NLP Innovators at MentalRiskES 2026: A Winning Reranker-Based Approach for Therapist Response Selection

Asifa Bibi[1], Tayyab Latif[1], Sabur Butt[2,*], Alexander Gelbukh[1] and Grigori Sidorov[1]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[2]Institute for the Future of Education (IFE), Tecnológico de Monterrey, Mexico

Abstract

This paper outlines the NLP Innovators system for Task 2 of the MentalRiskES 2026 shared task, where the goal is to choose the therapist response that is most relevant, supportive, and appropriate for the patient's message and the ongoing conversation. The system is given the current patient message and three candidate therapist responses at each round, and it is required to select the best response based on the context of the dialogue. We tackle the problem as a multiple-choice response selection problem and test reranker-based models that compare the dialogue context to each candidate response. We submit reranker systems: BGE (BAAI/bge-reranker-v2-m3), Qwen (Tomaarsen/Qwen3-Reranker-4B-seq-cls), mMiniLM (cross-encoder/mmarco-mMiniLMv2-L12-H384-v1). The official results reveal that our Qwen reranker performed the best in the competition, with an overall accuracy of 0.393 and a Macro F1 score of 0.392. The findings suggest that rerankers with stronger contextual understanding are well suited to therapist response selection and that majority-voting ensembles may still fall short of Qwen.

Keywords

Mental Health NLP, Response Selection, Reranking, Spanish Dialogue Systems, Therapist Response Ranking, MentalRiskES 2026

1. Introduction

The increasing prevalence of mental health concerns around the world has created a growing demand for accessible, scalable, and reliable support systems [1]. In this context, Natural Language Processing (NLP) has emerged as a promising tool for developing intelligent systems capable of assisting mental health professionals. However, applying NLP in sensitive domains such as mental health requires not only technical accuracy but also careful consideration of ethical, contextual, and safety aspects [2]. This challenge has motivated shared evaluation campaigns such as MentalRiskES 2026, organized within IberLEF 2026, which focuses on advancing computational methods for mental health risk detection and intervention in conversational settings [3, 4].

MentalRiskES 2026 provides a structured benchmark for evaluating NLP systems on real-world mental health scenarios, particularly in multilingual and multi-turn dialogue environments [3]. The shared task emphasizes the importance of context-aware modeling, where systems must go beyond isolated message understanding and instead capture the evolving dynamics of user interactions. By simulating realistic conversations between users and therapists, the task encourages the development of models that can support decision-making processes in clinical or semi-clinical settings.

MentalRiskES 2026 [5] included two tasks. Task 1, titled Early Symptom Detection in Therapeutic Conversations, focused on identifying mental health symptoms from therapist–patient dialogues by estimating patient responses to standardized clinical questionnaires such as PHQ-9, CompACT-10, and GAD-7. Task 2, titled Therapist Response Selection, required systems to choose the most appropriate therapist reply from three candidate responses. In this work, we participated only in Task 2, focusing on the selection of suitable therapist responses in Spanish therapeutic conversations.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
abibib25@cic.ipn.mx (A. Bibi); tayyablatif80@gmail.com (T. Latif); saburb@tec.mx (S. Butt); gelbukh@cic.ipn.mx. (A. Gelbukh); sidorov@cic.ipn.mx (G. Sidorov)

In Task 2, the most appropriate response is the therapist reply that best matches the patient's current message, follows the ongoing conversation, and reflects expert-defined therapeutic practice. Such a response should be relevant, supportive, empathetic, and clinically sensible rather than merely similar in wording. If a system selects an inappropriate reply, it may interrupt the flow of the conversation, fail to address the patient's concern, or offer guidance that is not helpful for the situation. In a real mental health setting, repeated poor responses could reduce trust, leave the patient feeling misunderstood, and potentially weaken the quality of support they receive [5].

To address the response selection setting of Task 2, we used models that could directly compare the dialogue context with the three candidate therapist replies. We selected `BAAI/bge-reranker-v2-m3` [6] for semantic relevance-based ranking and `tomaarsen/Qwen3-Reranker-4B-seq-cls` [7] for its stronger contextual and multilingual reranking capability. In addition, we designed an ensemble system that combined BGE, Qwen4B, and `cross-encoder/mmarco-mMiniLMv2-L12-H384-v1` [8] through majority voting, allowing us to examine whether combining complementary rerankers could produce more stable response selection.

We investigate reranking methods for selecting therapist responses in Spanish therapeutic conversations. Our study contributes a practical response-selection pipeline for an interactive multi-turn setting and provides a direct comparison of these approaches. In the official evaluation, the Qwen reranker delivered the strongest result among our systems and ranked first in the shared task, reaching 0.393 accuracy and 0.392 Macro F1.

2. Task Description

Task 2 was designed as a therapist response selection task. In each round, the system received the latest patient message along with three possible therapist responses, as shown in Figure 1. The objective was to select the response that best matched the therapeutic guidelines provided for the task.

Figure 1: Illustration of Task 2 therapist response selection. Given the latest patient message, the system chooses the appropriate therapist response from three candidate options (English version).

Figure 1 shows the response selection process used in Task 2. The system receives the latest patient message, evaluates the candidate therapist responses, and selects the response that best matches the dialogue context and expert-defined therapeutic criteria.

This task was different from a simple offline classification problem because the evaluation was interactive. Instead of receiving the whole test set at once, the data was collected from the official server step by step. After receiving the messages for the current round, the system had to submit its prediction before the next round became available. This made the task more challenging, as the model needed to keep track of the conversation history and use previous turns to understand the current situation more accurately. In our experiments, the server returned several active dialogue sessions

simultaneously, and each session continued over multiple rounds. Therefore, maintaining the history of each session was an important part of our approach [9]. Task 2 was evaluated through an interactive server-based setting in both the trial and test phases [3]. The trial phase was provided to let teams test their systems under the same round-by-round conditions used in the final evaluation, while the test phase delivered unseen data progressively through the server. After each round, teams had to submit their selected response for every active session before receiving the next round. Performance was assessed by comparing the system's choices with the expert-selected responses, using agreement- and accuracy-based measures, while each submission also included efficiency information such as runtime, resource use, and emissions.

3. Dataset

The Task 2 dataset is composed of Spanish multi-turn therapeutic conversations that reflect realistic exchanges between patients and therapists. Due to the sensitive nature of mental health data, the dataset was not made publicly downloadable. Instead, the organizers provided access through the official evaluation server, which delivered the conversations round by round during the trial and test phases. The dataset did not include identifiable information such as patient names, exact locations, or other personal details that could reveal the identity of the participants. The patient information was anonymized, and the interaction setting was kept confidential to protect privacy [5].

Table 2 summarizes the main structure of the Task 2 dataset used in our experiments. The trial phase will involve 19 rounds, and this will enable the participants to test and prove their systems. The test stage is more extended and consists of 83 rounds, during which several sessions are run simultaneously, and predictions are made one after another. In our experiments, there were multiple active sessions per round, which formed a stream of instances rather than a fixed test set.

The dataset was delivered through an interactive server rather than as a single static file. Each round presents the system with the current patient message and three candidate responses (option 1, option 2, and option 3). One of these answers is the one that is deemed the most suitable according to the judgment of the experts, and the challenge is to find it. Each of them is linked to a session ID and a round number, which enables the dialogue to be traced through time. One of the main features of this data is that it is context-dependent. Table 1 presents an example instance from the Task 2 dataset. One of the main challenges of this dataset was its context-dependent nature. Each response must be selected not only from the current patient message but also from the earlier turns of the conversation, since the full dialogue history was not provided again at every round. In addition, the candidate responses can be very close in meaning, making the task harder than simple text matching.

Table 1
Example of a Task 2 dataset instance showing its structure and a real-style English dialogue sample.

Attribute	Example Value	Description
Session ID	S01	Unique identifier of the conversation session
Round	1	Current step in the multi-turn dialogue
Patient Message	*Hello. Well, I feel very nervous, and I cannot stop thinking about things.*	Current message expressed by the patient in English
Option 1	*I understand. What kind of things are making you feel this nervous?*	Therapist response focused on exploring the issue further
Option 2	*I understand. Is this the reason why you decided to seek psychological help?*	More general and less context-specific response
Option 3	*I understand. That feeling can be very uncomfortable. Can you tell me what is worrying you?*	Empathetic and context-aware therapeutic response
Correct Label	Option 3	Expert-selected best response
Task Type	Response Selection	Model must choose the most appropriate response
Context Dependency	Yes	Requires understanding of previous dialogue turns

Table 2
Overview of the Task 2 dataset showing the number of rounds, sessions, and candidate responses per instance.

Phase	Rounds	Sessions per Round	Candidates per Instance	Task Type
Trial	19	≈9	3	Response Selection
Test	83	≈9	3	Response Selection

4. Submissions

4.1. Trial Phase

During the trial phase, we evaluated five systems: BAAI/bge-reranker-v2-m3 [6], tomaarsen/Qwen3-Reranker-4B-seq-cls [7], cross-encoder/mmarco-mMiniLMv2-L12-H384-v1 [8], tomaarsen/Qwen3-Reranker-8B-seq-cls [10], and a three-model ensemble (BAAI/bge-reranker-v2-m3 + tomaarsen/Qwen3-Reranker-4B-seq-cls + cross-encoder/mmarco-mMiniLMv2-L12-H384-v1). The BGE model was used as a semantic ranker, while Qwen4B and Qwen8B were tested due to their stronger language understanding capacity. The mMiniLM cross-encoder was included as a lightweight comparison model. The choice of models was guided by both task suitability and efficiency considerations. Because the evaluation followed an interactive round-based format, the system needed to generate predictions quickly while also recording computational metadata such as runtime and resource consumption. For this reason, we prioritize compact reranking models instead of larger LLM-based approaches, aiming to maintain a practical balance between contextual understanding and computational efficiency.

Finally, the ensemble system combined the selected model outputs to produce a more stable prediction. The trial data consisted of 19 rounds. Since the saved trial files contained only the predicted options and computational details of the models, such as runtime, energy use, emissions, and hardware information, but did not include the gold answers, trial-phase scores such as Cohen's kappa and precision could not be calculated [11].

Without gold labels, we could not rank the trial models using numerical performance scores or confirm which one selected the best responses more often. For this reason, the trial data was used mainly for pipeline validation and qualitative observation of model behavior. During the trial phase, we also examined whether the models showed a strong preference for any single option, since a more balanced prediction pattern was considered desirable for the final selection. The final models were therefore chosen based on their relevance to the task, their complementary strengths, and their comparatively balanced response behavior, allowing us to compare a semantic reranker, a context-aware reranker, and a combined ensemble strategy during the official test phase.

Table 3 reports the predictions produced by each model during the 19-round trial phase, while Tables 4 and 5 summarize the resource usage and hardware configuration. Since the trial files did not include gold labels, these values describe model behavior and computational metadata rather than accuracy-based scores.

4.2. Test Phase

For the final test phase, we submitted three systems for therapist response selection. Run 0 used the BGE Reranker, selected for relevance-based ranking between the dialogue context and candidate responses. Run 1 used the Qwen Reranker, chosen for its stronger contextual understanding. Run 2 used an ensemble system that combined multiple model decisions to improve prediction stability.

The test data were provided through the official server in an interactive round-based format. At each round, the server returned the active sessions, including the session ID, round number, patient message, and three candidate therapist responses. The system selected one response from the three options and submitted predictions separately for Run 0, Run 1, and Run 2. Each prediction was submitted using one of the required labels: `option_1`, `option_2`, or `option_3`.

Each submission also included emissions and resource metadata, such as runtime, CPU energy, GPU energy, RAM energy, and total energy consumed. After every submission, the system requested the

Table 3
Trial-phase predictions generated by the five response-selection systems across 19 rounds.

ID	Round	BGE	Qwen4B	mMiniLM	Qwen8B	Ensemble
trial	1	option_3	option_1	option_1	option_1	option_1
trial	2	option_1	option_2	option_1	option_2	option_1
trial	3	option_2	option_3	option_2	option_3	option_2
trial	4	option_3	option_2	option_1	option_1	option_3
trial	5	option_1	option_2	option_1	option_2	option_1
trial	6	option_2	option_2	option_2	option_1	option_2
trial	7	option_1	option_3	option_1	option_1	option_1
trial	8	option_2	option_1	option_3	option_3	option_2
trial	9	option_1	option_3	option_1	option_2	option_1
trial	10	option_3	option_3	option_2	option_2	option_3
trial	11	option_3	option_1	option_2	option_1	option_3
trial	12	option_1	option_2	option_2	option_1	option_2
trial	13	option_1	option_1	option_2	option_1	option_1
trial	14	option_3	option_1	option_1	option_2	option_1
trial	15	option_2	option_3	option_2	option_1	option_2
trial	16	option_2	option_3	option_3	option_2	option_3
trial	17	option_2	option_2	option_2	option_3	option_2
trial	18	option_1	option_3	option_3	option_1	option_3
trial	19	option_2	option_3	option_3	option_1	option_3

Table 4
Resource and emissions metadata recorded during the trial-phase prediction process.

Rounds	Duration	Emissions	CPU Energy	GPU Energy	RAM Energy	Energy Consumed
19	0.00794	3.09E-08	5.55E-08	0	1.01E-08	6.57E-08

Table 5
Hardware configuration used during the trial-phase prediction process.

Resource	Value
CPU Count	12
GPU Count	1
RAM Total Size	83.4737 GB
CPU Model	Intel Xeon CPU @ 2.20GHz
GPU Model	NVIDIA A100-SXM4-40GB
Country ISO Code	SGP

next round from the server. This process continued until the server returned no further active sessions, marking the end of the test phase. This format was used for each official run. The system submitted Run 0, Run 1, and Run 2 separately at every round, then checked the server for the next batch of messages. When the server returned an empty dictionary, the evaluation process was considered complete.

5. Experimental Design

As shown in Figure 2, the proposed system was designed as a round-based pipeline to match the server-driven evaluation setting of Task 2. For every round, the system collected the incoming data from the official server and combined the stored dialogue history with the current patient message to form the model input. The three response options were then evaluated, the highest-ranked response was selected, and the final output was converted into the required JSON format for submission across Run 0, Run 1, and Run 2.

The code shows how the system preserved dialogue context during prediction. For each session, it extracted the current patient message and response options, combined them with the stored history, and passed the full query to BGE, Qwen4B, and the ensemble model. Each prediction was formatted with the session ID, round number, and selected label. Afterward, the selected response was saved back

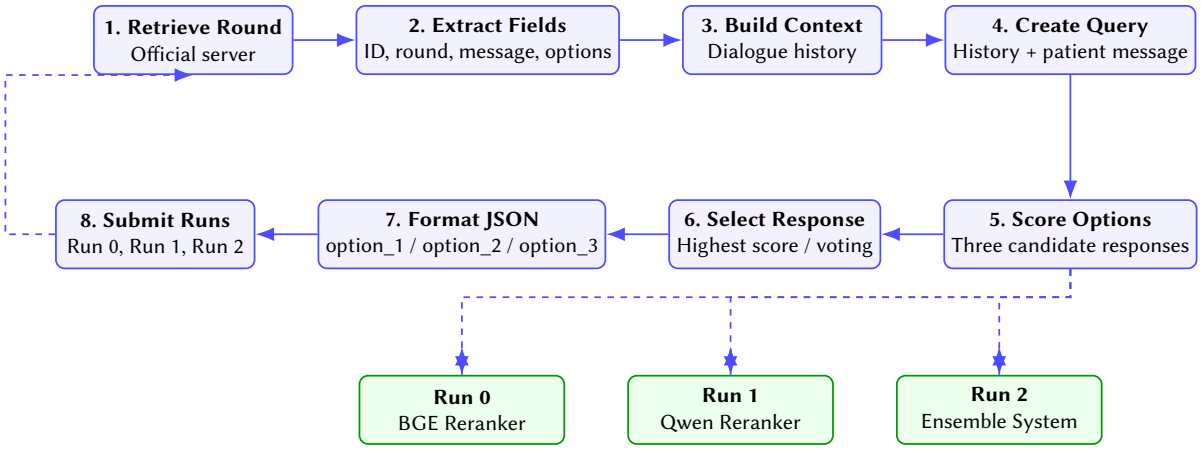

Figure 2: Workflow of the proposed interactive therapist response selection system for MentalRiskES 2026 Task 2. The system retrieves each round from the server, builds the dialogue context, scores the three candidate therapist responses using the submitted runs, formats the prediction, submits the outputs to the server, and repeats the process until no further rounds are returned.

into the same session history so that the next round could use the updated conversation context.

The evaluation followed the interactive, round-by-round protocol defined by the official MentalRiskES 2026 server [9]. At each step, the system processed the newly released dialogue instance, produced its decision, and submitted the prediction before the next round became available. The submitted output was evaluated against the expert-selected answer provided by the organizers. Performance was reported using accuracy and Macro F1, allowing the evaluation to reflect both the overall correctness of the system and its consistency across the available response labels. In addition to prediction results, the submission files included resource-related metadata, such as execution time, energy consumption, and hardware details, as required by the shared task. Figure 3 shows the history-tracking code used in our Task 2 pipeline.

5.1. Run 0: BGE Reranker

Run 0 used BAAI/bge-reranker-v2-m3. This model was selected because the task naturally fits a reranking formulation: the system receives one dialogue context and three candidate responses, and the model must rank the candidates according to their suitability. BGE was used to measure how well a strong semantic re-ranker could identify the most relevant therapist response.

5.2. Run 1: Qwen Reranker

Run 1 used tomaarsen/Qwen3-Reranker-4B-seq-cls. This model was included because response selection in this task requires more than surface-level similarity. The selected response must fit the emotional tone of the patient, the direction of the conversation, and the therapeutic context. Qwen was chosen because of its stronger contextual modeling capacity, which made it suitable for comparing subtle differences between the candidate therapist responses.

5.3. Run 2: Ensemble System

Run 2 used an ensemble system composed of three models:

- **BAAI/bge-reranker-v2-m3**
- **tomaarsen/Qwen3-Reranker-4B-seq-cls**
- **cross-encoder/mmarco-mMiniLMv2-L12-H384-v1**

The ensemble was designed to reduce dependence on a single model. Each model scored the three candidate responses independently and selected its preferred option. The final ensemble prediction was

then decided through majority voting. This allowed the system to combine the semantic strength of BGE, the contextual capacity of Qwen, and the lightweight cross-encoder behavior of mMiniLM.

Figure 3 shows the code used to preserve dialogue history during the Task 2 prediction process.

Dialogue History Tracking Code for Task 2

```python
def build_task2_predictions_full_history_live(messages, history):
    run_predictions = [[], [], []]

    option_name_map = {
        0: "option_1",
        1: "option_2",
        2: "option_3"
    }

    for session_id, row in messages.items():
        round_number = int(row["round"])
        patient_input = row["patient_input"].strip()

        options = [
            row["option_1"].strip(),
            row["option_2"].strip(),
            row["option_3"].strip()
        ]

        prior_history = history.get(session_id, [])
        full_query = "\n".join(prior_history + [f"Paciente: {patient_input}"])

        pred_bge, pred_qwen, pred_ensemble = predict_task2_live_three_runs(
            full_query, options
        )

        preds = [pred_bge, pred_qwen, pred_ensemble]

        for run_idx in range(3):
            run_predictions[run_idx].append({
                "id": session_id,
                "round": round_number,
                "prediction": option_name_map[int(preds[run_idx])]
            })

        selected_response = options[pred_ensemble]
        history.setdefault(session_id, []).extend([
            f"Paciente: {patient_input}",
            f"Terapeuta: {selected_response}"
        ])

    return run_predictions, history
```

Figure 3: Dialogue history tracking code used in the Task 2 response-selection pipeline.

5.4. Summary of Submitted Runs

Table 6: Summary of the three official systems submitted for MentalRiskES 2026 Task 2.

Run	Model/System	Decision Strategy
Run 0	BGE Reranker	Highest scoring option
Run 1	Qwen Reranker	Highest scoring option
Run 2	BGE + Qwen + mMiniLM	Majority voting

Overall, the methodology was built around the interactive nature of the shared task. Instead of treating each message as an isolated example, the system maintained session-level context, scored all candidate therapist responses with reranking models, and submitted predictions round by round through the official server. This design allowed the system to work directly within the live evaluation setup while preserving as much conversational context as possible.

6. Results

The results are presented in two parts. First, we present a summary of the predictions made by the models used in the development of the trial phase. Second, we show the official test results from the MentalRiskES 2026 Task 2 leader-board. The trial phase was used primarily to examine the behavior of the model and to validate the submission pipeline, as only model predictions were available locally and no gold labels. Therefore, the main quantitative evaluation is based on the official test results.

Accuracy was computed by the official evaluation server by comparing each submitted prediction with the expert-selected label for the same session and round. Since the gold labels were not released, we could not perform a detailed error analysis. To keep the analysis transparent, we report the official scores together with prediction distributions and baseline comparisons.

We tested five systems during the trial phase: BGE, Qwen4B, mMiniLM, Qwen8B, and an ensemble system. The distribution of selected options over the 19 trial rounds is presented in Table 7. The models demonstrated various prediction patterns. Qwen8B was more likely to choose option 1, and Qwen4B was more likely to choose option 3. The ensemble system gave a more even distribution among the three options, indicating that the voting process helped to minimize the individual bias of individual models.

Table 7
Distribution of predicted options generated by each model across the 19 trial rounds.

Model/System	Option 1	Option 2	Option 3
BGE	7	7	5
Qwen4B	5	6	8
mMiniLM	7	8	4
Qwen8B	10	6	3
Ensemble	7	6	6

We ran three times for the official test phase. The reranker used by Run 0 is BGE, the reranker used by Run 1 is Qwen reranker, and the reranker used by Run 2 is ensemble system of BGE, Qwen4B, and mMiniLM. The official results are given in Table 8. The best performance was obtained by the Qwen reranker among the three systems submitted. It achieved an overall accuracy of 0.393 and a Macro F1 score of 0.392 in the shared task. The ensemble system had an accuracy of 0.320 and a Macro F1 score of 0.321, whereas the BGE reranker had an accuracy of 0.296 and a Macro F1 score of 0.296.

Table 8
Official test results of the three submitted NLP Innovators runs for MentalRiskES 2026 Task 2.

Rank	Team/System	Run	Accuracy	Macro Precision	Macro Recall	Macro F1
1	NLP Innovators	1	0.393	0.392	0.392	0.392
10	NLP Innovators	2	0.320	0.320	0.321	0.321
19	NLP Innovators	0	0.296	0.296	0.297	0.296

The official results indicate a clear difference between the systems submitted. The Qwen reranker obtained higher official scores than both the BGE reranker and the ensemble system. This suggests that the Qwen model was better able to understand the context and therapeutic meaning of the candidate answers. For this task, the semantic similarity is not the only factor to consider when selecting the correct therapist response; the emotional direction of the conversation and the selection of a supportive, coherent, and clinically appropriate response are also important.

The ensemble system performed better than the BGE-only run, but not as well as the reranker of Qwen. This implies that simple majority voting was not enough to outperform the best single model. A potential explanation is that the final ensemble result might have been affected by weaker model decisions, which could have diminished the benefit of Qwen. A weighted ensemble, in which stronger models are given more weight, might be more effective for future work. We also checked our best run against the official baselines in the leaderboard. As presented in Table 9, our system based on Qwen achieved better performance than the random baseline and the Gemma baseline. The random baseline had an accuracy of 0.363 and Macro F1 of 0.361, and the Gemma baseline had an accuracy of 0.180 and Macro F1 of 0.176. Our best system exceeded these baselines, showing that a reranker-based approach is effective for this response selection task [5].

Table 9
Comparison of the best NLP Innovators run with the official baselines.

System	Accuracy	Macro F1
NLP Innovators Run 1	0.393	0.392
Random Baseline	0.363	0.361
Gemma Baseline	0.180	0.176

7. Discussion

Official results show that Task 2 required more than a simple match between the patient message and the candidate response. The selected response had to fit the patient's emotional state, the previous dialogue context, and the therapeutic direction of the conversation. This made the task difficult, especially when the response options were close in meaning.

Among our submitted systems, the Qwen reranker achieved the best official performance. Since the gold labels were not released, we could not examine individual errors or make a detailed case-level analysis. However, the results show that Qwen ranked the candidate responses more effectively in this setting than the other submitted systems.

The ensemble system performed better than the BGE-only run but did not outperform Qwen. A possible explanation is that the ensemble used simple majority voting, where all models were treated equally. In future work, a weighted ensemble could be explored so that stronger models contribute more to the final decision.

Overall, the results suggest that therapist response selection benefits from models that can capture both semantic relevance and dialogue-level context. Although the analysis is limited by the absence of gold labels, the official scores, prediction distributions, and baseline comparison provide a clear view of the submitted systems' behavior.

Acknowledgments

The work was done with partial support from grants 20260626 (G.S.), 20260643 (A.G.) by Secretary of Research and Posgraduate Studies (SIP) of Instituto Politécnico Nacional, Mexico

Declaration on Generative AI

The authors used generative AI tools during the preparation of this manuscript for grammar checking, spelling correction, sentence refinement, and support in preparing visual material. The methodology, experiments, results, discussion, and conclusions were produced and validated by the authors. All AI-assisted material was reviewed and edited before submission, and the authors take full responsibility for the final content of the paper.

References

[1] World Health Organization, Over a billion people living with mental health conditions: Services require urgent scale-up, https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up, 2025. Accessed: 2026-05-10.

[2] M. Rahsepar Meadi, T. Sillekens, S. Metselaar, A. van Balkom, J. Bernstein, N. Batelaan, Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review, JMIR Mental Health 12 (2025) e60432. URL: https://mental.jmir.org/2025/1/e60432. doi:10.2196/60432.

[3] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2026: Early detection of mental disorders risk in Spanish, volume 77, 2026.

[4] 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.

[5] A. Montejo-Ráez, A. M. Mármol-Romero, M. D. Molina-González, F. M. Plaza-del Arco, M. R. García-Viedma, M. Hernández-López, F. Suárez-Maroto, A. Ureña-López, MentalRiskES 2026: Early detection of mental disorders risk in Spanish, https://sites.google.com/view/mentalriskes2026, 2026. Fourth edition, IberLEF 2026 shared task.

[6] BAAI, Baai/bge-reranker-v2-m3, https://huggingface.co/BAAI/bge-reranker-v2-m3, 2024. Hugging Face model repository.

[7] T. Aarsen, tomaarsen/qwen3-reranker-4b-seq-cls, https://huggingface.co/tomaarsen/Qwen3-Reranker-4B-seq-cls, 2025. Hugging Face model repository.

[8] Sentence Transformers, cross-encoder/mmarco-mminilmv2-l12-h384-v1, https://huggingface.co/cross-encoder/mmarco-mMiniLMv2-L12-H384-v1, 2021. Hugging Face model repository.

[9] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2026: Early Detection of Addiction Risk in Spanish, Procesamiento del Lenguaje Natural 77 (2026). In press.

[10] T. Aarsen, tomaarsen/qwen3-reranker-8b-seq-cls, https://huggingface.co/tomaarsen/Qwen3-Reranker-8B-seq-cls, 2025. Hugging Face model repository.

[11] P. K. Mantell, A. Baumeister, S. Ruhrmann, A. Janhsen, C. Woopen, Attitudes towards risk prediction in a help seeking population of early detection centers for mental disorders—a qualitative approach, International journal of environmental research and public health 18 (2021) 1036.