-29	20-29	20-29	60-69	40-49	60-69
Race/ Ethnicity	Mixed	Asian	Asian	Asian	White	White	South Asian
Native Language	English	Cantonese	English	Mandarin	English	English	English
Student Status	Yes	Yes	Yes	Yes	No	No	No
Employment Status	N/A	N/A	N/A	N/A	Full-Time	Full-Time	Self
Highest Education	Undergrad.	Secondary	Secondary	Secondary	Graduate	Graduate	Graduate
Country of Residence	Canada	Canada	Canada	China	Canada	Canada	Canada
Country of Birth	Canada	China	Canada	China	Canada	Canada	India
Training in Linguistics	No	No	No	No	No	No	No
Training in MI	No	No	No	No	Yes	Yes	Yes

[1] Engineering graduate student with no formal training in MI.
[2] Engineering undergraduate student with no formal training in MI.
[3] Motivational Interviewing Network of Trainers (MINT) member since 2009; Motivational Interviewing Treatment Integrity (MITI) coding trained; extensive training and coaching experience.
[4] Introductory-Intermediate-Advance MI training; MINT member since 2014; MI supervision; MITI training.
[4] Clinician-scientist and educator; extensive MI training and supervision experience; MINT member.

Table B.1: Demographic Information of Annotators and MI Experts

C Comparison to Consensus Labels: All Results

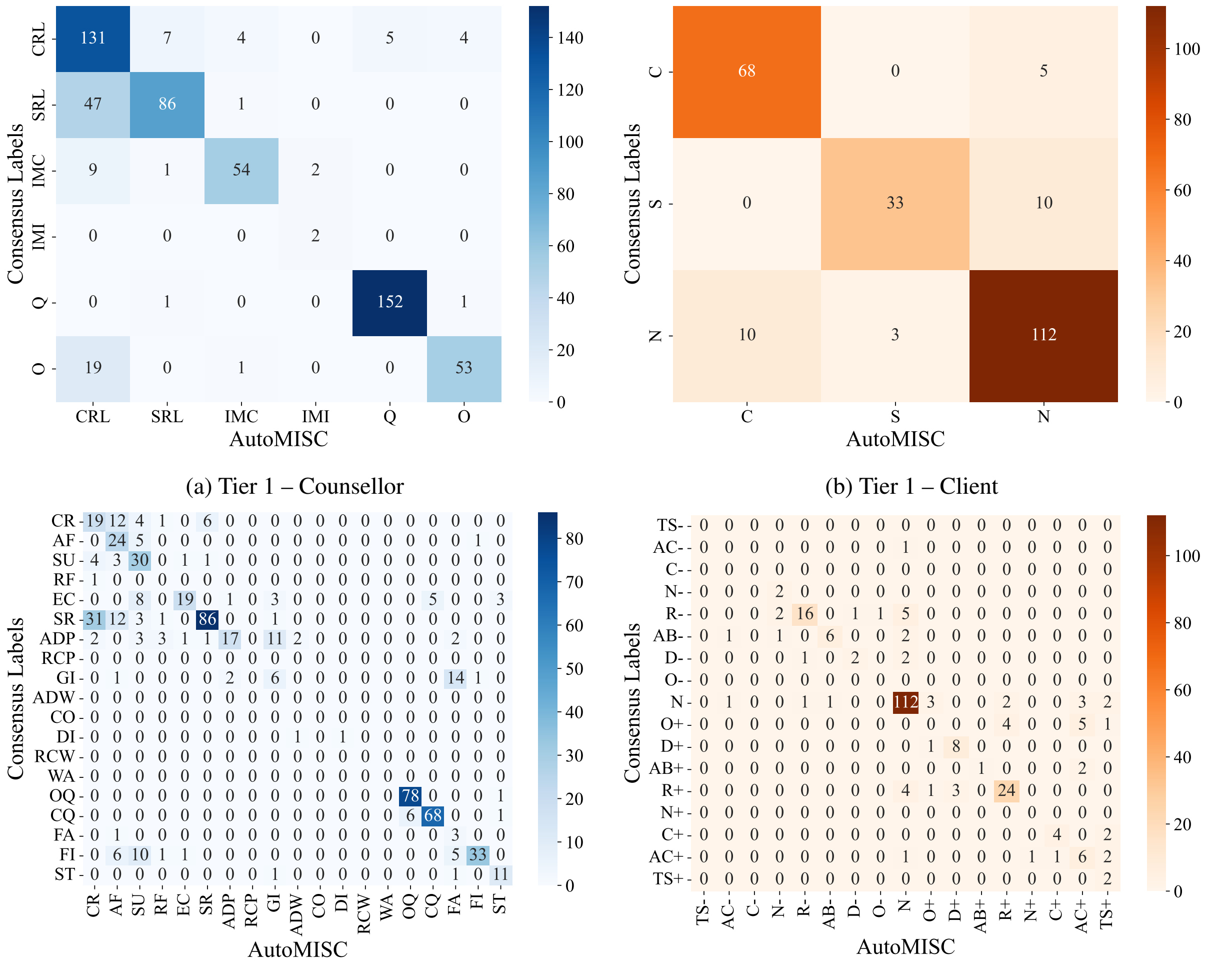
This section contains the complete results from the experiments described in Section 4.3. Table C.1 lists the numerical classification performance results for all models across all classification approaches and all context window sizes. Figure C.1 plots all macro F1 and accuracy scores for them. Figure C.2 show the confusion matrices of *AutoMISC*'s best performing configuration, GPT-4.1 with three context volleys, using the hierarchical classification approach.



(a) Macro F1 Score

(b) Accuracy

Figure C.1: Accuracy and F1 score across all configurations on consensus labels ($n = 821$).



(a) Tier 1 – Counsellor

(b) Tier 1 – Client

(c) Tier 2 – Counsellor

(d) Tier 2 – Client

Figure C.2: Confusion matrices for each speaker and tier, comparing *AutoMISC*'s predictions to the consensus annotations on ten transcripts from the smoking cessation study ($n = 821$).

Model	Class. Structure	Context Volleys	T1 Couns. F1	T1 Couns. Acc.	T1 Client F1	T1 Client Acc.	T2 Couns. F1	T2 Couns. Acc.	T2 Client F1	T2 Client Acc.	T2 Overall F1	T2 Overall Acc.
GPT-4.1	hier.	0	0.54	70	0.82	83	0.36	64	**0.44**	73	0.39	67
		1	0.63	76	0.85	86	0.38	66	0.41	73	0.39	68
		2	0.78	**82**	0.83	85	0.39	**70**	0.40	71	0.39	**70**
		3	0.80	**82**	0.87	88	**0.42**	68	0.41	**76**	**0.42**	**70**
		4	0.77	81	0.86	87	0.38	67	0.39	72	0.39	69
		5	0.77	81	**0.89**	**90**	0.38	67	0.41	75	0.39	69
		10	0.79	81	0.86	88	0.37	66	0.38	73	0.37	68
		20	**0.83**	79	0.86	88	0.37	64	0.40	74	0.38	67
	flat	0	–	–	–	–	0.32	56	0.40	66	0.34	59
		1	–	–	–	–	0.34	57	0.40	68	0.36	60
		2	–	–	–	–	0.38	62	0.41	68	0.39	63
		3	–	–	–	–	0.37	60	0.40	70	0.38	63
		4	–	–	–	–	0.38	60	0.41	68	0.39	62
		5	–	–	–	–	0.38	60	0.41	69	0.39	63
		10	–	–	–	–	0.38	61	0.41	69	0.39	63
		20	–	–	–	–	0.37	58	0.41	69	0.38	62
GPT-4o	hier.	0	0.54	69	0.83	84	0.36	61	0.42	71	0.38	64
		1	0.62	75	0.85	86	0.39	65	0.41	69	0.39	66
		2	0.76	81	0.85	85	0.37	66	0.41	69	0.38	67
		3	0.76	80	0.85	86	0.39	66	0.42	70	0.40	67
		4	0.76	80	0.85	85	0.36	64	0.43	69	0.38	66
		5	0.78	81	0.84	84	0.36	64	0.38	66	0.37	65
		10	0.77	81	0.85	85	0.38	64	0.39	67	0.38	65
		20	0.74	79	0.85	85	0.38	63	0.39	66	0.38	64
	flat	0	–	–	–	–	0.38	57	0.40	64	0.39	59
		1	–	–	–	–	0.41	57	0.42	65	0.41	60
		2	–	–	–	–	0.41	61	0.41	65	0.41	62
		3	–	–	–	–	0.39	62	0.41	63	0.40	62
		4	–	–	–	–	0.38	61	0.43	65	0.39	62
		5	–	–	–	–	0.39	61	0.41	60	0.39	61
		10	–	–	–	–	0.39	60	0.39	59	0.39	60
		20	–	–	–	–	0.36	58	0.39	59	0.37	58
Qwen3-30b-a3b	hier.	0	0.54	69	0.77	78	0.28	55	0.35	63	0.30	57
		1	0.56	71	0.79	79	0.27	55	0.35	63	0.29	57
		2	0.62	73	0.73	73	0.27	55	0.32	56	0.28	55
		3	0.59	73	0.73	73	0.26	56	0.28	58	0.27	56
		4	0.61	71	0.77	77	0.26	51	0.32	59	0.28	53
		5	0.57	68	0.76	76	0.26	50	0.32	56	0.28	52
		10	0.59	69	0.78	78	0.25	48	0.34	59	0.28	51
		20	0.58	68	0.77	78	0.25	48	0.34	59	0.28	51
	flat	0	–	–	–	–	0.29	53	0.31	55	0.29	53
		1	–	–	–	–	0.31	52	0.33	52	0.31	52
		2	–	–	–	–	0.29	51	0.32	49	0.30	50
		3	–	–	–	–	0.28	50	0.31	47	0.29	49
		4	–	–	–	–	0.29	50	0.32	48	0.30	50
		5	–	–	–	–	0.30	51	0.32	50	0.30	51
		10	–	–	–	–	0.29	52	0.31	50	0.30	51
		20	–	–	–	–	0.29	52	0.30	50	0.29	51
Gemma-3-12b	hier.	0	0.54	65	0.73	76	0.29	52	0.41	60	0.32	54
		1	0.60	71	0.80	81	0.30	54	0.40	59	0.33	56
		2	0.62	72	0.77	78	0.28	52	0.38	57	0.31	53
		3	0.60	69	0.77	78	0.27	50	0.39	58	0.30	52
		4	0.60	68	0.76	76	0.27	48	0.34	54	0.29	50
		5	0.58	67	0.76	76	0.26	48	0.35	54	0.29	50
		10	0.55	61	0.75	76	0.23	41	0.34	54	0.26	45
		20	0.57	66	0.73	72	0.23	43	0.34	50	0.27	45
	flat	0	–	–	–	–	0.28	50	0.37	53	0.31	51
		1	–	–	–	–	0.27	51	0.40	52	0.30	51
		2	–	–	–	–	0.28	54	0.37	46	0.31	51
		3	–	–	–	–	0.29	53	0.35	46	0.31	51
		4	–	–	–	–	0.29	55	0.34	45	0.30	52
		5	–	–	–	–	0.28	53	0.31	42	0.29	50
		10	–	–	–	–	0.27	52	0.29	40	0.28	49
		20	–	–	–	–	0.28	52	0.30	39	0.28	48

Table C.1: Macro F1 score and accuracy (%) across all models and configurations ($n = 821$ consensus labels).

D Supplementary Validation Experiments

D.1 Comparison to AnnoMI

As a secondary form of validation, we compare *AutoMISC*'s labels (using our best-performing configuration) against those from the AnnoMI dataset (Wu et al., 2023). This dataset contains 133 MI conversations professionally transcribed and coded under a custom volley-level coding scheme by experienced MI practitioners. Each volley in the dataset has up to three counsellor codes (drawn from questions, reflections, and therapist input categories) and a single client code indicating Change Talk (C), Sustain Talk (S), or Neutral Talk (N). Although inspired by MITI/MISC, it differs significantly from the MISC coding used in this work. To make a direct comparison between *AutoMISC* and the AnnoMI codes, the AnnoMI codes were transformed in the following ways:

1. *AutoMISC* Tier 1 utterance-level **client codes** are aggregated across each volley through a majority vote. Ties are broken using the hierarchy C>S>N. The resulting aggregated labels are compared to AnnoMI's single client label per volley using Cohen's κ, accuracy, and a confusion matrix.
2. For **counsellor codes**, an AnnoMI volley-level label is considered matched if for each counsellor code there exists at least one corresponding utterance-level code in *AutoMISC*'s annotations for that volley, according to the mapping shown in Table D.1. A volley-level match occurs only if all AnnoMI codes are covered.

AnnoMI Code	Mapped MISC 2.5 Codes
Question: open	{OQ}
Question: closed	{CQ}
Reflection: simple	{SR}
Reflection: complex	{CR, RF, AF}
Therapist input: information	{GI}
Therapist input: advice	{ADP, ADW}
Therapist input: options	{ADP, ADW, EC, ST}
Therapist input: negotiation	{ADP, ADW, EC, ST, RCP, RCW, WA, CO, DI}
None of the above	{FA, FI, SU}

Table D.1: Mapping from AnnoMI counsellor labels to MISC 2.5 codes used by *AutoMISC*.

With this mapping the *AutoMISC* client coding achieves a Cohen's $\kappa = 0.51$ (which is considered 'moderate' agreement) and an accuracy of 77% over $n = 4817$ volleys. Figure C.2 gives the confusion matrix betwen the C, S, and N codes between *AutoMISC* and AnnoMI.

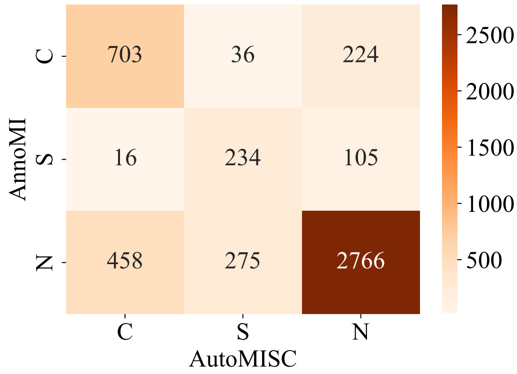


Figure D.1: Confusion matrix comparing *AutoMISC* and AnnoMI client codes (aggregated to volley-level C/S/N).

The counsellor code accuracy is 65% over $n = 4882$ volleys.

D.2 Distinguishing High/Low Quality on the HLQC Dataset

The High Low Quality Counselling (HLQC) dataset (Pérez-Rosas et al., 2019) contains 258 transcribed MI sessions rated as either *high* or *low* quality by expert MI practitioners. HLQC does not include fine-grained behavioural codes for a direct comparison with *AutoMISC*. However, the binary quality rating offers an opportunity to assess whether *AutoMISC* 's outputs align with expert judgments at the session level, using the following process: *AutoMISC* is run on the HLQC dataset using the best-performing configuration, and the three MISC summary scores described in subsection 3.1 are produced. These are used to predict binary counselling quality by training a logistic regression classifier using leave-one-out cross-validation (LOOCV).

Predictor(s)	Acc. (%)	F1	AUC
%MIC	**87**	**0.90**	0.933
R:Q	70	0.79	0.741
%CT	75	0.80	0.729
All Combined	86	0.89	**0.940**

Table D.2: LOOCV classification performance for predicting binary session-level MI quality on HLQC using summary scores derived from *AutoMISC* ($n = 258$).

As shown in Table D.2, the %MIC summary score is the most predictive individual feature, achieving 87% accuracy and an AUC of 0.93. Combining all three summary scores yields the an overall accuracy of 86% accuracy and an AUC of 0.94. These results are consistent with those reported in the original HLQC study, where handcrafted MITI-derived features achieved 83–87% accuracy (Pérez-Rosas et al., 2019).

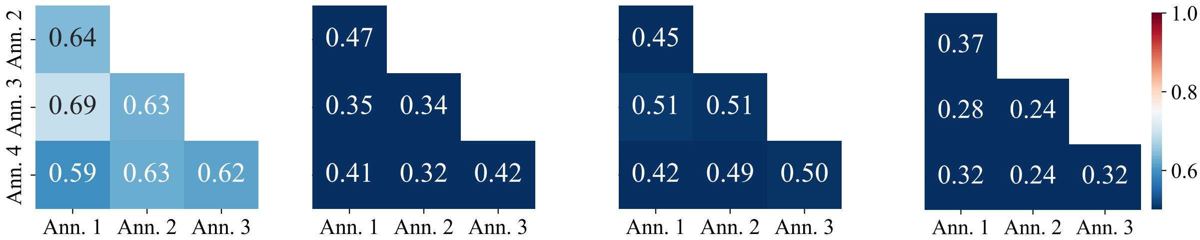
These results demonstrate that *AutoMISC*'s summary scores can serve as evaluators of counselling quality. This highlights the potential for applications of automated coding in MI quality assessment.

E Comparison to Consensus Labels: HLQC Subset

This section contains the complete results from the experiments described in Section 5.4. Based on the label distribution in the HLQC dataset from our experiment in Appendix D.2, we selected a larger and more balanced subset of 10 conversations ($n = 1924$ utterances) for manual annotation to perform this additional validation experiment. Figure E.1 shows the pairwise Cohen's Kappa between annotators and overall Fleiss' Kappa. We then repeated the automated annotation experiments described in Section 5.1 across all models and configuration parameters. The full numerical results are listed in Table E.1, with all macro F1 and accuracy scores visualized in Figure E.2. Figure E.3 show the confusion matrices for *AutoMISC*'s best performing configuration, GPT-4.1 with five context volleys, using the hierarchical classification approach.

We note that, unlike the smoking cessation chatbot transcripts, HLQC is comprised of audio transcriptions of live MI sessions. These include frequent interruptions, filler words, overlapping speech, and transcription errors, such as swapped speaker roles, resulting in a more "noisy" dataset that was more difficult to annotate (we tried to correct these errors to the best of our ability). This was reflected in the lower Fleiss' Kappas: the target of 0.6 was not met for either T2 counsellor codes ($\kappa = 0.47$) or T2 client codes ($\kappa = 0.3$), as shown in Figure E.1.

The automated annotation performance also differed from the chatbot study. The best counsellor accuracy is 14% lower (56% vs 70%), whereas the client accuracy is 5% higher. The macro F1 scores decreased by 0.07 on T2 counsellor codes (0.42 to 0.35) and 0.09 on T2 client codes (0.41 vs 0.32). The decrease in counsellor accuracy is expected, as the HLQC subset contains a more balanced distribution of MI-consistent and MI-inconsistent behaviours, in contrast to the chatbot transcripts which rarely contained MI-inconsistent utterances. The higher client accuracy can be attributed to the substantial increase in filler speech and small talk which is inherent to real speech. The reductions in macro F1 score are consistent with the increased noise and transcription artifacts discussed above. We observe similar trends to those in Section 5.2.1: counsellor accuracy/F1 score generally improves as the number of context volleys increases, up to a point of diminishing returns. Unlike in the chatbot study, this trend also appeared for client codes, likely due to the greater variability and noise in spoken dialogue.



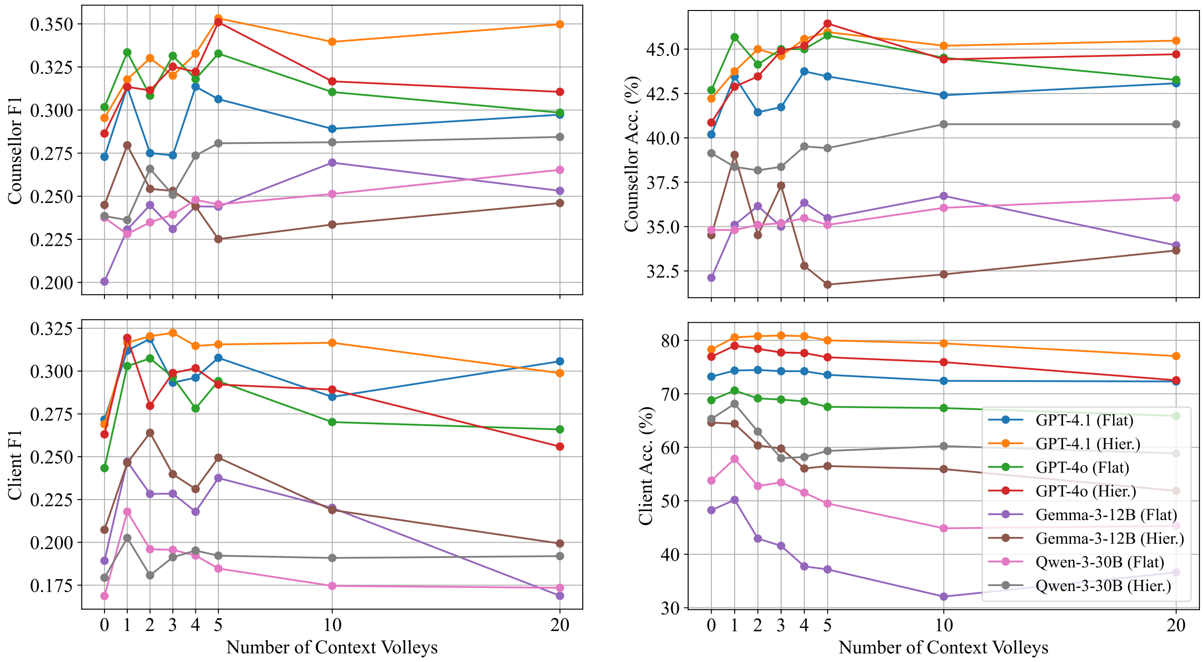
T1 Counsellor
(a) Fleiss' $\kappa = \mathbf{0.63}$

T1 Client
(b) Fleiss' $\kappa = \mathbf{0.39}$

T2 Counsellor
(c) Fleiss' $\kappa = \mathbf{0.47}$

T2 Client
(d) Fleiss' $\kappa = \mathbf{0.30}$

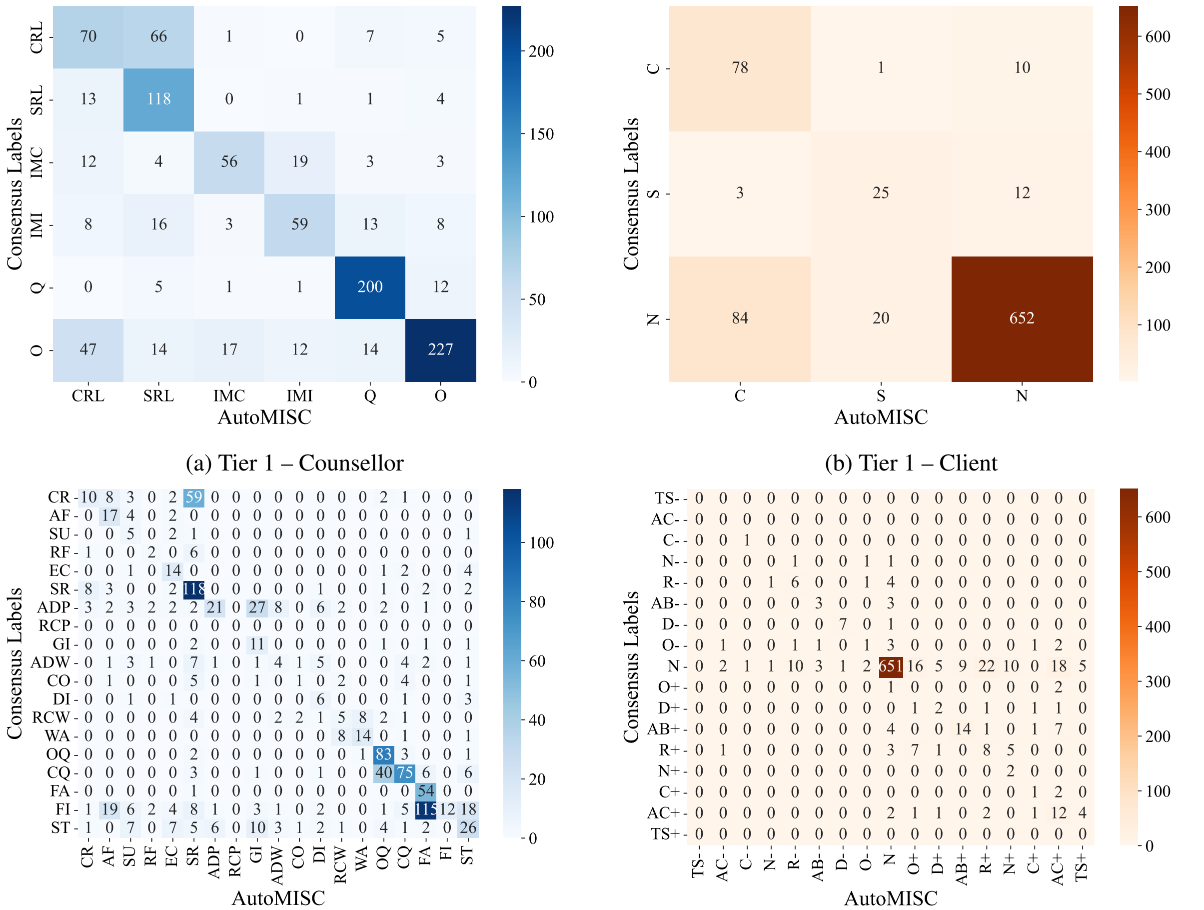
Figure E.1: Pairwise Cohen's Kappa (and Fleiss' Kappa between all annotators) on HLQC subset ($n = 1924$).



(a) Macro F1 Score

(b) Accuracy

Figure E.2: Accuracy and F1 score across all configurations on HLQC subset ($n = 1924$).



(a) Tier 1 – Counsellor

(b) Tier 1 – Client

(c) Tier 2 – Counsellor

(d) Tier 2 – Client

Figure E.3: Confusion matrices for each speaker and tier, comparing *AutoMISC*'s predictions to the consensus annotations on the subset of HLQC ($n = 1924$).

Model	Class. Structure	Context Volleys	T1 Couns.		T1 Client		T2 Couns.		T2 Client		T2 Overall	
			F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
GPT-4.1	hier.	0	0.54	65	0.64	82	0.30	42	0.27	78	0.29	53
		1	0.63	68	0.67	84	0.32	44	0.32	81	0.32	55
		2	0.64	68	0.70	86	0.33	45	0.32	81	0.33	56
		3	0.65	69	0.70	86	0.32	45	0.32	81	0.32	55
		4	0.65	69	0.71	86	0.33	46	0.31	81	0.33	56
		5	0.67	70	0.70	85	0.35	46	0.32	80	0.34	56
		10	0.68	71	0.71	85	0.34	45	0.32	79	0.33	55
		20	0.68	71	0.68	83	0.35	45	0.30	77	0.33	55
	flat	0	–	–	–	–	0.27	40	0.27	73	0.27	50
		1	–	–	–	–	0.31	43	0.31	74	0.31	53
		2	–	–	–	–	0.28	41	0.32	74	0.29	51
		3	–	–	–	–	0.27	42	0.29	74	0.28	51
		4	–	–	–	–	0.31	44	0.30	74	0.31	53
		5	–	–	–	–	0.31	43	0.31	74	0.31	52
		10	–	–	–	–	0.29	42	0.28	72	0.29	51
		20	–	–	–	–	0.30	43	0.31	72	0.30	52
GPT-4o	hier.	0	0.54	64	0.62	81	0.29	41	0.26	77	0.28	51
		1	0.63	68	0.66	83	0.31	43	0.32	79	0.32	53
		2	0.64	68	0.67	84	0.31	43	0.28	78	0.30	54
		3	0.65	69	0.67	83	0.33	45	0.30	78	0.32	55
		4	0.65	70	0.67	83	0.32	45	0.30	78	0.32	55
		5	0.67	71	0.67	83	0.35	46	0.29	77	0.33	55
		10	0.64	68	0.65	81	0.32	44	0.29	76	0.31	54
		20	0.62	68	0.61	78	0.31	45	0.26	73	0.29	53
	flat	0	–	–	–	–	0.30	43	0.24	69	0.28	50
		1	–	–	–	–	0.33	46	0.30	71	0.32	53
		2	–	–	–	–	0.31	44	0.31	69	0.31	51
		3	–	–	–	–	0.33	45	0.30	69	0.32	52
		4	–	–	–	–	0.32	45	0.28	69	0.31	52
		5	–	–	–	–	0.33	46	0.29	68	0.32	52
		10	–	–	–	–	0.31	45	0.27	67	0.30	51
		20	–	–	–	–	0.30	43	0.27	66	0.29	50
Qwen3-30b-a3b	hier.	0	0.50	59	0.51	71	0.24	39	0.18	65	0.22	47
		1	0.51	59	0.53	73	0.24	38	0.20	68	0.23	47
		2	0.52	60	0.51	69	0.27	38	0.18	63	0.24	45
		3	0.54	59	0.49	64	0.25	38	0.19	58	0.23	44
		4	0.54	60	0.50	64	0.27	40	0.20	58	0.25	45
		5	0.55	60	0.50	65	0.28	39	0.19	59	0.25	45
		10	0.57	63	0.51	66	0.28	41	0.19	60	0.25	46
		20	0.58	64	0.49	65	0.28	41	0.19	59	0.26	46
	flat	0	–	–	–	–	0.24	35	0.17	54	0.22	40
		1	–	–	–	–	0.23	35	0.22	58	0.23	42
		2	–	–	–	–	0.23	35	0.20	53	0.22	40
		3	–	–	–	–	0.24	35	0.20	53	0.23	41
		4	–	–	–	–	0.25	35	0.19	52	0.23	40
		5	–	–	–	–	0.25	35	0.18	49	0.23	39
		10	–	–	–	–	0.25	36	0.17	45	0.23	39
		20	–	–	–	–	0.27	37	0.17	45	0.24	39
Gemma-3-12b	hier.	0	0.55	61	0.54	74	0.24	35	0.21	65	0.23	43
		1	0.62	67	0.56	73	0.28	39	0.25	64	0.27	46
		2	0.55	57	0.56	72	0.25	35	0.26	60	0.26	42
		3	0.59	64	0.54	69	0.25	37	0.24	60	0.25	44
		4	0.55	57	0.54	69	0.24	33	0.23	56	0.24	40
		5	0.53	55	0.54	68	0.23	32	0.25	56	0.23	39
		10	0.58	62	0.51	63	0.23	32	0.22	56	0.23	39
		20	0.57	61	0.47	58	0.25	34	0.20	52	0.23	39
	flat	0	–	–	–	–	0.20	32	0.19	48	0.20	37
		1	–	–	–	–	0.23	35	0.25	50	0.24	40
		2	–	–	–	–	0.24	36	0.23	43	0.24	38
		3	–	–	–	–	0.23	35	0.23	42	0.23	37
		4	–	–	–	–	0.24	36	0.22	38	0.24	37
		5	–	–	–	–	0.24	35	0.24	37	0.24	36
		10	–	–	–	–	0.27	37	0.22	32	0.26	35
		20	–	–	–	–	0.25	34	0.17	37	0.23	35

Table E.1: Macro F1 score and accuracy (%) across all configurations on the HLQC subset ($n = 1924$ utterances).

F Correlation Experiment Supplementary Material

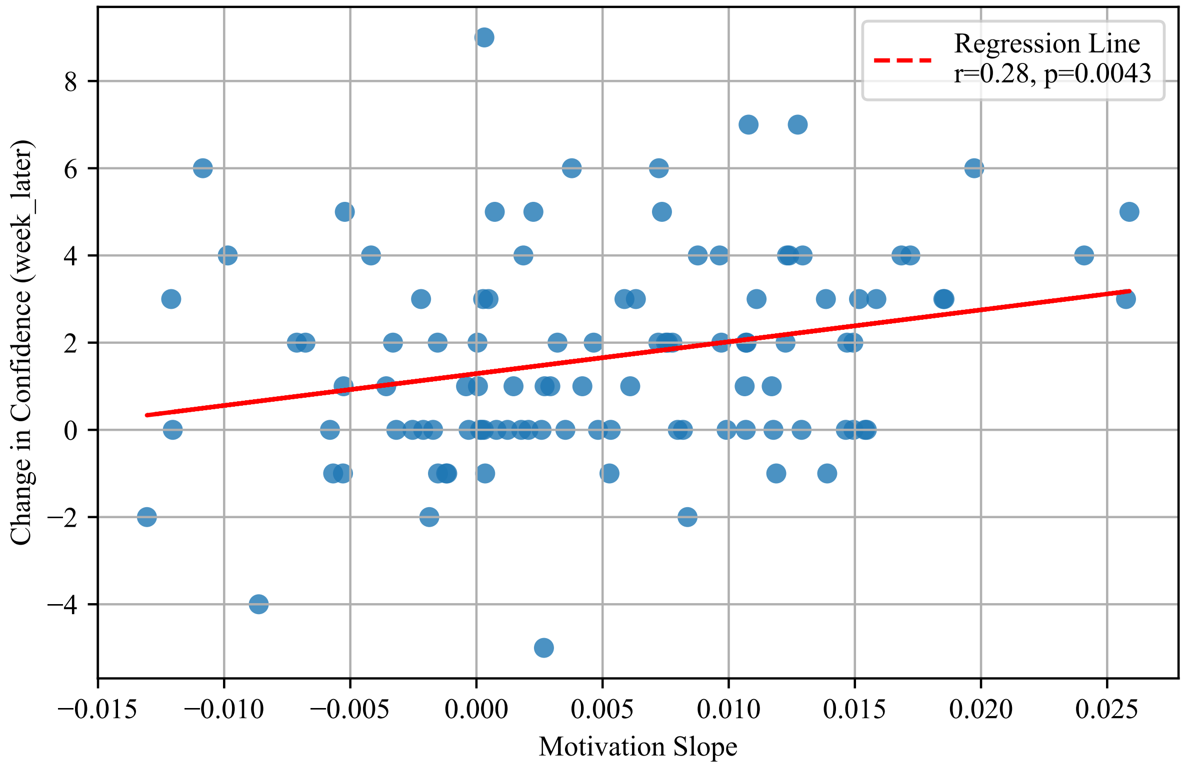


Figure F.1: Client motivation trajectory slope vs. change in client self-reported confidence to quit smoking one week after the session ($n = 106$).