Roman_lab at MIRROR@IberLEF 2026: Style-Aware Synthetic Data Generation for Motivational Interviewing

Dimitris Bilianos[1,*], Katerina Florou[1]

[1]National and Kapodistrian University of Athens, 1 Nikolaou Politi Street, University Campus, 15772 Zografou, Athens, Greece

Abstract

This paper describes the participation of the Roman_lab team in the MIRROR shared task at IberLEF 2026, which focused on behavioral coding classification for Motivational Interviewing. The shared task comprised three binary classification subtasks based on synthetic conversational turns: Simple Reflection vs. Complex Reflection (SR/CR), Open Question vs. Closed Question (OQ/CQ), and Persuasion vs. Giving Information (PE/GI). Our approach relied on synthetic dataset generation using instruction-tuned, low-resource large language models. We first conducted local validation experiments using DistilBERT classifiers trained on automatically generated datasets in order to compare prompting strategies and generation models. Initial experiments with Gemma-2B showed moderate performance, while the introduction of stylistic conditioning in prompts substantially improved local validation scores. We subsequently compared multiple instruction-tuned models, namely Gemma-2B, Qwen2.5-3B, and Ministral-3B-Instruct, using style-aware prompting. Although several models achieved competitive local results, Gemma-2B with stylistic prompting demonstrated the strongest and most consistent performance and was therefore selected for generating all final datasets. The final system incorporated synthetic data generation conditioned on conversational topic, speaker emotion, persona, and interactional style. Following disappointing leaderboard performance in the OQ/CQ subtask, we conducted additional multilingual experiments. Translating the generated English OQ/CQ dataset into Spanish produced a measurable improvement over the original English version, while fully native Spanish data generation yielded lower performance. These findings suggest that linguistic alignment between synthetic training data and hidden evaluation data may significantly affect downstream performance in low-resource Motivational Interviewing classification settings. Roman_lab achieved the highest score among all participating systems in the Reflection subtask, with an F1 score of 0.5948.

Keywords

Motivational Interviewing, Style-Aware Synthetic Data Generation, Behavioral Coding

1. Introduction

Motivational Interviewing (MI) is a collaborative method of counseling clients, viewed as a useful intervention strategy in the treatment of lifestyle problems and disease [1]. MI seeks to strengthen an individual's motivation and commitment towards behavioral change and it has been widely adopted in healthcare [2], psychotherapy [3], education [4], and behavioral intervention settings.

Central to the MI framework is the quality of interaction between counselor and client, particularly the communicative strategies used to encourage self-reflection, autonomy, and behavioral change [5].

In order to study and evaluate such interactions, researchers frequently employ behavioral coding frameworks that categorize conversational turns according to their communicative function. One of the most widely used frameworks is the Motivational Interviewing Treatment Integrity (MITI) coding scheme, which distinguishes between different clinician behaviors such as reflections, questions, persuasive statements, and informational responses. Automated behavioral coding has become an important research direction within Natural Language Processing, as it can aid the analysis of therapeutic interactions and facilitate the development of decision-support tools.

The MIRROR shared task [6] at IberLEF 2026 [7] addressed this problem by focusing on the automatic classification of conversational turns in MI settings. Participants were required to generate synthetic datasets and train models for three binary classification subtasks: Simple Reflection vs. Complex

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

dbilianos@ill.uoa.gr (D. Bilianos); kathyflorou@ill.uoa.gr (K. Florou)

0000-0001-8407-870X (D. Bilianos); 0009-0001-3062-5728 (K. Florou)

Reflection (SR/CR), Open Question vs. Closed Question (OQ/CQ), and Persuasion vs. Giving Information (PE/GI). System performance was evaluated on hidden test sets provided by the organizers.

A central challenge of the shared task was the strict low-resource setting. Each submitted dataset was required to contain between 100 and 200 labeled conversational turns, forcing teams to carefully balance dataset diversity, label quality, and generalization capacity under extremely limited training conditions. As a result, the task emphasized the importance of synthetic data generation strategies, prompt engineering, and efficient use of language models.

Our team approached the synthetic conversational data generation task through a methodology combining prompt engineering with controlled variation in conversational style, emotional framing, and participant personas, aiming to increase dataset diversity within the competition's strict size constraints. We additionally conducted multilingual experiments and we implemented a lightweight local evaluation pipeline based on DistilBERT classifiers trained on the generated datasets. We report the best-performing system in the Reflection (SR/CR) subtask of MIRROR@IberLEF2026.

2. Task Description

The MIRROR shared task at IberLEF 2026 focused on the automatic classification of conversational behaviors in MI interactions [6]. Participants were required to construct synthetic datasets for three binary classification subtasks derived from the MITI framework:

- **Simple Reflection vs. Complex Reflection (SR/CR):** distinguishing between reflections that primarily restate client content and reflections that infer deeper meaning or emotional subtext.
- **Open Question vs. Closed Question (OQ/CQ):** distinguishing between questions that encourage elaborated responses and questions that restrict the response space to short or predefined answers.
- **Persuasion vs. Giving Information (PE/GI):** distinguishing between persuasive interventions aimed at influencing client behavior and informational statements intended primarily to provide knowledge or clarification.

These datasets were subsequently used by the organizers to fine-tune pretrained classification models, which were then evaluated on held-out test sets.

System ranking was based on classification performance across the three subtasks using macro F1-score as the primary evaluation metric.

3. Methodology

3.1. Synthetic Dataset Generation

Our approach relied on the generation of synthetic conversational data using the instruction-tuned large language model Gemma-2-2B-it.[1] The model was selected due to its relatively small computational footprint and strong instruction-following capabilities, making it suitable for iterative experimentation under limited computational resources.

Synthetic examples were generated through prompt-based conditioning designed to simulate realistic MI interactions. Each generated instance consisted of a client utterance paired with two clinician responses corresponding to the target labels of each subtask. For example, in the Reflection task, each instance included a client statement alongside both a Simple Reflection (SR) and a Complex Reflection (CR). The same paired generation strategy was applied to the Questions (OQ/CQ) and Information (PE/GI) subtasks.

To increase conversational diversity and reduce repetitive outputs, prompts were constructed using randomized scenario seeds combining four dimensions:

[1] https://huggingface.co/google/gemma-2-2b-it

- **Topics:** e.g., smoking cessation, alcohol use, stress at work, diabetes management.
- **Emotions:** e.g., ambivalent, frustrated, anxious, hopeful.
- **Personas:** e.g., college student, retired patient, middle-aged parent.
- **Interactional styles:** e.g., warm and empathetic, collaborative and curious, direct but supportive.

The generation pipeline instructed the model to produce outputs in structured JSON format in order to facilitate automatic extraction and filtering. Each generated item contained three fields: a client statement, and two clinician responses corresponding to the paired labels of the target task. Listing 1 presents a simplified version of the prompting structure used during generation. In practice, the same template was instantiated with randomly sampled topics, emotions, personas, and interactional styles to produce diverse synthetic conversations for each subtask.

Listing 1: Simplified prompting structure used for synthetic data generation.

```
Generate a motivational interviewing example
in valid JSON format only.

Style: collaborative and curious
Topic: smoking cessation
Emotion: ambivalent
Persona: college student

{
"client_statement": "...",
"open_question": "...",
"closed_question": "..."
}
```

Following generation, outputs underwent automatic validation and filtering procedures. We employed regular-expression-based JSON extraction to recover structured outputs from the generated text and discarded malformed or incomplete generations. Additional filtering criteria removed examples containing placeholder text, excessively short utterances, or unusually long responses. Only validated examples were included in the final datasets submitted to the shared task.

3.2. Stylistic Conditioning

A central component of our methodology involved the incorporation of stylistic conditioning during synthetic data generation. In addition to controlling for topic, emotion, and persona, prompts explicitly specified an interactional style intended to shape the tone and communicative characteristics of the generated clinician responses. Examples of stylistic conditions included *warm and empathetic*, *collaborative and curious*, *gentle and reflective*, and *direct but supportive*.

The introduction of stylistic variables was motivated by the hypothesis that increasing stylistic diversity in synthetic conversational data may improve downstream generalization performance; style-conditioned prompting encouraged the language model to produce responses exhibiting variation in lexical choice, sentence structure, emotional framing, and interactional tone.

Our focus on stylistic variation was also informed by previous work in stylometry and AI-generated text analysis, as prior research has demonstrated that language models exhibit identifiable stylistic patterns and linguistic fingerprints [8, 9].

Empirically, the inclusion of stylistic conditioning substantially improved local validation performance during preliminary experiments. In the OQ/CQ task, datasets generated with style-aware prompting consistently outperformed datasets generated without stylistic variation across multiple instruction-tuned models. The effect appeared especially beneficial for the Reflection subtask (SR/CR), where distinctions between simple and complex reflections frequently depend on nuanced emotional framing and discourse structure rather than surface-level lexical cues alone.

3.3. Local Validation Setup

In order to compare generation strategies prior to submission, we implemented a lightweight local validation pipeline based on DistilBERT sequence classification models.[2] The primary objective of these experiments was not to maximize absolute classification performance, but rather to evaluate the relative quality and separability of the synthetic datasets generated under different prompting configurations and language models.

For each subtask, generated conversational pairs were converted into a single input sequence by concatenating the client utterance and clinician response. Labels were mapped to binary numerical values corresponding to the target behavioral coding categories. The resulting datasets were then divided into training and validation subsets using a stratified train/validation split in order to preserve label balance across partitions.

The classifier architecture employed `DistilBertForSequenceClassification` with two output labels. To reduce computational requirements and stabilize training under the small-data setting, the encoder parameters of DistilBERT were frozen during training, allowing only the classification head to be optimized. Training was conducted for multiple epochs using the HuggingFace `Trainer` framework.

Performance was evaluated using both accuracy and F1-score metrics. These local experiments were primarily used comparatively to determine whether specific prompting strategies, stylistic conditioning approaches, or generation models produced more discriminative synthetic datasets prior to leaderboard submission.

To evaluate the effectiveness of the generation strategies, we compared several instruction-tuned LLMs using the local DistilBERT validation pipeline on the OQ/CQ task. The evaluated generators included Gemma-2B-it, Gemma-2B-it with stylistic conditioning, Qwen2.5-3B-Instruct with stylistic conditioning, and Ministral with stylistic conditioning.

The experiments showed that stylistic conditioning substantially improved downstream classification performance. In particular, Gemma-2B-it combined with style-aware prompting achieved the highest local validation performance, improving macro F1 from 0.71 to 0.94. Based on these results, the Gemma + style configuration was selected for the final generation of all three shared-task datasets.

Table 1
Local validation results on the OQ/CQ task.

Model	Accuracy	F1
Gemma-2B	0.69	0.71
Gemma-2B + Style	0.94	0.94
Qwen2.5-3B + Style	0.88	0.86
Ministral + Style	0.88	0.88

3.4. Multilingual Experimentation

During the initial submission phase, the synthetic datasets generated in English produced competitive results for the Reflection and Information subtasks, but considerably lower performance for the Questions task, despite positive validation results obtained during local evaluation.

To investigate this discrepancy, we conducted an additional multilingual experiment focusing exclusively on the OQ/CQ dataset. Rather than generating entirely new data, the existing English synthetic dataset was automatically translated (cf. [10]) into Spanish using the `deep-translator` Python library.[3] The translated dataset preserved the original labels and conversational structure while adapting the linguistic surface form to Spanish. The translated Spanish version of the dataset yielded improved leaderboard performance in the Questions subtask. At the same time, we also experimented with direct Spanish synthetic data generation using Gemma-2B-it and Spanish-language prompts modeled after the

[2] [https://huggingface.co/distilbert/distilbert-base-uncased
[3] https://pypi.org/project/deep-translator/

organizers' examples, but this configuration achieved a slightly lower leaderboard score compared to the translated dataset. Consequently, the translated Spanish dataset was retained for the final submission of the OQ/CQ task.

Given that MIRROR was organized within the IberLEF evaluation campaign, these findings suggest that language adaptation may influence downstream performance in behavioral coding tasks.

4. Experimental Results

Table 2 presents the official evaluation results obtained by the *Roman_lab* team in the MIRROR@IberLEF2026 shared task.

Table 2
Official leaderboard results for the Roman_lab system.

Task	Macro F1
Information (PE/GI)	0.7369
Questions (OQ/CQ)	0.8311
Reflection (SR/CR)	0.5948
Average	0.7209

Our system achieved the highest performance among all submitted systems on the Reflection subtask, obtaining a macro F1 score of 0.5948.

Strong results were observed in the Information subtask, where synthetic generation strategies appeared sufficient to capture distinctions between persuasive and informational clinician responses. Similarly, the Reflection task benefited substantially from stylistic diversification and the generation of emotionally nuanced responses, leading to the best-performing system in this category.

For the Questions task, the multilingual experimentation described in Section 3.4 demonstrated that adapting the synthetic data to Spanish improved performance substantially, with a macro F1 score of 0.8311.

Local validation experiments indicated that stylistic conditioning consistently improved the separability of the generated datasets, particularly for the OQ/CQ task. However, these gains transferred unevenly to the official evaluation, where the Questions task performance remained worse compared to other participants. In contrast, the Reflection task achieved the highest leaderboard performance among all participating systems, suggesting that stylistic diversification is especially beneficial for response generation tasks where conversational tone and discourse structure play a central role.

5. Discussion

The results highlight the strengths and limitations of style-aware synthetic data generation for behavioral coding in MI. A key observation is the strong performance on the Reflection subtask, where our system achieved the highest score among all participating teams. This suggests that stylistic conditioning can be effective for tasks that rely on nuanced reformulation and emotional attunement. In such cases, variation in tone and reflection style may contribute to improved generalization in downstream classifiers trained on small datasets.

The Information subtask also showed robust performance, indicating that synthetic generation can reliably capture more explicit pragmatic distinctions, such as persuasive versus informational intent. In contrast, the Questions subtask proved substantially more difficult. The gap persisted despite strong local validation results during development, suggesting potential mismatches between local evaluation conditions and the hidden test distribution.

An additional finding concerns multilingual adaptation. The observed improvement in the Questions subtask after translating the dataset into Spanish indicates that linguistic surface form can have a

measurable impact on model performance. However, these findings are based on controlled experiments and are not easy to interpret.

6. Conclusion

This paper presented a style-aware synthetic data generation approach for behavioral coding in Motivational Interviewing at MIRROR@IberLEF2026. We made use of an instruction-tuned language model with structured prompting, scenario-based sampling, and stylistic conditioning, and we generated labeled datasets for three behavioral coding subtasks under strict data constraints.

Our results demonstrate that stylistic diversification plays an important role in improving downstream classification performance, particularly for reflection-related behaviors; Our system obtained the best performance on the Reflection subtask.

Declaration on Generative AI

During the preparation of this work, the authors used Google and OpenAI models for language editing and table generation. After using these tools, the authors reviewed, validated and edited the content accordingly and take full responsibility for the publication's content.

References

[1] S. Rubak, A. Sandbæk, T. Lauritzen, B. Christensen, Motivational interviewing: a systematic review and meta-analysis, British Journal of General Practice 55 (2005) 305.

[2] K. M. Emmons, S. Rollnick, Motivational interviewing in health care settings: opportunities and limitations, American Journal of Preventive Medicine 20 (2001) 68–74.

[3] H. A. Swartz, A. Zuckoff, N. K. Grote, H. N. Spielvogle, S. E. Bledsoe, M. K. Shear, E. Frank, Engaging depressed patients in psychotherapy: integrating techniques from motivational interviewing and ethnographic interviewing to improve treatment participation, Professional Psychology: Research and Practice 38 (2007) 430–439.

[4] L. Snape, C. Atkinson, The evidence for student-focused motivational interviewing in educational settings: a review of the literature, Advances in School Mental Health Promotion 9 (2016) 119–139.

[5] W. R. Miller, The evolution of motivational interviewing, Behavioural and Cognitive Psychotherapy 51 (2023) 616–632.

[6] L. J. Arellano, C. Olachea, J. Piette, H. J. Escalante, L. Villaseñor-Pineda, M. M. 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. Á. 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.

[8] G. K. Mikros, A. Koursaris, D. Bilianos, G. Markopoulos, Ai-writing detection using an ensemble of transformers and stylometric features, in: Proceedings of IberLEF@SEPLN, 2023.

[9] D. Bilianos, Identifying the linguistic fingerprints of generative ai: A comparative analysis of gemma-2b and mistral-7b textual outputs, 2025. URL: https://hal.science/hal-05279301/.

[10] D. Bilianos, G. K. Mikros, Sentiment analysis in cross-linguistic context: How can machine translation influence sentiment classification?, Digital Scholarship in the Humanities 38 (2023) 23–33.