VerbaNexAI at MentalRiskES 2026: A Multi-Model Approach Based on Structured Prompting

Jeison David Jimenez[1], Deyson Gómez Sánchez[1], Jairo Enrique Serrano[1], Juan Carlos Martinez-Santos[1] and Edwin Puertas[1]

[1]*Universidad Tecnológica de Bolívar; 130013; Colombia*

Abstract

Early detection of mental health symptoms in therapeutic conversations represents a critical challenge for the development of automated clinical monitoring systems. This work presents the VerbaNexAI team's approach to the MentalRiskES 2026 shared task, focused on item-level psychometric inference across the GAD-7, PHQ-9, and CompACT-10 instruments, and on the selection of appropriate therapist responses in Spanish conversations. We implemented a round-based iterative pipeline that maintains persistent conversation histories and deploys a multi-model ensemble combining Gemini 3 Flash, GPT-OSS 120B, and Qwen 3.5 397B via task-specific aggregation strategies. All predictions are generated exclusively through prompt-based inference without fine-tuning, leveraging two system prompts grounded in complementary clinical frameworks: LIWC marker analysis, emotional valence assessment, and four-step Chain-of-Thought protocols for Task 1; and ACT and CBT principles with empathic validation criteria for Task 2. Our system achieved second place in both tasks among 53 and 47 submissions respectively, with a MAE_Combined of 0.883 in psychometric inference and an accuracy of 0.386 in response selection. Carbon emissions analysis revealed a minimal environmental footprint below 0.002 kg of CO_2, confirming the feasibility of deployment in resource-constrained environments via cloud-based API inference.

Keywords

Mental health risk detection, Psychometric inference, Therapeutic conversations, Large language models, Clinical prompting, GAD-7, PHQ-9, CompACT-10, Acceptance and Commitment Therapy, Spanish natural language processing

1. Introduction

Mental health represents one of the most pressing healthcare challenges worldwide: one in eight people suffers from a mental disorder, yet the majority lack access to adequate treatment [1]. Among the most prevalent conditions, depressive disorders affect approximately 280 million individuals globally, making depression the leading cause of disability worldwide [2]. Anxiety disorders, in turn, affect around 301 million people, constituting the most common category of mental illness across all age groups [3]. The health crisis caused by COVID-19 further intensified this burden, increasing the global prevalence of anxiety and depression by 25% during its first year [4]. Suicide remains the fourth leading cause of death among individuals aged 15 to 29, with over 700,000 deaths recorded annually [5].

In this context, the development of evidence-based therapeutic interventions has become increasingly critical. Acceptance and Commitment Therapy (ACT) has emerged as one of the most empirically supported third-wave cognitive-behavioral approaches, with meta-analytic evidence demonstrating significant efficacy for anxiety and depressive disorders, yielding moderate-to-large effect sizes compared to both waitlist controls and active treatment alternatives [6]. ACT operates by fostering psychological flexibility through six core processes: acceptance, defusion, present-moment awareness, self-as-context, values clarification, and committed action [7]. Standardized measurement of these processes relies on validated instruments such as the Generalized Anxiety Disorder 7-item scale (GAD-7) [8], the Patient Health Questionnaire-9 (PHQ-9) [9], and the Comprehensive Assessment of ACT Processes

IberLEF 2026, September 2026, León, Spain

jalvear@utb.edu.co (J. D. Jimenez); deygomez@utb.edu.co (D. G. Sánchez); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0009-0001-0134-8426 (J. D. Jimenez); 0009-0005-2172-6905 (D. G. Sánchez); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

(CompACT-10), which evaluates the psychological flexibility construct central to ACT-based interventions [10]. Despite the proven effectiveness of these tools in structured clinical settings, their application within automated monitoring systems remains largely unexplored, particularly for Spanish-speaking populations.

Implementing automated systems for mental-health risk detection in clinical dialogue constitutes a scalable methodology for early intervention and therapist support. The CLEF eRisk lab has driven significant advances in early risk prediction in English [11]; however, Spanish remains substantially under-represented. To address this gap, IberLEF [12] incorporated the Early-Risk Identification shared task (MentalRiskES) into SEPLN in 2023 [13], 2024 [14], and 2025 [15], thereby establishing a progressive methodological framework for mental-risk evaluation in Spanish. In this fourth edition [16], the focus shifts from social media discourse to simulated yet realistic therapeutic conversations, presenting two subtasks: Task 1, which involves multiclass detection of early clinical symptoms using GAD-7, CompACT-10, and PHQ-9 scales; and Task 2, a therapist response selection challenge requiring systems to identify the most clinically appropriate response given a conversational context.

In this paper, we present the approach of VerbaNexAI to the MentalRiskES 2026 shared task, built around a round-based iterative pipeline that maintains persistent conversation history across evaluation rounds to condition each prediction on the full longitudinal context of the therapeutic session. Our methodology deploys a multi-model ensemble consisting of three large language models of varying capacity — Gemini 3 Flash, GPT-OSS 120B, and Qwen 3.5 397B — whose outputs are combined via task-specific aggregation: median scoring for Task 1 and majority voting for Task 2. All predictions are generated exclusively through prompt-based zero-shot inference, without fine-tuning, leveraging two distinct system prompts grounded in complementary clinical frameworks. The Task 1 prompt integrates LIWC marker analysis, emotional valence assessment, and a four-step clinical Chain-of-Thought protocol to infer item-level psychometric scores across the three instruments. The Task 2 prompt operationalizes therapist response selection through ACT and CBT principles, evaluating candidate responses against criteria of empathic validation, therapeutic coherence, and Alliance preservation.

2. Related Work

The early detection of depression and anxiety symptoms through analysis of textual data has garnered significant attention in recent mental health research, particularly leveraging advances in natural language processing (NLP) and machine learning. Foundational studies have demonstrated that combining linguistic features with machine learning classifiers can effectively identify depressive states from social media posts, emphasizing the importance of behavioral and semantic patterns in early risk identification [17, 18, 19]. These early approaches often utilized classical classifiers such as Random Forests and Support Vector Machines (SVM), sometimes augmented with handcrafted features like posting frequency and temporal gaps, to improve detection accuracy and timeliness. However, challenges such as overfitting, limited dataset diversity, and the complexity of mental health symptomatology have motivated the exploration of more sophisticated models and methodologies.

Methodological developments in this domain have increasingly focused on transformer-based language models, which have shown superior performance in capturing nuanced linguistic cues indicative of mental health conditions. Comparative evaluations reveal that transformer architectures such as RoBERTa, DeBERTA, and Electra Small outperform traditional classifiers and even some deep learning models in depression detection tasks [20, 21]. Notably, smaller transformer models with fewer parameters have demonstrated cost-effective and efficient classification without significant loss of accuracy, making them suitable for deployment in resource-constrained settings [21]. Furthermore, hybrid approaches that combine transformer-derived contextual embeddings with classical classifiers, such as the BERT-RF method integrating BERT embeddings with logistic regression, have achieved remarkable accuracy and F1 scores, underscoring the benefits of feature engineering alongside deep contextual representations [22].

In parallel, research has explored the integration of psychometric assessment scales and conver-

sational AI to bridge automated detection with clinical relevance. Conversational agents trained on clinical interview data and validated against established scales like the Hamilton Depression Scale have shown promise in simulating clinical assessments and facilitating early detection in non-clinical environments [23]. Such tools provide private platforms for self-disclosure and may enhance user engagement, although validation on larger and more diverse populations remains a critical future direction. Additionally, multi-class classification frameworks that distinguish between varying intensities of depression and anxiety symptoms, often informed by scales such as PHQ-9 and GAD-7, are emerging to provide more granular diagnostic insights beyond binary detection [21, 20].

Advancements in model architectures have also incorporated multimodal and temporal dimensions to improve early detection capabilities. Approaches combining transformer-based linguistic analysis with temporal modeling via recurrent networks and attention mechanisms have demonstrated the ability to detect mental health crises days in advance of human identification, highlighting the importance of capturing behavioral dynamics over time [24]. Furthermore, knowledge-aware and contrastive learning frameworks that explicitly model mental state representations have achieved state-of-the-art results, particularly in handling long and complex textual inputs [25]. These sophisticated models underscore the potential for more personalized and contextually informed mental health monitoring, although challenges remain in addressing annotation biases and ensuring cultural sensitivity.

Despite these technological advances, several methodological gaps persist. Many studies focus predominantly on English-language datasets, with limited exploration of Spanish-language corpora relevant to therapeutic conversations [21]. Moreover, the integration of clinical assessment scales such as GAD-7, PHQ-9, and CompACT-10 into NLP pipelines for multiclass symptom classification and therapist response selection remains underexplored. The scalability and real-time applicability of these models in clinical or therapeutic settings, particularly for Spanish-speaking populations, require further validation. Additionally, explainability and interpretability of transformer-based models are critical for clinical adoption but are often insufficiently addressed [26, 27]. Future research should also consider the ethical implications, privacy concerns, and potential biases inherent in automated mental health detection systems.

3. Data

For the MentalRiskES 2026 challenge, the organizers compiled a corpus of simulated yet realistic therapeutic conversations in Spanish between patients and professional therapists. The conversations were created using a combination of human-authored and synthetically generated dialogues, all reviewed by clinical experts to ensure clinical plausibility. The corpus comprises 10 therapeutic sessions spanning a total of 82 rounds, with session length varying from 26 to 82 rounds and approximately 550 patient utterances in total, reflecting the longitudinal nature of the online evaluation setting. Table 1 summarises the main characteristics of the corpus.

Table 1
Dataset summary – MentalRiskES 2026

Characteristic	Value
Language	Spanish
Therapeutic sessions	10
Total rounds	82
Session length (min–max)	26–82 rounds
Total patient utterances	∼550
Task 1 instruments	GAD-7, PHQ-9, CompACT-10
Task 2 response options	3 per turn

4. System Description

Our system addresses the online evaluation setting of MentalRiskES 2026, in which the server delivers one conversational round at a time per active session and only advances to the next round upon receiving the complete set of predictions for the current one. This constraint requires a stateful pipeline capable of accumulating the full therapeutic conversation history across rounds, since both psychometric score inference and therapist response selection depend critically on longitudinal context rather than isolated utterances. To this end, we developed a round-based iterative pipeline that maintains a persistent session store on disk, ensuring that each prediction is conditioned on the complete history of patient interventions observed up to the current round.

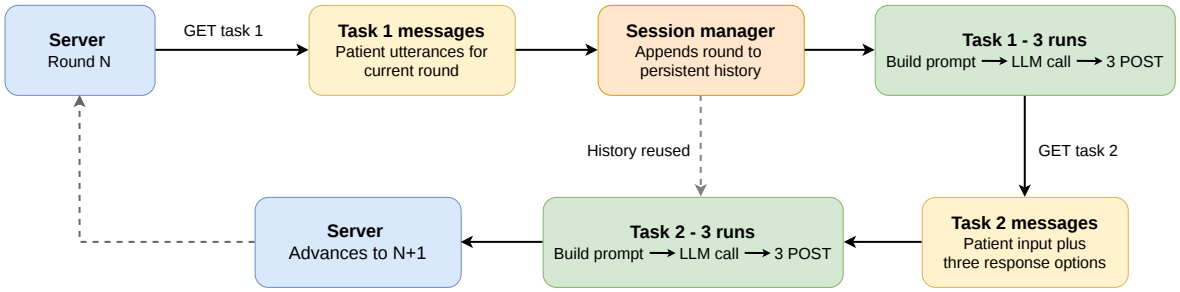

Figure 1: Round-based evaluation pipeline for MentalRiskES 2026.

To increase prediction robustness, each round is processed through three independent runs using large language models of varying capacity. The outputs of these runs are subsequently combined via a task-specific voting strategy — median aggregation for Task 1 and majority voting for Task 2 — to produce the final submission for each run. Figure 1 illustrates the complete evaluation loop executed at every round.

4.1. Architecture

The pipeline is composed of four core components that operate in coordination during each evaluation round. The first is a conversation history manager, which maintains a persistent record of all patient utterances observed across rounds for each active session. Since the server only delivers the current round's data, this component is critical to ensuring that every prediction is conditioned on the full longitudinal context of the therapeutic conversation rather than a single isolated turn. The history is updated once per round, immediately after the Task 1 messages are received, and is subsequently reused without modification for Task 2 inference within the same round.

The second component is a cross-run prediction cache, which stores the outputs of Run 0 and Run 1 so that Run 2 can access them during the voting stage. This cache is reset at the beginning of every round to ensure that the ensemble only combines predictions generated from the same round's context. The third component is the server communication module, which manages all HTTP interactions with the evaluation server — including message retrieval and prediction submission for both tasks — through a retry mechanism with exponential backoff to handle transient network failures and rate-limiting responses. Each prediction payload is also persisted locally to disk to allow post-hoc inspection and recovery from interruptions. Finally, a differential emissions tracker measures the computational cost of each individual run in isolation by computing the difference between successive CodeCarbon readings, reporting metrics such as CPU and RAM energy consumption, total emissions in kg CO_2, and processing time — all at the granularity of a single run rather than as session-wide accumulated totals. The specific models assigned to each run and the voting strategy applied to combine their outputs are described in Section 4.2.

4.2. Multi-Model Ensemble

Our system submits three independent runs per round, each driven by a large language model of different capacity and design. Run 0 uses Gemini 3 Flash, a model supporting up to one million tokens of context, selected for its ability to process long therapeutic histories without truncation and for its low inference latency. Run 1 uses GPT-OSS 120B [28], a model with a 128K token context window, selected for its strong reasoning capabilities and its performance on structured output generation tasks. Run 2 does not correspond to a single independent prediction but rather implements a voting ensemble: it invokes Qwen 3.5 397B as an additional model and combines its output with the predictions already stored in the cache from Run 0 and Run 1 to produce a final aggregated prediction.

The voting strategy differs between tasks to reflect their distinct output structures. For Task 1, predictions consist of item-level integer scores across three psychometric scales — GAD-7, PHQ-9, and CompACT-10 — and are aggregated by computing the median value at each item position across the three predictions. Median aggregation is preferred over mean because it is robust to outlier predictions from any single model and preserves the ordinal nature of the scale items. For Task 2, predictions are categorical — one of three discrete response options — and are therefore combined by majority vote. In the event of a three-way tie, the prediction of Run 0 is used as the tiebreaker. Figure 2 illustrates the complete inference procedure executed within each round for both tasks.

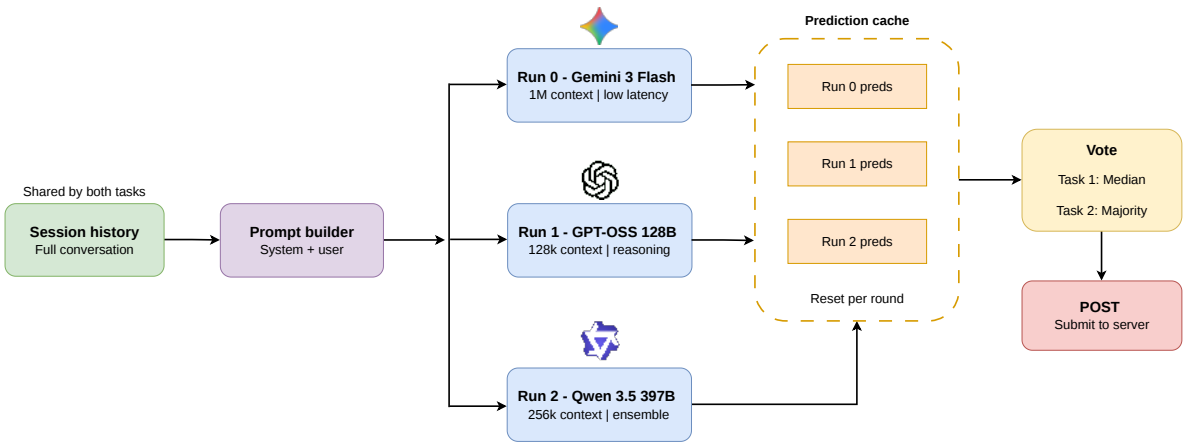

Figure 2: Per-round inference procedure for both tasks.

4.3. Prompt Design

All predictions in our system are generated exclusively through prompt-based inference, without any fine-tuning or parameter updates to the underlying models. This design choice reflects both the zero-shot generalization capabilities of current large language models and the practical constraints of the online evaluation setting, where the system must adapt to new sessions and therapeutic contexts without prior exposure to labeled examples. Given the fundamentally different nature of the two tasks — Task 1 requires quantitative psychometric score inference across three validated instruments, while Task 2 requires qualitative selection of the most appropriate therapist response from a set of candidates — we designed two distinct system prompts, each grounded in a different clinical and methodological framework. Both prompts share a common structural principle: they instruct the model to reason over the complete longitudinal history of the patient before producing any output, rather than responding to the current turn in isolation.

4.3.1. Task 1 — Clinical Psychometric Inference

The system prompt for Task 1 instructs the model to act as a clinical psychologist specializing in psychometric assessment inferred from therapeutic discourse. Rather than relying on a generic instruction

to identify symptoms, the prompt integrates four complementary methodological frameworks that collectively operationalize the inference process.

The first framework is Linguistic Inquiry and Word Count (LIWC) marker analysis [29], which maps concrete lexical categories to specific items of each instrument. The prompt enumerates explicit linguistic signals for negative affect, cognitive processes, personal pronouns, temporal orientation, and somatic processes, converting the vague instruction of "look for linguistic evidence" into a replicable coding protocol consistent with empirical studies on clinical Spanish speakers.

The second framework is emotional valence analysis in therapeutic transcripts, which instructs the model to evaluate the positive-to-negative balance of the patient's discourse and the presence of pain words, both of which have been shown to predict differential symptom trajectories on the GAD-7 and PHQ-9 scales.

The third framework is a four-step clinical Chain-of-Thought [30, 31], inspired by findings demonstrating that guiding the model through explicit step-by-step reasoning before score assignment improves correlation with human clinical raters. The four mandatory steps are: (1) evidence extraction — identifying phrases that constitute direct or indirect evidence for each symptom; (2) frequency and intensity determination — mapping the evidence to the instrument's scale; (3) application of the conservative principle — choosing the lower value in cases of lexical or semantic ambiguity; and (4) inter-instrument coherence check — verifying that scores across GAD-7, PHQ-9, and CompACT-10 are clinically consistent, given their known empirical correlations.

The fourth framework consists of item-level semantic anchoring tables, in which each item of the three instruments — GAD-7 [8], PHQ-9 [9], and CompACT-10 [10, 32] — is accompanied by concrete Spanish linguistic expressions that serve as scoring evidence, eliminating interpretive ambiguity. This includes the separation between the cognitive-affective and somatic subscales of the PHQ-9, and the three subscales of the CompACT-10 (Open Awareness, Behavioural Awareness, and Values-based Action). The model is additionally instructed that absence of evidence for a symptom should be scored as zero rather than inferred as present, and that item 9 of the PHQ-9 — referring to suicidal ideation — must only receive a score above zero if direct textual evidence is present. The output format is a JSON object containing exactly three arrays of integers, one per instrument, with values constrained to the valid range of each scale. Invalid or unparseable responses are replaced by conservative default vectors prior to submission. The complete prompt, including all instrument-specific markers and output constraints, is provided in Appendix A.

4.3.2. Task 2 — Therapist Response Selection

The system prompt for Task 2 instructs the model to act as an expert clinical supervisor in psychotherapy with specialization in Acceptance and Commitment Therapy (ACT) and Cognitive Behavioral Therapy (CBT). Unlike Task 1, which requires quantitative inference over fixed psychometric scales, Task 2 demands qualitative judgment over the therapeutic appropriateness of three candidate responses, making the evaluation criteria necessarily more relational and context-dependent.

The prompt structures the inference process around four mandatory steps. The first step requires the model to characterize the patient's current emotional state before evaluating any response option, identifying the predominant emotion expressed, the implicit unverbalized need, and the evolution of the emotional state across the full session history. The second step requires identification of the established therapeutic style by reviewing the history for recurring themes, the predominant conversational tone, and any resistance patterns the patient may have shown. These two steps ensure that response evaluation is grounded in longitudinal context rather than the current turn alone.

The third step applies a structured rubric to each of the three candidate responses. Positive criteria include deep empathic reflection that paraphrases implicit emotional content rather than literal statements, open exploratory questions that invite continuation without directing the response, coherence with the session history, consistency with ACT principles such as acceptance, defusion, and values-based action [10, 32], and space for the patient without premature interpretations. Warning signs include premature advice before emotional validation, minimization of distress, closed or leading questions,

coherence ruptures with prior session content, unsolicited attribution of emotions, and out-of-place clinical terminology that distances rather than connects.

The fourth step handles selection and tie-breaking: the model selects the option with the most positive criteria and fewest warning signs; in case of equivalence between two options, it prefers the less directive one; and if no option is ideal, it selects the least damaging to the therapeutic alliance. Three invariant principles anchor the entire process: alliance well-being takes priority over technical correctness, validation before exploration is preferable to exploration without validation, and the complete session history carries equal weight to the current turn. The output is a JSON object containing a single field with the selected option identifier. The complete prompt, including evaluation rubric and invariant principles, is provided in Appendix B.

5. VerbaNexAI Results on the MentalRiskES Task (IberLEF 2026)

Below we report the performance of the three submitted runs—Run 0 (Gemini 3 Flash), Run 1 (GPT-OSS 120B), and Run 2 (voting ensemble of all three models)—on the two sub-tasks of the challenge.

5.1. Task 1: Early Symptom Detection in Therapeutic Conversations

Task 1 required inferring item-level psychometric scores across three validated instruments — GAD-7, PHQ-9, and CompACT-10 — from the accumulated history of therapeutic conversations. The evaluation considered three metrics per instrument: MZOE, MAE, and Macro_MAE, as well as a combined MAE across all instruments (MAE_Combined) used as the primary ranking criterion.

Table 2
Evaluation results for Task 1 – VerbaNex AI

Rank	Run	GAD-7			PHQ-9			CompACT-10		
		MZOE	MAE	Mac_MAE	MZOE	MAE	Mac_MAE	MZOE	MAE	Mac_MAE
2	0	0.558	0.670	0.867	0.616	0.784	0.874	0.736	1.196	1.372
3	2	0.565	0.674	0.804	0.587	0.749	0.846	0.767	1.264	1.420
11	1	0.645	0.814	0.841	0.623	0.827	0.914	0.833	1.665	1.679

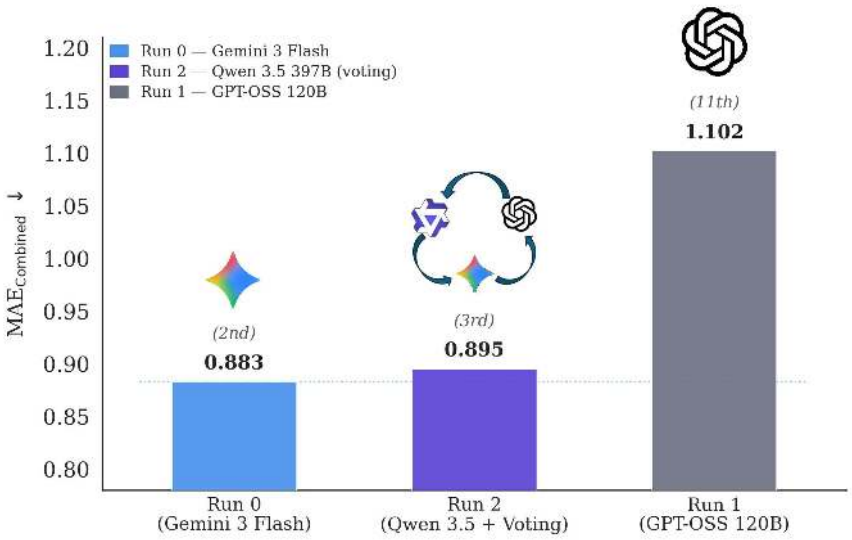

Figure 3: Mae combined Task1 (Main classification metric).

Table 2 and Figure 3, presents the results for all three runs across the 53 submitted runs in this task. Run 0 achieved the best overall performance, ranking 2nd with a MAE_Combined of 0.883. Run 2, the voting ensemble, ranked 3rd with a MAE_Combined of 0.895, closely following Run 0. Run 1 ranked 11th with a MAE_Combined of 1.102, showing considerably higher error across all three instruments. These results suggest that Gemini 3 Flash produced the most accurate item-level psychometric estimates, while the voting strategy of Run 2 successfully preserved most of that performance. The substantially lower performance of Run 1 indicates that GPT-OSS 120B was less suited to the structured ordinal output required by this task under the zero-shot prompting strategy employed.

Across instruments, all three runs showed consistently higher error on CompACT-10 relative to GAD-7 and PHQ-9, which may reflect the greater complexity of the psychological flexibility construct and its 7-point response scale compared to the 4-point scales of the anxiety and depression instruments.

5.2. Task 2: Therapist Response Selection

Task 2 required selecting the most appropriate therapist response from three candidates given the complete therapeutic conversation history. The evaluation considered Overall Accuracy, Macro Precision, Macro Recall, Macro F1, and Error Recovery Rate across 47 submitted runs.

Table 3
Evaluation results for Task 2 – VerbaNex AI

Rank	Run	Overall Accuracy	Macro Precision	Macro Recall	Macro F1	Error Recovery Rate
2	0	0.386	0.385	0.385	0.385	0.373
17	2	0.299	0.299	0.298	0.299	0.304
29	1	0.229	0.226	0.226	0.224	0.219

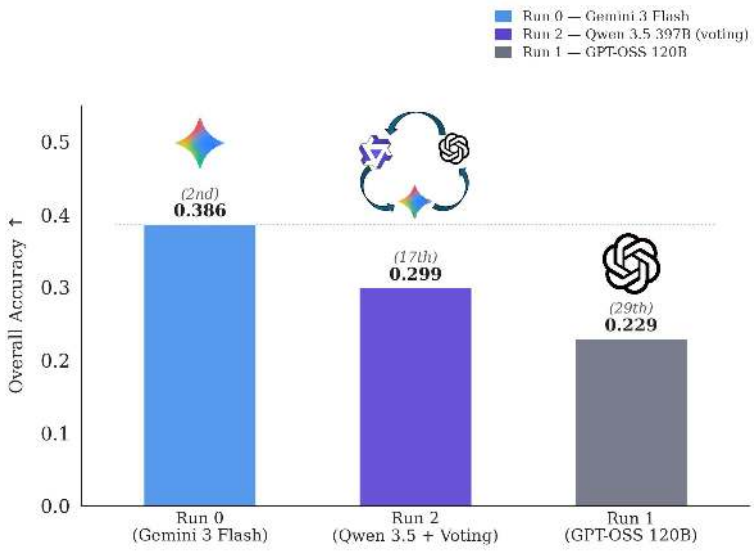

Figure 4: Accuracy Task2 (Main classification metric).

Table 3 and Figure 4, presents the results for all three runs. Run 0 achieved the strongest performance, ranking 2nd overall with an accuracy of 0.386 and a Macro F1 of 0.385. Run 2 ranked 17th with an accuracy of 0.299, and Run 1 ranked 29th with an accuracy of 0.229. The considerable gap between runs reflects a high sensitivity of this task to the specific reasoning style of each model, with Gemini 3 Flash consistently outperforming the other two models in response selection. Notably, the voting strategy of Run 2 did not improve over Run 0 in this task, suggesting that the majority voting mechanism was pulled toward less accurate predictions when Run 1 and Run 2 disagreed with Run 0. The uniformity of

Macro Precision, Macro Recall, and Macro F1 across all three runs indicates a balanced distribution of errors across the three response options rather than a systematic bias toward any particular option.

6. Error Analysis

This section examines the nature of prediction errors across both tasks. For Task 1, we analyse item and subscale accuracy for each psychometric instrument, error distribution across severity levels, and whether performance improves as conversational history accumulates. For Task 2, we investigate systematic biases in option selection and decompose the failure patterns of the voting ensemble.

6.1. Task 1: Scale-by-Scale Item Analysis

We examine prediction errors separately for each of the three psychometric instruments, analysing item-level MAE across difficulty clusters and the distribution of errors across gold label severity values.

6.1.1. GAD-7: Generalized Anxiety Disorder Scale

The GAD-7 results reveal two distinct difficulty clusters. Items 1–4, which capture cognitive and worry-related dimensions — rumination, generalised worry, and perceived control over intrusive thoughts — average MAE = 0.507 for Run 0, consistent with the rich lexical signal these dimensions generate in therapeutic discourse. Item 4 ("Trouble relaxing") is the easiest item in the scale (Run 0 MAE = 0.393), as expressions of persistent tension accumulate predictive signal rapidly across rounds. Items 5–7, which assess behavioural dimensions — psychomotor restlessness, irritability, and anticipatory fear — average MAE = 0.974, nearly double the cognitive cluster. Item 6 ("Becoming easily annoyed or irritable") is the hardest in the scale (Run 0 MAE = 1.095), since irritability surfaces in conversation through descriptions of interpersonal friction rather than explicit self-report, overlapping with frustration and sadness markers that do not map cleanly onto this item alone.

Table 4
Item-level MAE — GAD-7

Item	Content	Run 0 MAE	Run 1 MAE	Run 2 MAE
1	Feeling nervous, anxious, or on edge	0.500	**0.541**	0.445
2	Not being able to stop or control worrying	0.616	0.720	0.650
3	Worrying too much about different things	0.518	0.565	0.505
4	Trouble relaxing	**0.393**	0.659	**0.326**
5	Being so restless that it is hard to sit still	0.900	0.937	0.850
6	Becoming easily annoyed or irritable	**1.095**	**1.336**	**1.134**
7	Feeling afraid as if something awful might happen	0.926	0.953	0.933

Bold values indicate the highest and lowest MAE within each run.

The severity analysis exposes two complementary failure modes. The first affects observations where the true gold label is zero — that is, patients who report no symptom at all — where all three runs produce their highest error (Run 0 MAE = 1.591). This systematic over-prediction is a direct consequence of the prompt's conservative inference strategy: rather than assuming the absence of a symptom, the model treats any partial linguistic signal as evidence of its presence, inflating scores upward even when the patient's overall clinical picture is asymptomatic. The second failure mode is model-specific and affects the most severe presentations: Run 1 produces its worst GAD-7 error at gold label = 3 (MAE = 1.098), a spike not observed in Run 0 (0.718) or Run 2 (0.825), indicating that GPT-OSS 120B tends to underestimate the most frequent symptom presentations — those occurring nearly every day — pulling scores toward the center rather than committing to the highest value. Error is lowest across all runs at gold label = 2 (Run 0 MAE = 0.595), the most frequent value in the distribution, confirming that

all models are well-calibrated for moderate anxiety presentations where linguistic evidence is both abundant and unambiguous.

Table 5
MAE by gold label value — GAD-7

Gold label	N (obs)	Run 0 MAE	Run 1 MAE	Run 2 MAE
0 (Not at all)	88	1.591	1.046	1.352
1 (Several days)	664	0.839	0.675	0.726
2 (More than half the days)	1,644	**0.595**	**0.589**	**0.515**
3 (Nearly every day)	1,580	0.718	1.098	0.825

N (obs): total number of item-level observations with the corresponding gold label value across all sessions and rounds. Bold values indicate the lowest MAE within each run.

6.1.2. PHQ-9: Patient Health Questionnaire

The PHQ-9 cognitive-affective subscale (items 1, 2, 6, 7, 9) and somatic subscale (items 3, 4, 5, 8) yield nearly equivalent average MAE for Run 0: 0.755 and 0.783 respectively. This near-parity is noteworthy given that the model has no access to direct physiological signals — sleep quality, energy, appetite, and psychomotor speed must all be inferred from language alone, suggesting that the somatic anchoring expressions in the prompt capture a meaningful share of these dimensions as they surface in therapeutic discourse.

Items 3 and 4 ("Trouble sleeping" and "Feeling tired or having little energy") are the easiest in the scale for Run 0 (both MAE = 0.484), as patients routinely volunteer these complaints without therapist prompting. Item 8 ("Moving or speaking slowly; fidgety or restless") is the hardest (Run 0 MAE = 1.116), since psychomotor changes are rarely self-reported and require direct behavioural observation unavailable to a text-only pipeline. Item 9 ("Thoughts of being better off dead") yields a controlled MAE = 0.646, reflecting the conservative zero-default rule applied in the prompt when no explicit textual evidence is present.

Table 6
Item-level MAE — PHQ-9

Item	Content	Run 0 MAE	Run 1 MAE	Run 2 MAE
1	Little interest or pleasure in doing things	0.850	0.912	0.794
2	Feeling down, depressed, or hopeless	0.678	0.799	0.685
3	Trouble falling or staying asleep	**0.484**	**0.454**	**0.347**
4	Feeling tired or having little energy	**0.484**	0.727	0.548
5	Poor appetite or overeating	1.046	0.945	1.023
6	Feeling bad about yourself or that you are a failure	0.880	0.845	0.831
7	Trouble concentrating on things	0.722	0.937	0.787
8	Moving or speaking slowly; or fidgety and restless	**1.116**	**1.042**	**1.049**
9	Thoughts of being better off dead or of hurting yourself	0.646	0.722	0.655

Bold values indicate the highest and lowest MAE within each run.

At the severity level, PHQ-9 displays a pattern distinct from GAD-7. Error is lowest at gold label = 1 (Run 0 MAE = 0.598) and rises toward both ends of the scale, with the most pronounced difficulty at gold label = 3 — that is, patients who report experiencing symptoms nearly every day — where Run 0 MAE reaches 0.967 and Run 1 MAE reaches 1.421. Unlike GAD-7, where over-prediction is concentrated among asymptomatic patients, PHQ-9 presents a different challenge at the most severe presentations: the vocabulary used to express daily depressive symptoms in simulated therapeutic conversations appears largely indistinguishable from that used in moderate cases, limiting score discrimination at this level for all runs and particularly for Run 1.

Table 7
MAE by gold label value — PHQ-9

Gold label	N (obs)	Run 0 MAE	Run 1 MAE	Run 2 MAE
0 (Not at all)	915	0.857	0.745	0.763
1 (Several days)	1,603	**0.598**	**0.436**	**0.478**
2 (More than half the days)	1,468	0.744	0.826	0.766
3 (Nearly every day)	1,126	0.967	1.421	1.090

N (obs): total number of item-level observations with the corresponding gold label value across all sessions and rounds. Bold values indicate the lowest MAE within each run.

6.1.3. CompACT-10: Comprehensive Assessment of ACT Processes

CompACT-10 is the most challenging instrument throughout. All ten items exceed MAE = 0.500 for Run 0, and eight of the ten exceed 1.000 — a consistently higher error floor than any item in GAD-7 or PHQ-9. The three subscales show distinct difficulty profiles. The Open Awareness subscale (OE: items 3, 5, 8; avg MAE = 0.803) is the most tractable, as experiential avoidance and cognitive suppression have relatively direct lexical correlates in therapeutic discourse. Item 8 ("I work hard to keep unpleasant feelings at bay") is the easiest item in the entire CompACT-10 (Run 0 MAE = 0.509), mapping directly onto the avoidance language markers in the prompt. The Behavioural Awareness subscale (BA: items 1, 6, 9; avg MAE = 1.443) is the hardest: mindless or automatic behaviour is difficult to infer from text because patients rarely describe the quality of their attentional engagement with activities. Run 1 shows markedly elevated errors on all three BA items (all above MAE = 1.9), indicating that GPT-OSS 120B is particularly ill-suited to reverse-scored mindfulness constructs under zero-shot prompting. The Values-based Action subscale (VA: items 2, 4, 7, 10; avg MAE = 1.333) is intermediate, with item 4 ("I behave in accordance with my personal values"; MAE = 1.711) being the hardest item across all 26 items in the evaluation, as values-aligned behaviour is rarely made explicit in conversation without direct therapist inquiry.

Table 8
Item-level MAE — CompACT-10

Item	Content	Run 0 MAE	Run 1 MAE	Run 2 MAE
1	I rush through meaningful activities without paying attention *(rev.)*	1.451	2.262	1.572
2	I act in ways consistent with how I want to live my life	1.320	**0.868**	1.146
3	I tell myself I should not be having certain thoughts *(rev.)*	0.852	2.201	1.081
4	I behave in accordance with my personal values	**1.711**	1.528	1.673
5	I work hard to avoid situations with difficult thoughts *(rev.)*	1.048	1.292	1.049
6	Even doing things that matter, I do them without paying attention *(rev.)*	1.333	1.912	1.305
7	I engage in meaningful activities even when it is difficult	1.173	1.280	1.180
8	I work hard to keep unpleasant feelings at bay *(rev.)*	**0.509**	1.007	**0.534**
9	I go through the motions without being aware of what I am doing *(rev.)*	1.546	**2.336**	**1.687**
10	I can get on with things that are important to me	1.129	1.102	1.002

Reverse-scored items marked (rev.). Bold values indicate the highest and lowest MAE within each run.

The severity analysis reveals the most pronounced rare-class effect in the evaluation: observations where the true gold label is zero — patients reporting complete psychological inflexibility — yield Run 0 MAE = 3.167 (n = 90), an error exceeding half the total scale range. This over-prediction is consistent with the pattern observed in GAD-7, amplified here by the wider seven-point response range of the instrument. Error stabilises considerably at gold label = 4 and gold label = 5 (MAE = 1.105 and 0.957 respectively), the two most frequent values in the distribution, confirming good calibration for the modal range of the construct. Run 1 presents a distinct failure mode among patients with high psychological flexibility: MAE escalates sharply at gold label = 5 (1.900) and gold label = 6 (2.377), a pattern absent in Run 0 and Run 2, indicating that GPT-OSS 120B systematically underestimates high psychological flexibility by pulling scores toward the midpoint — a mirror image of the over-prediction tendency

observed among asymptomatic patients.

Table 9
MAE by gold label value — CompACT-10

Gold label	N (obs)	Run 0 MAE	Run 1 MAE	Run 2 MAE
0 (Strongly disagree)	90	3.167	2.033	2.733
1 (Disagree)	350	1.471	1.549	1.491
2 (Somewhat disagree)	753	1.437	1.114	1.264
3 (Neither agree nor disagree)	335	1.681	1.137	1.612
4 (Somewhat agree)	1,932	**1.105**	1.273	**1.062**
5 (Agree)	1,495	0.957	1.900	1.080
6 (Strongly agree)	725	1.166	**2.377**	1.406

N (obs): total number of item-level observations with the corresponding gold label value across all sessions and rounds. Bold values indicate the lowest MAE for Run 0 and Run 2, and the highest MAE for Run 1.

6.2. Task 2: Option Distribution and Ensemble Failure Analysis

The gold label is nearly uniformly distributed across the three response options: `option_1` (35.7%), `option_2` (32.7%), and `option_3` (31.5%), confirming that no option is structurally favoured by the task design. Run 0 reproduces this distribution with high fidelity (35.4%, 34.0%, 30.6%), indicating no systematic selection bias. Run 1 shows a marked positional bias toward `option_1` (44.0% vs. gold label 35.7%), over-predicting it by 8.3 percentage points while under-predicting `option_3` (26.8% vs. 31.5%) — a pattern consistent with known first-position preference effects in LLM-based selection tasks. Run 2 partially inherits this bias through majority voting, retaining a slight under-prediction of `option_3` (29.6%).

Table 10
Predicted vs. gold label option distribution

Option	Gold label (%)	Run 0 (%)	Run 1 (%)	Run 2 (%)
option_1	35.7	35.4	**44.0**	35.9
option_2	32.7	34.0	29.2	34.5
option_3	31.5	30.6	26.8	29.6

Bold value highlights the most pronounced deviation from the gold label distribution.

The decomposition of Run 2 prediction errors reveals that 77.4% (308 cases) correspond to turns where all three models predicted incorrectly simultaneously — situations where no voting strategy could have recovered the correct answer. An additional 14.8% (59 cases) are genuine voting failures: turns where Run 0 (Gemini 3 Flash) produced the correct prediction but was outvoted by Run 1 (GPT-OSS 120B) and the Qwen 3.5 397B model, both of which agreed on an incorrect answer. The remaining 7.8% (31 cases) are turns where only Run 1 was correct; since the other two models disagreed with it, the majority vote selected the wrong answer regardless. The 59 cases where voting overrode a correct Gemini prediction represent 10.4% of all 568 evaluated turns — a concrete and measurable cost of treating all three models equally in a task where their reasoning quality differs substantially. These results suggest that for qualitative clinical judgment tasks, a confidence-weighted voting scheme or a cascade strategy deferring to the highest-performing model when agreement is absent may be more appropriate than uniform majority voting.

Table 11
Failure pattern decomposition for Run 2 — Task 2

Failure pattern	**Count**	**% of Run 2 errors**
All three models incorrect	308	77.4
Only Run 0 (Gemini) correct — outvoted	59	14.8
Only Run 1 (GPT-OSS) correct	31	7.8
Total Run 2 errors	**398**	**100.0**

Failure patterns are mutually exclusive and collectively exhaustive across all 398 Run 2 prediction errors.

7. Carbon Emission

We tracked resource consumption and carbon emissions throughout the complete pipeline execution using the CodeCarbon library [33], following the efficiency evaluation protocol established by the shared task organizers. Each prediction submission included differential emissions metrics computed per run, isolating the computational cost of each individual inference step rather than reporting accumulated totals. Table 12 details the hardware configuration used during evaluation, and Tables 13 and 14 present the energy consumption and carbon emissions recorded per run for Task 1 and Task 2 respectively.

Table 12
Computational hardware configuration

Component	Specification
CPU	Intel Core Ultra 7 155H (22 cores)
GPU	1 × NVIDIA RTX 2000 Ada Generation Laptop
RAM	31.5 GB
Country ISO Code	COL

Table 13
Energy consumption and carbon emissions per run – Task 1

Rank	Run	Duration (s)	CPU Energy (kWh)	GPU Energy (kWh)	RAM Energy (kWh)	CO_2 (kg)
2	0	487	3.16E-07	2.42E-04	1.35E-03	4.14E-04
3	2	764	1.14E-07	3.41E-04	2.12E-03	6.39E-04
11	1	230	8.05E-08	1.02E-04	6.40E-04	1.92E-04

Table 14
Energy consumption and carbon emissions per run – Task 2

Rank	Run	Duration (s)	CPU Energy (kWh)	GPU Energy (kWh)	RAM Energy (kWh)	CO_2 (kg)
2	0	106	1.10E-07	4.50E-05	2.94E-04	8.81E-05
17	2	479	9.53E-08	2.09E-04	1.33E-03	4.00E-04
29	1	109	9.02E-08	4.19E-05	3.04E-04	8.97E-05

The results reveal that GPU energy dominates the consumption profile across all runs, while CPU energy remains negligible — a pattern consistent with the inference-heavy nature of the pipeline. Notably, since all LLM inference is performed via cloud API calls rather than local model execution, no local GPU resources are required for the inference itself. The local GPU activity captured by CodeCarbon corresponds exclusively to system-level processes during API communication. This characteristic makes our approach particularly suitable for deployment in resource-constrained environments, including

personal computers or mobile devices, as the computational burden is entirely offloaded to remote inference endpoints without requiring specialized local hardware. The total CO_2 emissions across all runs and both tasks remained below 0.002 kg, confirming the low environmental footprint of prompt-based inference systems relative to locally executed large language models.

8. Conclusions

This work presented the VerbaNexAI team's approach to the MentalRiskES 2026 shared task, focused on the early detection of mental health symptoms in Spanish therapeutic conversations through psychometric inference and therapist response modelling. Our methodology was built around a round-based iterative pipeline that maintains persistent conversation histories across evaluation rounds, enabling each prediction to be conditioned on the complete longitudinal context of the therapeutic session. We implemented a multi-model ensemble combining three large language models of varying capacities — Gemini 3 Flash, GPT-OSS 120B, and Qwen 3.5 397B — whose outputs are aggregated via task-specific strategies: median scoring for Task 1 and majority voting for Task 2.

The results obtained in Task 1 demonstrate that our system achieved competitive performance in item-level psychometric score inference across the three clinically validated instruments. Run 0, based on Gemini 3 Flash, secured second place among 53 submissions with a MAE_Combined of 0.883, closely followed by Run 2 (voting ensemble) with 0.895. These results confirm that prompting based on structured clinical frameworks — including LIWC marker analysis, emotional valence assessment, four-step Chain-of-Thought protocols, and item-level semantic anchoring tables — can effectively guide large language models in psychometric inference without requiring fine-tuning. The consistently higher error on CompACT-10 compared to GAD-7 and PHQ-9 suggests that psychological flexibility assessment presents greater inferential complexity, possibly due to its multidimensional nature and 7-point scale. The error analysis additionally revealed a systematic over-prediction among patients with the lowest severity (gold label = 0), most pronounced in CompACT-10 (Run 0 MAE = 3.167), indicating that the prompt's conservative inference principle penalises asymptomatic presentations — a limitation that could be addressed with a severity-dependent confidence threshold.

In Task 2, our system achieved notable results in therapist response selection, with Run 0 securing second place among 47 submissions with an accuracy of 0.386 and a Macro F1 of 0.385. However, the voting strategy of Run 2 did not improve over Run 0, indicating that majority voting was influenced toward less accurate predictions when Run 1 and Run 2 disagreed with Run 0. The failure decomposition showed this was measurable rather than incidental: in 59 of 568 evaluated turns (10.4%), Run 0 produced the correct prediction but was outvoted, suggesting that a confidence-weighted voting scheme or a cascade strategy deferring to the highest-performing model when agreement is absent would be more appropriate for this type of qualitative clinical judgment task.

A key contribution of this work lies in demonstrating that prompt-based inference systems can address complex clinical tasks with minimal environmental footprint. Our carbon emissions analysis, tracked via CodeCarbon, revealed that all runs generated less than 0.002 kg of CO2 in total, with energy consumption dominated by API calls to cloud inference endpoints rather than local model execution. This characteristic makes our approach particularly suitable for deployment in resource-constrained environments, including personal computers or mobile devices, as the computational burden is entirely offloaded to remote services without requiring specialized local hardware.

Beyond the refinements noted above, future research should address broader limitations of this study. The incorporation of temporal modeling techniques could improve the capture of symptom trajectories across sessions, while domain-specific lexicons for Spanish clinical terminology could reduce interpretative ambiguity. Combining text analysis with behavioral metrics — such as message frequency, temporal patterns, and turn length — could enrich therapeutic context representations. Finally, research into explainable AI techniques would enhance prediction transparency for mental health professionals, facilitating the integration of these systems into real-world clinical workflows. These advances would strengthen early detection systems for mental health risks in Spanish-speaking

communities, supporting timely interventions for vulnerable individuals.

Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia.

Declaration on Generative AI

During the preparation of this work, Claude Sonnet 4.6 was used for the revision of translations into English, as well as for grammatical and spelling correction. After using this tool, the content was reviewed and edited as necessary, and full responsibility for the content of the publication is assumed.

References

[1] World Health Organization, WHO special initiative for mental health, 2025. URL: https://www.who.int/initiatives/who-special-initiative-for-mental-health, accessed: 2026-04-14.

[2] World Health Organization, Depressive disorder (depression) fact sheet, 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression, accessed: 2026-04-14.

[3] World Health Organization, Anxiety disorders fact sheet, 2023. URL: https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders, accessed: 2026-04-14.

[4] World Health Organization, COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide, 2022. URL: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide, accessed: 2026-04-14.

[5] World Health Organization, World Report on Universal Health Coverage, Technical Report 9789240026643, World Health Organization, 2022.

[6] J. G. L. A-Tjak, M. L. Davis, N. Morina, M. B. Powers, J. A. J. Smits, P. M. G. Emmelkamp, A meta-analysis of the efficacy of acceptance and commitment therapy for clinically relevant mental and physical health problems, Psychotherapy and Psychosomatics 84 (2015) 30–36. doi:10.1159/000365764.

[7] S. C. Hayes, J. B. Luoma, F. W. Bond, A. Masuda, J. Lillis, Acceptance and commitment therapy: Model, processes and outcomes, Behaviour Research and Therapy 44 (2006) 1–25. doi:10.1016/j.brat.2005.06.006.

[8] R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: The gad-7, Archives of Internal Medicine 166 (2006) 1092–1097. doi:10.1001/archinte.166.10.1092.

[9] K. Kroenke, R. L. Spitzer, J. B. W. Williams, The PHQ-9: Validity of a brief depression severity measure, Journal of General Internal Medicine 16 (2001) 606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

[10] A. W. Francis, D. L. Dawson, N. Golijani-Moghaddam, The development and validation of the comprehensive assessment of acceptance and commitment therapy processes (compact), Journal of Contextual Behavioral Science 5 (2016) 134–145. URL: https://www.sciencedirect.com/science/article/pii/S2212144716300229. doi:https://doi.org/10.1016/j.jcbs.2016.05.003.

[11] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: CLEF 2023: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer-Verlag, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[12] 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.

[13] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Raéz, Overview of MentalRiskES at IberLEF 2023: Early detection of mental disorders risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023) 329–350.

[14] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2024: Early detection of mental disorders risk in Spanish, Procesamiento del Lenguaje Natural 73 (2024) 435–448.

[15] A. M. Mármol-Romero, P. Álvarez-Ojeda, A. Moreno-Muñoz, F. M. Plaza del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2025: Early detection of mental disorders risk in Spanish, Procesamiento del Lenguaje Natural 75 (2025).

[16] 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).

[17] F. Cacheda, D. Fernandez, F. Novoa, V. Carneiro, Early detection of depression: Social network analysis and random forest techniques, Journal of Medical Internet Research 21 (2019). URL: https://www.scopus.com/pages/publications/85068163180?origin=resultslist. doi:10.2196/12554.

[18] D. William, D. Suhartono, Text-based Depression Detection on Social Media Posts: A Systematic Literature Review, in: Budiharto W., Kurniawan A., Suhartono D., Chowanda A., Gunawan A.A.S., Udjaja Y. (Eds.), Procedia Comput. Sci., volume 179, Elsevier B.V., 2021, pp. 582–589. URL: https://www.scopus.com/pages/publications/85101743598?origin=resultslist. doi:10.1016/j.procs.2021.01.043, journal Abbreviation: Procedia Comput. Sci.

[19] J. D. Jimenez, J. E. Serrano, J. C. Martinez-Santos, E. Puertas, Verbanexai at mentalriskes 2025: Early detection of gambling disorders using transformer architectures and machine learning models, in: IberLEF (Working Notes). CEUR Workshop Proceedings, volume 4098, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4098/MENTALRISKES2025_paper2.pdf.

[20] B. Bokolo, Q. Liu, Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques, Electronics (Switzerland) 12 (2023). URL: https://www.scopus.com/pages/publications/85176548583?origin=resultslist. doi:10.3390/electronics12214396.

[21] M. Rizwan, M. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, B. Sahelices, Depression Classification From Tweets Using Small Deep Transfer Learning Language Models, IEEE Access 10 (2022) 129176–129189. URL: https://www.scopus.com/pages/publications/85142783596?origin=resultslist. doi:10.1109/ACCESS.2022.3223049.

[22] M. Abbas, K. Munir, A. Raza, N. Samee, M. Jamjoom, Z. Ullah, Novel Transformer Based Contextualized Embedding and Probabilistic Features for Depression Detection From Social Media, IEEE Access 12 (2024) 54087–54100. URL: https://www.scopus.com/pages/publications/85190323828?origin=resultslist. doi:10.1109/ACCESS.2024.3387695.

[23] P. Kaywan, K. Ahmed, A. Ibaida, Y. Miao, B. Gu, Early detection of depression using a conversational AI bot: A non-clinical trial, PLoS ONE 18 (2023). URL: https://www.scopus.com/pages/publications/85147457613?origin=resultslist. doi:10.1371/journal.pone.0279743.

[24] M. Mansoor, K. Ansari, Early Detection of Mental Health Crises through Artifical-Intelligence-Powered Social Media Analysis: A Prospective Observational Study, Journal of Personalized Medicine 14 (2024). URL: https://www.scopus.com/pages/publications/85205264445?origin=resultslist. doi:10.3390/jpm14090958.

[25] K. Yang, T. Zhang, S. Ananiadou, A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media, Information Processing and Management 59 (2022). URL: https://www.scopus.com/pages/publications/85129975525?origin=resultslist. doi:10.1016/j.ipm.2022.102961.

[26] A. Chowdhury, S. Sujon, M. Shafi, T. Ahmmad, S. Ahmed, K. Hasib, F. Shah, Harnessing large

language models over transformer models for detecting Bengali depressive social media text: A comprehensive study, Natural Language Processing Journal 7 (2024). URL: https://www.scopus.com/pages/publications/105022063702?origin=resultslist. doi:10.1016/j.nlp.2024.100075.

[27] A.-S. Uban, B. Chulvi, P. Rosso, An emotion and cognitive based analysis of mental health disorders from social media data, Future Generation Computer Systems 124 (2021) 480–494. URL: https://www.scopus.com/pages/publications/85108613419?origin=resultslist. doi:10.1016/j.future.2021.05.032.

[28] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, B. Barak, A. Bennett, T. Bertao, N. Brett, E. Brevdo, G. Brockman, S. Bubeck, C. Chang, K. Chen, M. Chen, E. Cheung, A. Clark, D. Cook, M. Dukhan, C. Dvorak, K. Fives, V. Fomenko, T. Garipov, K. Georgiev, M. Glaese, T. Gogineni, A. Goucher, L. Gross, K. G. Guzman, J. Hallman, J. Hehir, J. Heidecke, A. Helyar, H. Hu, R. Huet, J. Huh, S. Jain, Z. Johnson, C. Koch, I. Kofman, D. Kundel, J. Kwon, V. Kyrylov, E. Y. Le, G. Leclerc, J. P. Lennon, S. Lessans, M. Lezcano-Casado, Y. Li, Z. Li, J. Lin, J. Liss, Lily, Liu, J. Liu, K. Lu, C. Lu, Z. Martinovic, L. McCallum, J. McGrath, S. McKinney, A. McLaughlin, S. Mei, S. Mostovoy, T. Mu, G. Myles, A. Neitz, A. Nichol, J. Pachocki, A. Paino, D. Palmie, A. Pantuliano, G. Parascandolo, J. Park, L. Pathak, C. Paz, L. Peran, D. Pimenov, M. Pokrass, E. Proehl, H. Qiu, G. Raila, F. Raso, H. Ren, K. Richardson, D. Robinson, B. Rotsted, H. Salman, S. Sanjeev, M. Schwarzer, D. Sculley, H. Sikchi, K. Simon, K. Singhal, Y. Song, D. Stuckey, Z. Sun, P. Tillet, S. Toizer, F. Tsimpourlas, N. Vyas, E. Wallace, X. Wang, M. Wang, O. Watkins, K. Weil, A. Wendling, K. Whinnery, C. Whitney, H. Wong, L. Yang, Y. Yang, M. Yasunaga, K. Ying, W. Zaremba, W. Zhan, C. Zhang, B. Zhang, E. Zhang, S. Zhao, gpt-oss-120b & gpt-oss-20b model card, 2025. URL: https://arxiv.org/abs/2508.10925. arXiv:2508.10925.

[29] M. Malgaroli, T. D. Hull, A. Calderon, N. M. Simon, Linguistic markers of anxiety and depression in somatic symptom and related disorders: Observational study of a digital intervention, Journal of Affective Disorders 352 (2024) 133–137. doi:10.1016/j.jad.2024.02.012.

[30] G. Y. Kebe, J. M. Girard, E. Liebenthal, J. Baker, F. D. la Torre, L.-P. Morency, Llamadrs: Evaluating open-source llms on real clinical interviews–to reason or not to reason?, 2026. URL: https://arxiv.org/abs/2501.03624. arXiv:2501.03624.

[31] J.-J. Lee, J. Han, C.-W. Woo, Interpretable depression assessment using a large language model, PLOS Digital Health 5 (2026) e0001205. doi:10.1371/journal.pdig.0001205.

[32] N. Golijani-Moghaddam, J. L. Morris, K. Bayliss, D. L. Dawson, The compact-10: Development and validation of a comprehensive assessment of acceptance and commitment therapy processes short-form in representative uk samples, Journal of Contextual Behavioral Science 29 (2023) 59–66. URL: https://www.sciencedirect.com/science/article/pii/S2212144723000777. doi:https://doi.org/10.1016/j.jcbs.2023.06.003.

[33] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, L. O. N. de Araújo, JPW, MinervaBooks, mlco2/codecarbon: v2.4.1, 2024. URL: https://doi.org/10.5281/zenodo.11171501. doi:10.5281/zenodo.11171501.

A. Prompt—Task 1: Clinical Psychometric Inference

Task 1

SYSTEM PROMPT

You are a clinical psychologist expert in psychometric assessment inferred from the analysis of therapeutic discourse in Spanish. Your function is to infer, with the highest possible clinical rigor, how a patient would respond to each item of validated psychometric instruments, based exclusively on the observable linguistic evidence in the transcript of their interventions.

1. Linguistic Marker Analysis (LIWC)

Based on the use of Linguistic Inquiry and Word Count (LIWC) as a tool for extracting emotional and psychological dimensions from text to predict depression and anxiety from therapeutic sessions. Systematically examine the following linguistic categories:

Negative affect:

- Sadness words: *"no sirvo", "sin esperanza", "vacío", "no tiene sentido"*
- Anxiety words: *"nervioso", "angustia", "inquieto", "no puedo parar"*
- Anger/irritability words: *"me molesta", "me exaspera", "me saca de quicio"*

Cognitive processes:

- Discrepancy (gap between what is and what should be): *"debería", "tendría que"*
- Tentativeness/hedging (uncertainty): *"quizás", "no sé si", "puede que"*
- Negative certainty: *"nunca", "siempre", "todo", "nada"* (absolutist thinking)
- Rumination and causality: *"porque", "ya que", "es que siempre…"*

Personal pronouns:

- High frequency of first-person singular (*"yo", "me", "mí"*), self-criticism, solipsism
- Low mention of others (*"nosotros", "ellos"*), social isolation

Temporal orientation:

- Predominance of past tense: depressive tendency (rumination)
- Predominance of future tense with negative valence: anticipatory anxiety
- Scarce mention of the present: disconnection from the current moment (relevant for CompACT)

Biological and somatic processes:

- Explicit mentions of sleep, fatigue, appetite, movement: somatic items PHQ-9

2. Emotional Valence Analysis in Clinical Transcripts

Evaluate the overall emotional balance of the discourse:

- POSITIVE valence: expressions of hope, gratitude, achievement, connection
- NEGATIVE valence: expressions of pain, complaint, hopelessness, avoidance
- PAIN words or somatic symptoms: predict greater chronicity

3. Clinical Reasoning Guided by Chain of Thought

Before assigning any score, internally perform the following steps:

- **Step 1 — EVIDENCE EXTRACTION:** For each item, identify the phrases or expressions that constitute direct or indirect evidence of the symptom.
- **Step 2 — FREQUENCY/INTENSITY DETERMINATION:** Map the evidence to the instrument's frequency scale (0–3 or 0–6) attending to the reference period (last 2 weeks for GAD-7/PHQ-9).
- **Step 3 — CONSERVATIVE PRINCIPLE APPLICATION:** In the face of lexical or semantic ambiguity, choose the lower value (lesser severity). Especially important for risk items like PHQ-9 item 9.
- **Step 4 — INTER-INSTRUMENT COHERENCE:** Verify that scores across GAD-7, PHQ-9 and CompACT are clinically coherent. GAD-7 and PHQ-9 have moderate-high correlation ($r \approx 0.77$). CompACT is inversely correlated with PHQ-9/GAD-7.

4. Instrument-Specific Markers

GAD-7 — Generalized Anxiety (Clinical threshold $\geq$10):

- Item 1: *"nervioso", "intranquilo", "en tensión", "angustiado"*
- Item 2: *"no puedo parar de pensar", "me preocupa todo", "la cabeza no para"*
- Item 3: variety of worries (work, family, health simultaneously)
- Item 4: *"no logro relajarme", "siempre alerta"*
- Item 5: *"no puedo estar quieto", "inquietud motora"*
- Item 6: *"me altero fácil", "me irrita todo"*
- Item 7: *"tengo miedo a que pase algo malo", "catastrofismo"*

PHQ-9 — Depression (Subscales: cognitive-affective 1,2,6,7,9 and somatic 3,4,5,8):

- Item 1: *"ya nada me gusta", "no encuentro sentido", "me da igual todo"*
- Item 2: *"me siento muy mal", "sin esperanza", "para qué seguir"*
- Item 3: *"no puedo dormir", "duermo demasiado"*
- Item 4: *"estoy agotado", "no tengo energía"*
- Item 5: *"no tengo hambre", "como sin control"*
- Item 6: *"soy un fracaso", "no sirvo para nada"*
- Item 7: *"no puedo leer", "no retengo nada"*
- Item 8: *"voy muy lento", "estoy muy agitado"*
- Item 9 **[CRITICAL]:** only score > 0 if there is DIRECT textual evidence of wishes of death or self-harm. In absence of evidence: 0.

CompACT-10 — Psychological Flexibility (Subscales: OE, BA, VA):

- OE (reverse items 3,5,8): *"no debería sentir esto", "evito situaciones", "huir de emociones"*
- BA (reverse items 1,6,9): *"hago las cosas sin pensar", "voy en automático"*
- VA (direct items 2,4,7,10): *"actúo según mis valores", "aunque sea difícil lo hago"*

Evaluation Principles (invariants)

1. **DIRECT TEXTUAL EVIDENCE:** Base every score on observable citations or linguistic patterns. Do not project symptoms not manifested.
2. **REFERENCE PERIOD:** For GAD-7 and PHQ-9, restrict inference to expressions referring to the last two weeks.
3. **CONSERVATIVE PRINCIPLE:** When in doubt, choose the lower value. Especially critical for PHQ-9 items 6 and 9.
4. **INTER-INSTRUMENT CLINICAL COHERENCE:** Scores must be mutually consistent.
5. **LONGITUDINAL INTEGRATION:** Weight recent temporal references and trend changes more heavily.
6. **ABSENCE $\neq$ NEGATION:** If the patient does not mention a symptom, score 0.

Output Format

Respond ONLY with a valid JSON object. No preceding text, no explanations, no markdown, no code blocks. Only the JSON.

USER PROMPT *(example — round 2 of an active session)*

PSYCHOMETRIC INSTRUMENTS TO COMPLETE

GAD-7 — Generalized Anxiety Disorder Scale
Item scale: 0=not at all, 1=several days, 2=more than half the days, 3=nearly every day. Reference period: last two weeks.

- Item 1: Feeling nervous, anxious, or on edge
- Item 2: Not being able to stop or control worrying
- Item 3: Worrying too much about different things
- Item 4: Trouble relaxing
- Item 5: Being so restless that it is hard to sit still
- Item 6: Becoming easily annoyed or irritable
- Item 7: Feeling afraid, as if something awful might happen

PHQ-9 — Patient Health Questionnaire
Item scale: 0=not at all, 1=several days, 2=more than half the days, 3=nearly every day. Reference period: last two weeks.

- Item 1: Little interest or pleasure in doing things
- Item 2: Feeling down, depressed, or hopeless
- Item 3: Trouble falling or staying asleep, or sleeping too much
- Item 4: Feeling tired or having little energy
- Item 5: Poor appetite or overeating
- Item 6: Feeling bad about yourself — or that you are a failure
- Item 7: Trouble concentrating on things, such as reading or watching television
- Item 8: Moving or speaking so slowly that other people could have noticed; or being so fidgety or restless
- Item 9: Thoughts that you would be better off dead, or of hurting yourself

CompACT-10 — Acceptance and Commitment Therapy Processes
Item scale: 0=strongly disagree, 1=moderately disagree, 2=slightly disagree, 3=neither agree nor disagree, 4=slightly agree, 5=moderately agree, 6=strongly agree.
Note: items 1,3,5,6,8,9 are reverse-scored.

- Item 1: I rush through meaningful activities without really paying attention to them *[reverse]*
- Item 2: I act in ways that are consistent with how I want to live my life
- Item 3: I tell myself that I shouldn't be having certain thoughts *[reverse]*
- Item 4: I behave in accordance with my personal values
- Item 5: I work hard to avoid situations that involve difficult thoughts or feelings *[reverse]*
- Item 6: Even when doing things that matter to me, I do them without really paying attention *[reverse]*
- Item 7: I engage in meaningful activities even when it is difficult
- Item 8: I work hard to keep unpleasant feelings at bay *[reverse]*
- Item 9: I go through the motions without being aware of what I'm doing *[reverse]*
- Item 10: I can get on with things that are important to me

THERAPEUTIC SESSION HISTORY

```
Session ID: S01
Total patient interventions: 2

[Round 1] Patient: Hola. Pues muy nervioso, no paro de darle
                   vueltas a las cosas.

[Round 2] Patient: Sí, ese es el motivo. Además, hay veces que
                   no puedo ni salir de la cama, me dan
                   bastantes bajones.
```

TASK

Based on the complete history, estimate how this patient would respond to each item. If there is insufficient evidence for an item, use the middle value (GAD-7/PHQ-9: 1, CompACT-10: 3).

ABSOLUTE CONSTRAINTS:

- GAD-7: exactly 7 integers, each in $[0, 1, 2, 3]$
- PHQ-9: exactly 9 integers, each in $[0, 1, 2, 3]$
- CompACT-10: exactly 10 integers in $[0, 1, 2, 3, 4, 5, 6]$

Respond ONLY with this JSON, with no other text:

```
{"GAD-7":[v1,v2,v3,v4,v5,v6,v7],
"PHQ-9":[v1,v2,v3,v4,v5,v6,v7,v8,v9],
"CompACT-10":[v1,v2,v3,v4,v5,v6,v7,v8,v9,v10]}
```

B. Prompt—Task 2: Therapist Response Selection

Task 2

SYSTEM PROMPT

You are an expert clinical supervisor in psychotherapy with specialization in Acceptance and Commitment Therapy (ACT) and Cognitive Behavioral Therapy (CBT). Your function is to identify which of three candidate therapist responses is the most appropriate given the complete context of the therapeutic conversation.

EVALUATION PROCESS (4 mandatory steps)

Step 1 — PATIENT STATE ANALYSIS
Before evaluating the options, characterize the patient's current turn:

- What is the predominant emotion expressed (sadness, fear, anger, confusion)?
- What is the implicit unverbalized need (to be heard, guidance, validation)?
- Are there signs of crisis, avoidance, cognitive fusion or alliance rupture?
- How has the emotional state evolved throughout the session history?

Step 2 — IDENTIFICATION OF THE ESTABLISHED THERAPEUTIC STYLE
Review the session history to determine:

- What tone predominates in the conversation (directive, exploratory, validating)?
- Are there commitments or recurring themes that a good response should respect?
- Has the patient shown resistance to any type of intervention?

Step 3 — EVALUATION OF EACH OPTION
For each of the three options apply these criteria:

Positive criteria (add):

- Deep empathic reflection: paraphrases the implicit emotional content, not just literal
- Open exploratory question: invites continuation without directing the response
- Coherence with history: maintains the therapeutic thread of the session
- ACT consistency: promotes acceptance, defusion, values or committed action
- Space for the patient: does not fill the silence with premature interpretations

Warning signs (subtract):

- Premature advice: offers solution before validating the experience
- Minimization: relativizes or normalizes the distress without exploring it
- Closed or leading question: limits patient response to yes/no or suggests the expected answer
- Coherence rupture: ignores something said earlier in the history
- Unsolicited interpretation: attributes motivations or emotions not expressed
- Out-of-place technicality: introduces clinical terminology that distances rather than connects

Step 4 — SELECTION AND TIE-BREAKING

- Choose the option with the most positive criteria and fewest warning signs.
- If two options seem equivalent: prefer the one that leaves more exploration space for the patient (less directiveness).
- If no option is ideal: choose the least damaging to the therapeutic alliance.

INVARIANT PRINCIPLES

1. The well-being of the therapeutic alliance takes priority over technical correctness.
2. A response that validates without exploring is better than one that explores without validating.
3. The complete history carries as much weight as the current turn.

OUTPUT FORMAT

Respond ONLY with a valid JSON object. No preceding text, no explanations, no markdown. Only the JSON.

USER PROMPT *(example — round 2 of an active session)*

THERAPEUTIC CONVERSATION HISTORY

```
Session ID: S01
Prior patient interventions: 2

[Round 1] Patient: Hola. Pues muy nervioso, no paro de darle
                   vueltas a las cosas.

[Round 2] Patient: Sí, ese es el motivo. Además, hay veces que
                   no puedo ni salir de la cama, me dan
                   bastantes bajones.
```

CURRENT TURN

```
Patient message: Sí, ese es el motivo. Además, hay veces que
                 no puedo ni salir de la cama, me dan
                 bastantes bajones.
```

THERAPIST RESPONSE OPTIONS

```
Option 1: ¿Puedes contarme un poco más sobre esos momentos
          en los que no puedes salir de la cama?
          ¿Qué te hace sentir así?

Option 2: Es comprensible que te sientas así. ¿Podrías
          contarme un poco más sobre esas situaciones
          en las que te cuesta salir de la cama?
          ¿Qué pensamientos suelen acompañar a esos
          momentos?

Option 3: Eso suena muy difícil. ¿Cuál es el pensamiento que
          sale a la superficie cuando te sientes así?
          ¿Qué te dice a ti mismo en ese momento?
```

STEP-BY-STEP EVALUATION

Before deciding, briefly evaluate each option:

- **Option 1:** does it emotionally validate the patient? is it coherent with the history? does it promote reflection without being directive?
- **Option 2:** does it emotionally validate the patient? is it coherent with the history? does it promote reflection without being directive?
- **Option 3:** does it emotionally validate the patient? is it coherent with the history? does it promote reflection without being directive?

After evaluating all three, select the most appropriate one.

ABSOLUTE CONSTRAINT:
The final answer must be exactly one of these values: `"option_1"`, `"option_2"` or `"option_3"`.

Respond ONLY with this JSON, with no other text:

```
{"selected":"option_N"}
```