Procesamiento del Lenguaje Natural, Revista nº 77, septiembre de 2026, pp 605–617 recibido 30-06-2026 revisado 16-07-2026 aceptado 25-07-2026

Overview of MIRROR at IberLEF 2026: Motivational Interviewing Response Rating via synthetic cOnversational tuRns

Resumen de la Tarea MIRROR en IberLEF 2026: Evaluación de Respuestas en Entrevistas Motivacionales Mediante Turnos Conversacionales Sintéticos

Luis J. Arellano,[1] **Carlos Olachea,**[1] **John D. Piette,**[2] **Hugo Jair Escalante,**[1,3] **Luis Villaseñor-Pineda,**[1] **Manuel Montes-y-Gómez,**[1] **Delia Irazú Hernández-Farías**[1]
[1]Laboratorio de Tecnologías del Lenguaje, INAOE
[2]School of Public Health,
University of Michigan
[3]Department of Computer Science,
University of Texas at El Paso
{arellano.luis, ca_olachea, hugojair, villasen, mmontesg, dirazuhf}@inaoep.mx
jpiette@med.umich.edu

Abstract: This paper presents the overview of the MIRROR shared task, organized in the framework of IberLEF 2026. The aim of the task is to promote research on Generative Artificial Intelligence methods for producing synthetic conversation turns that can improve the performance of models trained to recognize Motivational Interviewing *Behavior Codes*. We formulated the task as data-centric, asking participants to generate datasets for fine-tuning pre-trained models for distinguishing *Behavior Codes* as follows: (1) *Simple Reflection vs. Complex Reflection*; (2) *Open Question vs. Closed Question*; and (3) *Persuasion vs. Giving Information*. A total of six teams submitted their generated datasets for the final evaluation. Two teams outperformed the baseline, and novel sample generation strategies were proposed in the context of this competition. An extensive analysis of the results is presented, proving the potential impact of Generative Artificial Intelligence strategies for robust synthetic data augmentation in the context of Motivational Interviewing analysis.
Keywords: Motivational interviewing analysis, generation of synthetic conversational turns, behavior code classification, data-centric competition.

Resumen: Este artículo presenta el resumen de la tarea MIRROR organizada en el marco de IberLEF 2026. El objetivo de la tarea es promover la investigación acerca de métodos de Inteligencia Artificial Generativa para la producción de turnos conversacionales sintéticos que permitan mejorar el rendimiento de modelos entrenados para reconocer *Códigos de Comportamiento* en el contexto de Entrevistas Motivacionales. La tarea se formuló desde una perspectiva centrada en los datos, solicitando a los participantes la generación de conjuntos de datos destinados al ajuste fino de modelos preentrenados para distinguir entre las siguientes categorías de *Códigos de Comportamiento*: (1) *Reflexión Simple* vs. *Reflexión Compleja*; (2) *Pregunta Abierta* vs. *Pregunta Cerrada*; y (3) *Persuasión* vs. *Proporcionar Información*. Un total de seis equipos enviaron los conjuntos de datos generados para la fase final de evaluación. Dos de ellos superaron el rendimiento de la línea base, proponiendo estrategias novedosas en el contexto de esta competencia. Se presenta además un análisis exhaustivo de los resultados, que demuestra el impacto potencial de las estrategias basadas en Inteligencia Artificial Generativa para la ampliación sintética de datos en el ámbito del análisis de Entrevistas Motivacionales.
Palabras clave: Análisis de entrevistas motivacionales, generación de turnos conversacionales sintéticos, clasificación de códigos de comportamiento, competencia centrada en datos.

ISSN 1135-5948 DOI 10.26342-2026-77-42 © 2026 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Motivational Interviewing (MI), a client-centered counseling approach, has become a powerful and widely adopted clinical approach to promote behavior change among patients with self-management or lifestyle behavior challenges including increasing physical activity, moderating alcohol consumption or overcoming an addiction to nicotine. Unlike more directive approaches, MI emphasizes a collaborative, empathetic conversational style that respects the client's autonomy and helps resolve ambivalence. Although MI has proven to be effective, assessing the MI skills of clinicians is a challenging and resource-intensive task. These skills are traditionally evaluated using the *Motivational Interviewing Treatment Integrity (MITI* [1]*)* rubric (Moyers et al., 2024), which consists of two main components: *global scores* (e.g., Cultivating Change Talk, Softening Sustain Talk, Partnership, and Empathy) assigned on a five-point Likert scale to capture an overall impression of the session, and *behavior counts* (e.g., Simple and Complex Reflections, Open and Closed Questions, Giving Information, and Persuasion) that tally specific techniques at the utterance level to determine adherence to MI-consistent practices.

Collecting real clinical conversations is hindered by privacy and confidentiality constraints. Furthermore, manually using live MI experts to annotate these conversations with *Behavior Codes* (BCs) is costly, time-consuming, and requires specialized expertise. Consequently, building reliable models for the automatic assessment of clinician communication skills using MITI criteria is extremely challenging due to the scarcity of high-quality data representing realistic clinician-patient conversations.

It is in this context that we introduce the ***M**otivational **I**nterviewing **R**esponse **R**ating via synthetic c**O**nversational tu**R**ns* (MIRROR) shared-task. MIRROR was part of the IberLEF2026 program (Bonet-Jover, González-Barba, and Chiruzzo, 2026), and it aims to evaluate the capacity of Generative Artificial Intelligence (GenAI) methods to create useful synthetic conversational data for training models assessing MI. The ultimate goal is to substantially improve the performance of models trained to recognize BCs in the context of MI for Spanish.

Unlike conventional shared tasks, where training and test data are fixed, and participants compete by designing better models, MIRROR follows a data-centric formulation: model, training routine, and hidden evaluation data are controlled by the organizers, while participants only control the data used to adapt the model. Accordingly, MIRROR participants are expected to produce diverse high-quality datasets representing a range of clinical conversation topics to enhance the performance of a frozen baseline model used for BCs classification. This design aligns with recent data-centric AI benchmarks, where progress is measured through the utility, comparability, and reproducibility of submitted data artifacts, rather than through changes in the model architecture (Mazumder et al., 2023).

The remainder of this paper is organized as follows. The next section provides a review of related work on automated methods for analyzing MIs. Section 3 describes the shared task in detail. Section 4 provides a summary of the data submitted by the participating teams, the official results, and a descriptive analysis of the corpora generated. Finally, our conclusions and future work directions are presented in Section 5.

2 Related work

Computational methods are increasingly used to support the automatic coding and analysis of Motivational Interviewing transcripts. Early deep learning approaches, such as BERT-based models, have shown that clinician and client behaviors can be classified with good accuracy in sensitive settings such as chat-based suicide prevention helplines. This suggests that automated behavioral coding can provide useful, objective feedback for evaluating counseling interventions (Pellemans et al., 2024).

More recently, Large Language Models (LLMs) have expanded this line of work. Systems such as AutoMISC can automatically parse MI conversations and classify counselor and client behaviors using coding frameworks like MISC 2.5, reaching up to 76% accuracy in fine-grained categories and 87% accuracy when predicting session-level counseling quality (Ali et al., 2025). LLMs have also

[1] `https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf`

shown strong potential for judging empathic communication (Kumar et al., 2026). However, broader clinical evaluations show that current models do not yet match expert counselors, especially for complex skills such as deep empathy, nuanced reasoning, and therapeutic judgment (Nguyen et al., 2024). Related approaches align LLMs with MI strategies, either by making the model predict the strategy before generating a response (Sun et al., 2024), by learning strategies from a few expert examples (Xie et al., 2024), or by adapting LLMs to other therapeutic frameworks such as Cognitive Behavioral Therapy (Kim et al., 2025).

Another active research direction is the use of LLMs to generate synthetic counseling dialogues, where simulated clients and counselor agents create therapeutic conversations that can later be used to fine-tune dialogue models (Qiu and Lan, 2024). An approach exploring the use of multi-agent systems that control MI strategies during the conversation and evaluate the generated dialogue using both general language metrics and MI-specific measures was proposed by (Meng et al., 2026). Synthetic MI agents can also include explicit strategy constraints, client-state tracking, and motivation-focused exploration to encourage change talk (Yang et al., 2025). Fully generative MI chatbots have shown high MI fidelity, increased client confidence to change, and good perceived empathy (Mahmood et al., 2025).

Despite all of this progress, the computational analysis of MI remains constrained by a persistent scarcity of suitable data, largely owing to the confidentiality and ethical restrictions surrounding real therapy recordings. Prior efforts to alleviate this bottleneck came from (Pérez-Rosas et al., 2016), who built the Motivational Interviewing Dataset, comprising 22,719 clinician utterances annotated with ten MITI BCs. The same group later assembled a complementary resource of 259 web-sourced MI conversations annotated for micro-skills such as questions and reflections, and used it to characterize what distinguishes effective counseling (Pérez-Rosas et al., 2019). In another recent effort, AnnoMI (Wu et al., 2022) released 133 expert-annotated, publicly available MI demonstrations with utterance-level BCs, which has since become a common reference corpus for the task. Valuable as these resources are, they share important limitations: most are sourced from demonstration or role-play material rather than authentic clinical encounters, they remain modest in scale, and, critically for our purposes, they are exclusively in English. The sole existing Spanish resource, MIDAS (Gunal et al., 2025), comprises only 74 web-sourced conversations annotated for just two micro-skills (questions and reflections). While an important first step, its modest size and narrow annotation scope leave the broader space of MI behavior codes largely uncovered for Spanish. This gap motivates MIRROR, which explores synthetic data generation as a means of building richer MI BC resources for Spanish without compromising client privacy.

3 Task description

Participants are required to generate training data for three binary classification tasks, each one aiming to distinguish between two semantically similar BCs. The training data comprises conversation turns between the client (i.e., a patient) and clinician. Submissions are used for fine-tuning pretrained models approaching the three classification tasks. The average performance of the fine-tuned models is used as the evaluation metric.

It is worth mentioning that a successful submission is not merely a fluent synthetic dialogue. It must be a compact training set that transfers to unseen clinician-client interactions, preserves the distinctions defined by the MITI rubric, and avoids collapsing into repetitive templates. This is especially relevant for LLM-based data augmentation, where generated examples must be evaluated not only by surface quality, but also by downstream usefulness, label consistency, diversity, and domain fidelity (Tan et al., 2024; Ding et al., 2024). The remainder of this section describes in detail the MIRROR task.

3.1 Data generation setting

The basic unit of the task is a *conversation volley or turn*: a short segment that begins with a client's utterance and ends with the clinician's complete response (examples are provided in Section 3.3 and Table 3). Each instance is represented by three fields: (i) the client utterance, (ii) the clinician response, and (iii) the target BC. Participants were asked to generate three independent CSV files, one per subtask, each containing be-

tween 100 and 200 labeled volleys. Each turn was expected to contain no more than 100 words, including both speakers. The generated data had to follow the MITI definitions and was encouraged to include varied patient backgrounds, health-related topics, and types of professionals. Although the target application is clinical interviewing, participants were allowed to generate examples from different but related scenarios if they believed these examples could improve transfer to the hidden evaluation data.

3.2 Reference models and data

Reference data, withheld from participants, were used to construct the evaluation models. The reference dataset comprised 5780 annotated conversation turns derived from both genuine and synthetic MI-related sources, subsequently translated into Spanish for the challenge. Text inputs were represented by concatenating client and clinician turns according to the following template:
`Client: {client_utterance} [SEP] Clinician: {clinician_utterance}`.

To ensure a rigorous assessment of model generalization, the data partitioning procedure departed from a conventional random split and adopted an AnnoMI-only holdout strategy. Specifically, 20% of the genuine samples were reserved as the final ground-truth test partition and were never exposed during base-model training. These samples were drawn exclusively from the AnnoMI dataset (Wu et al., 2022). The remaining 80% of genuine samples, together with all synthetic instances formed the training set used for model development.

The evaluation base model was initialized from a BETO-family Spanish Transformer checkpoint[2] and further trained on the training set using 5-fold stratified cross-validation. Within each fold, 80% of the training set was used for training and the remaining 20% for validation. Training was conducted for up to 5 epochs with a batch size of 16, using the AdamW optimizer, a learning rate of 2×10^{-5}, and a 10% linear warmup schedule. Early stopping was applied with a patience of 2 epochs, using validation loss as the monitoring criterion. The model iteration achieving the highest validation Macro F1 score was selected as the definitive base evaluation model.

During the competition, participants had access only to the task definition, label descriptions, formatting constraints, and development-phase platform feedback. This feedback enabled trial assessments of submitted synthetic datasets during the development phase, without revealing the held-out test labels or the final evaluation data. Such separation operationalized the data-centric nature of the task: participants optimized the utility of synthetic data under restricted visibility, whereas the platform quantified whether the submitted data improved a fixed classifier on unseen examples.

For the final shared-task evaluation, the augmented synthetic datasets submitted by participants were used to adapt the pretrained base classifier via Low-Rank Adaptation (LoRA) (Hu et al., 2021), thereby avoiding full model fine-tuning. LoRA was configured for sequence classification with rank $r = 8$, $\alpha = 32$, and a dropout probability of 0.1. Tokenization used truncation with a maximum sequence length of 128 tokens and dynamic padding. The classifier was adapted for 5 epochs with a per-device batch size of 16, a learning rate of 2×10^{-4}, and a weight decay of 0.01. All hyperparameters, tokenization settings, training epochs, and evaluation code were fixed across submissions. Consequently, leaderboard differences reflect the relative utility of the submitted datasets under identical model, adaptation, and evaluation conditions. The definitive shared-task performance was computed exclusively on the sequestered test partition.

3.3 BC Classification Subtasks

As previously mentioned, the MIRROR shared task involves generating synthetic datasets for three specific pairs of BCs. Each subtask is formulated as a binary classification problem. To illustrate the behavior codes, the following examples were extracted directly from the ground-truth evaluation partition (AnnoMI translation) used in the shared task:

- **Subtask 1 - Reflections**: *Simple Reflection (SR) vs. Complex Reflection (CR)*. SR convey understanding by repeating or rephrasing the client's statement without adding new meaning. CR add substantial meaning, emphasis, affect, or interpretation.

[2] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased`

Example SR:
Client: "*Mmm, sí. Esto no es lo que quiero para mi vida.*" ("*Hmm, yeah. This isn't what I want for my life.*)
Clinician: "*Esto no es lo que habías imaginado*" ("*This isn't what you imagined*")

Example CR:
Client: "*Sí. Bebo unas cuatro tazas al día.*" ("*Yes. I drink about four cups a day.*")
Clinician: "*De acuerdo. ¿Y eso te parece bien?*" ("*Okay. And you're fine with that?*")

- **Subtask 2 - Questions**: *Open Question (OQ) vs. Closed Question (CQ).* OQ invite elaboration, opinion, or emotional expression, whereas CQ can be answered with a simple "*yes*", "*no*", or similarly brief factual response.

  Example OQ:
  Client: "*De acuerdo.*" ("*Okay.*)
  Clinician: "*¿Pero cuáles son tus objetivos a largo plazo?*" ("*But, What are your long-term goals?*")

  Example CQ:
  Client: "*En realidad no recuerdo mucho de esa noche.*" ("*Indeed, I don't really remember much about that night.*")
  Clinician: "*Ah, ¿eso fue un apagón?*" ("*Oh, was that a power outage?*")

- **Subtask 3 - Information**: *Persuasion (PE) vs. Giving Information (GI).* GI provides neutral education, feedback, or professional opinion without pressure. PE is an MI-inconsistent behavior that attempts to change the client's opinion or behavior through logic, authority, or advice without emphasizing autonomy.

  Example GI:
  Client: "*Me gusta caminar.*" ("*I like to walk.*")
  Clinician: "*Yo también. Caminar es un gran ejercicio.*" ("*Me too. Walking is great exercise.*")
  Example PE:
  Client: "*Es decir, hasta ahora todo ha estado bien. No me ha causado ningún problema.*" ("*In other words, everything has been fine so far. It hasn't caused me any problems.*")
  Clinician: "*Bueno, solo ha pasado un mes.*" ("*Well, it's only been a month.*")

We chose these three subtasks because each pair of behaviors represents linguistically and semantically similar utterances. For example, *Giving Information* is a common and expected clinical role, but if that information takes on an unduly authoritative or argumentative tone, it becomes *Persuasion*. Similarly, both *Complex* and *Simple reflections* have the goal of communicating understanding and reflecting back patients' utterance, but the two types of reflections are differentiated by somewhat subtle differences in tone, framing, and emphasis.

3.4 Phases and Evaluation

The evaluation was conducted through a dedicated challenge platform[3] and was divided into two phases:

- **Development**. Participants uploaded their generated datasets and received feedback from the platform. For each subtask, the submitted data was used to fine-tune the fixed classifier, and the resulting model was evaluated on a hidden validation split.
- **Final.** Participants selected their final generated datasets. These submissions were evaluated on a held-out test split not used for development feedback.

Teams were allowed up to 50 attempts during the development phase, and a maximum of 5 submissions was allowed in the final phase. For each valid submission and each subtask, the platform trained independent classifiers using the participant-generated examples for that subtask. At each evaluation run, the platform concatenated the client and clinician fields to form the classifier's input, while the submitted BC was used as the supervised label. In order to obtain stable performance estimates, models were fined tuned five times for each submission, and average performance was used. The leading evaluation measure was macro-averaged F1 across the three subtasks.

4 Overview of Participants' Submissions

4.1 Systems descriptions

A total of six teams submitted results for the final phase of the MIRROR shared task. However, only five of them provided detailed descriptions of their proposed systems. Table 1 presents a summary of the submitted solutions.

UNISON-GOATS (Robles-Robles et al., 2026) developed a prompt-driven synthetic data generation pipeline; the system comprises an instruction-tuned Mixture-

[3] https://www.mirror-iberlef.lat/

Rank	Team	LLM	Language	Pair	Conditioning	Filtering	Elements of novelty
1	UNISON-GOATS	Gemma 4 26B-A4B-it	SP	No	Demographics, topic	Yes	Code-contrastive prompts with a diversity matrix to encourage thematic diversity.
3	UMUTeam	Qwen3-4B/8B (LoRA)	EN	No	Medical topic	Yes	Continual fine-tuning with medical corpora and AnnoMI.
4	UCDICSNLP TEAM	ChatGPT API	EN	No	Manual (AnnoMI)	Yes	Human-in-the-loop generation and selection.
5	STEAM UAZ	GLM-5.1	SP	Yes	–	No	Agent-based framework simulating therapist and client interactions.
6	ROMAN LAB	Gemma-2B, Qwen2.5-3B, Mistral	EN/SP	Yes	Interaction style	Yes	Multilingual generation via translation and style conditioning.

Table 1: Summary of the synthetic data generation strategies employed by participating teams. **Language** indicates the language in which conversation volleys were generated, either SP for Spanish and EN for English. **Pair** indicates whether two clinician utterances were generated by fixing a single client utterance (client-reuse). **Conditioning** refers to the use of specific features conditioned in the prompt. **Filtering** indicates whether a post-generation filtering step was performed.

of-Experts language model developed entirely through prompt engineering, combining three complementary strategies: a stratified diversity matrix spanning demographic and clinical axes to counter distribution collapse and repetitive phrasing; code-contrastive few-shot prompting that supplies both the target code definition and its paired contrast code to force generation of boundary-clear examples; and a two-stage curation pipeline combining a format-based hard filter for question-type turns with semantic deduplication.

STEAM UAZ (Gámez-Medina, 2026) presented a multi-agent synthetic data pipeline using a two-agent simulator in which one agent plays a resistant client (adopting an informal Mexican sociolect and displaying denial, cognitive dissonance, or fatigue) while the other plays a clinician constrained to respond within a single MITI behavior category, with categorical constraints enforced at the level of syntax, punctuation, and discourse function.

Roman Lab (Bilianos and Florou, 2026) presented a style-aware synthetic data generation approach that relies on instruction prompting with small-scale, low-resource language models, generating each instance as a client utterance paired with two clinician responses corresponding to the two target labels of a given subtask, with prompts conditioned on randomized combinations of conversational topic, speaker emotion, persona, and interactional style. A set of multilingual experiments is reported.

UMUTeam (Pan et al., 2026) presented a data-centric synthetic generation pipeline using the AnnoMI dataset as seed data, reformatting it into the three target MIRROR binary pairs. They fine-tuned Qwen3-4B and Qwen3-8B with LoRA on a label-conditioned format in which each example is wrapped with an explicit natural-language definition of its behavior code. Synthetic examples are generated by sampling a health-related topic and a target code, followed by automatic filtering of malformed, generic, duplicated, or label-inconsistent outputs.

UCDICSNLPTEAM (Wang and Younus, 2026) presented a human-AI few-shot generation pipeline. The approach uses the AnnoMI dataset as a source of seed conversations, with one author manually selecting high-quality transcript segments for each target behavior category, and applies few-shot prompting through the ChatGPT API to generate labeled synthetic turns for the three MIRROR pairs. Its distinctive element is the incorporation of *human-in-the-loop* supervision at two stages—prompt design and final output review—intended to remove noisy or inappropriate clinician responses and ensure ethical alignment.

Discussion Two trends were common across participating approaches: the use of post-generation filtering to remove noisy or malformed turns, and the conditioning of turn generation on contextual variables,

Rank	Team	Information	Questions	Reflections	Avg.
1	UNISON-GOATS	0.7728 ± 0.0066	0.8559 ± 0.0042	0.5811 ± 0.0175	**0.7366** ± 0.0064
2	OpenCodice	0.7466 ± 0.0183	**0.8663** ± 0.0045	0.5942 ± 0.0140	0.7357 ± 0.0061
3	UMUTeam	0.7593 ± 0.0039	0.8464 ± 0.0027	0.5875 ± 0.0159	0.7311 ± 0.0054
4	UCDICSNLP	0.7603 ± 0.0112	0.8438 ± 0.0047	0.5869 ± 0.0153	0.7303 ± 0.0062
5	STEAM UAZ	**0.7749** ± 0.0055	0.8382 ± 0.0052	0.5667 ± 0.0131	0.7266 ± 0.0056
-	Baseline	0.7491 ± 0.0122	0.8535 ± 0.0051	0.5685 ± 0.0128	0.7237 ± 0.0046
6	ROMAN LAB	0.7369 ± 0.0135	0.8311 ± 0.0066	**0.5948** ± 0.0168	0.7209 ± 0.0068

Table 2: Macro F1 scores and average performance by team (5 runs were performed for each task and team). Best scores per column are highlighted in bold.

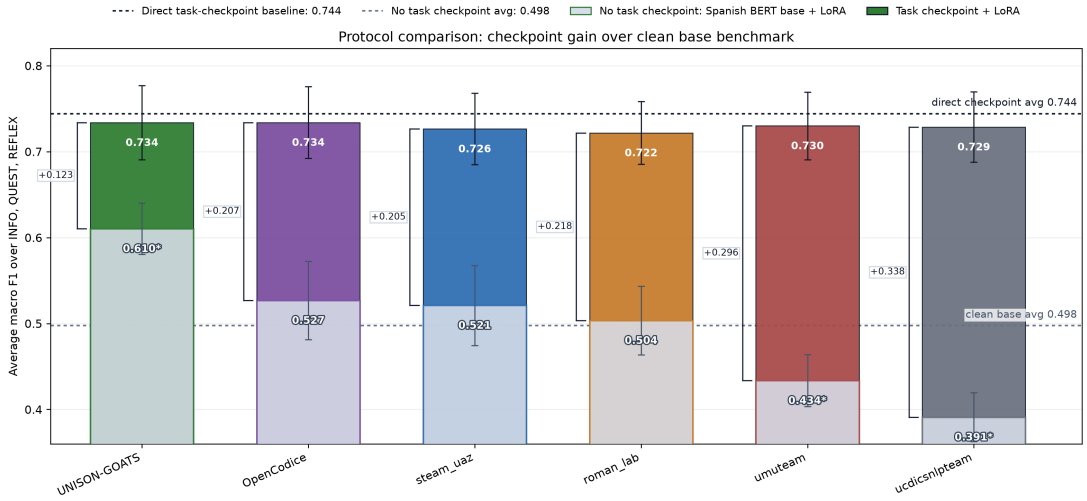


Figure 1: Comparison of participant datasets under two evaluation protocols. Light-colored bars show average macro-F1 when LoRA adaptation starts from a model without a task-specific checkpoint. Dark-colored bars show average macro-F1 when adaptation starts from a checkpoint of a model already pretrained with data of the corresponding task (i.e., the shared task scenario).

though teams differed in the information they conditioned on (see column **Conditioning** in Table 1). Generation was carried out in both Spanish and English, and in some cases, turns were generated in English and subsequently translated into Spanish. Two teams relied on AnnoMI samples as seeds for their methods. The Gemma and Qwen families were the most widely used LLMs, in all cases in their smaller variants. Teams' approaches differed in how they constructed paired examples for the tasks. OpenCodice, STEAM UAZ, and Roman Lab frequently reused the same client utterance with two different clinician responses (see column 5 in Table 1). This strategy can make label contrasts explicit and may help the classifier to detect response-level differences. Overall, we consider that the challenge fostered the development of interesting methodologies that should prove valuable to researchers and practitioners working on the generation of synthetic conversational data.

4.2 Evaluation campaign results

Table 2 summarizes the official results. UNISON-GOATS team obtained the best overall score with an average macro-F1 of 0.7366. However, the differences with the other teams were rather small, the relative difference between participants ranked 1st and 6th is of ≈ 2%. STEAM UAZ achieved the highest score (0.7749) for the *Information* BC, and OpenCodice achieved the highest score (0.8663) on *Questions*, while Roman Lab was the best performer in *Reflections*. Interestingly, the top ranked team did not obtained the best performance in any of the individual tasks but it achieved regular performance across all of them.

With the aim to better appreciate the differences in performance among participants, we performed another evaluation contrasting the scenario adopted in Table 2 (i.e., fine-tuning a BETO model pretrained with BC samples) with a scenario where the model was not pretrained with BC samples. The goal

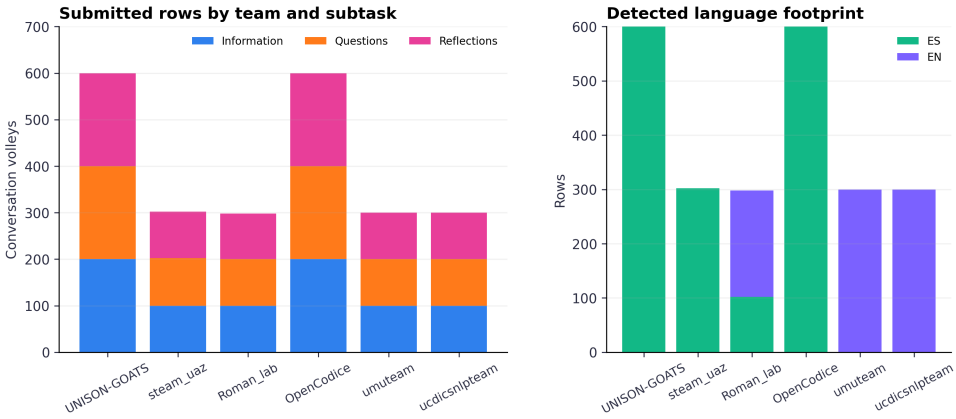


Figure 2: Landscape of submitted synthetic data. The left plot reports the number of samples submitted by the teams. The right plot shows the detected language by team.

was to determine to what extent the high-quality samples used in the former scenario, *disguise* the actual usefulness of data from participants. Figure 1 shows the results of this experiment.

Dark-colored bars are consistent with the results in Table 2. However, performance differences increase drastically when the base model did not use BC samples (light-colored bars). These results provide a more reliable estimation on the quality of submissions: marginal improvements in the original shared task setting hide the actual usefulness of datasets.

4.3 Analysis

Figure 2 shows global statistics of the submissions from all of the participants. The left plot in Figure 2 shows the number of samples that each team submitted. UNISON-GOATS and OpenCodice each submitted 600 interaction turns (200 per BC task), they both achieved the two best results in the final evaluation. However, the remaining teams submitted roughly half as many samples and were close in performance. In particular, STEAM UAZ and Roman Lab achieved very similar performance to OpenCodice in Figure 1, but using less samples. Therefore, compact datasets with clear label contrast and useful variation can be more valuable than larger ones. The right plot in Figure 2 shows the language detected in the submissions for each of the teams. As expected, better performance was obtained by participants that provided samples in Spanish. While the performance of submissions in English is somewhat competitive for the official results (Table 2), the evaluation from Figure 1 (light bars) shows that teams that submitted samples in English had a considerable decrement in performance for the no task checkpoint setting. Therefore suggesting that the language is critical for obtaining competitive performance.

A basic lexical analysis revealed that, across teams, *Information* samples had the most dense clinician language (mean lexical density 0.48), which is consistent with their explanatory biomedical content. *Questions* were the shortest clinician turns (15.70 tokens on average), but, as expected, *open questions* were longer and syntactically deeper than *closed questions* (18.10 vs. 13.19 tokens. *Reflections* also showed the expected gradient: *complex reflections* were longer and more syntactically elaborate than *simple reflections* (28.73 vs. 19.34 tokens; tree depth 6.00 vs. 4.57).

Figure 3, shows word clouds corresponding to synthetic data generated by UNISON-GOATS for the clinician utterance for each BC. In the case of *Simple Reflection*, the most salient word is "*Entonces*" ("*Then*"), which is often used as a discourse marker for denoting an outcome connector, while "*Parece*" ("*It seems*") is associated with expressing an opinion or judgment. In the case of *Questions*, the word "*Cómo*" ("*How*") is the most frequent one for *open questions*, while terms like "*último*"/"*última*" ("*last*") and "*día*" ("*day*") serve to express a direct and most focused question. Finally, terms like "*Debe*" ("*must*"), "*salud*" ("*health*"), and "*si*" ("*if*") emerge as common in *Persua-*



Figure 3: Word clouds (frequency) for the submission of UNISON-GOATS. Clinician utterances were considered for all of the BC categories.

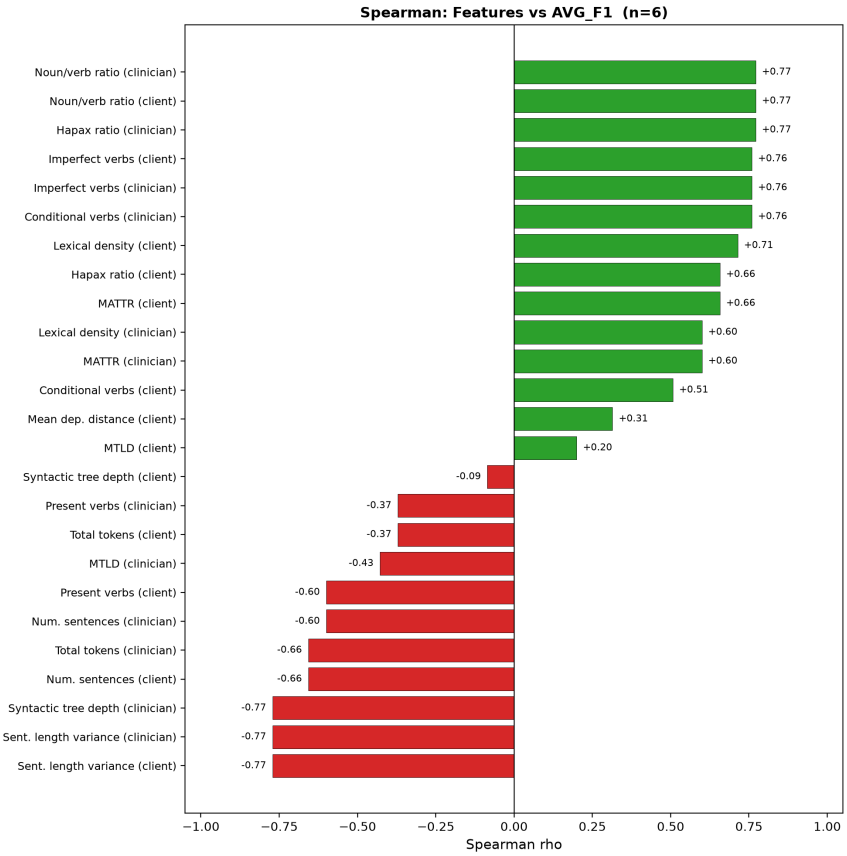


Figure 4: Spearman correlations between extracted linguistic features and average macro-F1 across the six participant submissions. Green bars indicate positive correlations with performance, whereas red bars indicate negative correlations. MATTR stands for Moving Average Type Token Ratio and MTLD for Measure of Textual Lexical Diversity.

sion, while more objective words like "*puede*" ("*can/may*"), "*influir*" ("*influence*"), "*niveles*" ("*levels*"), and "*sistema*" ("*system*") appear in the *Given Information* category.

Table 3 presents representative examples of generated texts for each BC category. For *Information* and *Persuasion*, quality depends on whether the utterance stays neutral or becomes directive. For *Questions*, quality depends on whether the response genuinely opens the patient's reasoning or restricts the answer to a factual choice. For *Reflections*, the most informative contrast is between a literal restatement and an inferred meaning that captures emotion, ambivalence, or motivation.

Figure 4 reports an exploratory correlation analysis between speaker-specific linguistic features and the average macro-F1 obtained by each submission distinguishing between the clinician and client utterances. Since the analysis includes only six teams, the correlations should be interpreted descriptively rather than as statistically conclusive evidence. The strongest positive asso-

BC	Client excerpt	Clinician excerpt
GI	A veces dudo si la dosis de medicación que estoy tomando es la adecuada.	La dosificación de los anticoagulantes se ajusta habitualmente basándose en pruebas periódicas de coagulación sanguínea.
	Sometimes I wonder if the dose of medication I'm using is the proper one.	*The dosage of anticoagulants is typically adjusted based on periodic blood coagulation tests.*
PE	He dejado de tomar la medicación estas últimas semanas porque me siento con energía.	Es necesario que retome su tratamiento de inmediato para evitar complicaciones mayores.
	I've stopped taking my medication for the last few weeks, because I feel more energetic.	*It is essential that you return to your treatment immediately to avoid further complications.*
OQ	Mi glucosa salió en 180 pero yo me siento perfectamente bien.	¿Qué has notado al comparar cómo te sientes con lo que dicen los estudios?
	My blood sugar reading was 180, but I feel perfectly fine.	*What have you observed when comparing how you feel with study findings?*
CQ	Mi glucosa salió en 180 pero yo me siento perfectamente bien.	¿Te tomaste la glucosa en ayunas o después de comer?
	My blood sugar reading was 180, but I feel perfectly fine.	*Did you take the glucose test on an empty stomach or after eating?*
SR	Últimamente he estado tomando la medicación casi todos los días y siento que por fin estoy agarrando el ritmo.	Has estado tomando la medicación casi todos los días y sientes que por fin estás agarrando el ritmo.
	Lately, I've been taking my medication almost every day, and I feel like I'm finally getting into a steady rhythm.	*You've been taking your medication almost every day, and you feel like you're finally getting into a steady rhythm.*
CR	Como enfermera, sé que mi presión alta no es algo menor. Empecé a caminar y a bajar la sal.	Aunque entiendes muy bien los riesgos de la hipertensión por tu trabajo, te sientes esperanzada porque ya estás haciendo cambios concretos.
	As a nurse, I understand that my high blood pressure is no minor issue. I started walking and cutting back on salty foods.	*Although you understand the risks of high blood pressure because of your occupation, you feel hopeful because you're already making concrete changes.*

Table 3: Representative examples selected from participant submissions. Examples were chosen because they are in Spanish, remain concise, match the target behavior-code definition, and make the intended label contrast clear.

ciations correspond to noun–verb ratios, hapax ratios, imperfect and conditional verb usage, and lexical-density measures. In contrast, longer or more structurally variable turns, represented by features such as token counts, number of sentences, sentence-length variance, and tree depth, tend to correlate negatively with performance. This pattern supports the qualitative observation that useful MIRROR examples tend to be concise and lexically varied rather than simply longer or syntactically more complex.

5 Conclusions and Future Work

This paper describes an overview of MIRROR, a data-centric shared task focused on generating synthetic conversational turns for improving behavior-code classification in Motivational Interviewing. The evaluation campaign showed that synthetic data can be useful, but its impact depends strongly on the generation strategy and on the evaluation protocol. The best-performing submissions emphasized clear label contrasts, Spanish linguistic alignment, compact examples, and post-generation filtering. The post-evaluation analysis also showed that a task-specific checkpoint can compress the differences among submissions, whereas a clean-base protocol makes synthetic-data quality easier to distinguish.

Future work should expand the evaluation beyond aggregate macro-F1 by incorporating more systematic qualitative and clinical validation of generated turns. In particular, expert review could help assess rubric fidelity, conversational plausibility, and potentially harmful or directive responses that automatic metrics may miss. Further editions of MIRROR should also consider larger participant pools, additional MITI behavior codes, multi-turn context, and more stable evaluation protocols for measuring when synthetic data genuinely improves generalization rather than merely adapting to a strong pretrained checkpoint.

Acknowledgements

We thank SECIHTI-Mexico for partially supporting this work under scholarships 4057199, and 4026056.

References

Ali, S., J. Zhu, A. Guo, X. N. Ye, Q. Gu, J. Wolff, C. Cooper, O. C. Melamed, P. Selby, and J. Rose. 2025. Automated Coding of Counsellor and Client Behaviours in Motivational Interviewing Transcripts: Validation and Application. In A. Zechariah, B. Krishna S, D. Misra Sharma, H. Mary Thomas, J. Mammen, P. Krishnamurthy, and V. Mujadia, editors, *NLP-AI4Health*, pages 25–54, Mumbai, India, December. Association for Computational Linguistics.

Bilianos, D. and K. Florou. 2026. Style-Aware Synthetic Data Generation for Motivational Interviewing. 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*, CEUR Workshop Proceedings. CEUR-WS.org.

Bonet-Jover, A., J. Á. González-Barba, and L. Chiruzzo. 2026. 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*.

Ding, B., C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, and S. Joty. 2024. Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand, August. Association for Computational Linguistics.

Gámez-Medina, J. M. 2026. Synthetic Motivational Interviewing Dialogues in Mexican Spanish for Response Generation and MITI Code Classification at MIRROR@ IberLEF 2026. In *Personal communication*.

Gunal, A. E., B. Yi, J. D. Piette, R. Mihalcea, and V. Perez-Rosas. 2025. Examining Spanish Counseling with MIDAS: a Motivational Interviewing Dataset in Spanish. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 866–872, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv.org*. https://arxiv.org/abs/2106.09685.

Kim, Y., C.-H. Choi, S. Cho, J. yong Sohn, and B.-H. Kim. 2025. Aligning large language models for cognitive behavioral therapy: a proof-of-concept study. *Frontiers in Psychiatry*, Volume 16 - 2025.

Kumar, A., N. Poungpeth, D. Yang, E. Farrell, B. L. Lambert, and M. Groh. 2026. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8(2):173–185, Feb.

Mahmood, Z., S. Ali, J. Zhu, M. Abdelwahab, M. Y. Collins, S. Chen, Y. C. Zhao, J. Wolff, O. Melamed, N. Minian, M. Maslej, C. Cooper, M. Ratto, P. Selby, and J. Rose. 2025. 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, jul. Association for Computational Linguistics.

Mazumder, M., C. Banbury, X. Yao, B. Karlaš, W. G. Rojas, S. Diamos, G. Diamos, L. He, A. Parrish, H. R. Kirk, J. Quaye, C. Rastogi, D. Kiela, D. Jurado, D. Kanter, R. Mosquera, J. Ciro, L. Aroyo, B. Acun, L. Chen, M. S. Raje, M. Bartolo, S. Eyuboglu, A. Ghorbani, E. Goodman, O. Inel, T. Kane, C. R. Kirkpatrick, T.-S. Kuo, J. Mueller, T. Thrush, J. Vanschoren, M. Warren, A. Williams, S. Yeung, N. Ardalani, P. Paritosh, L. Bat-Leah, C. Zhang, J. Zou, C.-J. Wu, C. Coleman, A. Ng, P. Mattson, and V. J. Reddi. 2023. DataPerf: Benchmarks for

Data-Centric AI Development. *ArXiv.org*. https://arxiv.org/abs/2207.10062.

Meng, Q., M. Chen, D. Liu, Y. Mo, Y. Su, X. Sun, K. Hindriks, and J. Pei. 2026. StoryMI: Steerable Multi-Agent Therapeutic Dialogue Generation. *ArXiv.org*. https://openalex.org/W7162817958.

Moyers, T. B., L. N. Rowell, J. K. Manuel, D. Ernst, and J. M. Houck. 2024. The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. *Journal of substance abuse treatment*, 65:36–42.

Nguyen, V. C., M. Taher, D. Hong, V. K. Possobom, V. Gopalakrishnan, E. Raj, Z. Li, H. J. Soled, M. L. Birnbaum, S. Kumar, and M. D. Choudhury. 2024. Do Large Language Models Align with Core Mental Health Counseling Competencies? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.22446.

Pan, R., P. J. Vivancos-Vicente, J. Gómez-Navalón1, T. Bernal-Beltrán, and J. A. García-Díaz. 2026. UMUTeam at MIR-ROR@IberLEF 2026: Synthetic Motivational Interviewing Data Generation with Qwen3 and Medical Continual Pretraining. 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*, CEUR Workshop Proceedings. CEUR-WS.org.

Pellemans, M., S. Salmi, S. Mérelle, W. Janssen, and R. 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. *Journal of Medical Internet Research*, 26:e53562, August.

Pérez-Rosas, V., R. Mihalcea, K. Resnicow, S. Singh, and L. An. 2016. Building a Motivational Interviewing Dataset. In K. Hollingshead and L. Ungar, editors, *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA, June. Association for Computational Linguistics.

Pérez-Rosas, V., X. Wu, K. Resnicow, and R. Mihalcea. 2019. What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy, July. Association for Computational Linguistics.

Qiu, H. and Z. Lan. 2024. Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.15787.

Robles-Robles, G. S., M. A. Castro-Lerma, J. J. Ramirez-Ramirez, and J. A. Flores-Briones. 2026. Prompt-Driven Synthetic Data Generation for MITI Behavior Code Classification: A Diversity-Aware Few-Shot Approach. 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*, CEUR Workshop Proceedings. CEUR-WS.org.

Sun, X., X. Y. Tang, A. E. Ali, Z. Li, P. Ren, J. de Wit, J. Pei, and J. A. Bosch. 2024. Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.06527.

Tan, Z., D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, C. Lu, and H. Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA, November. Association for Computational Linguistics.

Wang, C. and A. Younus. 2026. A Human-AI Approach to Few-Shot Learning for Motivational Interviewing-based Counseling. 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*, CEUR Workshop Proceedings. CEUR-WS.org.

Wu, Z., S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, and D. Riboni. 2022. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Xie, Z., B. P. Majumder, M. Zhao, Y. Maeda, K. Yamada, H. Wakaki, and J. 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, August. Association for Computational Linguistics.

Yang, Y., P. Achananuparp, H. Huang, J. Jiang, K. P. Leng, N. G. Lim, C. T. S. Ern, and E. Lim. 2025. CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria, July. Association for Computational Linguistics.