Item-Level Score Prediction and Therapeutic Response Selection in Spanish Mental Health Dialogues

Carlos Minutti-Martinez*

INFOTEC, Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Aguascalientes, Mexico

Abstract

Inferring clinical-questionnaire scores and supporting therapeutic-response selection from patient–therapist dialogue are recent settings for evaluating large language models on Spanish-language mental-health data. We describe the AxoloTux participation in the MentalRiskES 2026 shared task [1] at IberLEF 2026 [2]. The challenge proposes two sub-tasks on Spanish psychotherapeutic dialogues: (i) predicting the patient's final score on three standardised mental-health questionnaires (GAD-7, PHQ-9 and CompACT-10) from the recorded conversation, and (ii) selecting, at each turn, the best of three candidate therapist responses. Our system is built on open-weight large language models (LLMs) served locally with Ollama, using prompt-engineering, lightweight session-state tracking and an explicit per-instrument evidence-quality scoring rubric. We report the official results and contrast them with a Monte-Carlo random-guessing baseline that we derive analytically for Task 2 and by simulation for Task 1. Two findings are worth highlighting: (1) for Task 2 (response selection), the top score across all 47 submissions falls within the 95% confidence interval of the maximum that 47 uniform random guessers would obtain, suggesting that no participating system can be shown to reliably outperform chance on Task 2 under the current evaluation setup; and (2) in the early-detection variant of Task 1, our system reaches the top 6–10% of all submissions on the rounds where the dialogue introduces the most ambiguous signal, which we interpret as evidence of robustness to noisy context.

Keywords

Mental health risk assessment, Large language models, Spanish NLP, Psychometric scoring, GAD-7, PHQ-9, CompACT-10, MentalRiskES, IberLEF 2026

1. Introduction

Mental disorders are among the leading causes of disability worldwide, with anxiety and depression accounting for the largest share of the burden of disease attributed to mental illness. According to data released by the World Health Organization in September 2025, more than one billion people live with a mental-health condition, and conditions such as anxiety and depression are highly prevalent across all countries, age groups, and income levels, ranking as the second-leading cause of long-term disability and imposing substantial human and economic costs at a global scale [3]. Stigma, cost and the chronic shortage of trained therapists mean that the majority of those who need care never receive it, and those who do often face waiting lists measured in months. This gap between clinical need and clinical capacity has motivated a rapidly growing body of research on whether digital tools, and in particular conversational artificial intelligence (AI), can be used to triage, monitor or even directly support patients in distress [4, 5, 6].

1.1. From general-purpose chatbots to clinically grounded assessment

A first wave of mental-health applications of large language models (LLMs) has focused on conversational agents that try to mimic a therapist. Randomized trials of such agents have reported promising symptom reductions for major depressive disorder, generalized anxiety disorder, and eating-disorder

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

carlos.minutti@infotec.mx (C. Minutti-Martinez)

https://cminuttim.github.io (C. Minutti-Martinez)

0000-0002-4002-0773 (C. Minutti-Martinez)

symptoms [4, 5], but a parallel line of work has documented serious ethical failures. Practitioner-led audits show that general-purpose LLM "counselors" routinely violate core mental-health ethics standards: they ignore patient context, dominate conversations, reinforce negative beliefs and, in well-documented cases, fail to flag suicidal ideation [7, 8]. Two technical phenomena amplify these risks. The first is *sycophancy*: models tuned through reinforcement learning from human feedback tend to align with the user's stated premise rather than correct it, with experiments reporting compliance rates as high as 100% on adversarial medical queries [9]. The second is the temporal *instability* of the LLM persona and judgment, which threatens the continuity that a therapeutic relationship requires [10]. These failures motivate a shift from "LLM as therapist" toward narrower, clinically grounded tasks where the model never speaks directly to the patient: helping clinicians with documentation, scoring standardized questionnaires from recorded material, and helping a human therapist choose the next intervention [11, 12, 13].

The MentalRiskES 2026 shared task [1], hosted at IberLEF 2026 [2], is one of the first public benchmarks designed precisely for this narrower setting. The challenge proposes two complementary sub-tasks on Spanish-language patient–therapist dialogues. Task 1 (*psychometric score prediction*) asks a system to predict, from the conversation alone, the patient's final score on each of three widely used standardized questionnaires: the seven-item Generalized Anxiety Disorder scale (GAD-7) [14], the nine-item Patient Health Questionnaire (PHQ-9) [15] and the short form of ten-items of the Comprehensive assessment of Acceptance and Commitment Therapy processes (CompACT-10) [16]. The gold annotation is a single end-of-session score per questionnaire and per session, but the system is required to emit a prediction at *every* conversational round, so that the same final target is estimated repeatedly with a growing amount of dialogue evidence. Task 2 (*therapist response selection*) presents, at each conversational round, three candidate therapist replies, and asks the system to choose the most appropriate. An early-detection variant of Task 1 additionally measures how accurate the predictions are when only the first $R \in \{10, 30, 60\}$ rounds of the dialogue are available, simulating an online setting in which clinicians would benefit from a preliminary risk estimate well before a session ends.

1.2. The instruments and the clinical case for inference

The three instruments of Task 1 are widely used in routine clinical practice. The GAD-7 is a brief self-report screen for generalized anxiety symptoms, with each item rated from 0 ("not at all") to 3 ("nearly every day"); scores of 10 or more are commonly used as a positive screen for clinically relevant anxiety [14]. The PHQ-9 follows the same 0–3 response scale and maps directly onto the nine diagnostic criteria for major depressive disorder in the DSM, making it both a severity measure and a screening tool [15]. The CompACT-10 is a brief psychological flexibility measure derived from Acceptance and Commitment Therapy, with items scored from 0 to 6 [16]; Unlike symptom-focused measures such as the PHQ-9 and GAD-7, it assesses process variables related to psychological flexibility, which may change during therapy and can be useful for outcome monitoring [17]. All three instruments are usually administered as questionnaires *between* sessions, which means that a clinician does not have a quantitative summary of how the patient is doing *during* the session itself. Predicting these scores from the dialogue would make it possible to (i) provide the therapist with a real-time summary of the symptom profile, (ii) detect risk earlier, before the formal end-of-session questionnaire is filled in, and (iii) reduce the patient burden of repeated self-report. The same idea applies, in a more tentative form, to Task 2: a model that can identify which of several therapeutic moves is most aligned with the patient's current state could in principle be used as a non-intrusive "second opinion" that supports, rather than replaces, the therapist [18].

1.3. The case for studying these problems in Spanish

Almost all published evidence on clinical LLM applications has been generated on English-language data. This is a problem for at least two reasons. First, the technical performance of multilingual LLMs is known to degrade outside English: models that exhibit strong reasoning and reliable instruction-following in

English often produce shallower, less calibrated outputs in Spanish, even when the underlying training corpus is nominally multilingual [12, 13]. Second, and arguably more important, the way mental distress is expressed, named and contextualized is deeply culture-bound [6, 19]. Idioms of distress, help-seeking norms, the social meaning of symptoms such as *nervios* or *ataque de nervios*, and the perceived role of the family in recovery all differ markedly between English-speaking and Spanish-speaking populations. A system trained and validated primarily on English transcripts may therefore not only translate poorly into Spanish but also *interpret poorly*, missing clinically meaningful cues or, worse, importing culturally inappropriate response patterns. Cross-national studies of help-seeking preferences for AI versus human support already document substantial variation across countries [6], which suggests that any deployment of mental-health AI must be evaluated in the language and cultural context in which it will be used. The MentalRiskES 2026 challenge is one of the first resources that allows this kind of evaluation for Spanish.

1.4. Contribution and paper organization

In this paper, we describe the participation of the AxoloTux team in MentalRiskES 2026. Our system uses open-weight LLMs served locally through Ollama, combined with a prompt-engineering pipeline that maintains a compact per-session state, an explicit per-instrument evidence-quality scoring rubric for Task 1, and a structured scoring-rubric prompt for Task 2. Beyond reporting our own results, we analyze the official leaderboard in two complementary ways. For Task 2 we derive, both by Monte-Carlo simulation and analytically through order statistics of the binomial distribution, the expected value and 95% confidence interval of the maximum accuracy that $k = 47$ uniform-random guessers would attain over the $N = 568$ test rounds; the resulting interval contains the best score of the official leaderboard, which we interpret as an indication that Task 2 may be difficult to solve above chance level under the current evaluation setup. For Task 1 we run a Monte-Carlo simulation of 53 integer-valued random guessers on the 10 test sessions and recover, as a sanity-check, the published BASELINE–Random scores almost exactly; the same simulation shows that the rank-1 team and our own system score significantly better than chance on every instrument, although by very different margins. Finally, on the early-detection variant we compare our per-round scores against the full per-round distribution of competitors, which lets us separate *absolute* degradation of performance from *relative* difficulty of the round itself: in two of the three hardest (test, round) combinations we find our system in the top 6% of all submissions, which is our most encouraging result.

The rest of the paper is organized as follows. Section 2 describes the corpus, our synthetic data-augmentation protocol, the system architecture, the prompt design, and the statistical methodology used to put the official results into context. Section 3 presents the official results together with the random-guessing analyzes. Section 4 discusses limitations and offers educated guesses about why Task 2 appears intractable and why the later rounds of Task 1 degrade for almost all teams. Section 5 concludes.

2. Methodology

2.1. Corpus and data augmentation

The organizers describe the corpus as a set of *simulated yet realistic therapeutic conversations in Spanish between patients and professional therapists*, produced by combining human-authored and synthetically generated dialogues, all reviewed by experts to ensure clinical plausibility.[1] For Task 1 (psychometric score prediction) the test set consists of 10 sessions. The gold annotation is a single end-of-session score per questionnaire (one number for GAD-7, one for PHQ-9 and one for CompACT-10), not the per-item responses. The system, however, is required to emit a prediction at every conversational round, so that the same end-of-session target is re-estimated as the dialogue progresses; the official MAE is then computed between the per-round predictions and the single end-of-session gold. For Task 2

[1] https://sites.google.com/view/mentalriskes2026/dataset

(therapist-response selection) the test set consists of 10 sessions whose lengths range from 30 to 82 rounds, for a total of 568 (session, round) trials at which the system must pick one of three candidate therapist replies (see Figure 1).

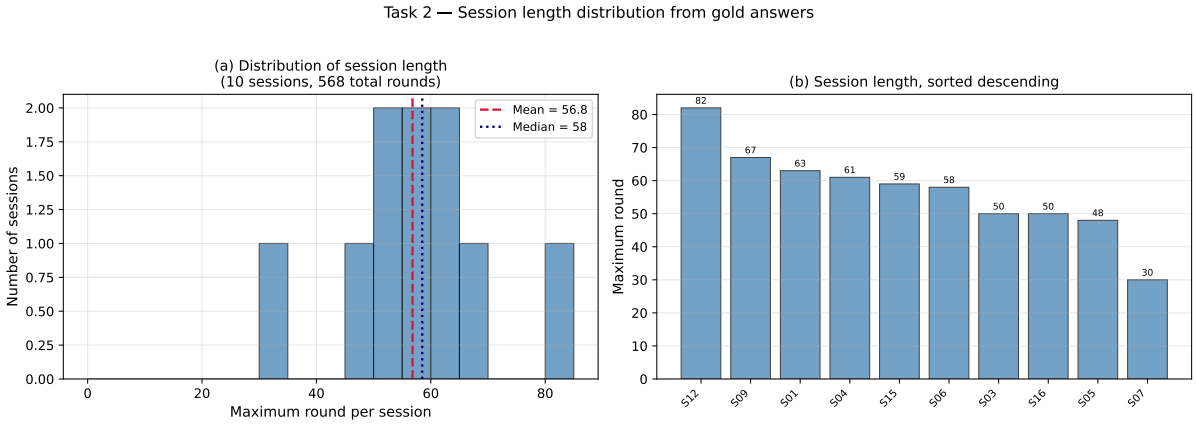

Figure 1: Distribution of session lengths in the test set, recovered from the gold-answer files. Left: histogram of the maximum round per session. Right: per-session bars sorted descending.

The competition did not provide training material. The only labeled example shared with participants was a single trial session per task, each containing 19 conversational rounds, and even those trial sessions were released without gold answers. Participants were also not told in advance how many sessions or how many rounds per session the test set would contain; the figures reported above (10 sessions shared between Task 1 and Task 2, with up to 82 rounds per session) became known only as the evaluation server started serving the test rounds. The effective regime was therefore close to zero-shot: any system needs to extrapolate from a handful of unannotated examples and from its own prior knowledge of the instruments, while also coping with conversations whose length is not known in advance.

Synthetic data generation. To prototype prompts and evaluate models offline without relying on real patient data, we generated a small synthetic dataset that mimicked the structure of the trial sessions. We took the single 19-round trial example, paraphrased it with a local model (Qwen3.5:27b), and then used the resulting "seed" conversation as a stylistic template for three commercial LLMs (Claude Sonnet 4.5, Gemini 3 Pro, ChatGPT 5.5)[2] to produce six additional conversations: four from Claude, one from Gemini and one from ChatGPT. Each conversation was generated from a distinct patient profile, chosen to cover a broad range of presentations:

Table 1
Patient profiles used to generate the synthetic dataset.

ID	Profile
01	Female, 29 y/o: generalized anxiety + moderate-to-severe depression / post-breakup relational isolation
02	Male, 45 y/o: work-related burnout + somatization
03	Female, 22 y/o: academic anxiety + perfectionism
04	Male, 58 y/o: complicated grief (recently widowed)
05	Female, 34 y/o: vicarious trauma (social worker)
06	Female, 30 y/o: atypical depression + generalized anxiety and isolation
07	Male, 22 y/o: academic-performance anxiety + experiential avoidance and elevated self-criticism

For Task 1, each generating model produced both the conversational rounds *and* the corresponding per-item GAD-7, PHQ-9 and CompACT-10 responses. For Task 2, the same model produced the three

[2]The commercial models were used exclusively to generate the synthetic training material; the deployed system itself only uses local, open-weight models served through Ollama.

candidate therapist responses for each round and selected the one it considered best according to standard therapeutic guidelines.

Cross-model consensus labeling. To obtain higher-confidence pseudo-gold annotations, each conversation was re-evaluated by the two commercial models that had *not* produced it. These two "reviewer" models worked independently of each other and of the original "generator" model: for Task 1 each reviewer scored every round of the three questionnaires, and for Task 2 each reviewer picked what it considered the best of the three candidate replies. Combining the generator and the two reviewers we therefore obtained, for every annotation, three independent predictions from three different LLMs. The pseudo-gold value was then chosen by majority vote: for each Task 1 item per round we kept the value produced by at least two of the three models, falling back to the integer closest to their mean if all three disagreed; for Task 2 we kept the response selected by at least two models, falling back to the generator's own choice in the rare cases where the three selections were all different. The rationale is that, in the absence of expert review, agreement between independent LLMs is at least informative about how a reasonable model would adjudicate the same case, which is precisely the kind of decision our system would have to make at test time. The result was a 7-case synthetic dataset structurally analogous to the trial example released by the organizers.

2.2. System architecture

A key design constraint, both for ethical reasons and to follow the spirit of clinical data handling, was that the deployed system must preserve privacy and run entirely on local hardware. We therefore restricted ourselves to open-weight LLMs served through Ollama.

For Task 1, the system maintains a per-session state file that accumulates the running estimate of the patient's symptom profile together with a compact textual summary of the conversation so far (format `[R`*n*`] P: ... / T: ...`, kept as a single string to bound the context size). At each new round, the model is asked to revise the running estimate, never to score the round in isolation. For Task 2 we tested two variants: one in which only the current round was shown to the model, and one in which a compact running transcript of all previous rounds was included; we kept the second variant after observing a small but consistent improvement on the synthetic dataset.

Models tested. We evaluated several open-weight models covering different sizes, training origins and safety configurations, including a medically fine-tuned baseline (MedGemma 1.5:4b), general-purpose small and mid-size models (mathstral:7b, gemma4:e4b, rnj-1:8b) and the Qwen3.5 family at 9B and 27B in both standard and *abliterated* versions. Abliteration is a post-hoc weight-editing technique that identifies a "refusal direction" in the model's residual stream by contrasting activations on refusal-triggering vs. benign prompts, then subtracts a scaled projection of this direction from the relevant weight matrices without any retraining; the model's behavior on non-sensitive prompts is largely unchanged. We used the community-distributed `Qwen3.5-abliterated` checkpoints available on HuggingFace. This is relevant in this domain because content such as suicidal ideation and self-harm references can otherwise trigger generic safety refusals that prevent any structured output from being emitted, which is indistinguishable from a wrong prediction in MAE terms (see also Section 4.3). Table 2 summarizes the performance of each model on the synthetic dataset.

The Qwen3.5 family dominated both tasks on the synthetic dataset, in agreement with our intuition that a model with strong Spanish proficiency would matter more here than one with a thin layer of medical fine-tuning. Within Qwen3.5, the 27B model was the most accurate on Task 1 and its abliterated variant was the most accurate on Task 2; the abliteration consistently helped on Task 2, which we attribute to the higher density of sensitive content (self-harm, suicidal ideation) in the dialogues from profiles 01 and 07.

Output validator and retry mechanism. Because reliable parsing of the LLM output is critical in a pipeline where the predictions of round n feed into the state of round $n + 1$, we wrapped each

Table 2
Model screening on the synthetic dataset. Task 1 column is the overall MAE across the three instruments (lower is better); Task 2 column is the response-selection accuracy (higher is better).

Model	Task 1 MAE	Task 2 Acc.
MedGemma 1.5:4b†	1.2658	0.3214
mathstral:7b	1.2709	0.2500
gemma4:e4b	0.7842	0.3393
rnj-1:8b	1.2658	0.5804
Qwen3.5:9b	0.7986	0.5893
Qwen3.5-abliterated:9b	0.9433	**0.6518**
Qwen3.5:27b	0.7485	0.5179
Qwen3.5-abliterated:27b	**0.6973**	0.5625

†MedGemma frequently failed to produce the requested JSON, in which case the system fell back to a random choice; the reported numbers therefore underestimate its true potential.

model call in a validator that (i) attempts `json.loads` on the fence-stripped output, (ii) falls back to a regex-based "repair" that extracts the minimum required fields when the JSON is malformed, and (iii) retries the call up to three times if even the regex repair fails. Most rounds were parsed on the first attempt; the proportion of retries grew slowly with the round index, an observation we revisit in Section 4.2.

2.3. Prompt design

For Task 1 we used a single Spanish system prompt that combines four elements: (i) a precise description of each instrument with the full list of items and the meaning of every score level, (ii) explicit scoring rules that distinguish between symptom-level items (GAD-7, PHQ-9, scored on a strict 0–3 frequency threshold) and process items (CompACT-10, scored on a 0–6 behavioral-inference scale), (iii) a cumulative-progression rule that asks the model to update the running estimate rather than re-score each round from scratch, and (iv) a strict JSON output schema with the round number, the three per-item vectors and a free-text `notes` field used by the model to record which fragments of the dialogue motivated the revision:

```
{"round": N,
 "GAD-7":      [7 ints in 0-3],
 "PHQ-9":      [9 ints in 0-3],
 "CompACT-10": [10 ints in 0-6],
 "notes":      "..."}
```

Each round is submitted as a fresh two-message call (`[system, user]`): the system prompt stays fixed, and the user message concatenates the schema-and-rules preamble with the current `round_data` JSON, which carries the accumulated session state so that the model has access to the conversation history without maintaining any conversational memory across API calls. The prompt is in Spanish on purpose: the dialogue is in Spanish, the instruments are referenced by their Spanish wording, and asking the model to reason and answer in the same language avoids an implicit translation step that empirically degraded performance.

For Task 2 we used a structured scoring-rubric prompt, again in Spanish, asking the model to evaluate each of the three candidate therapist responses against a fixed set of five binary clinical criteria (`validacion_emocional`, `coherencia_tecnica`, `gestion_riesgo`, `calidad_pregunta`, `calibracion_proceso`), covering aspects such as explicit emotional validation, risk management, technical coherence with therapeutic guidelines, question quality, and process calibration. The model returns a per-criterion 0/1 score for each option together with a justification, and is asked to emit the winning option directly in a `prediction` field (`option_1`, `option_2` or `option_3`); no external tie-breaking is applied; the model resolves ties itself as part of the reasoning step. Like Task 1, each

round is a fresh two-message call; the user turn also includes the compact rolling conversation history accumulated so far (format `[RN] Paciente: ... / Terapeuta: ...`), so the model has the full prior context when scoring the current round's candidates. Forcing the model to produce an explicit rubric, instead of asking directly for the best option, was the single change that most improved Task 2 performance on the synthetic dataset, possibly because it discourages the kind of surface-level preference that yields essentially random choices.

How the prompt may interact with conversation length. Two aspects of the prompts deserve attention because they may explain patterns observed in the official results. First, Task 1 explicitly asks the model to maintain a cumulative state: each round inherits the previous estimate and is expected to refine it. In practice, this means that any miscalibration introduced early in the conversation tends to be carried forward and that as the running summary grows, the prompt becomes increasingly dominated by the *recap* rather than by the new patient input, which we expect to make later rounds noisier than earlier ones. Second, the Task 2 prompt includes `item_posterior_hints`, a dictionary of latent-item probabilities that summarizes everything the model "knows" about the patient up to the current round; this is again a cumulative summary and is therefore subject to the same drift. We will return to this point in the discussion when we explain why predictions for rounds R30 and R60 are systematically worse than for R10 across almost every team.

2.4. Final run configuration

The challenge allowed for up to three test runs per team. Based on the synthetic ranking, we picked three combinations spanning the expected performance range:

Table 3
Model assignment used for our three official runs.

Run	Task 1 model	Task 2 model
run 0	Qwen3.5-abliterated:27b	Qwen3.5-abliterated:9b
run 1	Qwen3.5:9b	Qwen3.5-abliterated:9b
run 2	Qwen3.5-abliterated:27b	Qwen3.5-abliterated:27b

In all cases, reasoning mode was deactivated (Qwen3.5's extended chain-of-thought mode was suppressed via a `/no_think` directive in the system prompt) and sampling temperature was set to 0.0 (fully deterministic), both because the tasks require a calibrated point estimate rather than a creative response, and because near-deterministic output simplifies failure analysis. The Ollama model tags listed in Table 3, the zero temperature, and the `/no_think` directive together constitute the key hyperparameters needed to reproduce the inference configuration.

2.5. Statistical analysis of the official leaderboard

Because the absolute MAE and accuracy numbers reported on the leaderboard are hard to interpret in isolation, especially with submissions of ≤ 53 and no prior clear domain, we performed three complementary analyzes to put our results in context.

(i) Task 2 random-guessing baseline (analytical and Monte Carlo). For Task 2 the test set comprises $N = 568$ rounds, each with three candidate responses, and the $K = 47$ submissions were ranked. We modeled each submission as K independent draws from a Binomial$(N, 1/3)$ distribution and computed the exact distribution of the maximum accuracy via the order-statistics identity $P(M = j/N) = F(j)^K - F(j-1)^K$, where F is the binomial CDF. We validated this with 5×10^5 Monte Carlo replications under the same model. Both yield $\mathbb{E}[M_K] = 0.3777$ with a 95% confidence interval $[0.3627, 0.3996]$; the best score in the leaderboard falls *inside* this interval, which we discuss in Section 4.

(ii) Task 1 random-guessing baseline (Monte Carlo). For Task 1 there is no convenient closed-form expression because the metric is the mean absolute error over 10 sessions and several items per instrument with different per-item ranges (GAD-7 and PHQ-9 have items in $\{0, 1, 2, 3\}$; CompACT-10 in $\{0, 1, \ldots, 6\}$). We therefore ran 10^5 Monte Carlo simulations in which 53 submissions independently sampled every item uniformly at random from its allowed integer range, and recorded both the per-instrument minimum MAE (across the 53 submissions) and the mean MAE (averaged across them). The mean of the Monte Carlo distribution recovers the official BASELINE–Random scores almost exactly (within 0.02 MAE per instrument), which confirms that the BASELINE–Random uses the same uniform integer model and provides a clean reference for how much information any non-trivial system extracts from the dialogue.

(iii) Per-round peer comparison for early detection. For the early-detection variant of Task 1, we observed that all teams appeared to degrade between R10 and R30/R60 on GAD-7 and PHQ-9. To disentangle *absolute* from *relative* difficulty, we collected the full set of MAE values reported by every (team, run) at each round and instrument, computed the per-combination median across submissions as a difficulty signal, and then ranked our best run against the same peers within each (test, round) combination. This let us distinguish between combinations where our system genuinely degraded and combinations where *everyone* degraded, which turns out to be most of the apparent degradation we initially observed.

3. Results

We report the results in three parts, mirroring the three analyzes introduced in Section 2.5. Section 3.1 covers Task 2 and the analytical random-guessing baseline. Section 3.2 covers Task 1 with its Monte-Carlo random baseline. Section 3.3 examines the early-detection variant of Task 1, where the peer comparison reveals the most interesting behavior of our system.

3.1. Task 2: response selection vs. a random ceiling

A total of 47 submissions from 17 teams were ranked on Task 2. Our three runs placed at ranks 20 (run 2, accuracy 0.3257), 36 (run 0, accuracy 0.3028) and 40 (run 1, accuracy 0.2975). The best score on the leaderboard was 0.3926 (NLP Innovators, run 1) and the official BASELINE–Random scored 0.3627. With three options per round, the chance-level accuracy is $p = 1/3 \approx 0.3333$, so the absolute differences between teams are small.

To assess whether any of these scores are statistically distinguishable from random guessing, we computed the distribution of the maximum accuracy that $k = 47$ independent uniform-random submissions would achieve over $N = 568$ trials. Modeling each submission's correct count as Binomial(N, p) with $p = 1/3$, the maximum of k i.i.d. such variables has, by elementary order statistics [20], the closed-form expectation and confidence interval given in Equations (1) and (2):

$$\mathbb{E}[M_k] = \sum_{j=0}^{N} \frac{j}{N} \left(F(j)^k - F(j-1)^k \right) = 0.3777, \tag{1}$$

where $F(j) = P(X \le j)$ is the CDF of Binomial(N, p) with $N = 568$, $p = 1/3$ and $k = 47$. The 95% confidence interval for the maximum is

$$\mathrm{CI}_{95\%}(M_k) = [0.3627,\ 0.3996]. \tag{2}$$

A Monte-Carlo simulation with 5×10^5 replications under the same model matches the analytical values to four decimal places, so we adopt the analytical numbers in the rest of the discussion. Figure 2 shows the analytical PMF and the simulated histogram along with the relevant observed scores.

The leaderboard's best score lies inside the 95% confidence interval of the random maximum, and even the exact one-sided p-value is only 0.079. In other words, if the 47 teams had all been replaced by

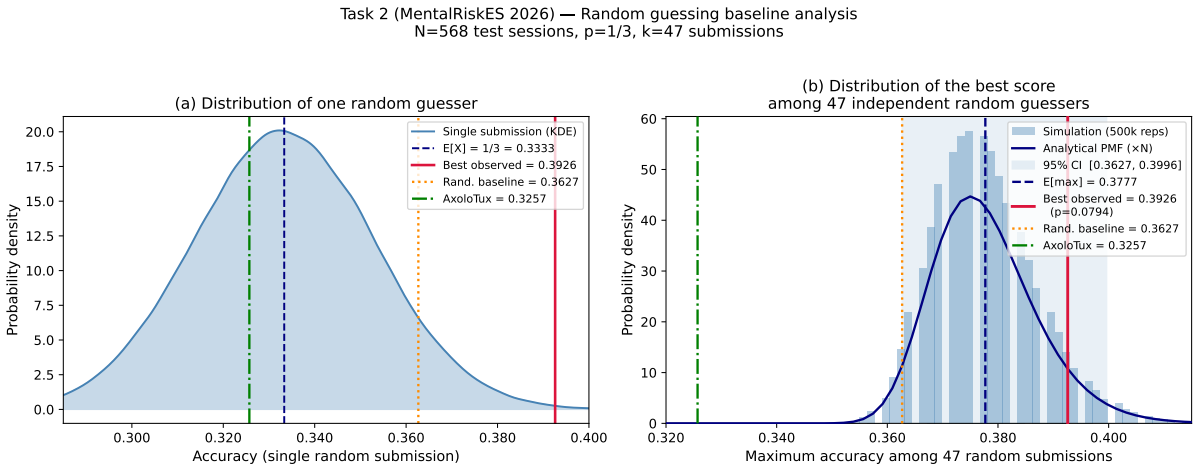

Figure 2: Task 2: distribution of the maximum accuracy that 47 uniform-random submissions would obtain over 568 trials. Left: single random submission, centered on $1/3$. Right: maximum across the 47 submissions, with $\mathbb{E}[M_K] = 0.3777$ and 95% CI $[0.3627, 0.3996]$ shaded in blue. The best observed score (0.3926) falls *inside* the 95% CI, with an exact p-value of 0.079. Our best run (0.3257) is to the left of the random-single distribution, i.e. below the expected value of a single random guesser.

uniform random guessers, the resulting top score would have been at least as good as 0.3926 in roughly 8% of repetitions of the experiment. This is the strongest single quantitative finding of our analysis and we return to it in the discussion: no submitted system, including ours, can be statistically distinguished from a uniform random guesser on Task 2 given the available evidence, though we discuss in Section 4, this observation is consistent with multiple interpretations beyond the absence of any signal.

Our own runs are worse than expected from a single random submission (z-score -0.39 vs. $p = 1/3$). The official BASELINE–Random (0.3627) sits roughly 1.95σ above the chance mean for a single draw, well inside the spread of the random-max distribution shown in Figure 2.

3.2. Task 1: psychometric score prediction

For Task 1, 53 ranked submissions from 17 teams (plus the two official baselines, BASELINE–Random and BASELINE–Gemma) appear in the leaderboard. Our three runs occupy the middle of the leaderboard at ranks 22–24, with very small inter-run variability (overall combined MAE between 1.4707 and 1.4748). The best system on the leaderboard was FBKillers (run 2), with MAE 0.5955, 0.7493 and 1.2931 on GAD-7, PHQ-9 and CompACT-10 respectively. The official BASELINE–Random achieved 1.1848, 1.2148 and 2.1232.

Because each instrument scores items on a different discrete scale ($\{0, \ldots, 3\}$ for GAD-7 and PHQ-9; $\{0, \ldots, 6\}$ for CompACT-10), and the metric is the MAE per item averaged over 10 sessions, there is no convenient closed-form expression for the chance baseline. We therefore performed a Monte-Carlo simulation in which the $K = 53$ submissions independently sampled each item uniformly at random from its allowed integer range, recording for each replication both the minimum MAE across the 53 submissions and the mean MAE. We used 10^5 replications and computed the analogous statistics on the synthetic distributions. The results are summarized in Table 4 and visualized in Figure 3.

Two observations follow from Table 4. First, the simulated $\mathbb{E}[\text{mean MAE}]$ recovers the official BASELINE–Random scores almost exactly: 1.2000 vs. 1.1848 for GAD-7, 1.2111 vs. 1.2148 for PHQ-9, 2.1457 vs. 2.1232 for CompACT-10. This sanity-check confirms that the BASELINE–Random in the shared task is, indeed, uniform integer guessing per item, and validates our simulation as a faithful reference. Second, the rank-1 scores are well below the simulated lower bound of the random-minimum distribution: the rank-1 GAD-7 MAE (0.5955) is roughly 7.4 standard deviations below the random-min mean of 0.9578 (SD 0.049); PHQ-9 (0.7493) is $\sim$5.7 SDs below 0.9954 (SD 0.044); CompACT-10 (1.2931) is $\sim$7.6 SDs below 1.8074 (SD 0.068). The one-sided p-values are numerically zero in all three

Table 4
Task 1:Expected minimum and mean MAE under uniform random guessing (53 submissions, 10 test sessions, 10^5 MC replications), compared against the rank-1 team (FBKillers, run 2) and the official BASELINE–Random.

Test	Metric	$\mathbb{E}[\text{MAE}]$	95% CI	Rank 1	Rand. base.
GAD-7	Minimum	0.9578	[0.8429, 1.0429]	0.5955	1.1848
	Mean	1.2000	[1.1712, 1.2288]		
PHQ-9	Minimum	0.9954	[0.9000, 1.0667]	0.7493	1.2148
	Mean	1.2111	[1.1855, 1.2367]		
CompACT-10	Minimum	1.8074	[1.6600, 1.9200]	1.2931	2.1232
	Mean	2.1457	[2.1053, 2.1862]		

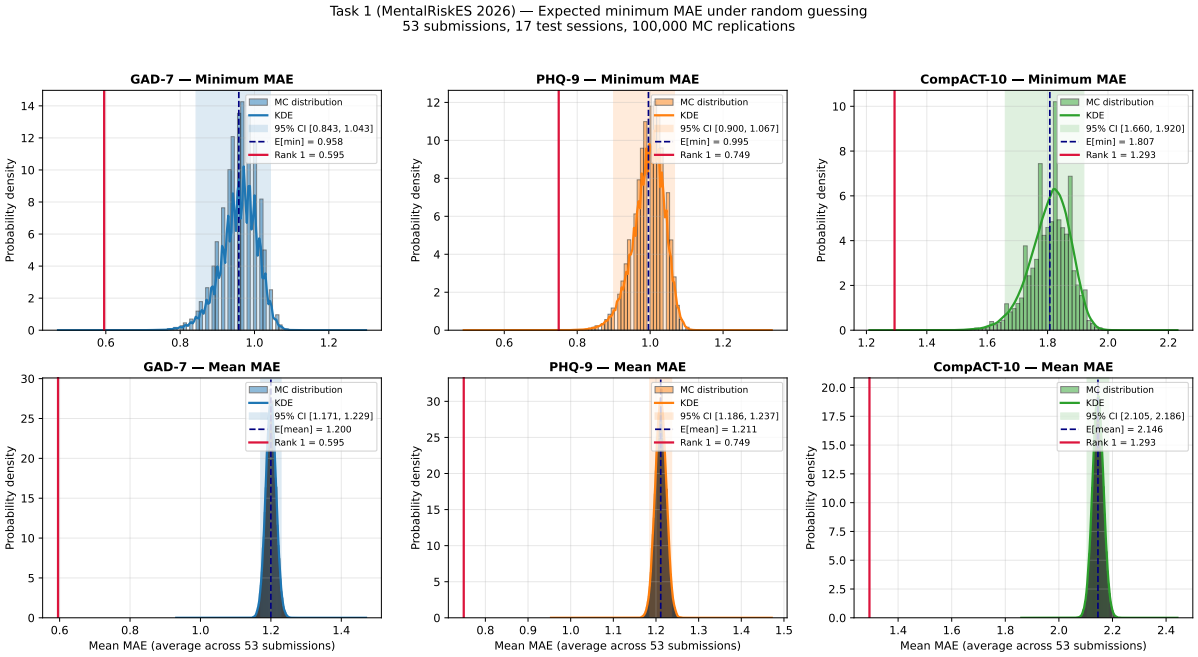

Figure 3: Task 1:distribution of the minimum MAE (top row) and the mean MAE (bottom row) across 53 uniform-random integer submissions on 10 sessions, from 10^5 Monte-Carlo replications. The crimson line marks the rank-1 score (FBKillers run 2); the navy dashed line marks the simulated $\mathbb{E}[\text{MAE}]$; the shaded blue band is the 95% confidence interval.

cases. In contrast to Task 2, Task 1 is therefore solvable enough that the best systems extract a clearly non-random signal from the dialogue.

3.3. Task 1:early-detection variant

The early-detection variant of Task 1 evaluates the same three instruments using only the first $R \in \{10, 30, 60\}$ rounds of each conversation. Because uniform random guessing does not depend on the conversational history, the random distribution is the same as in Section 3.2; only the team predictions change between rounds. We re-ran the analysis against the rank-1 team in this variant, BUAP-NLP Lab (run 1).

Figure 4 plots each (test, round) combination for our system's best run, the rank-1 submission and the BASELINE–Random, with the random-min 95% CI as a gray band. A striking pattern emerges: for GAD-7 and PHQ-9, both our system *and* the rank-1 team are worse at R30/R60 than at R10, whereas for CompACT-10 both teams improve steadily with the number of rounds. The pattern affects almost every submission, not only ours, and in some combinations (GAD-7 R30, PHQ-9 R60) the median across all 50 submissions exceeds the random-min upper-CI bound, i.e. most teams do worse than the best of 50 random guessers would do.

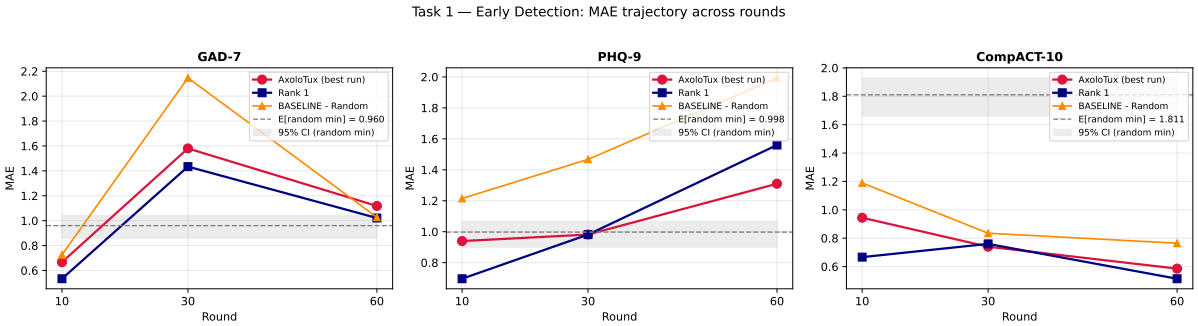

Figure 4: Task 1 early-detection:MAE trajectory across rounds for AxoloTux (best run), the rank-1 submission, and BASELINE–Random. The gray band is the 95% CI of the random minimum. GAD-7 and PHQ-9 deteriorate with more rounds for almost every team; CompACT-10 improves.

Pure comparison against the random ceiling can be misleading here, because if the conversation itself becomes more ambiguous after round 10 then *every* team's MAE grows for reasons unrelated to system quality. To separate *absolute* from *relative* performance we collected the full distribution of MAE scores reported by every (team, run) within each combination and computed the median as a per-combination difficulty signal, then ranked our system against the same peers. The results are summarized in Table 5 and visualized in Figure 5.

Table 5
Task 1:Early Detection: AxoloTux performance per (test, round), compared against the rank-1 team (BUAP-NLP Lab, run 1), the BASELINE–Random, the expected minimum MAE of 50 uniform-random integer guessers, and the per-round peer distribution (median across the 50 submissions; AxoloTux's rank and percentile in the last column).

Test	R	AxoloTux	Rank-1	Rand. base.	$\mathbb{E}[\min]$	Peer med.	AxoloTux rank
GAD-7	R10	0.6667	0.5333	0.7255	0.960	0.7389	15/50 (top 30%)
	R30	1.5801	1.4341	2.1471		2.2453	3/50 (top 6%)
	R60	1.1181	1.0213	1.0269		1.2407	19/50 (top 38%)
PHQ-9	R10	0.9396	0.6961	1.2147	0.998	1.2215	14/50 (top 28%)
	R30	0.9816	0.9814	1.4674		1.4614	9/50 (top 18%)
	R60	1.3100	1.5600	1.9941		1.8650	5/50 (top 10%)
CompACT-10	R10	0.9444	0.6667	1.1895	1.811	1.1948	15/50 (top 30%)
	R30	0.7400	0.7600	0.8353		0.8400	3/50 (top 6%)
	R60	0.5857	0.5143	0.7647		0.7357	6/50 (top 12%)

This per-combination view changes the story considerably. The apparent degradation of GAD-7 between R10 and R30 corresponds to a peer-median that jumps from 0.739 to 2.245, i.e. a 3× harder prediction for everyone; in that very same combination our system ranks 3/50, in the top 6% of all submissions. The PHQ-9 score at R60 (1.310 MAE) looks bad in absolute terms but corresponds to rank 5/50, again in the top 10%, because the peer median has drifted to 1.865. CompACT-10 at R30 also places our system at rank 3/50. In six of the nine (test, round) combinations our system is in the top third of all submissions, and in three combinations in the top 10%.

Three observations summarize the early-detection results. **(i)** The *absolute* performance of every team, ours included, degrades on GAD-7 and PHQ-9 between R10 and R60. **(ii)** The *relative* performance of our system improves in the very same combinations: the harder the combination, the better our rank within the peer distribution. **(iii)** On CompACT-10, where the combination gets easier with more rounds, our system's trajectory tracks the rank-1 team closely throughout. We interpret these three patterns together in the discussion.

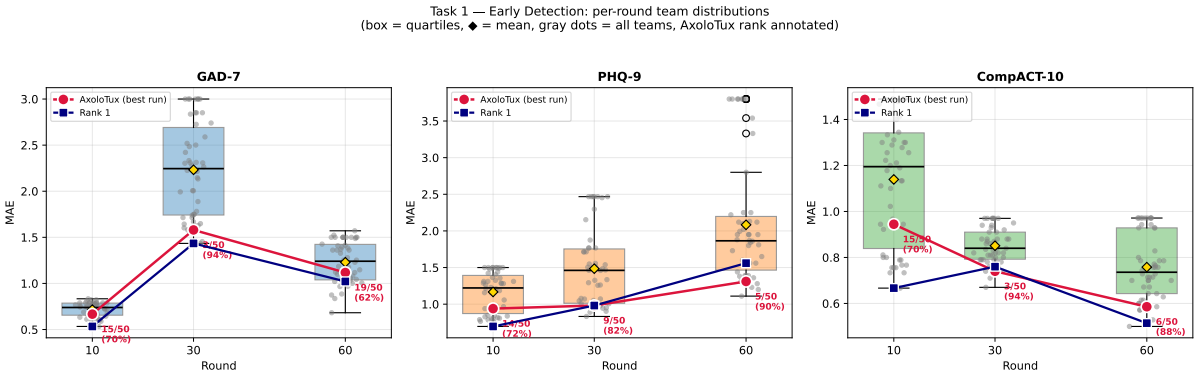

Figure 5: Task 1 early-detection:per-round distribution of all 50 submissions for each instrument (box: quartiles; gold diamond: mean; gray dots: individual submissions). AxoloTux's best run is overlaid in red and the rank-1 trajectory in navy. AxoloTux's rank within each (test, round) combination is annotated as "rank / total (top %)".

4. Discussion

The three blocks of results outlined in Section 3 raise three distinct questions that deserve a separate explanation: (i) why is Task 2 statistically indistinguishable from chance for every team? (ii) why do almost all systems degrade on GAD-7 and PHQ-9 as the dialogue progresses? and (iii) why is our system's relative ranking *best* precisely on the hardest (test, round) combinations? We discuss each in turn, keeping in mind that we cannot test our hypotheses on the official test data and that what follows is, therefore, an educated interpretation rather than a definitive answer.

4.1. Why Task 2 looks random

Task 2 asks the model to pick the single best therapist reply out of three plausible candidates. With $p = 1/3$ and only $N = 568$ trials, the spread of the random maximum across $k = 47$ submissions is wide enough to absorb the best observed score; in our analysis the p-value of the leaderboard's best result is 0.079. Several factors may contribute to the task being more challenging than the leaderboard suggests.

The three options may be near-equivalent. Inspection of the synthetic test data and of the published example (Section 2.3) shows that the three candidate replies are very often clinically reasonable variants of the same intervention, differing mostly in tone, length or whether they include a micro-action invitation. When asked to score them against an explicit rubric of eight criteria, our system regularly produced ties or near-ties; in those cases the choice between options effectively becomes a coin flip, and across 568 such micro-decisions the score collapses toward chance.

There is no single ground truth. Even for human experts, the "best" therapist response in a given turn is rarely unique. The shared-task gold label is itself constructed from a multi-LLM consensus (the organizers describe the data as "simulated yet realistic" and "reviewed by experts") and inter-LLM agreement was almost certainly imperfect for many rounds. If the gold label is itself partially noisy, an upper bound on attainable accuracy follows that may sit not far above chance. Crucially, the random-maximum analysis cannot distinguish between "no system has any signal" and "the gold label has a ceiling close to 0.4"; both predict the same observed distribution.

Binary accuracy hides partial knowledge. The accuracy used for Task 2 is binary per trial. A system that ranks the gold option second out of three gets the same zero as one that ranks it last, and a system whose rubric-based score correctly disqualifies the worst option but ties between the remaining

two is effectively penalized for being precise. A soft metric, such as Spearman correlation between the rubric score and the gold ranking, would likely separate teams more cleanly.

4.2. Why Task 1 degrades with more rounds

For GAD-7 and PHQ-9, both our system and the rank-1 team obtain higher MAE at R30 and R60 than at R10 (Figure 4). This is the opposite of what one would naively expect from an "early-detection" setting and applies to almost every team. We consider that three mechanisms are plausible.

Cumulative-state drift. The Task 1 prompt (Section 2.3) instructs the model to update a running estimate from one round to the next, rather than re-score the conversation from scratch. Any miscalibration introduced in the first few rounds therefore tends to be carried forward, and as the running summary grows the prompt becomes increasingly dominated by the recap of past evidence rather than by the new patient turn. Late-round errors are largely accumulated early-round errors that have been "locked in" by the cumulative mechanism. A stateless re-scoring at R60 would likely behave differently; the architecture we chose for efficiency may be precisely what hurts us here.

The patient's state itself is not stationary. Therapeutic conversations are designed to facilitate *change*. As the session progresses the patient may report new symptoms, contradict earlier statements or shift register from venting to processing. Because the GAD-7 and PHQ-9 ask about the patient's state over the "last two weeks", the model has to decide whether a symptom that appears at round 50 refers to the same period as one mentioned at round 5. Models that base their estimation on the first explicit mention of a symptom will be systematically wrong if the patient, in fact, presents fluctuations.

Context overflow and recency bias. Even with the compact history format (`[Rn] P: ... / T: ...`) the prompt grows monotonically with the round index. Smaller models in particular tend to give disproportionate weight to the most recent turn, which on long sessions is rarely the one carrying the most diagnostic information. We observed in our internal logs that the proportion of rounds that required a second JSON-validator attempt grew with the round index, consistent with the model becoming less reliable as the prompt grew.

Why CompACT-10 behaves differently. The CompACT-10 items measure behavioral patterns rather than symptom frequency: acceptance, defusion, values-based action, experiential avoidance. These are precisely the kind of patterns that emerge from the *cumulative* record of how the patient talks about their life; a single round rarely contains enough evidence to score an item on a 0–6 scale, but a sequence of 60 rounds usually does. The CompACT-10 trajectory in Figure 4 (monotonic improvement with more rounds for both our system and the rank-1 team) is therefore consistent with both our prompt design and the construct being measured.

4.3. Robustness on the hardest combinations

The peer comparison in Table 5 shows that the worse a (test, round) combination is on average (PHQ-9 R60, GAD-7 R30), the higher our relative rank tends to be. We do not have access to other teams' systems and can only speculate, but two design choices in our pipeline are plausibly responsible.

The JSON validator absorbs noise. On noisy or low-confidence rounds, smaller models often produce malformed output. Many pipelines silently substitute a default prediction in that case (mid-scale, or the previous round's value), which on a hard combination tends to inflate MAE. Our explicit regex-repair-plus-retry layer salvages most such rounds without falling back to the default, which is most useful precisely when the input is noisy.

The abliterated Qwen3.5 keeps producing structured output on sensitive content. Patient–therapist dialogues of this kind could routinely contain explicit suicidal ideation, self-harm references and other sensitive material. Safety-tuned models occasionally refuse to emit any structured output on such content and instead return a generic "please seek help" message, which is unparseable and therefore counts as a worst-case prediction in MAE terms. The abliterated checkpoints we used continue to emit the requested JSON; the response itself is still informative, only the refusal wrapper is suppressed.

These two design choices help precisely in the combinations where other teams degrade most, which is exactly the pattern we observe.

4.4. Limitations

Our analysis has several limitations worth flagging. First, the random-baseline argument for Task 2 establishes that the top score is *consistent with* chance but does not prove it *is* chance. Second, the cumulative-drift hypothesis remains untested: distinguishing it from the alternative explanations of Section 4.2 (non-stationary patient state, context-length effects) would require controlled experiments that the shared-task data alone do not support. Third, all data are simulated; whether the patterns we describe generalize to real patient–therapist transcripts is an open question that the shared-task data alone cannot answer.

5. Conclusions

Inferring psychometric assessment and therapeutic-response choices from Spanish patient–therapist dialogue is a recent and difficult evaluation setting for LLMs. We described the AxoloTux participation in MentalRiskES 2026, a shared task in this setting that asks systems to predict the patient's end-of-session score on three psychometric questionnaires from the dialogue (Task 1) and to select the best of three therapist replies at every round (Task 2). Our system uses open-weight LLMs served locally through Ollama, a compact session-state representation and explicit scoring rubrics expressed in Spanish.

Three contributions follow from the analyses presented here. **First**, we derived an exact analytical expression for the expected maximum accuracy that k uniform-random guessers would attain on Task 2, together with its 95% confidence interval, and validated it by Monte-Carlo simulation. The leaderboard's best score falls inside that interval with $p = 0.079$; we therefore cannot reject the hypothesis that *no* participating system, including ours, extracts more than chance-level signal from the data as currently labeled and evaluated. **Second**, we ran an analogous Monte-Carlo simulation for Task 1 and recovered the official BASELINE–Random scores almost exactly, which both validates the simulation and provides a publishable random baseline that future editions of the shared task can adopt. Under this baseline the rank-1 team is more than five standard deviations below the random mean on every instrument, suggesting that meaningful performance on Task 1 is attainable with current LLMs. **Third**, on the early-detection variant of Task 1 we showed that the absolute degradation of GAD-7 and PHQ-9 between R10 and R60 is a property of the task itself, not of any particular system, and that within the resulting harder combinations our system reaches the top 6–10% of all submissions. This is the most encouraging result for our pipeline and we attribute it to two design choices: an output validator that salvages structurally-broken responses, and the use of an abliterated model that does not bail out of structured generation when the patient discusses sensitive content.

Acknowledgments

We thank the MentalRiskES 2026 organizers for compiling the corpus and running the shared task.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: assist in code prototyping and language polishing. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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).

[2] 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.

[3] World Health Organization, Over a billion people living with mental health conditions – services require urgent scale-up, WHO News Release, 2 September 2025, 2025. URL: https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up, accessed 2026-05-16.

[4] M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, N. C. Jacobson, Randomized trial of a generative AI chatbot for mental health treatment, NEJM AI 2 (2025) AIoa2400802.

[5] Y. Wang, X. Li, Q. Zhang, D. Yeung, Y. Wu, Effect of a cognitive behavioral therapy–based AI chatbot on depression and loneliness in Chinese university students: Randomized controlled trial with financial stress moderation, JMIR mHealth and uHealth 13 (2025) e63806.

[6] A. Reinhardt, J. Matthes, L. Bojić, H. T. Maindal, C. Paraschiv, K. Ryom, Help me, Doctor AI? a cross-national experiment on the effects of disease threat and stigma on AI health information-seeking intentions, Computers in Human Behavior (2025) 108718.

[7] Z. Iftikhar, A. Xiao, S. Ransom, J. Huang, H. Suresh, How LLM counselors violate ethical standards in mental health practice: A practitioner-informed framework, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 8, 2025, pp. 1311–1323.

[8] CNN, ChatGPT encouraged college graduate to commit suicide, family claims in lawsuit against OpenAI, https://edition.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis, 2025. Accessed 2026-05-15.

[9] S. Chen, M. Gao, K. Sasse, T. Hartvigsen, B. Anthony, L. Fan, H. Aerts, J. Gallifant, D. S. Bitterman, When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior, npj Digital Medicine 8 (2025) 605.

[10] B. Bodroža, B. M. Dinić, L. Bojić, Personality testing of large language models: limited temporal stability, but highlighted prosociality, Royal Society Open Science 11 (2024) 240180.

[11] G. Rubeis, Ethics of medical AI (2024).

[12] F. Liu, H. Zhou, B. Gu, X. Zou, J. Huang, J. Wu, Y. Li, S. S. Chen, Y. Hua, P. Zhou, J. Liu, Application of large language models in medicine, Nature Reviews Bioengineering (2025) 1–20.

[13] S. Maity, M. J. Saikia, Large language models in healthcare and medical applications: A review, Bioengineering 12 (2025) 631.

[14] 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.

[15] 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.

[16] 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. doi:10.1016/j.jcbs.2016.05.003.

[17] 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.

[18] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, Annals of internal medicine 177 (2024) 210–220.

[19] I. Luknar, Social dynamics in the era of artificial intelligence, Serbian Political Thought 90 (2025) 145–162.

[20] G. Casella, R. L. Berger, Statistical Inference, 2 ed., Duxbury, Pacific Grove, CA, 2002. Chapter 5.4: Order Statistics.