UMUTeam at MIRROR@IberLEF 2026: Synthetic Motivational Interviewing Data Generation with Qwen3 and Medical Continual Pretraining

Ronghao Pan[1], Pedro José Vivancos-Vicente[2], Jorge Gómez-Navalón[1], Tomás Bernal-Beltrán[1] and José Antonio García-Díaz[1,*]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain

Abstract

This paper describes the UMUTeam participation in the MIRROR at IberLEF 2026 task, which focuses on generating synthetic Motivational Interviewing exchanges to improve behavior-code classification. Our approach uses AnnoMI as the seed dataset and reformats it into three target binary pairs: Simple Reflection vs. Complex Reflection (`sr_cr`), Open Question vs. Closed Question (`oq_cq`), and Persuasion vs. Giving Information (`pe_gi`). We fine-tune Qwen3-4B and Qwen3-8B models using LoRA and generate balanced synthetic examples conditioned on explicit behavior-code definitions. In addition, we compare base models with variants that were continual-pretrained on Spanish medical-domain resources, including clinical practice guidelines, CoWeSe, and SMC, to evaluate the effect of cross-lingual medical adaptation. The best result was obtained by the Qwen3-8B base model, with an average F1 score of 40.67%. The results show that the approach was most effective for the information-related category, achieving an `INFO_F1` of 0.7721, while question and reflection categories remained more challenging. Overall, our results suggest that label-controlled synthetic data generation is a promising strategy for low-resource Motivational Interviewing classification, but that medical continual pretraining and larger model size alone are not sufficient to guarantee better downstream performance.

Keywords

Synthetic Data Generation, Continual Pretraining, Large Language Models, Behavior Code Classification

1. Introduction

Motivational Interviewing (MI) is a client-centred counselling approach designed to support behavioural change by helping clients explore ambivalence, strengthen motivation, and articulate their own reasons for change [1]. It is widely used in health-related domains such as smoking cessation, alcohol reduction, medication adherence, weight management, chronic disease management, and mental health support [2]. Unlike directive counselling strategies, MI emphasises collaboration, empathy, autonomy, and reflective listening [1]. For this reason, the quality of an MI session depends not only on the topic being discussed, but also on the linguistic behaviour of the clinician during the interaction.

A central aspect of MI assessment is the identification of clinician Behaviour Codes. These codes describe the communicative function of a clinician utterance, for example whether the clinician asks an open question, asks a closed question, provides information, persuades the client, or reflects the client's previous statement [3]. Such labels are useful because they provide a structured way to evaluate whether a conversation follows the principles of MI [3]. For instance, open questions and complex reflections are generally associated with a more exploratory and client-centred style, while persuasion may indicate a more directive style that can be less consistent with MI principles. Automatic recognition

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ronghao.pan@um.es (R. Pan); pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); jorge.gomezn@um.es (J. Gómez-Navalón); tomas.bernalb@um.es (T. Bernal-Beltrán); joseantonio.garcia8@um.es (J. A. García-Díaz)
https://www.vocali.net/ (P. J. Vivancos-Vicente)
0009-0008-7317-7145 (R. Pan); 0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0002-3651-2660 (J. A. García-Díaz)

of these behaviours can therefore support therapist training, supervision, feedback generation, and large-scale analysis of counselling conversations [4].

However, building robust automatic systems for MI Behaviour Code classification is challenging. Clinical and counselling conversations are sensitive, difficult to collect, and expensive to annotate. Expert annotation requires domain knowledge and careful interpretation, especially when distinguishing between closely related categories [3]. For example, the difference between a simple reflection and a complex reflection can be subtle: a simple reflection mainly repeats or lightly rephrases what the client said, whereas a complex reflection adds meaning, emotion, implication, or interpretation [3]. Similarly, the distinction between giving information and persuasion depends not only on the topic of the clinician's response, but also on whether the response is neutral or attempts to push the client toward a specific action. These subtle pragmatic differences make the task difficult for both human annotators and machine learning models [3].

The scarcity of annotated MI data motivates the use of synthetic data generation. Synthetic examples can be used to complement limited real datasets, increase label balance, and provide additional controlled training material for underrepresented or difficult classes. Recent advances in large language models make it possible to generate coherent dialogue turns that follow a desired structure and label definition. Nevertheless, synthetic data generation for counselling dialogue must be handled carefully. The generated examples should be realistic enough to resemble MI interactions, but also controlled enough to preserve the intended Behaviour Code. In this setting, quality is more important than volume: a small number of well-formed, label-consistent examples may be more useful than a large set of noisy generations [5].

MIRROR [6] at IberLEF 2026 [7] frames this problem as a data-centric shared task. Instead of asking participants to design a new classifier, the task focuses on generating synthetic conversational turns that can improve the performance of a fixed Behaviour Code classifier. This setup encourages participants to concentrate on data quality, label control, diversity, and formatting. The task is organised around three binary classification pairs: Simple Reflection vs. Complex Reflection (`sr_cr`), Open Question vs. Closed Question (`oq_cq`), and Persuasion vs. Giving Information (`pe_gi`). These pairs are particularly relevant because each one contains labels that are semantically close but pragmatically distinct.

In this paper, we describe the participation of UMUTeam in the MIRROR at IberLEF 2026 task. Our approach uses the public AnnoMI dataset as the starting point. AnnoMI [4] contains expert-annotated Motivational Interviewing dialogues and provides a suitable source of real counselling-style utterances. We convert the original annotations into the three target pairwise formats required by the MIRROR task: `sr_cr`, `oq_cq`, and `pe_gi`. Each instance is represented using four fields: topic, client utterance, clinician response, and Behaviour Code label. This structure allows the model to learn not only the surface form of clinician responses, but also the relationship between the client's statement, the conversational topic, and the intended clinician behaviour.

In addition to generating synthetic data, another objective of our work is to evaluate whether models that have undergone continual pretraining on medical-domain data are more effective for this task than their corresponding base models. Since the MIRROR task is situated in a health-related conversational setting, domain-adapted language models may be better equipped to generate clinically plausible and semantically coherent examples. For this reason, we test both base and medically continual-pretrained variants of the Qwen3 family, focusing specifically on Qwen3-4B and Qwen3-8B. This comparison allows us to analyse whether domain-specific pretraining contributes to better synthetic MI data generation, and whether this effect depends on model size.

To generate synthetic data, we fine-tune the selected Qwen3 [8] models using parameter-efficient LoRA adaptation. The models are trained on formatted examples that include the target label, a natural-language definition of the label, and a complete example. This label-conditioned format is intended to make the generation process more controllable. At inference time, the model receives a target Behaviour Code and a sampled health-related topic, and generates a complete synthetic example containing a client utterance and a clinician response. The generated outputs are then parsed and filtered using automatic quality-control rules to remove malformed, duplicated, overly short, or label-inconsistent examples.

The main contribution of this paper is a simple and reproducible pipeline for generating compact, balanced synthetic datasets for MI Behaviour Code classification, together with an analysis framework for comparing base and medically adapted Qwen3 models. The pipeline consists of three stages: first, reformatting AnnoMI into the MIRROR pairwise label structure; second, fine-tuning Qwen3-4B and Qwen3-8B models with LoRA; and third, generating and filtering label-controlled synthetic examples for the three target pairs.

2. Related Works

Motivational Interviewing (MI) is a counseling approach designed to support behavior change through collaboration, empathy, and client autonomy. In MI, the clinician's linguistic behavior plays a central role, since different response types can either encourage client elaboration or introduce a more directive style. For this reason, several coding schemes have been proposed to assess the quality and consistency of MI sessions [1]. Among them, the Motivational Interviewing Treatment Integrity (MITI) code is widely used to evaluate clinician behaviors such as questions, reflections, information giving, and persuasion [3].

Automatic MI behavior classification aims to identify these clinician behaviors from dialogue transcripts. This task is challenging because the distinctions between labels are often pragmatic rather than purely lexical. For example, a Simple Reflection and a Complex Reflection may both restate the client's previous utterance, but the latter adds interpretation, emotion, or implied meaning. Similarly, Giving Information and Persuasion may both contain health-related advice, but they differ in whether the clinician presents neutral information or attempts to push the client toward change. These subtle distinctions make MI behavior classification a difficult task for machine learning systems, especially in low-resource settings [9].

The development of automatic MI analysis systems depends on the availability of annotated counseling data. However, such data is difficult to obtain because counseling conversations are sensitive, require privacy protection, and must be annotated by trained experts [10, 11]. As a result, available MI datasets are relatively small compared with general-domain dialogue corpora.

AnnoMI [4] is one of the most relevant publicly available resources for this task. It contains expert-annotated Motivational Interviewing dialogues and provides utterance-level labels for client and therapist behavior. Because of its public availability and its focus on MI-style interactions, AnnoMI has been used as a starting point for research on automatic counseling dialogue analysis. In this paper, we use AnnoMI as the seed dataset and reformat its annotations into the three target MIRROR label pairs: `sr_cr`, `oq_cq`, and `pe_gi`.

Other MI resources, such as MI-TAGS [12], have also been developed for coding therapist behavior. However, access to these datasets may be restricted, which limits their direct use in shared-task settings. This motivates approaches that can make effective use of public data and generate additional synthetic examples while preserving the target label distinctions.

Synthetic data generation has become an increasingly common strategy for low-resource NLP tasks. When annotated data is scarce, generated examples can be used to increase training set size, improve class balance, and expose models to a wider range of linguistic patterns [13, 14]. This is especially relevant for specialized domains such as clinical dialogue, where real data is difficult to collect and annotation is costly.

Large language models are well suited for synthetic data generation because they can produce fluent and contextually coherent text under controlled prompting conditions [15]. However, synthetic data quality remains a critical concern. Generated examples may be fluent but mislabeled, repetitive, unrealistic, or biased toward patterns learned from the prompt [16]. In classification tasks, label consistency is particularly important: a synthetic example is only useful if the generated text clearly expresses the intended class.

For MI behavior classification, synthetic data generation requires both domain plausibility and pragmatic label control. A generated clinician response must sound like a realistic counseling utterance,

but it must also clearly match the target Behavior Code. For this reason, our approach uses label-conditioned generation with explicit natural-language definitions of each Behavior Code [17]. The generated outputs are then parsed and filtered to remove malformed, duplicated, or label-inconsistent examples.

Large language models have shown strong performance across many NLP tasks, including text generation, classification, and dialogue modeling. In clinical and medical NLP, LLMs are particularly attractive because they can encode broad linguistic knowledge and adapt to specialized terminology [18]. However, general-purpose LLMs may not always be optimal for medical or counseling contexts, where domain-specific language and safety-sensitive communication are important.

One way to improve domain suitability is continual pretraining. In continual pretraining, a base language model is further trained on domain-specific corpora before being adapted to a downstream task [19]. For medical NLP, this process can help the model better represent clinical vocabulary, health-related concepts, and domain-specific discourse patterns. In this paper, we evaluate this idea by comparing base Qwen3 models with Qwen3 variants that have undergone continual pretraining on medical-domain data. This comparison allows us to study whether medical domain adaptation improves synthetic MI data generation.

Our work combines ideas from MI behavior classification, synthetic data generation, medical-domain language modeling, and parameter-efficient fine-tuning. Unlike approaches that focus on building a new classifier, the task frames the problem as a data-centric task: participants must generate compact synthetic datasets that improve a fixed downstream classifier. Within this setting, our contribution is a reproducible pipeline that starts from AnnoMI, reformats the data into the required MIRROR label pairs, fine-tunes Qwen3 models with LoRA [20], and generates balanced synthetic examples for each target Behavior Code.

The main difference between our approach and standard data augmentation is that generation is explicitly conditioned on Behavior Code definitions. This design is intended to guide the model toward producing examples that are not only fluent but also label-consistent. In addition, by comparing base models with medical continual-pretrained models, our work explores whether domain adaptation improves synthetic data generation for health-related counseling dialogue.

3. Methodology

The methodology followed by UMUTeam is organized into four main stages: dataset construction, supervised fine-tuning dataset preparation, model adaptation, and synthetic data generation, as summarized in Figure 1. The overall objective is to generate compact and balanced synthetic datasets for the three MIRROR target pairs: Simple Reflection vs. Complex Reflection (`sr_cr`), Open Question vs. Closed Question (`oq_cq`), and Persuasion vs. Giving Information (`pe_gi`). In addition, we compare base Qwen3 models with their corresponding continual-pretrained variants on medical-domain data in order to assess whether medical domain adaptation improves the quality and usefulness of the generated examples.

3.1. Source Data

The main source dataset used in our experiments is AnnoMI, a public dataset of expert-annotated Motivational Interviewing dialogues. AnnoMI provides counseling-style conversations with utterance-level annotations, making it suitable for generating examples of clinician behaviors in health-related dialogue. We initially considered combining AnnoMI with MI-TAGS; however, access to MI-TAGS was requested and was not available at the time of preparing this pipeline. Therefore, the current version of the dataset construction process is based on AnnoMI.

The original AnnoMI corpus is distributed at the utterance level. To match the MIRROR task format, conversations were first reconstructed using the transcript and utterance identifiers. Subsequently, consecutive client–therapist utterance pairs were extracted, where each example consists of a client utterance followed by the immediate therapist response.

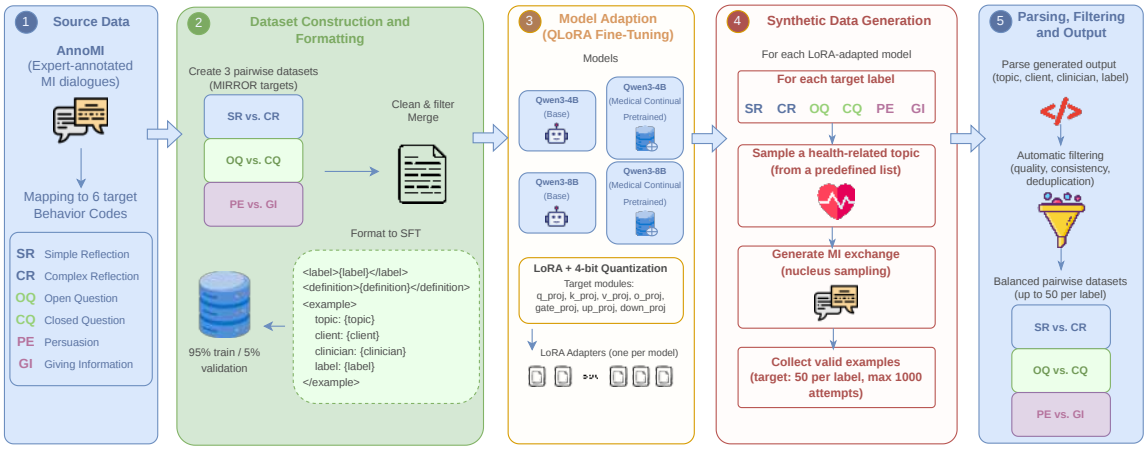

Figure 1: Overview of the UMUTeam approach for MIRROR synthetic data generation.

The therapist annotations associated with the response utterance were then mapped to the MIRROR target labels. Reflection subtypes were mapped to SR (Simple Reflection) and CR (Complex Reflection), while question subtypes were mapped to OQ (Open Question) and CQ (Closed Question). For the information-related task, therapist input subtypes were used to derive the target labels: information was mapped to GI (Giving Information), whereas advice and negotiation were mapped to PE (Persuasion). The options subtype was excluded due to its ambiguous correspondence with the MIRROR label definitions.

The resulting examples were grouped into the three binary datasets required by the task: `sr_cr.csv`, `oq_cq.csv` and `pe_gi.csv`. Each file contains examples for a single Behavior Code distinction and follows the same schema: `topic`, `client`, `clinician` and `label`.

The `topic` column represents the health or counseling topic of the exchange, `client` contains the client utterance, `clinician` contains the clinician response, and `label` contains the Behavior Code assigned to the clinician response.

3.2. Behavior Code Definitions

To ensure consistency across dataset preparation, fine-tuning, and generation, we defined explicit natural-language descriptions for each Behavior Code. These definitions were used both during dataset formatting and during synthetic generation. Simple Reflection is defined as a clinician response that repeats or lightly rephrases the client without adding much new meaning. Complex Reflection adds meaning, feeling, implication, or deeper interpretation. Open Question refers to a broad question that invites elaboration, whereas Closed Question refers to a yes/no or narrow factual question. Persuasion is defined as a clinician response that tries to convince, push, or pressure the client toward change. Giving Information refers to neutral facts, explanations, or education without pressure.

These definitions are especially important because several of the target pairs involve subtle pragmatic differences. For example, Simple Reflection and Complex Reflection may have similar surface forms, but Complex Reflection adds interpretation or emotional meaning. Similarly, Persuasion and Giving Information may both include health-related content, but Persuasion attempts to push the client toward a decision, while Giving Information remains neutral.

3.3. Dataset Construction and Formatting

After constructing the three pairwise datasets from AnnoMI, we prepared the resulting examples for supervised fine-tuning. The dataset preparation script reads the three CSV files `sr_cr.csv`, `oq_cq.csv`, and `pe_gi.csv`. It keeps only the columns `topic`, `client`, `clinician`, and `label`. Rows with missing or empty values are removed, and basic text cleaning is applied by replacing line breaks with spaces

and stripping unnecessary whitespace. After cleaning, all three pairwise files are concatenated into a single training dataset.

Each row is converted into a structured text format. The format includes the target label, the definition of the label, and the complete example:

Listing 1: Structured format used for supervised fine-tuning.

```
<label>{label}</label>
<definition>{definition}</definition>
<example>
    topic: {topic}
    client: {client}
    clinician: {clinician}
    label: {label}
</example>
```

This format was selected because it makes the generation task explicitly label-conditioned. The model is not only exposed to the final label, but also to a short explanation of what the label means. This helps the model learn the relationship between the Behavior Code and the clinician response. The XML-like tags also make the output easier to parse during generation.

The formatted examples are stored as a Hugging Face `Dataset` with a single `text` field. The dataset is split into training and validation partitions using a 95/5 split with a fixed seed. Finally, the resulting dataset is saved locally as `mirror_sft_dataset`, which is the input used in the model fine-tuning stage.

3.4. Prompt Design

In addition to the fine-tuning format, we also defined an explicit generation-oriented prompt format. The goal of the prompt is to instruct the model to generate a short Motivational Interviewing exchange under controlled conditions. The intended input structure includes the task, domain, topic, target behavior code, client profile, constraints, and output format:

Listing 2: Initial generation prompt design.

```
Task: Generate a motivational interviewing exchange.
Domain: medical counseling
Topic: smoking cessation
Target behavior code: CR
Client profile: adult patient, ambivalent, failed previous quit
   attempt
Constraints:
    - one client turn
    - one clinician turn
    - clinician response must clearly match CR
    - natural tone
    - medically plausible
    - no explicit mention of the label
Output format: JSON
```

The expected output is a structured JSON object containing the client utterance, the clinician response, and the target label:

Listing 3: Example of the expected generation output.

```
{
  "client": "I have tried several times, but when I get stressed I
     start smoking again.",
```

```
  "clinician": "Smoking seems to have become a way to cope with
     those tense moments, even though part of you is already tired
     of depending on it.",
  "label": "CR"
}
```

In the implemented generation script, this idea is adapted into a more compact tag-based prompt that follows the same label-conditioned logic used during fine-tuning. The generation prompt contains the target label, its definition, and a sampled health-related topic:

Listing 4: Compact tag-based prompt used during generation.

```
<label>{label}</label>
<definition>{definition}</definition>
<target_topic>{topic}</target_topic>
<example>
topic:
```

The model is then expected to complete the example with the topic, client utterance, clinician response, and label.

To illustrate the exact prompt used during generation, Listing 5 shows an example for the CR label. The same structure was used for the remaining labels, replacing the target label, its definition, and the sampled topic.

Listing 5: Example of the generation prompt used for the CR label.

```
CR
A Complex Reflection is a clinician response that goes beyond
    repeating the 'clients words.
It adds meaning, feeling, implication, or a deeper interpretation of
     what the client expressed.
<target_topic>smoking cessation</target_topic>
topic:
```

Given this prompt, the model was expected to complete the structured example by generating the topic, a client utterance, a clinician response matching the requested Behaviour Code, and the corresponding label. For example, a valid completion for this prompt would be:

Listing 6: Example of a valid generated completion.

```
topic: smoking cessation
client: I have tried to quit smoking before, but every time work
    gets stressful I end up buying cigarettes again.
clinician: Smoking has become a way to manage pressure at work, even
     though another part of you wants to stop relying on it.
label: CR
```

This prompt format was selected to keep generation close to the supervised fine-tuning format while still allowing the topic and clinician response to vary across generations.

3.5. Models

We evaluate two model sizes from the Qwen3 family: Qwen3-4B and Qwen3-8B. For each size, we consider two variants: the base model and a continual-pretrained model adapted on medical-domain data. The motivation for this comparison is to evaluate whether continual pretraining on medical data provides an advantage in a health-related dialogue generation task. Since MIRROR examples are situated in medical counseling scenarios, a model further adapted to medical-domain text may generate more plausible, domain-aware, and clinically coherent examples than a general base model.

Thus, the experimental comparison includes Qwen3-4B base, Qwen3-4B continual-pretrained on medical data, Qwen3-8B base, and Qwen3-8B continual-pretrained on medical data. All models are adapted using the same supervised fine-tuning pipeline and the same formatted MIRROR-style dataset, allowing the effect of model size and medical continual pretraining to be compared under similar conditions.

3.6. LoRA Fine-Tuning

To fine-tune the models efficiently, we use LoRA, a parameter-efficient fine-tuning method. This allows us to adapt large causal language models without updating all model parameters. The training script loads the formatted dataset from `mirror_sft_dataset`. The tokenizer is loaded from the selected model path. If the tokenizer does not define a padding token, the end-of-sequence token is used as the padding token.

The model is loaded in 4-bit quantization using BitsAndBytes. Before applying LoRA, the model is prepared for k-bit training. LoRA is then applied to the main attention and feed-forward projection modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The quantization and LoRA settings are summarized in Table 1.

Table 1
Quantization and LoRA configuration.

Parameter	Value
Quantization	
Quantization	4-bit
Quantization type	NF4
Double quantization	True
Compute dtype	bfloat16
Device mapping	auto
LoRA	
Rank r	16
LoRA alpha	32
LoRA dropout	0.05
Bias	none
Task type	Causal language modeling
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

The dataset is tokenized with a maximum sequence length of 512 tokens. Since the task is formulated as causal language modeling, the data collator uses standard language modeling without masked language modeling. The complete training hyperparameters are shown in Table 2.

After training, the LoRA adapter and tokenizer are saved to a model-specific output directory. The same training strategy is applied to Qwen3-4B and Qwen3-8B, both in their base and continual-pretrained variants.

3.7. Synthetic Data Generation

Generation is performed independently for each of the six labels: SR, CR, OQ, CQ, PE, and GI. For each generation attempt, a topic is randomly sampled from a predefined list of health-related topics. The topic list includes smoking cessation, alcohol reduction, exercise, healthy eating, sleep habits, diabetes management, medication adherence, stress management, weight management, chronic pain, blood pressure, and mental health. The generation configuration uses nucleus sampling to encourage diversity. For each label, the script aims to collect 50 valid examples, with a maximum of 1000 attempts per label. The full generation configuration is summarized in Table 3.

Table 2
Training hyperparameters.

Parameter	Value
Per-device train batch size	2
Gradient accumulation steps	8
Effective batch size	16
Learning rate	2×10^{-4}
Number of training epochs	3
Warmup ratio	0.05
Weight decay	0.01
Logging steps	20
Save steps	250
Evaluation steps	250
Evaluation strategy	steps
Save total limit	2
Precision	fp16
Maximum sequence length	512
Data collator	Causal language modeling, `mlm=False`

Table 3
Synthetic data generation configuration.

Parameter	Value
Labels	SR, CR, OQ, CQ, PE, GI
Final examples per label	50
Candidate examples per label	50
Maximum attempts per label	1000
Batch size	32
Maximum new tokens	512
Temperature	0.9
Top-p	0.95
Repetition penalty	1.05
Decoding strategy	Nucleus sampling

3.8. Parsing and Filtering

Generated outputs are automatically parsed using regular expressions. The parser extracts the `topic`, `client`, `clinician`, and `label` fields. An example is accepted only if all required fields are present and the generated label matches the expected target label. If the model generates a different label from the one requested, the example is discarded.

The filtering process removes low-quality examples with empty fields, overly short client or clinician turns, generic client responses, malformed clinician responses, and duplicated client-clinician-label triples. The duplicate filter is applied using the lowercased `client`, lowercased `clinician`, and `label` values.

Table 4
Automatic filtering criteria for generated examples.

Criterion	Action
Missing topic, client, or clinician field	Discard
Generated label differs from target label	Discard
Client utterance shorter than 10 words	Discard
Clinician response shorter than 4 words	Discard
Generic client response	Discard
Malformed clinician response	Discard
Duplicate client-clinician-label triple	Discard

Generic client responses include short non-informative utterances such as “yes,” “no,” “okay,” “maybe,” and “I don’t know.” This filtering step is intended to improve the structural quality and label consistency of the final synthetic dataset.

4. Results

The final evaluation was carried out using the official MIRROR@IberLEF 2026 evaluation framework. The performance of each submission was reported in terms of F1 scores for the three target classification groups: information-related behaviors, question-related behaviors, and reflection-related behaviors. The final score was computed as the average of the three group-level F1 scores.

UMUTeam obtained an average F1 score of 0.7311 ± 0.0054 in the final phase, ranking third among the participating teams. The detailed results are shown in Table 5. Our best performance was obtained for the question-related category, with a QUEST_F1 of 0.8464 ± 0.0027. The information-related category also achieved a competitive score, with an INFO_F1 of 0.7593 ± 0.0039, close to the best-performing systems. The reflection-related category obtained a score of 0.5875 ± 0.0159, also placing UMUTeam among the top systems for this group.

Table 5
Official MIRROR@IberLEF 2026 final phase results.

Team	Rank	Information	Questions	Reflections	Avg.
UNISON-GOATS	1	0.7728 ± 0.0066	0.8559 ± 0.0042	0.5811 ± 0.0175	**0.7366 ± 0.0064**
OpenCodice	2	0.7466 ± 0.0183	**0.8663 ± 0.0045**	0.5942 ± 0.0140	0.7357 ± 0.0061
UMUTeam	**3**	**0.7593 ± 0.0039**	**0.8464 ± 0.0027**	**0.5875 ± 0.0159**	**0.7311 ± 0.0054**
ucdicsnlpteam	4	0.7603 ± 0.0112	0.8438 ± 0.0047	0.5869 ± 0.0153	0.7303 ± 0.0062
steam_uaz	5	**0.7749 ± 0.0055**	0.8382 ± 0.0052	0.5667 ± 0.0131	0.7266 ± 0.0056
roman_lab	7	0.7369 ± 0.0135	0.8311 ± 0.0066	**0.5948 ± 0.0168**	0.7209 ± 0.0068

The results show that our synthetic data generation pipeline was more effective for the pe_gi pair than for the oq_cq and sr_cr pairs. This suggests that the model was able to generate examples where the distinction between neutral information giving and persuasive language was clearer and more useful for the downstream classifier. In particular, INFO_F1 reached 0.7721, matching the score obtained by OpenCodice and remaining close to the best score in this category.

However, the lower QUEST_F1 and REFLX_F1 values indicate that generating useful examples for questions and reflections remains more challenging. The oq_cq distinction may be sensitive to subtle linguistic cues, since open and closed questions can share similar lexical patterns while differing in the type of response they elicit. Similarly, the sr_cr distinction requires identifying whether the clinician merely repeats the client’s meaning or adds interpretation, emotion, or implication. These pragmatic differences may be harder to capture with automatic generation and filtering alone.

The category-level results also provide an indirect view of the types of errors made by the downstream classifier trained on our generated data. Since the official evaluation did not provide instance-level predictions, we could not compute a full confusion matrix or manually inspect individual misclassified examples. However, the large differences between the three F1 scores suggest that the generated datasets did not provide the same level of discriminative signal for all Behaviour Code pairs.

The strongest result was obtained for the information-related pair, suggesting that the synthetic examples generated for PE and GI contained clearer surface and pragmatic cues. In contrast, the very low score for the question-related pair indicates that the generated OQ/CQ examples were likely not sufficiently useful for separating open and closed questions in the official test set. A possible explanation is that the generated questions may have relied on repetitive lexical templates rather than capturing the functional distinction between questions that invite elaboration and questions that request narrow factual answers. Similarly, the lower reflection score suggests that the generated SR/CR examples may not have consistently encoded the difference between simple reformulation and deeper interpretation. These observations indicate that the main classification errors were probably concentrated in the OQ/CQ

and SR/CR distinctions, where the labels depend on subtle pragmatic properties rather than only on explicit lexical markers.

Overall, the official results indicate that the proposed approach is a promising data-centric baseline, particularly for information-related behavior codes. Nevertheless, the gap with the highest-ranked systems shows that additional quality control and selection strategies are needed. Future improvements could include classifier-in-the-loop filtering, human validation of generated examples, more diverse prompt templates, and separate generation strategies for each behavior-code pair.

5. Discussion

The four submitted configurations obtained very similar results, as shown in Table 6. The best score was achieved by the Qwen3-8B base model, with an average F1 of 40.67%, followed by the medical continual-pretrained Qwen3-4B model with 40.63%, the Qwen3-4B base model with 40.61%, and the medical continual-pretrained Qwen3-8B model with 40.04%. These small differences suggest that the proposed pipeline was stable across model sizes and pretraining settings.

Table 6
Average F1 scores obtained by the evaluated model configurations.

Model configuration	Average F1 score
Qwen3-4B base	40.61%
Qwen3-4B medical continual-pretrained	40.63%
Qwen3-8B base	40.67%
Qwen3-8B medical continual-pretrained	40.04%

Increasing the model size from 4B to 8B did not lead to a clear improvement. Although Qwen3-8B base obtained the highest score, its advantage over Qwen3-4B base was only 0.06 percentage points. This indicates that, in this setting, the quality of the synthetic data may depend more on prompt design, label control, and filtering than on model size alone.

The effect of medical continual pretraining was also limited. For Qwen3-4B, it produced a very small improvement, from 40.61% to 40.63%. However, for Qwen3-8B, performance decreased from 40.67% to 40.04%. The continual-pretrained models were adapted using Spanish medical-domain resources, including clinical practice guidelines, the Corpus Web Salud Español (CoWeSe dataset [21], and the Spanish Medical Corpus (SMC) dataset [22]. Therefore, this experiment also involved a cross-lingual transfer setting, since the medical continual pretraining was mainly performed on Spanish data, while the MIRROR synthetic examples and the AnnoMI-based seed data were in English.

The small differences between base and medical continual-pretrained models lead to two main conclusions. First, domain-specific medical knowledge was not the limiting factor in this task. Although the examples are situated in health-related counseling scenarios, the target distinctions are primarily pragmatic and interactional. The model must learn whether a clinician response functions as a reflection, a question, information giving, or persuasion, rather than simply produce medically accurate content. Therefore, additional exposure to medical documents may improve terminology or factual plausibility, but it does not necessarily improve control over MI Behaviour Codes.

Second, the results suggest that the generation pipeline itself had a stronger influence than the underlying model variant. All four configurations used the same AnnoMI-derived training data, the same label definitions, the same generation format, and the same automatic filtering rules. As a result, the synthetic datasets generated by the different models may have been structurally similar, leading to similar downstream performance. This indicates that future improvements are more likely to come from better data selection, pair-specific prompts, label-aware filtering, and classifier-in-the-loop validation than from replacing the base model with a medically adapted version alone. This cross-lingual setting may have further reduced the benefits of continual pretraining. While medical concepts are often transferable across languages, the MIRROR task relies heavily on subtle linguistic and pragmatic distinctions expressed in English dialogue. Consequently, improvements in Spanish medical

representations may not directly translate into better generation of English Motivational Interviewing interactions.

These results suggest that Spanish medical continual pretraining did not consistently transfer to better English Motivational Interviewing data generation. One possible explanation is that the task requires not only medical plausibility, but also precise control over counseling behavior codes. Clinical guidelines and medical corpora may improve domain knowledge, but they do not necessarily teach the model how to distinguish between pragmatic categories such as OQ vs. CQ or SR vs. CR.

At the category level, the approach was most effective for the information-related pair. UMUTeam obtained an INFO_F1 of 0.7721, while the scores for questions and reflections were lower, with QUEST_F1 = 0.1431 and REFLX_F1 = 0.3046. This suggests that the distinction between Persuasion and Giving Information was easier for the generation pipeline, while Open vs. Closed Questions and Simple vs. Complex Reflections required more subtle pragmatic control.

Overall, the results show that the proposed approach is a competitive baseline for synthetic data generation in MIRROR, especially for information-related behavior codes. However, the small differences between model variants and the weaker performance on questions and reflections indicate that future improvements should focus on better quality control, classifier-in-the-loop filtering, and more specific prompts for each behavior-code pair.

Although the present ablation study compares four model configurations, additional pipeline components could also be investigated through local experimentation. For example, future analyses could evaluate the contribution of the natural-language Behavior Code definitions, compare different prompt formats, study the effect of the filtering strategy, or analyze the impact of generating different numbers of examples per label. These experiments were outside the scope of the official evaluation but represent promising directions for improving the generation pipeline.

6. Conclusion

In this paper, we presented the UMUTeam approach for the MIRROR at IberLEF 2026 task. Our system follows a data-centric strategy based on generating synthetic Motivational Interviewing exchanges for three behavior-code pairs: `sr_cr`, `oq_cq`, and `pe_gi`. Starting from the AnnoMI dataset, we reformatted the data into a unified structure with `topic`, `client`, `clinician`, and `label` fields, and used this format to fine-tune Qwen3 models with LoRA.

We evaluated four model configurations: Qwen3-4B base, Qwen3-4B medical continual-pretrained, Qwen3-8B base, and Qwen3-8B medical continual-pretrained. The best result was obtained by the Qwen3-8B base model, with an average F1 score of 40.67%. However, the differences among the four configurations were small, suggesting that model size and medical continual pretraining were not the main factors determining performance in this setting. In particular, the Spanish medical continual-pretrained models did not consistently improve the generation of English Motivational Interviewing examples, highlighting the difficulty of cross-lingual transfer from medical documents to counseling-style dialogue.

The results show that our approach was most effective for the information-related category, achieving an INFO_F1 of 0.7721. In contrast, lower scores were obtained for questions and reflections, indicating that these behavior-code pairs require more precise pragmatic control. Distinguishing OQ from CQ and SR from CR remains challenging because the differences often depend on subtle aspects of intention, interpretation, and conversational function.

Overall, the proposed pipeline provides a simple and reproducible baseline for synthetic data generation in low-resource Motivational Interviewing behavior classification. Future work should focus on improving the semantic quality of generated examples through classifier-in-the-loop filtering, pair-specific prompting strategies, semantic diversity control, and human validation. These improvements may help generate more label-consistent synthetic data and better support downstream behavior-code classifiers.

Acknowledgments

This research is part of the research project MedicaLLM (CPP2024-011574) funded by MICIU/AEI /10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] W. R. Miller, S. Rollnick, Motivational interviewing: Helping people change, Guilford press, 2012.
[2] S. Rubak, A. Sandbæk, T. Lauritzen, B. Christensen, Motivational interviewing: a systematic review and meta-analysis, The British journal of general practice 55 (2005) 305.
[3] T. B. Moyers, J. K. Manuel, D. Ernst, Motivational interviewing treatment integrity coding manual 4.1, Unpublished manual 1 (2014) 3.
[4] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. R. Recupero, D. Riboni, Anno-mi: A dataset of expert-annotated counselling dialogues, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6177–6181.
[5] A. Ng, Mlops: From model-centric to data-centric ai. deeplearning. ai, IEEE Spectr. (2021).
[6] L. J. Arellano, C. Olachea, J. Piette, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. I. Hernández-Farías, Overview of mirror at iberlef 2026: Motivational interviewing response & rating via synthetic conversational turns, Procesamiento del Lenguaje Natural 77 (2026).
[7] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026, pp. –.
[8] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[9] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, L. An, K. J. Goggin, D. Catley, 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, 2017, pp. 1128–1137.
[10] N. Flemotomos, V. R. Martinez, Z. Chen, T. A. Creed, D. C. Atkins, S. Narayanan, Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations, PloS one 16 (2021) e0258639.
[11] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, V. Srikumar, A comparison of natural language processing methods for automated coding of motivational interviewing, Journal of substance abuse treatment 65 (2016) 43–50.
[12] B. Cohen, M. Zisquit, S. Yosef, D. Friedman, K. Bar, Motivational interviewing transcripts annotated with global scores, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11642–11657. URL: https://aclanthology.org/2024.lrec-main.1017/.
[13] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, in: Findings of the association for computational linguistics: ACL-IJCNLP 2021, 2021, pp. 968–988.

[14] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 6382–6388.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[16] A. G. Møller, J. A. Dalsgaard, A. Pera, L. M. Aiello, Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks, arXiv preprint arXiv:2304.13861 2 (2023) 11.

[17] T. Schick, H. Schütze, Generating datasets with pretrained language models, in: Proceedings of the 2021 conference on empirical methods in natural language processing, 2021, pp. 6943–6951.

[18] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180.

[19] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8342–8360.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., Iclr 1 (2022) 3.

[21] C. P. Carrino, J. Armengol-Estapé, O. d. G. Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models, arXiv preprint arXiv:2109.07765 (2021).

[22] L. Dionis, G. Alvaro, M. Dylan, B. Daniel, SpanishMedicaLLM, https://huggingface.co/datasets/somosnlp/SMC, 2024. Dataset on Hugging Face.