PSI-UEX at MentalRiskES 2026: Zero-Shot Semantic Matching for Early Symptom Detection and Therapist Response Selection

Pablo Chengzu Ye-Chen[1,†], José Manuel Perea-Ortega[1,*,†]

[1]Department of Computer and Telematic Systems Engineering, Universidad de Extremadura, Badajoz, Spain

Abstract

This paper describes the participation of the PSI-UEX team in MentalRiskES 2026, a shared task focused on the analysis of Spanish therapeutic conversations in an online evaluation setting. We participated in both proposed tasks: Task 1, Early Symptom Detection in Therapeutic Conversations, and Task 2, Therapist Response Selection. Since no labeled training data were released for the official evaluation, our system was designed as a lightweight, zero-shot semantic matching approach. For Task 1, we represented each possible questionnaire response as a textual hypothesis and compared the patient's messages against these hypotheses using Spanish BERT embeddings and cosine similarity. The predictions were updated incrementally across rounds using run-specific global thresholds. For Task 2, we implemented a bi-encoder response selection system that encoded the dialogue history and the three candidate therapist responses, selecting the response with the highest semantic similarity score. We submitted three runs per task. In Task 1, the best PSI-UEX run was the moderate-threshold configuration, which achieved a MAE_Combined of 1.184 and ranked 17th. In Task 2, the best configuration was the holistic history representation, which achieved an Overall Accuracy of 0.343 and ranked 6th. The results suggest that simple semantic matching methods can provide robust baselines under low-resource conditions, although therapist response selection requires richer modeling of clinical intent and expert intervention strategies.

Keywords

Mental health, Early detection, Therapeutic conversations, Response selection, Semantic similarity, Spanish NLP, MentalRiskES

1. Introduction

The automatic analysis of mental health-related language has become an increasingly relevant research area within Natural Language Processing (NLP). Most previous work has addressed mental health detection as a retrospective classification problem [1, 2], where the complete set of messages written by a user is available before a decision is made. However, real-world clinical and conversational scenarios are naturally incremental: information is revealed progressively, and systems must operate under partial context while maintaining previous evidence [2, 3].

MentalRiskES 2026 [4, 5, 6] addresses this challenge through an online evaluation framework based on Spanish therapeutic conversations. The shared task simulates a real-time interaction where systems receive new information by rounds and must send predictions before obtaining the next set of messages. This setup introduces several practical difficulties that are not present in offline classification scenarios: participants must implement a client capable of interacting with the server, maintain user-level conversation histories, store intermediate predictions, validate submission payloads, and update predictions as new dialogue turns become available.

The PSI-UEX team participated in the two tasks proposed in MentalRiskES 2026. Task 1, Early Symptom Detection in Therapeutic Conversations, requires systems to infer a patient's responses to standardized psychometric questionnaires after each patient turn. Task 2, Therapist Response Selection, requires systems to select the most appropriate therapist response among three candidate options

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

pyechen@alumnos.unex.es (P. C. Ye-Chen); jmperea@unex.es (J. M. Perea-Ortega)

0000-0002-7929-3963 (J. M. Perea-Ortega)

according to expert judgment. Both tasks are framed in an online setting and use the same underlying therapeutic conversations.

Our participation was guided by three main constraints. First, no labeled training data were released for supervised model development. Second, the system had to operate incrementally and reliably during the official evaluation. Third, due to the limited development time and the need for robustness, we prioritized simple and interpretable semantic similarity methods over complex fine-tuned models. For this reason, both tasks were approached through Spanish BERT-based sentence representations and cosine similarity.

For Task 1, we formulated each possible answer to each questionnaire item as a textual hypothesis. Patient messages were encoded and compared with those hypotheses, and the system selected the response level whose hypothesis obtained the strongest evidence above a run-specific threshold. For Task 2, we implemented a retrieval-style response selection system based on a bi-encoder: the dialogue context and each candidate response were embedded separately, and the candidate with the highest similarity score was selected.

The main contributions of this paper are as follows. First, we describe a complete online system for participating in both MentalRiskES 2026 tasks. Second, we present a zero-shot, questionnaire-level semantic hypothesis method for early symptom estimation. Third, we evaluate three variants of a bi-encoder response selection system that differ in how the dialogue history is represented. Finally, we report the official results and discuss the limitations of semantic similarity methods in clinically grounded therapeutic response selection.

2. Task and Dataset Description

2.1. Task 1: Early Symptom Detection in Therapeutic Conversations

Task 1 focuses on the early identification of mental health symptoms expressed by a patient during a therapeutic dialogue. Instead of predicting abstract symptom labels directly, systems must infer how the patient would answer a set of standardized clinical questionnaires based on the conversation observed so far.

The questionnaires considered in the task are GAD-7, PHQ-9, and CompACT-10. GAD-7 consists of seven items related to anxiety symptoms, with answer values ranging from 0 to 3. PHQ-9 consists of nine items related to depressive symptoms, also with answer values ranging from 0 to 3. CompACT-10 consists of ten items related to psychological flexibility and Acceptance and Commitment Therapy processes, with answer values ranging from 0 to 6.

The task is therefore an item-level multiclass prediction problem. At each patient turn, the system must output three arrays of predictions: one for GAD-7, one for PHQ-9, and one for CompACT-10. Because predictions are required after each round, the system must update its estimates incrementally as the therapeutic conversation progresses.

The main official ranking metric for Task 1 is MAE_Combined, where lower values indicate better performance. Additional metrics include MZOE, MAE, Macro MAE, and early detection variants computed at different stages of the interaction.

2.2. Task 2: Therapist Response Selection

Task 2 evaluates systems as decision-support tools for therapist response selection. At each interaction step, the system receives the current patient message and three candidate therapist responses. The goal is to select the response considered most appropriate by experts.

This task is not generative: systems do not produce free-form therapist responses. Instead, they choose one candidate among three predefined options. This design makes the evaluation more controlled and reduces the risk of unsafe generated content in the mental health domain.

A key aspect of Task 2 is that the full dialogue history is not explicitly provided at every round. Systems must maintain and update the conversation history internally. This makes the task a multi-turn

response selection problem [7], where the relevance of a candidate response may depend on both the most recent patient message and the broader therapeutic context accumulated in previous rounds.

The official ranking metric for Task 2 is Overall Accuracy, where higher values indicate better performance. Other reported metrics include Macro Precision, Macro Recall, Macro F1, and Error Recovery Rate (ERR).

2.3. Dataset and Online Evaluation Protocol

The MentalRiskES 2026 corpus consists of simulated but realistic therapeutic conversations in Spanish between patients and professional therapists. The conversations were created using a combination of human-authored and synthetically generated dialogue, and were reviewed by experts to ensure clinical plausibility and coherence. The same set of conversations is used for both tasks.

The official data were delivered through an online server. In each round, the server returned the next available message for each active session. Sessions could disappear from later rounds once no further messages were available, and the server did not explicitly indicate the final message of a session. The evaluation ended when the server returned an empty response.

Each team could submit up to three runs for each task. Since PSI-UEX participated in both tasks, the system had to send six POST requests per round: three for Task 1 and three for Task 2. During the official evaluation, the PSI-UEX system completed 82 rounds starting from 10 sessions and did not encounter submission errors.

3. System Overview

3.1. General Pipeline

The PSI-UEX system was designed as an online, stateful pipeline. At each round, the system requested new data from the MentalRiskES server, updated the stored conversation history, computed predictions for each active session, validated the payload format, logged the output, and submitted one prediction file per task and run.

The system maintained separate state files for each task and run. This made it possible to preserve the current round, previous predictions, accumulated messages, and cached embeddings. This design was useful both for debugging and for robustness during the official evaluation, since the system could be restarted from the last stored round if necessary.

The pipeline can be summarized as follows: request the next round of messages from the server; update the stored conversation history for each active session; compute Task 1 predictions for each of the three runs; compute Task 2 predictions for each of the three runs; validate each generated payload; store logs, predictions, and intermediate state; submit the six required POST requests; and continue until the server returns an empty response.

3.2. Text Representation

Both tasks used the same text representation backbone: `dccuchile/bert-base-spanish-wwm-uncased`, commonly referred to as BETO [8]. BETO is a Spanish BERT model trained with whole-word masking and was selected because both tasks involve Spanish therapeutic conversations.

For each text segment, the system obtained an embedding representation using BETO. The resulting vectors were normalized so that cosine similarity could be computed efficiently as a dot product [9]. This representation was used in two ways: in Task 1, to compare patient messages against questionnaire-level hypotheses; and in Task 2, to compare the conversation context against candidate therapist responses.

The system also cached embeddings when possible. This was especially relevant for Task 2, where previously observed dialogue turns could be reused across rounds without recomputing the full conversation representation from scratch.

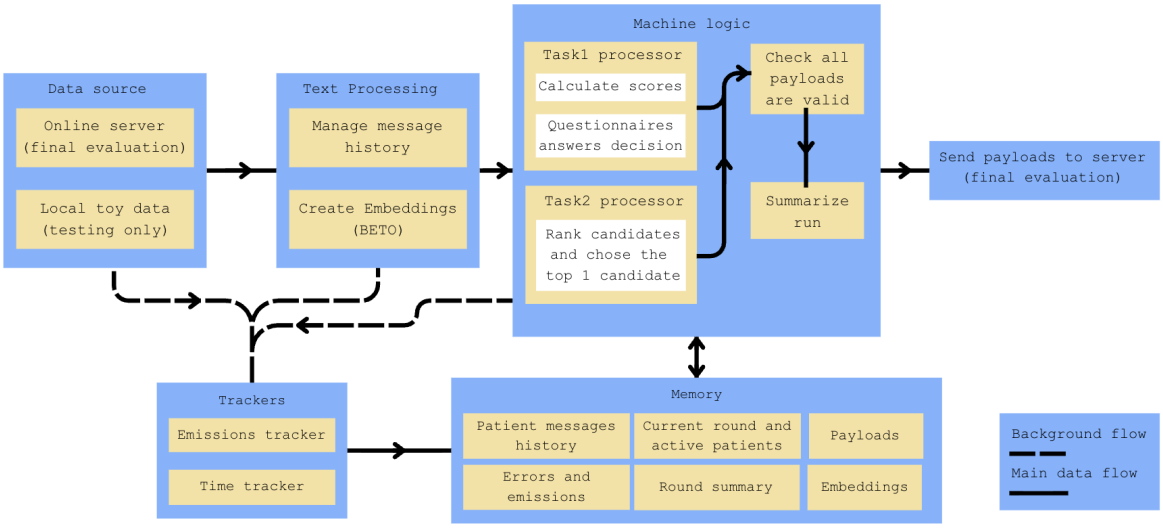

Figure 1: General architecture of the PSI-UEX system for MentalRiskES 2026. The system retrieves new messages from the server, updates the conversation state, computes task-specific predictions, validates payloads, stores logs, and submits one prediction file per task and run.

3.3. Task 1: Questionnaire-Level Semantic Hypotheses

Because no labeled training data were available, we treated Task 1 as a zero-shot semantic matching problem [10]. For each questionnaire item and each possible response level, we created a short textual hypothesis describing the semantic meaning of that level. For example, for an anxiety-related GAD-7 item, different hypotheses described occasional, frequent, or nearly constant nervousness.

At each round, the current patient message was encoded with BETO and compared against all hypotheses associated with each questionnaire item. The system then selected the response level whose hypothesis obtained the highest score, provided that the score exceeded a run-specific global threshold.

The method included three additional mechanisms. First, predictions were accumulated across rounds. The score obtained in the current round was combined with the previous score, assigning greater importance to recent evidence while preserving previously observed information. This reflected the incremental nature of the task. Second, the system implemented abstention [11]. If no candidate response level exceeded the threshold for a given item, the system kept the previous prediction. If no previous prediction existed, the initial answer was set to zero. This decision was motivated by the need to avoid updating questionnaire answers on the basis of weak semantic evidence. Third, the system limited abrupt changes between rounds. For GAD-7 and PHQ-9, a prediction could increase by at most one level per round. For CompACT-10, where the answer scale is larger, a prediction could increase by at most two levels per round. This rule was introduced to reduce sudden false positives produced by isolated messages.

The textual hypotheses used in Task 1 were drafted with the assistance of generative AI tools and manually reviewed before being integrated into the system. They were not learned from official labeled training data.

3.4. Task 2: Bi-Encoder Response Selection

Task 2 was addressed as a retrieval-based response selection problem[7]. We adopted a bi-encoder architecture because it is computationally simple and suitable for an online setting [12]. The dialogue context and each candidate therapist response were encoded separately using BETO. The system then selected the candidate with the highest cosine similarity to the context representation.

The main design question was how to represent the dialogue history. The system maintained the

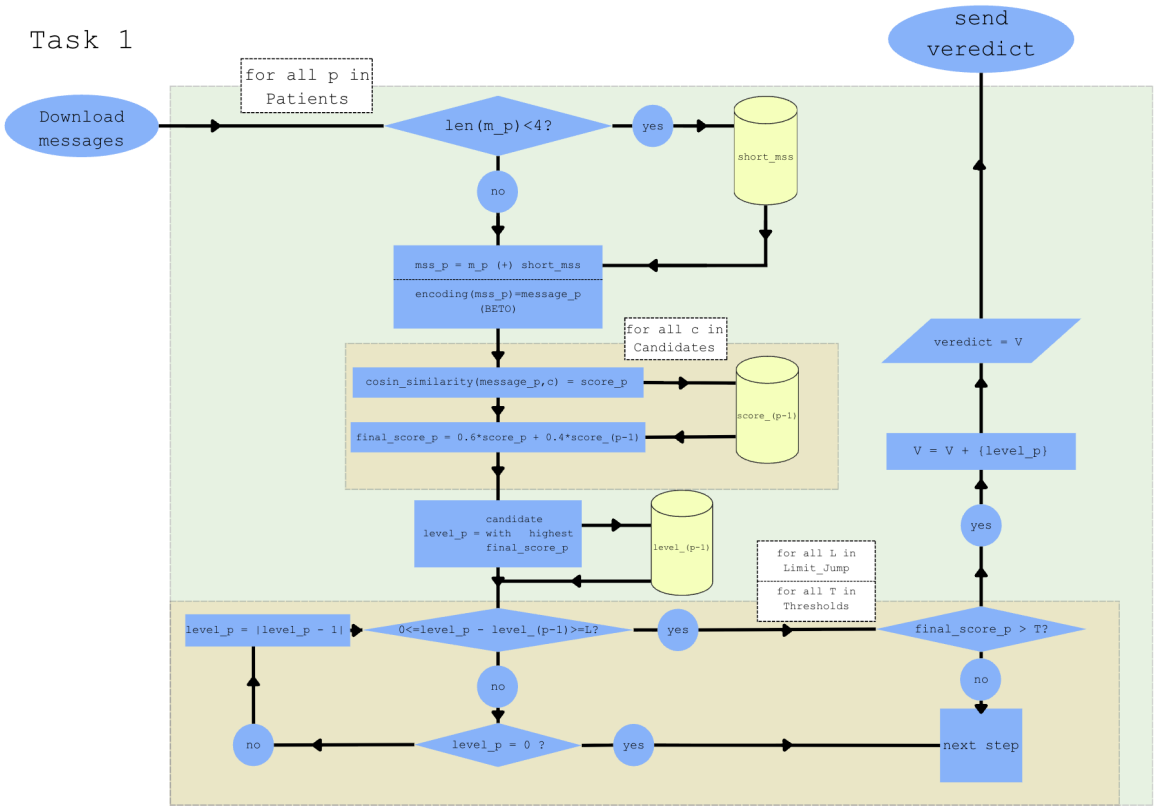

Figure 2: Task 1 workflow. The current patient message is encoded with BETO and compared against questionnaire-level hypotheses. Scores are accumulated across rounds and converted into item-level questionnaire predictions through run-specific thresholds.

conversation incrementally and added speaker markers before encoding messages, distinguishing patient and therapist turns. This was intended to preserve basic dialogue structure without requiring a more complex interaction model.

For the history representation, the system used temporally decayed averages of message embeddings [13]. More recent messages contributed more strongly than older ones. This made it possible to represent long histories without truncating text to a fixed maximum length and without recomputing a full cross-attention representation at every round.

Three variants were evaluated, corresponding to the three submitted runs. The first used a single holistic vector for the full conversation history. The second and third separated the last patient message from the previous history and combined their similarity scores with different weights.

4. Experiments and Results

4.1. Task 1: Submitted Runs and Results

For Task 1, the three submitted runs used the same semantic hypothesis method and differed only in the global threshold used to update questionnaire responses. The objective was to compare three levels of conservativeness when deciding whether a new response level should be accepted. Table 1 summarizes the submitted runs.

Run 0 was designed to update answers more easily, favoring early detection at the potential cost of more false positives. Run 1 was designed as a moderate configuration. Run 2 was the most conservative variant, requiring stronger semantic evidence before changing a questionnaire response.

Table 2 shows the official Task 1 results. The ranking metric is MAE_Combined, where lower values are better.

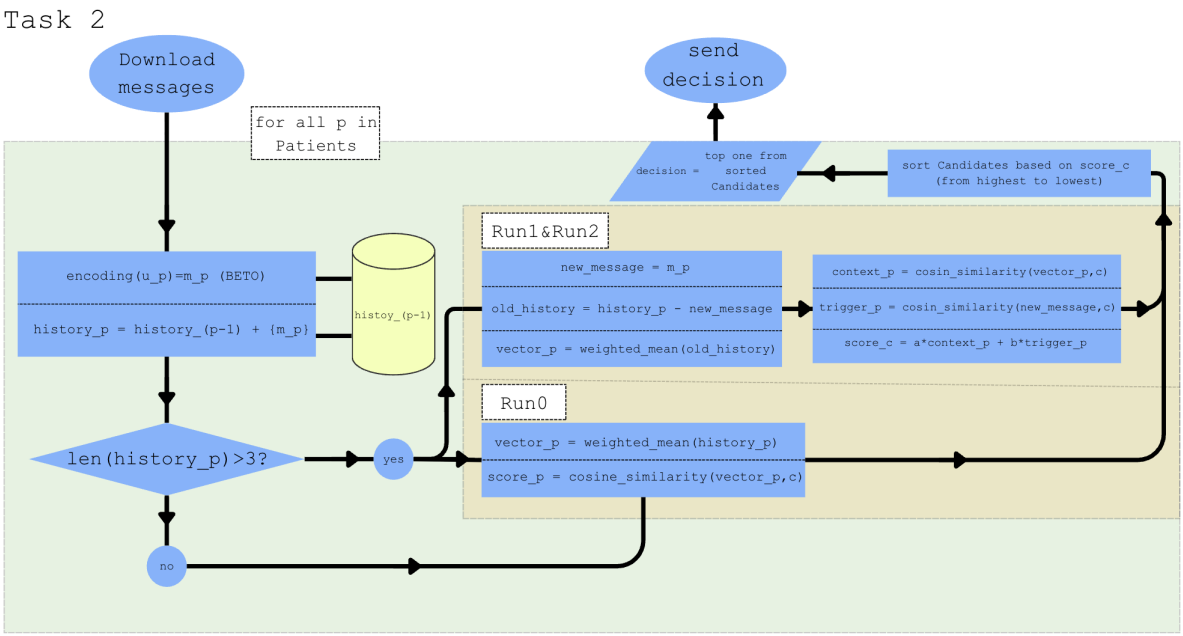

Figure 3: Task 2 workflow. The system maintains the dialogue history, encodes the current context and candidate therapist responses, and selects the candidate with the highest semantic similarity score.

Table 1
Task 1 submitted runs.

Run	Name	Configuration
Run 0	Permissive	Global threshold = 0.4
Run 1	Moderate	Global threshold = 0.6
Run 2	Conservative	Global threshold = 0.75

Table 2
Official Task 1 results for PSI-UEX together with the official competition baselines. Baseline rows are shaded.

System	Rank	MZOE GAD-7 / PHQ-9 / CompACT-10	MAE GAD-7 / PHQ-9 / CompACT-10	Macro MAE GAD-7 / PHQ-9 / CompACT-10	MAE_Combined
BASELINE - Gemma	4	0.515 / 0.599 / 0.806	0.582 / 0.724 / 1.700	0.854 / 0.762 / 1.913	1.002
PSI-UEX Run 1	17	0.598 / 0.771 / 0.780	0.764 / 1.230 / 1.558	1.278 / 1.265 / 1.911	1.184
PSI-UEX Run 0	19	0.597 / 0.776 / 0.781	0.760 / 1.239 / 1.579	1.279 / 1.274 / 1.975	1.192
BASELINE - Random	25	0.751 / 0.749 / 0.859	1.185 / 1.215 / 2.123	1.230 / 1.241 / 2.279	1.508
PSI-UEX Run 2	44	0.837 / 0.790 / 0.938	1.649 / 1.335 / 3.336	1.614 / 1.395 / 3.001	2.107

The best PSI-UEX submission for Task 1 was Run 1, with a MAE_Combined of 1.184 and a 17th position in the official ranking. Run 0 obtained a very similar result, with a MAE_Combined of 1.192. In contrast, Run 2 performed substantially worse, with a MAE_Combined of 2.107. This indicates that the most conservative threshold was not suitable for this task: requiring very strong evidence before updating questionnaire responses likely delayed or prevented necessary changes.

Compared with the official baselines, Run 1 and Run 0 outperformed the random baseline according to MAE_Combined, but remained behind the Gemma baseline in the global ranking. A more detailed inspection by questionnaire shows that Run 1 performed worse than Gemma on GAD-7 and PHQ-9, but obtained a lower MAE on CompACT-10. This suggests that the semantic hypothesis approach may have been better aligned with CompACT-10 items than with the anxiety and depression questionnaires.

The official early detection results are summarized in Table 3. To keep the comparison compact, we report the MAE_Combined values at the three available evaluation points: R10, R30, and R60.

The early detection results confirm the pattern observed in the global ranking. Run 1 was again the

Table 3
Task 1 early detection results for PSI-UEX together with the official competition baselines. Baseline rows are shaded.

System	Rank	R10	R30	R60
BASELINE - Gemma	17	1.113	1.950	0.657
PSI-UEX Run 1	18	1.182	1.690	0.757
PSI-UEX Run 0	20	1.199	1.700	0.786
BASELINE - Random	23	1.449	1.982	1.176
PSI-UEX Run 2	42	2.000	3.110	1.557

Table 4
Task 2 submitted runs.

Run	Name	Configuration
Run 0	Holistic	Single temporally decayed history vector
Run 1	Balanced	Previous history + last message; 0.3 old + 0.7 new
Run 2	Reactive	Previous history + last message; 0.1 old + 0.9 new

Table 5
Official Task 2 results for PSI-UEX together with the official competition baselines. Baseline rows are shaded.

System	Rank	Accuracy	Macro P	Macro R	Macro F1	ERR
BASELINE - random	3	0.363	0.361	0.361	0.361	0.342
PSI-UEX Run 0	6	0.343	0.343	0.343	0.343	0.350
PSI-UEX Run 1	13	0.310	0.311	0.309	0.310	0.306
PSI-UEX Run 2	16	0.303	0.303	0.302	0.302	0.305
BASELINE - Gemma	42	0.180	0.187	0.182	0.176	0.175

best PSI-UEX configuration, followed closely by Run 0. Run 2 was consistently worse. The results also show that the highest error appeared at R30 for all runs, while the R60 values were lower. This suggests that the system benefited from longer conversations, although the intermediate stage remained difficult.

The efficiency values for the best Task 1 run were low. Run 1 required an average raw time of 5.28 seconds per round and an average logic time of 0.16 seconds per round. The average emissions per round were 1.065e-05 kg CO_2eq, with an average energy consumption of 6.121e-05 kWh. Since the three runs only differed in the global threshold, the efficiency differences between them were negligible.

4.2. Task 2: Submitted Runs and Results

For Task 2, the three submitted runs differed in how the conversation history was represented and in the weight assigned to the last patient message. Table 4 summarizes the configurations.

Run 0 encoded the whole available history as a single temporally decayed vector and compared it with each candidate response. Run 1 separated the last patient message from the previous history and assigned greater weight to the last message. Run 2 followed the same structure but was even more reactive, assigning 90% of the final score to the most recent patient message.

The official Task 2 results are shown in Table 5. The ranking metric is Overall Accuracy, where higher values are better.

The best PSI-UEX submission for Task 2 was Run 0, which achieved an Overall Accuracy of 0.343 and ranked 6th. This result indicates that the holistic representation of the full dialogue history was more effective than explicitly separating the last message from the previous context. Run 1 and Run 2 performed worse, suggesting that assigning greater weight to the last patient message did not improve response selection.

Compared with the official baselines, Run 0 clearly outperformed the Gemma baseline but remained

below the random baseline. Since the task consists of selecting one option among three, the results are close to what would be expected from a random choice. This does not mean that the system failed operationally; rather, it suggests that semantic similarity alone is not sufficient to capture the expert criteria used to define the most appropriate therapist response.

The best Task 2 run required an average raw time of 5.25 seconds per round and an average logic time of 0.53 seconds per round. The average emissions per round were 1.055e-05 kg CO_2eq, with an average energy consumption of 6.080e-05 kWh. As in Task 1, the differences between runs were small because all variants shared the same embedding backbone and differed only in the scoring strategy.

5. Discussion

The Task 1 results indicate that threshold calibration was the most important factor among the evaluated variants. The moderate threshold used in Run 1 achieved the best global and early detection results, while the permissive threshold of Run 0 remained close. The conservative threshold of Run 2 was clearly worse, suggesting that excessive abstention is harmful in this setting. Since the task requires updating questionnaire responses as symptoms are progressively revealed, a high threshold may prevent the system from reacting to relevant evidence.

The semantic hypothesis strategy has several advantages. It is transparent, does not require labeled training data, and can be adapted to each questionnaire item. This is especially relevant for a shared task where no official labeled training set is available. However, the method also depends heavily on the quality of the manually reviewed hypotheses and on the ability of the embedding model to distinguish clinically meaningful differences between adjacent response levels. Some questionnaire levels are semantically close, and a generic sentence embedding model may not reliably separate them.

Another limitation of the Task 1 approach is its monotonic behavior. Since the system keeps previous predictions when no stronger evidence appears, early overestimations may persist until the end of the conversation. This design was intended to avoid reducing scores when the patient changes topic or provides short messages, but it may also produce overestimation. Conversely, limiting the maximum increase per round reduces abrupt false positives but may lead to underestimation when a message contains strong evidence for severe symptoms.

The Task 2 results show that the holistic representation of the full history was better than the variants that emphasized the last patient message. This is an important observation: although the most recent turn is often highly informative in dialogue systems, therapeutic response selection may require broader context. Expert-selected therapist responses may depend on previous topics, therapeutic goals, session structure, or intervention strategy rather than only on lexical or semantic similarity to the last patient message.

The modest Task 2 accuracy also highlights a central limitation of bi-encoder semantic matching. A response can be semantically close to the patient's message while still not being the best therapeutic intervention. For example, a candidate may repeat the patient's wording but fail to redirect the conversation appropriately, validate the patient's experience, or follow the intended clinical strategy. Capturing such distinctions likely requires richer interaction models, such as cross-encoders, rerankers, or models explicitly trained on therapeutic response preferences [14, 15].

From an engineering perspective, the system successfully completed the official online evaluation without submission errors. This confirms that simple methods can be useful when robustness, reproducibility, and online execution are major constraints. Nevertheless, future systems should combine this robustness with better calibration and stronger semantic-pragmatic modeling.

6. Conclusions and Future Work

This paper presented the PSI-UEX participation in MentalRiskES 2026. We submitted three runs for Task 1 and three runs for Task 2, all based on lightweight zero-shot semantic similarity methods using Spanish BERT embeddings.

For Task 1, we proposed a questionnaire-level semantic hypothesis approach. The best configuration was Run 1, which used a moderate global threshold and achieved a MAE_Combined of 1.184, ranking 17th. The results suggest that moderate evidence accumulation was more effective than either a permissive or a highly conservative threshold. The system outperformed the random baseline but did not reach the global performance of the Gemma baseline, although it performed better than Gemma on CompACT-10.

For Task 2, we implemented a bi-encoder response selection approach. The best configuration was Run 0, which represented the full dialogue history as a single temporally decayed vector and achieved an Overall Accuracy of 0.343, ranking 6th. This result outperformed the Gemma baseline but remained below the random baseline, indicating that semantic similarity alone is insufficient for modeling expert therapist response selection.

Future work should address several limitations. For Task 1, we plan to explore item-specific thresholds, supervised calibration when labeled development data become available, and controlled mechanisms for decreasing questionnaire scores when previous evidence is no longer supported. For Task 2, future systems should consider cross-encoder reranking, preference modeling, or clinically informed features that capture therapeutic intent rather than only semantic proximity. More systematic internal validation data would also be useful for calibrating thresholds and testing response selection strategies.

Declaration on Generative AI

During the development of the system, the authors used Gemini 3.1 Pro, developed by Google, to create auxiliary toy examples for pipeline testing and to draft the initial questionnaire-level hypotheses used in Task 1. The generated hypotheses were manually reviewed and edited by the authors before being integrated into the system. Gemini 3.1 Pro was not used to access official test labels, to train supervised models on official labeled data, or to directly generate any of the predictions submitted during the official evaluation.

Acknowledgments

This work has been 85% co-financed by the European Union, European Regional Development Fund, and the Regional Government of Extremadura (GR24182).

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Early Risk Prediction on the Internet, volume 16089 LNCS, 2026, pp. 242–265. URL: https://link.springer.com/10.1007/978-3-032-04354-2_15. doi:10.1007/978-3-032-04354-2_15.

[2] S. G. Burdisso, M. Errecalde, M. M. y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419303525. doi:10.1016/j.eswa.2019.05.023.

[3] H. Thompson, M. Errecalde, Early detection of depression and eating disorders in spanish: Unsl at mentalriskes 2023, 2023. URL: https://arxiv.org/abs/2310.20003. arXiv:2310.20003.

[4] MentalRiskES, Mentalriskes 2026, 2026. URL: https://sites.google.com/view/mentalriskes2026/home, accedido: 09-04-2026.

[5] 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, Procesamiento del Lenguaje Natural 77 (2026).

[6] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian

Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS. org, 2026.

[7] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, M. Zhou, A sequential matching framework for multi-turn response selection in retrieval-based chatbots, Computational Linguistics 45 (2019) 163–197. URL: https://direct.mit.edu/coli/article/45/1/163-197/1615. doi:10.1162/coli_a_00345.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. URL: https://arxiv.org/abs/2308.02976. arXiv:2308.02976.

[9] S. Zoupanos, S. Kolovos, A. Kanavos, O. Papadimitriou, M. Maragoudakis, Efficient comparison of sentence embeddings, in: Proceedings of the 12th Hellenic Conference on Artificial Intelligence, ACM, 2022, pp. 1–6. URL: https://dl.acm.org/doi/10.1145/3549737.3549752. doi:10.1145/3549737.3549752.

[10] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3912–3921. URL: https://www.aclweb.org/anthology/D19-1404. doi:10.18653/v1/D19-1404.

[11] A. Vazhentsev, I. Sviridov, A. Barseghyan, G. Kuzmin, A. Panchenko, A. Nesterov, A. Shelmanov, M. Panov, Uncertainty-aware abstention in medical diagnosis based on medical texts, 2025. URL: https://arxiv.org/abs/2502.18050. arXiv:2502.18050.

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3980–3990. URL: https://www.aclweb.org/anthology/D19-1410. doi:10.18653/v1/D19-1410.

[13] A.-M. Bucur, A. Cosma, P. Rosso, L. P. Dinu, It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers, volume 13980 LNCS, 2023, pp. 200–215. URL: https://link.springer.com/10.1007/978-3-031-28244-7_13. doi:10.1007/978-3-031-28244-7_13.

[14] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1, 2019, p. 4171–4186. URL: https://aclanthology.org/N19-1423.pdf.

[15] S. Humeau, K. Shuster, M. A. Lachaux, J. Weston, Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring, in: 8th International Conference on Learning Representations, ICLR 2020, 2020, pp. 1–13.