INSALyon at MentalRiskES 2026: ACTalk, a Two-Head LLM-based System for Psychometric Scoring and Therapist Response Selection in Spanish Therapeutic Dialogue

Diana Nurbakova[1,*]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69100 Villeurbanne, France

Abstract

We present ACTalk, our INSALyon submission to MentalRiskES 2026: a shared Llama-3.3-70B-Instruct-Turbo backbone with two task-specific heads. reACT, the Task 1 head, is a calibrated psychometric assessor that predicts PHQ-9, GAD-7, and CompACT-10 item vectors round-by-round through three calibration levels (per-item anchors, cross-instrument rules, and an agent-invocation layer). tACT, the Task 2 head, is an ACT-FM-grounded evaluator that scores three candidate therapist responses per turn against a per-candidate adaptation of the ACT Fidelity Measure. ACTalk placed Run 2 at rank 10 of 53 on Task 1 (MAE_Combined 1.063) and at rank 24 of 47 on Task 2 (accuracy 0.247). We also perform a post-hoc analysis by introducing a Gemma branch. This confirms that a model-and-prompt swap at the GAD-7 assessor closes most of the gap, while a bare-LLM Gemma 4 31B with anti-bias guardrails reached 0.470 accuracy on Task 2.

Keywords

mental-health NLP, psychometric assessment, PHQ-9, GAD-7, CompACT, Acceptance and Commitment Therapy, zero-shot LLM

1. Introduction

Automatic psychometric assessment from conversational text and automatic selection of clinically appropriate therapist responses are two tasks that, taken together, span the inferential work an NLP system needs to support a therapeutic-conversation tool. The first asks the system to recover patient-state estimates against standardised clinical instruments (PHQ-9, GAD-7, CompACT-10) as a session unfolds. The second asks the system to recognise the therapeutic move most consistent with the conversation's clinical trajectory and the therapist's stance. Both tasks operate over the same conversational substrate and plausibly share a substrate of clinical and linguistic competence, yet they reward different inferential operations: the first reads symptom severity, the second reads therapeutic intent. Whether a single LLM backbone can support both with task-specific calibration on top, and where such a backbone helps and hurts at the assessor and selector levels, is the empirical question this paper addresses.

The MentalRiskES 2026 shared task at IberLEF [1, 2] sets both tasks on Spanish-language therapeutic conversations, with no labelled training data provided. Task 1 (Early Symptom Detection in Therapeutic Conversations) asks participants to predict PHQ-9, GAD-7, and CompACT-10 item vectors round-by-round as the conversation unfolds. Task 2 (Therapist Response Selection) asks participants to select the most appropriate therapist response among three candidates at each round. The corpus consists of simulated yet expert-reviewed therapeutic conversations (§4.2), and the online emission protocol means both tasks must commit per-round predictions without access to future content. The zero-shot constraint, the online emission, and the joint psychometric-and-selection framing together motivate a shared backbone with task-specific heads rather than two unrelated systems.

We present ACTalk, our INSALyon submission to MentalRiskES 2026. ACTalk runs a shared Llama-3.3-70B-Instruct-Turbo backbone behind two task-specific heads. The Task 1 head, reACT, is a calibrated psychometric assessor with three calibration levels: per-item detection grounded in instrument anchors

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

diana.nurbakova@insa-lyon.fr (D. Nurbakova)

0000-0002-6620-7771 (D. Nurbakova)

(Level A), cross-instrument rules including distress-conditional rescaling of CompACT-10 against PHQ-9 and GAD-7 anchors (Level B), and an agent-invocation layer for edge cases (Level C). The Task 2 head, `tACT`, is an ACT-FM-grounded [3] evaluator that scores three candidate therapist responses against a per-candidate adaptation of the ACT Fidelity Measure rubric and selects one. Both heads share state through a Wasserstein-distance anomaly check on the per-instrument prediction sequence and through a common round-by-round event log. The full system, the persona-generation code used to build our simulated development cohort, and the post-hoc analysis scripts are released alongside this paper.

The contribution of this paper is diagnostic rather than competitive: a head-by-head account of where engineering helps a shared backbone and where it overreaches, grounded in a post-hoc Gemma branch that makes the architectural intervention point identifiable instrument-by-instrument and task-by-task. We also report a dev-vs-test rank disagreement across our three evaluation cohorts that motivates operational changes to our persona-generation protocol. We treat this finding as specific to our setting rather than as a transferable methodological claim (§8.6).

The rest of the paper is organised as follows. §2 introduces ACT therapy, the three clinical instruments and their severity bands, the ACT Fidelity Measure, and Theory of Mind as a methodological lens. §3 positions `ACTalk` against prior shared tasks and against the LLM-based zero-shot clinical-assessment literature. §4 formalises both tasks and describes the trial, simulated, and test cohorts. §5 specifies `ACTalk`'s two heads. §6 details the experimental setup and ablation grids. §7 reports the leaderboard results and per-instrument breakdown. §8 reports the post-submission Gemma branch, the consensus-failure analysis on Task 2, and the dev-vs-test rank disagreement. §9 interprets the two-head asymmetry, §10 concludes, and §11 bounds the scope of our claims.

2. Background

This section introduces the concepts our system depends on. We cover *Acceptance and Commitment Therapy* (ACT) and the hexaflex (§2.1), the three clinical instruments the system predicts (§2.2), the clinical-decision thresholds those instruments feed into (§2.3), the ACT Fidelity Measure that grounds our Task 1 evaluator (§2.4), and Theory of Mind (ToM) as a methodological lens we explored in our Task 1 solution (§2.5).

2.1. ACT Therapy and the Hexaflex

Acceptance and Commitment Therapy (ACT) [4] is a contextual cognitive-behavioural framework that operates over six core processes organised into the *hexaflex*:

1. *Defusion*: relating to one's thoughts as thoughts rather than as facts,
2. *Acceptance*: willingness to experience uncomfortable internal events without defence,
3. *Contact with the Present Moment* (*present-moment*): contact with the now,
4. *Self as Context* (*self-as-context*): a stable observing perspective distinct from one's content,
5. *Values*: chosen life directions,
6. *Committed Action*: behaviour aligned with values.

The construct ACT targets through these processes is *psychological flexibility*, namely the capacity to remain in contact with present experience while pursuing valued action despite distressing thoughts and feelings. ACT contrasts with symptom-reduction frameworks: a successful ACT intervention can increase psychological flexibility without reducing the patient's PHQ-9 or GAD-7 totals in the short term, and the inverse is also possible. This decoupling matters for our setting because the MentalRiskES 2026 Task 1 metric is computed across PHQ-9, GAD-7, and CompACT-10 jointly. The first two measure symptom severity, the third measures the protective process of psychological flexibility, and they may move in different directions within a single session.

2.2. PHQ-9, GAD-7, and CompACT-10 Instruments and Scoring

The Patient Health Questionnaire-9 (PHQ-9) [5], the *Generalised Anxiety Disorder 7-item scale (GAD-7)* [6], and the *Comprehensive Assessment of Acceptance and Commitment Therapy Processes (CompACT)* in its 10-item short form (CompACT-10) [7] are the three instruments scored on MentalRiskES 2026 Task 1. PHQ-9 is a 9-item depressive-symptom screener with each item on a $\{0, 1, 2, 3\}$ scale (frequency over the past two weeks), yielding totals in $[0, 27]$. GAD-7 is a 7-item anxiety screener with the same per-item scale, yielding totals in $[0, 21]$. CompACT-10 is a 10-item self-report of psychological flexibility with each item on a $\{0, 1, \ldots, 6\}$ scale, yielding totals in $[0, 60]$. Higher CompACT-10 indicates higher flexibility, namely the inverse polarity of PHQ-9 and GAD-7.

Spanish-language validation status differs across the three instruments. PHQ-9 and GAD-7 have established Spanish validations [8, 9, 10, 11], with Muñoz-Navarro et al. [9] recommending a ≥ 12 cut-off (rather than the standard ≥ 10) in Spanish primary-care populations. CompACT-10's Spanish validation status is partial: Giovannetti et al. [12] translated the parent CompACT-23 instrument into Spanish but we are not aware of a peer-reviewed Spanish-language validation of the 10-item short form. We note this as a limitation in §11.

Throughout this paper we make use of analytical item groupings within PHQ-9 and GAD-7 that organise the items by underlying symptom construct. For PHQ-9 we use the following categories:

- *Somatic*: items 3, 4, 5, 8;
- *Cognitive*: items 6, 7;
- *Affective*: items 1, 2, 9.

Similarly, for GAD-7 we use the following categories:

- *Somatic Anxiety*: items 1, 4, 5;
- *Cognitive Anxiety*: items 2, 3;
- *Emotional Reactivity*: items 6, 7.

Mind that these groupings are ours, grounded in DSM-5 symptom clusters rather than in a published factor analysis of either instrument. They are convenient for ablation and error analysis but do not constitute a psychometric claim. Table 1 gives concrete Spanish-language item-response examples at the mild, moderate, and severe ends of each scale. Each cell shows one patient turn that evidences a specific item on the target instrument at the target severity band, the gold total score and band for the session, and the ACTalk total predictions under our three submitted runs (Run 0 = A5-T3, Run 1 = A3-T2, Run 2 = A1-T2). Spanish is verbatim; English glosses are author translations. **CompACT-10 informal bands**: the instrument has no validated severity cutoffs in the literature (§2.3). For presentation only we use empirical thirds of the 0–60 range (low ≤ 20, medium 21–40, high ≥ 41). **Gap in the low-flexibility cell**: none of the ten released test sessions sit below CompACT-10 total 21. We document the gap rather than reaching for a non-test source. The table doubles as the first demonstration of ACTalk on real test data: cells where the prediction lands in the same band as the gold show the system aligned (S15 PHQ-9 moderate, S06 GAD-7 moderate, S16 CompACT-10 medium). Cells where it under-attributes (S01 PHQ-9 severe, S01 GAD-7 severe, S12 CompACT-10 high) surface the failure modes discussed in §7 and §8.

2.3. Clinical severity and treatment-decision thresholds

PHQ-9 totals are conventionally banded as 0–4 minimal, 5–9 mild, 10–14 moderate, 15–19 moderately severe, and 20–27 severe [5].

Following Spitzer et al. [6] and subsequent convention [13, 14], GAD-7 totals are banded as 0–4 (minimal), 5–9 (mild), 10–14 (moderate), and 15–21 (severe). Spitzer et al. originally specified only the three upper cutpoints (5, 10, 15). We assume that the additional 0–4 'minimal' label has been carried over from the PHQ-9 scheme of Kroenke et al. [5] and is now standard in clinical and epidemiological reporting.

CompACT-10 has no validated severity cutoffs, no minimal clinically important difference, and no normative bands [7]. It is interpreted as a continuous protective-process measure rather than as a measure of disorder severity.

Treatment-decision thresholds on top of these bands are not internationally agreed. NICE NG222 [15] dichotomises depression at PHQ-9 = 16, whereas the four-tier model adopted by the APA [16], CANMAT [17], the WHO mhGAP-IG [18], and Spain's GPC [19] already justifies active treatment with pharmacotherapy or structured psychotherapy at moderate totals ($\geq$ 10). Since the corpus is Spanish, we default to the Spanish GPC framing when severity bands have clinical-pathway implications. The clinical cost of a one-band misclassification is asymmetric: a minimal-mild boundary error has minimal consequence under any guideline, while a moderate-moderately-severe boundary error (14 vs 16) potentially shifts the patient between low-intensity self-help and combined active treatment. We return to this asymmetric-cost framing when interpreting ablation patterns in §9.

2.4. The ACT Fidelity Measure

The ACT Fidelity Measure (ACT-FM) [3] is a 25-item observer-rated rubric organised around four areas reflecting the Triflex model of psychological flexibility plus therapist stance: *Therapist Stance*, *Open Response Style* (covering acceptance and defusion), *Aware Response Style* (covering present-moment awareness and self-as-context), and *Engaged Response Style* (covering values and committed action). Each of the four areas contains both ACT-consistent and ACT-inconsistent items, since a clinician can exhibit both kinds of behaviour within the same session. The Triflex pairs map onto the six hexaflex processes (§2.1) in the standard ACT correspondence:

- *Open Response Style* pairs *defusion* and *acceptance*,
- *Aware Response Style* pairs *present-moment* and *self-as-context*,
- *Engaged Response Style* pairs *values* and *committed action*.

ACT-FM was designed to score a complete therapy session for fidelity to the ACT model, with raters watching or reading the full transcript.

We adapt ACT-FM as a per-candidate response-evaluation rubric: the evaluator scores each of the three candidate therapist responses at a given round against a subset of ACT-FM rubric items rather than scoring the complete session. The adaptation preserves the rubric's hexaflex anchoring and stance dimension but operates at one-turn granularity rather than session granularity. The full evaluator prompt is given in A.3.

2.5. Theory of Mind as a Methodological Lens

Theory of Mind (ToM) [20] is the cognitive capacity to attribute mental states (beliefs, desires, intentions, emotions) to oneself and to others, and to use those attributions to predict and explain behaviour. In a therapist-response-selection setting, ToM is plausibly load-bearing: choosing among three candidate responses requires modelling the patient's likely reaction to each one, and reactive modelling is precisely the operation ToM names. This conceptual fit motivated us to explore ToM-grounded framings in our system's evaluator alongside the functional-analysis framing inherited from ACT-FM.

We tested two ToM-grounded variants in our system's ablation grid (§6.2), labelled TOM-B (a belief-and-desire attribution framing layered on the B pipeline) and TOM-C (a contextualised stance-tracking framing layered on the B+ pipeline). The functional-analysis framing (FUNC) outperformed both ToM variants consistently across the development cohorts, which led to FUNC being submitted in all three runs. We discuss the empirical result in §7.6 and revisit the conceptual question of why ToM did not transfer in §9.2.

Table 1
Multi-severity instrument examples from the released test cohort.

	Mild / Low	Moderate / Medium	Severe / High
PHQ-9	**S16 / R23 / item 1 (anhedonia)** "Sí, es justo eso. No me gusta estar pensando eso en vez de disfrutar del cumpleaños de mi amigo y prestarle atención a él, que es para lo que estoy allí." *("Yes, that's exactly it. I don't like thinking about that instead of enjoying my friend's birthday and giving him my attention, which is what I'm there for.")* Gold 7 (mild). ACTalk 10/10/10 (Run 0 / 1 / 2).	**S15 / R6 / item 7 (concentration)** "Y realmente lo que me hace cambiar de tema, es el hecho de tener muchos pensamientos en mi cabeza a la vez mientras esa conversación está sucediendo, entonces de manera espontánea me sale decir otra cosa." *("And what really makes me change the topic is the fact of having many thoughts in my head at the same time while the conversation is happening, so spontaneously I end up saying something else.")* Gold 11 (moderate). ACTalk 5/5/5.	**S01 / R14 / item 2 (depressed mood)** "Pues no exactamente, porque lo he interiorizado, así que en el fondo es mi forma de ser. Pero diría el otro día cuando estábamos cenando, a mí no me gusta para nada el champiñón. Y a mis amigos sí, entonces pidieron una pizza llena de champiñón y pues claro, no podía decir que no me gustaba que pidieran otra, por si acaso se burlaban. O lo ignoraban." *("Well, not exactly, because I've internalised it, so deep down it's my way of being. But I'd say the other day when we were having dinner, I don't like mushrooms at all. And my friends do, so they ordered a pizza full of mushrooms and of course, I couldn't say I didn't like it and ask them to order another, just in case they made fun of me. Or ignored me.")* Gold 16 (moderately severe). ACTalk 11/11/11 (under-attributes).
GAD-7	**S07 / R15 / item 1 (nervousness)** "Pues me he sentido muy nerviosa, lo he pasado mal. ¿A qué te refieres con patrón?" *("Well, I've felt very nervous, I had a hard time. What do you mean by pattern?")* Gold 7 (mild). ACTalk 9/9/9.	**S06 / R13 / item 3 (excessive worry)** "Ahí soy feliz. Básicamente pienso 'ahora voy a descansar'. Y me olvido del mundo. Son los momentos en los que estoy bien, que son muy pocos. El problema es el resto del tiempo, que es cuando estoy amargado." *("There I'm happy. I basically think 'now I'm going to rest'. And I forget about the world. Those are the moments when I'm well, which are very few. The problem is the rest of the time, when I'm bitter.")* Gold 10 (moderate). ACTalk 6/8/8.	**S01 / R1 / item 2 (inability to control worry)** "Hola. Pues muy nervioso, no paro de darle vueltas a las cosas." *("Hello. Well, very nervous, I can't stop going over things in my head.")* Gold 16 (severe). ACTalk 10/10/10 (under-attributes).
CompACT-10	*No session in the released test cohort sits in the informal low-flexibility band (total 0–20).*	**S16 / R10 / item 5 (OtE avoidance, reverse-keyed)** "Ya, probablemente no... Y es verdad que para mí son muy importantes las relaciones sociales, me encanta conectar con la gente y tener intimidad con ella, vínculos que signifiquen algo para mí. Y probablemente yo misma hago que eso sea más difícil preguntándome todo el rato si lo que hago está bien." *("Right, probably not... And it's true that social relationships are very important to me, I love connecting with people and having intimacy with them, bonds that mean something to me. And I probably make that harder for myself by asking myself all the time whether what I'm doing is right.")* Gold 33 (medium). ACTalk 33/38/38.	**S12 / R21 / item 5 (OtE avoidance, reverse-keyed)** "Aún así tampoco sé muy bien cómo iniciar conversaciones, siento que se me da mejor escuchar que hablar y me da miedo aburrir a los demás. La cuestión es que me encantaría ser más extrovertida pero la vergüenza me acaba ganando y prefiero no hacer nada con tal de no quedar mal." *("Even so, I don't really know how to start conversations either; I feel I'm better at listening than at speaking and I'm afraid of boring others. The thing is, I'd love to be more extroverted but shame ends up winning and I'd rather do nothing than look bad.")* Gold 50 (high). ACTalk 32/35/35 (under-attributes).

3. Related Work

3.1. Online Prediction in Clinical NLP Shared Tasks

The eRisk campaign at CLEF [21] established online early-risk detection as a shared-task setting and has run annually since 2017 on user-content streams from social media. The campaign's evaluation framework rewards both decision quality and decision speed through the ERDE metric [22], later revised as $ERDE_o$ by Trotzek et al. [23] to address penalty-monotonicity issues. The canonical participant strategy on eRisk combines a content predictor with an explicit decision policy: UNSL's CPI+DMC framework [24] factorises the round-by-round decision into two heads, namely a content predictor evaluated at every round and a decision-maker policy trained to choose when to commit. Burdisso et al.'s SS3 [25] represents the alternative tradition of incremental linear classifiers with explicit per-term scores.

The CLPsych 2024 shared task [26] offered a complementary framing, namely LLM-based extraction of textual evidence supporting pre-assigned suicide-risk labels in Reddit user histories, under open-source LLM constraints. Top-performing systems combined supervised extractive scoring with generative summarisation [27] or retrieval-augmented generation with emotional-charge ranking [**?**].

In contrast to eRisk's ERDE-style scoring, MentalRiskES 2026 scores both tasks on the final-round output per session and does not penalise decision latency directly. The online constraint is implicit in the round-by-round prediction-emission protocol rather than weighted into the metric. Our submission accordingly emits at every round but does not use a CPI+DMC-style decision policy. We treat every round as a prediction round, and the leaderboard reads the last one.

3.2. Prior MentalRiskES Editions

The 2023, 2024, and 2025 editions of MentalRiskES [28, 29, 30] operated over Spanish Telegram conversations, with subtasks on eating-disorder, depression, anxiety, and gambling-risk detection. The dominant participant approaches across these editions were BERT-family encoders fine-tuned on the released training splits and prompt-based LLMs operating on accumulated conversation windows. ELiRF-VRAIN's Longformer-based long-context architecture [31] is representative of the encoder line on long sessions.

The 2026 edition introduces a methodological shift: the conversational substrate is now structured therapeutic dialogue (see §4.2), the primary target is psychometric score prediction rather than binary risk detection, and a second task asks for therapist response selection rather than user-state classification. In contrast to the social-media-monitoring framing of earlier editions, the 2026 setting requires the system to operate inside a clinical-conversation simulation and to recover scores derived from the patient profile assigned to the role-player, not labels assigned post-hoc to a social-media stream. Our submission is the first INSALyon entry to MentalRiskES and to our knowledge the first to address both 2026 tasks with a single shared backbone.

3.3. LLM-based Zero-Shot Clinical Assessment

A line of recent work uses LLMs to predict standardised clinical instrument scores from text in zero-shot or few-shot settings, mostly in English. The eRisk Task 1 traditions [21] target BDI-II prediction from social-media post histories, and the DAIC-WOZ-derived line [32] targets PHQ-9 prediction from clinical-interview transcripts. PRIMATE [33] contributes PHQ-9-grounded examples we reuse in our anchor pool. In contrast to all of these, the MentalRiskES 2026 setting we address is Spanish-language, online, and targets three instruments jointly (PHQ-9, GAD-7, and CompACT-10). CompACT-10 in particular has no established LLM-prediction baseline in any language to our knowledge, and the our pipeline addresses it through distress-conditional rescaling against PHQ-9 and GAD-7 anchors rather than through a CompACT-specific training signal.

3.4. Computational Work on ACT

Despite ACT's clinical adoption over two decades and the existence of the ACT Fidelity Measure as a session-level scoring rubric [3], we are not aware of prior computational work targeting automatic recognition of ACT processes in dialogue at scale. The nearest NLP traditions are motivational-interviewing fidelity coding [34] and emotional-support response selection [35], both of which share the move of scoring or selecting therapist turns against a clinical rubric. In contrast to these traditions, ACT-FM is constructed around six clinical processes (the hexaflex; see §2.1) and a stance dimension that motivational-interviewing rubrics do not include, and we adapt ACT-FM as a per-candidate response-evaluation rubric rather than as a session-level fidelity score. The bare-LLM Gemma result in §8.4 suggests this adaptation may have overridden the model's native conversational judgement on the present corpus, a finding we return to in §9.

4. Task and Data

4.1. Task Formulations

Both MentalRiskES 2026 tasks operate over the same input, namely a Spanish-language therapeutic conversation between a patient and a therapist, revealed round-by-round through a server loop. At each round r the participant receives the patient's latest message, and emits a per-round prediction without access to rounds $r+1, \ldots, R$. The conversation history up to round $r-1$ should be handled by the participants. The two tasks differ in the prediction target.

Task 1 (Early Symptom Detection in Therapeutic Conversations): given the conversation up to round r, extract the patient's current item-score vectors on three clinical instruments, namely PHQ-9 with item scores in $\{0,1,2,3\}^9$, GAD-7 with item scores in $\{0,1,2,3\}^7$, and CompACT-10 with item scores in $\{0,1,\ldots,6\}^{10}$. Predictions are emitted at every round. The leaderboard scores the final-round submitted prediction per session against the gold item vectors.

Task 2 (Therapist Response Selection): given the conversation up to round r and three candidate therapist responses $\{o_1, o_2, o_3\}$ presented at that round, select the option $o^\star \in \{o_1, o_2, o_3\}$ that best continues the conversation. The selection is emitted at round r and is scored against the therapist's actual continuation at that round.

Both tasks are zero-shot in the sense that the organisers provide no labelled training data. The only labelled material is the official trial conversation (§4.3).

4.2. Corpus Protocol and the Released-vs-Gold Gap

The MentalRiskES 2026 corpus consists of simulated yet expert-reviewed therapeutic conversations in Spanish. Five licensed psychologists interacted through a custom application with simulated patients role-played by psychology students, each assigned a patient profile with key symptoms derived from PHQ-9, GAD-7, and CompACT-10. During each session the application offered AI-generated response suggestions to the therapist at every turn, and the therapist could accept, edit, or replace any suggestion. The gold labels for Task 1 correspond to the patient profile assigned to the student role-player. The gold for Task 2 is the response the licensed psychologist actually chose (the option-1, option-2, or option-3 slot the therapist filled with their selected or edited text). All dialogues were reviewed by clinical experts for coherence and clinical validity prior to release. Mind that this protocol implies two methodological constraints: Task 1 gold reflects the assigned profile rather than a clinical interview, and Task 2 gold reflects licensed-psychologist editorial choices among AI-generated suggestions rather than blind clinical adjudication. We return to these constraints in §11.

The test cohort is released as 10 sessions (82 rounds total, mean 56.8 rounds per session). The severity distribution skews high: 7 of the 10 released sessions sit in the GAD-7 severe band (≥ 15), and the

PHQ-9 distribution covers all four upper bands. The official gold scoring set, however, contains 17 patients. Sessions S02, S08, S10, S11, S13, S14, and S17 are present in the gold but their conversational content was not released, so we could not run `ACTalk` on them. The leaderboard scored our submission as if default predictions had been emitted for these 7 unreleased sessions. Our descriptive statistics, ablations, and post-hoc analyses (§7–§8) operate on the released 10-session subset, and we flag this scope wherever a projection to the official ranking is reported.

4.3. Trial and Simulated Development Cohorts

Three evaluation cohorts informed our development. The official *trial* cohort is a single 19-round session (`Miguel`, moderate-severity profile: PHQ-9 total = 13, GAD-7 total = 14, CompACT-10 total = 33) generated under the same simulation protocol as the test cohort. To complement the trial with cases that span the severity range, we generated a *simulated* cohort of six personas at 15 rounds each (90 patient-rounds in total), with three anxiety profiles and three depression profiles spanning PHQ-9 totals in [4, 14] and GAD-7 totals in [3, 20], across mild, moderate, and high-distress CompACT-10 flexibility profiles (Appendix E.1; the tACT cohort adds a seventh short-session persona, in Appendix E.2). Personas were generated with Llama-3.3-70B-Instruct-Turbo at temperature 0.7, conditioned on a hand-authored profile sheet and a short conversational sketch. The therapist turns were generated in the same call. In contrast to the organiser's protocol, our simulated patients are LLM-generated rather than student-role-played, our therapist turns are LLM-generated without psychologist editing, and no clinical-expert review pass was performed. We treat our simulated cohort as a development convenience rather than as a clinical benchmark and flag its protocol gap with the test cohort throughout this paper. The third cohort is the *test* cohort described in §4.2, run by the organisers' server during the campaign's evaluation window. The persona-generation code is released with the rest of the implementation (§5.7).

4.4. External Resources

We use four external resources at prompt-engineering time, all as in-prompt anchors rather than as labelled training data: PRIMATE [33] for PHQ-9-grounded examples in `reACT`'s anchor pool, ESConv-ES [35] (Spanish ESConv) for emotional-support fragments used in tone calibration, MIDAS [36] for segment-level conversational labels anchoring the state-tracker terminology in `tACT`, and the ACT Fidelity Measure [3] as the rubric source for `tACT`'s evaluator (§5.3). No labelled training data is provided by the organisers in either task.

5. System Description: `ACTalk`

ACTalk is a zero-shot online pipeline that we developed for both MentalRiskES 2026 tasks. Its general arhitecture is depicted in Figure 1. **ACTalk** is built on a single backbone, namely the FP16 release of Llama-3.3-70B-Instruct served through the HuggingFace Inference API, and routes incoming rounds through a shared context-accumulation layer before branching into two task-specific heads. The Task 1 head, which we call **reACT**[1], performs repeated per-round psychometric scoring on PHQ-9, GAD-7, and CompACT-10. The Task 2 head, which we call **tACT**, selects among three candidate therapist responses at each round by scoring them against the ACT Fidelity Measure rubric. We describe the shared backbone first (§5.1), then each head in turn (§5.2, §5.3), then the submitted runs (§5.4), the full configuration space we explored (§5.5), the configurations we dropped along the way (§5.6), and the implementation details (§5.7).

[1]Our **reACT** (lowercase *re* + uppercase *ACT*) is not to be confused with the ReAct prompting framework of Yao et al. [37]: our name combines the prefix *re-* (repeated) with the framework name *ACT*, signalling repeated assessment against ACT-aware clinical instruments rather than reasoning-and-acting prompting.

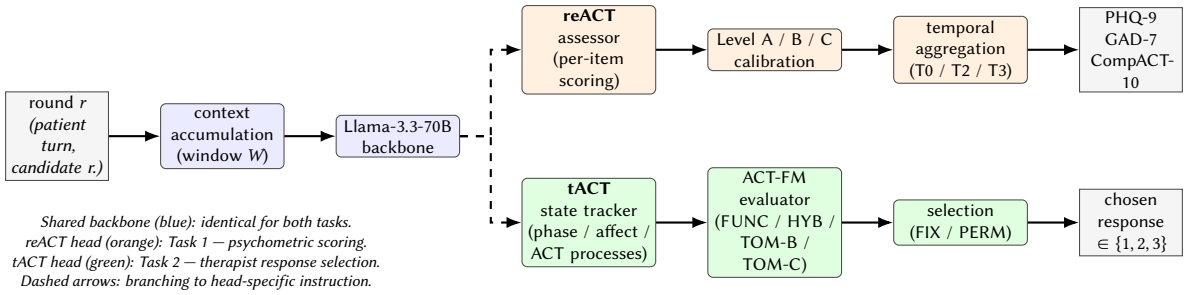

Figure 1: ACTalk architecture. Each incoming round is appended to the running conversation state and processed by a shared Llama-3.3-70B backbone with a head-specific instruction. The **reACT** head (Task 1) emits PHQ-9, GAD-7, and CompACT-10 item-score vectors through a three-tier calibration stack and a temporal-aggregation step. The **tACT** head (Task 2) maintains a therapeutic-state tracker, scores the three candidate responses against the ACT Fidelity Measure rubric, and selects one. The two heads share the context-accumulation layer and the LLM backbone but not the prompts, the post-processing rules, or the output format.

5.1. Shared Backbone: Context Accumulation and LLM

Both `ACTalk` heads process the conversation through a shared context-accumulation layer. At each round r, the new patient turn (and, on Task 2, the three candidate therapist responses) is appended to the running conversation state $C_r = \{(u_i, a_i)\}_{i=1}^{r}$, where u_i is the patient turn at round i and a_i is the therapist turn. The LLM assessor at the start of each head consumes a windowed slice of this state, namely the last W exchanges. Mind that the window W governs how much past context the head sees on each call. We set $W = 3$ as the default for all submitted runs and justify this choice via ablation in §6.

The shared backbone is the FP16 release of Llama-3.3-70B-Instruct served through the HuggingFace Inference API. Both heads run on the same model checkpoint with the same sampling parameters. Only the prompt and the post-processing layer differ.

5.2. `reACT`: Task 1 Head

At each round, the `reACT` assessor receives the windowed context C_r^W and produces a per-item score vector for each of PHQ-9 (9 items, 0–3 each), GAD-7 (7 items, 0–3 each), and CompACT-10 (10 items, 0–6 each), together with a brief textual justification per item. The prompt (full text in Appendix **??**) instructs the model to score from the patient's behavioural and affective evidence visible in C_r^W, with explicit anchoring through severity-labelled examples. One assessor call is issued per instrument, namely three LLM calls per round across the three instruments.

Level A — prompt anchors. The assessor prompt embeds severity-labelled examples for each item scale, namely one mild, one moderate, and one severe example per instrument, drawn from the trial conversation and from the PRIMATE corpus where available. The role of these anchors is to ground the item scale empirically. In our preliminary experiments, an un-anchored assessor tended to score conservatively (regression toward the middle of the scale), which the anchors counteract.

Level B — rule-based calibration. The assessor's output is passed through seven deterministic constraints that enforce psychometric coherence. We sketch the four most consequential of these:

- *Ordinal coherence within domain.* For each of the three item groupings we use (Somatic / Cognitive / Affective for PHQ-9 and Somatic Anxiety / Cognitive Anxiety / Emotional Reactivity for GAD-7, per §2), within-domain item scores must agree on direction: an isolated severe item against otherwise minimal scores in the same domain is flagged.
- *Between-round stability.* An item score may change by at most 2 between consecutive rounds in the absence of a corresponding textual event in C_r. The threshold is conservative. In practice the rule fires rarely.

- *Distress-conditional rescaling of CompACT-10.* CompACT-10 scores are rescaled against PHQ-9 and GAD-7 anchors: if the patient presents with PHQ-9 ≥ 15 or GAD-7 ≥ 15 and CompACT-10 sits above the cohort median, the rescaling pulls CompACT-10 downward toward the cohort-conditional expectation. This corrects an LLM floor effect we observed in early experiments, namely that the model tended to score CompACT-10 too high (over-attributing psychological flexibility) when the patient's distress was already severe.
- *Item-9 safety flag.* A non-zero PHQ-9 Item 9 (*Pensamientos de que estarías mejor muerto/a o de hacerte daño de alguna manera — thoughts that you would be better off dead or of hurting yourself in some way*) raises a separate safety flag alongside the total score. We do not gate any treatment recommendation on this flag (the system does not make treatment recommendations). We surface it to the deployment-layer evaluator as an explicit dual output, per the recommendation in §2.

The Level B rules are deterministic and do not add LLM calls.

Level C — LLM calibration agent. When Level B flags a violation that cannot be resolved deterministically (*e.g.*, two within-domain items disagreeing on direction with no clear majority), an LLM calibration agent is invoked. The agent receives the violating items, the relevant textual evidence from C_r^W, and the Level B rule that fired, and returns a revised score plus a justification. The Level C agent is conditional: in practice it is invoked on roughly 10% of rounds in the trial cohort and is skipped entirely when no Level B rule fires. Per-round LLM-call counts across the three submitted reACT runs therefore differ at the agent step but not at the assessor step: Run 2 (A1-T2) issues three calls per round; Run 1 (A3-T2) also issues three; Run 0 (A5-T3) issues three plus one conditional Level C call per Level B violation.

Temporal aggregation. The assessor produces a per-round score vector. The instrument predictions submitted for the session are obtained by aggregating across rounds. We define three aggregation strategies:

- **T0** (last-round): the prediction is the assessor's score at the final round of the session. Simple, but discards earlier evidence.
- **T2** (early-weighted): rounds 1–5 are doubled in the aggregation weight; later rounds contribute their nominal weight. The motivation is the recency-bias finding from our development cohorts: the assessor under-attributes severity in later rounds when the patient has shown within-session improvement, even if the baseline severity is high.
- **T3** (stability-adaptive for CompACT-10): identical to T2 for PHQ-9 and GAD-7. For CompACT-10, the aggregation weight is adapted per round to the inverse of the local variance across the last five rounds, namely rounds with stable scores receive higher weight than rounds with volatile scores. The intent is to track psychological flexibility as a process rather than as a symptom level.

Anomaly check across rounds. The early-weighted and stability-adaptive strategies above presuppose that each round's score vector is a faithful signal from the assessor. Mind that the assessor occasionally produces an anomalous per-round vector, namely one that disagrees sharply with the running session-level behaviour without a corresponding textual event in C_r. We detect these cases by computing the first Wasserstein distance $W_1^I(r)$ for instrument I at round r, between the per-item score distribution at round r and the running barycenter of the per-item distributions across rounds $1, \ldots, r$. The ground metric used in the W_1 computation is instrument-specific: a bifactor structure for PHQ-9, an anxiety-factor structure for GAD-7, and a hexaflex structure for CompACT-10. The general W_1 definition and its optimal-transport interpretation appear in §5.2 of the Background. Here we treat $W_1^I(r)$ operationally as a single scalar per (round, instrument) pair, computed with the implementation from the POT library [38].

The anomaly threshold is set at $\mu_{<r}^I + 2\sigma_{<r}^I$, where $\mu_{<r}^I$ and $\sigma_{<r}^I$ are the running mean and standard deviation of the W_1^I trajectory across the session up to (but not including) round r. To avoid spurious

firings while the running statistics are themselves unstable, the check is disabled for rounds $r \leq 3$ (a three-round warm-up). The two reACT instrument classes behave differently at the threshold by design. For PHQ-9 and GAD-7, the threshold-violating round is discarded from the T2 and T3 aggregations. Our motivation is that depression and anxiety severity rarely shift by the magnitude an anomalous round implies, so the anomaly is more plausibly an assessor artifact than a real change. For CompACT-10, the threshold-violating round is flagged but retained; the rationale is that psychological flexibility is more genuinely volatile within a session than depressive or anxious symptomatology, and discarding rounds would under-attribute the very volatility the T3 stability-adaptive aggregation is designed to track.

Table 2 reports the per-round W_1 values across the 19-round trial conversation for our run0_primary configuration (the canonical reACT pipeline). Six rounds fire the running threshold across the three instruments: PHQ-9 at rounds 11, 12, 17, 18, and 19; GAD-7 at round 14; CompACT-10 fires none. Two of these rounds carry distinct lessons. **Round 14 on GAD-7** produces the largest threshold excess on the session ($W_1 = 0.26$ against threshold 0.10, more than 2.5×), but the downstream effect on the T3 aggregate is a single-item delta. **Round 12 on PHQ-9** produces a modest threshold excess ($W_1 = 0.04$ against threshold 0.03) but the downstream T3 correction is a +6 shift in the PHQ-9 total sum, moving six of nine items back to their session-stable values. The reACT round-trace box that follows this subsection walks through round 12 in detail. Mind that the *detection sensitivity* of the anomaly check and its *correction visibility* downstream are not the same property: the former measures how decisively a round drifts from the running barycenter, the latter measures how much the discard-or-retain action matters for the aggregated session score. On the trial conversation T2 is unaffected by every firing round (the early-weighting step decay already anchors the T2 aggregate on rounds 1–5, making the discard a no-op on T2); the discard mechanism's visible work happens through T3. A full per-round trace for round 14 appears in Appendix D.

See an example of reACT round-trace in Appendix B.

5.3. tACT: Task 2 head

tACT selects among three candidate therapist responses at each round by scoring them against the ACT Fidelity Measure rubric. The head has two variants, namely a two-step pipeline (B) and a two-and-a-half-step pipeline (B+) that inserts a characterisation step between the state tracker and the evaluator.

State tracker. The state tracker maintains four pieces of evolving information across rounds: the therapeutic phase (initial / exploratory / consolidation), the patient's current emotional state, the ACT processes that have been active in the conversation so far (defusion, acceptance, present-moment awareness, self-as-context, values, committed action), and any metaphors the therapist has introduced. At each round, the tracker updates from the windowed context C_r^W and emits a structured state object that the downstream evaluator consumes (prompt in Appendix A.3). One LLM call per round.

Characterisation step (B+ only). In the B+ variant, an intermediate LLM call produces a brief characterisation of what the patient needs at this round (*e.g., "el paciente está atascado en la fusión cognitiva, necesitaría una intervención de defusión más que un nuevo encuadre" — "the patient is stuck in cognitive fusion and would benefit from a defusion intervention rather than a new reframe"*). This characterisation is then passed to the evaluator together with the state object. The B variant skips this step entirely, passing the state object directly to the evaluator.

Evaluator. The evaluator scores each of the three candidate responses against the ACT-FM rubric. Each response is rated on a subset of ACT-FM items (§2), and the per-candidate scores are aggregated into a single utility for selection. We considered four framings of the evaluator's instruction:

- **FUNC** (functional analysis): the evaluator is instructed to score each candidate purely against the ACT-FM rubric, asking what ACT process the response advances and how faithfully.

Table 2
Per-round Wasserstein distance $W_1^I(r)$ across the 19-round trial conversation, run0_primary configuration. W_1 values are ground-metric-weighted (bifactor for PHQ-9, anxiety-factor for GAD-7, hexaflex for CompACT-10), with the running barycenter of rounds $1, \ldots, r$ as the reference distribution. The † marker identifies rounds firing the *running* threshold $\mu^I_{<r} + 2\sigma^I_{<r}$ used by reACT in flight; the session-level threshold (last row of statistics) is computed once at the end of the session and is reported for completeness; note that round 18 on PHQ-9 exceeds the running threshold but lies at the session-level threshold. Round 1 is the degenerate self-reference $W_1 = 0$. Rounds 2 and 3 constitute the warm-up window during which firing is disabled. Firing rounds in PHQ-9 and GAD-7 are discarded from the T2 and T3 aggregations; CompACT-10 firing rounds (none on this trial) would be flagged but retained.

Round	PHQ-9 W_1	GAD-7 W_1	CompACT-10 W_1
1	0.00	0.00	0.00
2	0.03	0.09	0.05
3	0.02	0.07	0.06
4	0.01	0.06	0.01
5	0.01	0.08	0.02
6	0.01	0.07	0.02
7	0.01	0.07	0.01
8	0.01	0.04	0.02
9	0.01	0.03	0.01
10	0.01	0.01	0.01
11	**0.03** †	0.06	0.05
12	**0.04** †	0.03	0.03
13	0.01	0.07	0.04
14	0.01	**0.26** †	0.02
15	0.01	0.04	0.04
16	0.01	0.06	0.04
17	**0.06** †	0.12	0.01
18	**0.06** †	0.09	0.01
19	**0.05** †	0.09	0.03
Session μ (rounds 2–19)	0.02	0.07	0.03
Session σ (rounds 2–19)	0.02	0.05	0.02
Session-level $\mu + 2\sigma$	0.06	0.18	0.06
Rounds firing running threshold	11, 12, 17, 18, 19	14	none

- **HYB** (functional + ToM): the evaluator additionally reasons about the patient's likely reaction to each candidate, combining ACT-FM scoring with a Theory-of-Mind layer.
- **TOM-B** (ToM-only on B): the evaluator reasons only about the patient's reaction, without explicit ACT-FM rubric scoring.
- **TOM-C** (ToM-only on B+): TOM-B with the characterisation step.

Selection. The selection step picks the candidate with the highest aggregated utility. We considered two selection strategies:

- **FIX** (fixed ordering): the three candidates are presented to the evaluator in their original order $(1, 2, 3)$ and scored as such.
- **PERM** (permutation voting): the three candidates are scored under three cyclic orderings, namely $(1, 2, 3)$, $(2, 3, 1)$, and $(3, 1, 2)$. The selected candidate is the one that wins under the majority of the three orderings, with the original-order ($(1, 2, 3)$) result used as the tiebreak. The intent is to debias the evaluator against position effects.

Per-round LLM-call counts. The three submitted `tACT` runs differ in their per-round LLM-call counts. Run 1 (B/FUNC/FIX) issues two calls per round, namely one state-tracker update and one

Table 3
ACTalk submitted runs. All runs use Spanish prompting, FP16 Llama-3.3-70B-Instruct served through the HuggingFace Inference API, and the $W = 3$ context window. The notation *Ak-Tℓ* for reACT denotes calibration level k ($k \in \{0, 1, 2, 3, 4, 5\}$) and temporal aggregation ℓ ($\ell \in \{0, 2, 3\}$). The notation *B{,+}/FRAMING/SELECTION/Ww* for tACT denotes pipeline variant, evaluator framing, selection strategy, and window size.

Head	Run	Configuration	Rationale
reACT	Run 0	A5-T3 (full stack, stability-adaptive CompACT)	full pipeline as designed
reACT	Run 1	A3-T2 (anchors + rules, early-weighted)	reduced agent-LLM dependency
reACT	Run 2	A1-T2 (anchors only, early-weighted)	minimum-engineering baseline
tACT	Run 0	B/FUNC/PERM/W3	variance-minimised via position debiasing
tACT	Run 1	B/FUNC/FIX/W3	default configuration
tACT	Run 2	B+/HYB/FIX/W3	most engineered: characterisation + ToM

evaluator scoring. Run 0 (B/FUNC/PERM) issues four calls per round (one state-tracker update plus three evaluator scorings, one per ordering). Run 2 (B+/HYB/FIX) issues three calls per round (one state update, one characterisation, one evaluator scoring).

An example of tACT round-trace is given in Appendix B.

5.4. Submitted runs

We submitted three runs per task. Table 3 lists them with their configurations and the rationale for each.

The three reACT runs include: Run 0 invokes the full three-tier calibration stack with the stability-adaptive aggregation, Run 1 drops the Level C agent, and Run 2 drops both the rules and the agent. The choice across the three tACT runs is less monotonic: Run 0 minimises variance against position effects (PERM debiasing) on the simpler B pipeline; Run 1 is the same B/FUNC pipeline without PERM. Run 2 adds the characterisation step and the HYB framing. We discuss the engineering rationale behind each choice and what test results showed about the choice in §7 and §8.

5.5. Configuration Matrices

The submitted runs are a subset of a larger design space we explored in development. We summarise this space in two matrices, one per head. The matrices serve two purposes, namely to make the submitted runs' location in the design space explicit, and to make the cells we ablated visible alongside the cells we submitted.

Table 4
reACT configuration matrix: calibration level × temporal aggregation. **Bold** = submitted run. ⋆ = best on test among submitted runs (see §7). *ablated* = run in development on trial and simulated cohorts. *diagnostic* = un-anchored configuration explored only as an ablation, never a submission candidate. — = not run.

	T0 last-round	**T2** early-weighted	**T3** stab.-adap. CompACT
A0 (none)	ablated	ablated	—
A1 (anchors)	ablated	**Run 2** ⋆	—
A2 (rules only)	diagnostic	diagnostic	—
A3 (anchors + rules)	ablated	**Run 1**	ablated
A4 (agent only)	diagnostic	diagnostic	—
A5 (full stack)	ablated	ablated	**Run 0**

Table 5
tACT configuration matrix: pipeline variant × evaluator framing. **Bold** = submitted run (Run 0 = PERM, Run 1 = FIX, Run 2 = FIX). ⋆ = best on test among submitted runs. The secondary axes (ES/EN, W1/W3/W5, FIX/PERM) collapsed to a single value in submitted runs: all use Spanish prompting and $W = 3$. The FIX/PERM choice is part of the run-level rationale in Table 3.

	FUNC	HYB	TOM-B	TOM-C
B (2-step)	**Run 0** + **Run 1**	ablated	ablated	ablated
B+ (2.5-step)	ablated	**Run 2** ⋆	—	—

5.6. Configurations explored and dropped

In contrast to the configurations that reached submission, the ACTa1k design space also contained several configurations that we considered and dropped. For each, we sketch the empirical or methodological reason the configuration was abandoned. Here we describe our negative-evidence layer of our development process: configurations that were dropped as we judged sound at the time, but which our post-submission analysis (§8) revisits in several cases.

Theory-of-Mind framings TOM-B and TOM-C (tACT). These framings asked the evaluator to reason about the patient's likely reaction to each candidate response without explicit ACT-FM rubric scoring. Despite the conceptual fit (response selection is a clear ToM-load task: the therapist must model the patient's reaction), the FUNC framing consistently outperformed TOM-B and TOM-C on both trial and simulated cohorts. Our working hypothesis is that the ACT-FM rubric already encodes the relevant ToM signal in functional-analytic form, and that adding an explicit ToM layer introduces noise rather than discriminative power. We return to this hypothesis in §8.

English prompting (EN). Prompts were written and tested in both English and Spanish. The Spanish version consistently outperformed the English version on the Spanish-language corpus, by margins that varied across heads but were positive across all configurations we tested. All submitted runs use Spanish prompting.

Context windows $W = 1$ and $W = 5$. We tested three context window sizes. With $W = 1$ the assessor (or evaluator) lost cross-turn anchoring and tended to over-react to the most recent turn (the tACT ablation on $W = 1$ lost roughly 10 pp accuracy on the trial cohort relative to $W = 3$); with $W = 5$ the additional past context introduced irrelevant content (*e.g.*, early-session pleasantries) without improving discrimination on the current round (roughly −5 pp on the same cohort). $W = 3$ was the empirical sweet spot across both heads.

Calibration tiebreaker on tACT (CAL). We considered an additional tACT configuration in which the FUNC evaluator's per-candidate utilities were post-processed by a calibration tiebreaker that rescaled scores against an expected per-position distribution. The tiebreaker was intended to address the same position-bias concern as PERM but with one LLM call per round instead of three. On the trial cohort, the B+/FUNC/FIX/W3/CAL configuration lost roughly 5.6 pp accuracy relative to the equivalent configuration without the tiebreaker. We dropped the tiebreaker on this basis and did not include it in any submitted run.

Diagnostic-only cells of the reACT matrix (A2 and A4). The configurations A2 (rules only, un-anchored) and A4 (agent only, un-anchored) were never candidates for submission. We designed them specifically to isolate the contribution of the Level A prompt anchors under each higher-level calibration tier. The expectation, confirmed in the ablation, was that removing the anchors causes the assessor to score conservatively regardless of which Level B or Level C machinery sits downstream.

Table 6
ACTalk inference configuration for submitted runs. The sampling parameters (*temperature*, *top-p*, *seed*, *stop sequences*) are identical across runs and across heads. Only the per-call *max new tokens* budget and the time-out differ, reflecting the longer chain-of-thought reasoning needed for CompACT-10 scoring on reACT versus the shorter selection rationale on tACT. Full reproducibility parameters (retry back-off schedule, rate-limit delays, JSON-parser fallbacks) are in §6.

Setting	reACT (Task 1)	tACT (Task 2)
Provider	HuggingFace Inference API	HuggingFace Inference API
Model checkpoint	`meta-llama/Llama-3.3-70B-Instruct` (FP16)	`meta-llama/Llama-3.3-70B-Instruct` (FP16)
Temperature	0.1	0.1
Top-*p*	provider default (unset)	provider default (unset)
Max new tokens	8192	4096
Stop sequences	none	none
Random seed	not set	not set
Streaming	no	no
JSON mode	not requested	not requested
Time-out (per call)	180 s	300 s
Retry attempts	5	5
Fallback provider	TogetherAI `Llama-3.3-70B-Instruct-Turbo`	TogetherAI `Llama-3.3-70B-Instruct-Turbo`
LLM calls / round	3 (Run 2 / Run 1); 3 + 0–1 conditional (Run 0)	2 (Run 1); 4 (Run 0); 3 (Run 2)

5.7. Implementation Notes

ACTalk is implemented as a Python client that follows the shared-task server-loop protocol: each round, the client polls the server for the new patient turn (and, on Task 2, the candidate responses), invokes the relevant head, and submits the head's output. JSON parsing of the LLM output tolerates Markdown code fences, embedded JSON via outermost-brace bracket matching, Spanish/English key-name variants, accent-insensitive matching, and truncated outputs via close-bracket repair. Malformed outputs that survive a bounded retry policy fall back to a safe default (a score of zero on each item for reACT, a uniform random choice for tACT). In practice, the retry policy resolves the malformed-output cases within one or two attempts.

Provider, Model, and Sampling Parameters. Table 6 summarises the LLM provider and the sampling parameters used for the submitted runs. The full reproducibility table (retry policy, time-out budgets, all unset parameters) appears in §6.

Code release. We release the full ACTalk implementation (prompts, configuration files, the synthetic-persona generation code for our development cohorts, and the analysis scripts used in §8) at https://github.com/diana-nurbakova/depression-detection.

6. Experimental Setup

The setup for both ACTalk heads is shared at the system level (§5) and the ablation level: a single inference configuration applied across both reACT and tACT, two ablation grids (one per head) run on the same trial and simulated development cohorts, and a single live evaluation pass on the released test cohort. We document the configuration in three parts. §6.1 specifies the inference parameters in full; §6.2 describes the ablation grids and the cross-cohort design; and §6.3 discloses the deployment-time truncation that affected our submitted runs and explains how the leaderboard scored against it.

Table 7
Full inference configuration for the submitted ACTalk runs. All parameters listed as *unset* take the provider default; we list them explicitly so that reproducibility runs do not inadvertently override them. The implementation-notes summary in Table 6 of §5.7 is a compressed view of this table for readers who do not need the full record.

Parameter	reACT (Task 1)	tACT (Task 2)
Provider and checkpoint		
Provider	HuggingFace Inference API	HuggingFace Inference API
Model	`meta-llama/Llama-3.3-70B-Instruct`	`meta-llama/Llama-3.3-70B-Instruct`
Precision	FP16	FP16
Fallback provider	TogetherAI	TogetherAI
Fallback variant	`Llama-3.3-70B-Instruct-Turbo` (FP8)	`Llama-3.3-70B-Instruct-Turbo` (FP8)
Sampling		
Temperature	0.1	0.1
Top-p	provider default (unset)	provider default (unset)
Max new tokens	8192	4096
Stop sequences	none	none
Frequency penalty	unset	unset
Presence penalty	unset	unset
`logit_bias`	unset	unset
Random seed	not passed	not passed
Streaming	no	no
`response_format` / JSON mode	not requested	not requested
Retry and time-out		
Time-out (per call)	180 s	300 s
Retry attempts	5	5
HTTP 429 back-off	$\max(10, 2^{\text{attempt}+1})$ s	$\max(10, 2^{\text{attempt}+1})$ s
Other transient back-off	2^{attempt} s	2^{attempt} s
Rate-limit delay	2.0 s between calls	2.0 s between calls
Pipeline runtime		
Context window W	3 (last 3 exchanges)	3 (last 3 exchanges)
Parallelism	1 (sequential per round)	1 (sequential per round)
LLM calls per round	3 (Run 2 / Run 1); 3 + 0–1 conditional (Run 0)	2 (Run 1); 4 (Run 0); 3 (Run 2)
Other components		
Wasserstein implementation	POT [**?**]	—
Anomaly threshold	$\mu^I_{<r} + 2\sigma^I_{<r}$, warm-up $r \leq 3$	—
JSON parser fallbacks	fences, accent-insensitive keys, close-bracket repair	fences, accent-insensitive keys, close-bracket repair

6.1. Reproducibility parameters

Table 7 reports the inference configuration for the submitted runs in full. The configuration is identical across reACT and tACT at the level of provider, checkpoint, temperature, top-p, stop sequences, and sampling-seed policy; the two heads differ only in their per-call maximum-token budget (8192 for reACT to accommodate the longer CompACT-10 chain-of-thought, 4096 for tACT) and in their per-call time-out (180 s for reACT, 300 s for tACT to absorb the three-ordering PERM evaluator in Run 0).

6.2. Ablation grids

Two ablation grids drove our run selection, one per ACTalk head. Both were run on the 19-round trial conversation and on a synthetic simulated cohort that we generated to extend the severity range

covered by trial alone. The grid structures are visible in the configuration matrices of §5.5. Here, we describe how those grids were exercised and on what data.

reACT grid. The reACT grid crosses six calibration levels with three temporal aggregation strategies, namely {A0, A1, A2, A3, A4, A5} × {T0, T2, T3}. Cells A2 and A4 are diagnostic-only (§5.6): they isolate the contribution of the Level A prompt anchors under each higher-level calibration tier and were never candidates for submission. The remaining sixteen cells were exercised on the 19-round trial conversation and on six simulated personas of 15 rounds each, spanning PHQ-9 totals in $[4, 14]$ and GAD-7 totals in $[3, 20]$ across a range of CompACT-10 flexibility profiles. The trial cohort yields a single point per configuration; the simulated cohort provides six independent observations per configuration.

tACT grid. The tACT grid is wider. Its primary axes are the pipeline variant {B, B+} and the evaluator framing {FUNC, HYB, TOM-B, TOM-C}. Its secondary axes are the prompt language {ES, EN}, the context window $\{W = 1, W = 3, W = 5\}$, and the selection strategy {FIX, PERM}. In practice we exercised twelve primary configurations (the eight B ∪ B+ cells × four framings, with the secondary axes collapsed to the submitted-run values on cells that reached submission and varied only on the cells we ablated). Each configuration was run on 18 trial-conversation rounds and on seven simulated sessions totalling 87 rounds.

Synthetic-persona simulator. The simulator that generated the simulated cohort uses the same Llama-3.3-70B backbone but with a higher sampling temperature (0.7 versus 0.1 for inference) to introduce naturalistic variation in patient turns.

6.3. Truncation Disclosure

We disclose a deployment-time bug in our submission client. A configuration default, `--max-rounds=30`, was carried over unchanged from the trial-conversation runs into the live evaluation client and caused our submission to terminate after round 30 of each test-cohort session. The released test cohort comprises ten sessions of 82 rounds in total. Under the bug, our submitted predictions cover rounds 1 through 30 of each session and no rounds beyond round 30. The bug was identified and fixed after submission; the post-fix client raises a hard error when the cap would be reached during a live session. We note the bug here, in the experimental record, rather than in the limitations section, because it is an artefact of how we ran the system, not a limitation of the system itself.

The leaderboard scored our submission as if a default prediction had been emitted for every round we did not submit, namely as if the system had ceased contributing useful signal from round 31 onwards. This is the meaning of our official metrics in §7: they reflect the truncated submission, not the system's behaviour on the full 82-round cohort. To separate the question of *what the leaderboard measured* from the question of *what the system would have produced*, we ran a full-replay analysis after submission, re-executing the same ACTalk configuration on all 82 rounds of each session through a different provider (DeepInfra). The full-replay analysis appears in §8. In summary, the per-task deltas between the leaderboard metrics and the full-replay metrics are non-zero but small (on the order of 0.06 MAE on reACT GAD-7 Run 2, for example), confirming that the leaderboard scored only submitted rounds and that the missing rounds did not contain configuration-changing signal.

7. Results

We report what we observed on the released test cohort, in three movements. First, we describe the cohort itself: its session counts, its round counts, its gold-label distributions, and one structural asymmetry between the released cohort and the gold cohort that affects how leaderboard scores should be read (§7.1). Second, we report the headline numbers from the official leaderboard on both tasks (§7.2) and unpack the result on each task: per-instrument structure on Task 1 (§7.3), community-wide

Table 8
Released test cohort, per-session descriptive statistics and gold-label totals. *Rounds* is the number of patient turns in the session; *Words* is the total patient-side word count. PHQ-9 band thresholds per Kroenke et al. (2001); GAD-7 band thresholds per Spitzer et al. (2006); CompACT-10 has no validated severity bands (§2.3). Seven additional sessions (S02, S08, S10, S11, S13, S14, S17) are present in the official gold-label set but were not released to participants; the leaderboard treats our submission as emitting defaults for those sessions.

Session	Rounds	Words	PHQ-9 / band	GAD-7 / band	CompACT-10
S01	63	2022	16 / mod. severe	16 / severe	37
S03	50	1453	22 / severe	18 / severe	38
S04	61	2168	13 / moderate	15 / severe	44
S05	48	2593	15 / mod. severe	16 / severe	40
S06	58	2677	16 / mod. severe	10 / moderate	35
S07	30	986	2 / minimal	7 / mild	22
S09	67	2183	21 / severe	19 / severe	45
S12	82	2276	11 / moderate	17 / severe	50
S15	59	2005	11 / moderate	16 / severe	36
S16	50	2154	7 / mild	14 / moderate	33

difficulty on Task 2 (§7.4). Third, we summarise what the ablation grids told us about the submitted-run choices on each head (§7.5, §7.6). Interpretation and post-submission analysis are deferred to §8 and §9.

7.1. Exploratory data analysis of the released test cohort

The released test cohort contains ten conversational sessions (S01, S03, S04, S05, S06, S07, S09, S12, S15, S16) totalling 568 patient turns across 82 round positions. Mean session length is 56.8 rounds; the shortest session (S07) is 30 rounds, the longest (S12) is 82 rounds. Mean patient words per session is 2052 (median 2161), and mean patient words per round is 36.1.

Gold-label distributions, Task 1. The cohort is heavily weighted toward moderate-to-severe presentations. On PHQ-9, two sessions sit in the *severe* band (total ≥ 20), three in *moderately severe* (15–19), three in *moderate* (10–14), one in *mild* (5–9), and one in *minimal* (≤ 4). On GAD-7 the skew is sharper: seven of ten sessions land in the *severe* band (≥ 15), two in *moderate* (10–14), one in *mild* (5–9), and none in *minimal*. On CompACT-10 the gold totals span 22–50 on the 0–60 scale; using the informal-thirds convention from §2.3 (0–20 low, 21–40 medium, 41–60 high), six sessions sit in the medium band of psychological flexibility and four in the high band, with none in the low band. Table 8 reports the per-session gold totals.

Gold-label distributions, Task 2. The Task 2 gold is the index of the correct therapist response among the three candidates at each round. The 568 patient-rounds are distributed near-uniformly across the three classes, namely 35.7% gold=1, 32.8% gold=2, and 31.5% gold=3. The near-uniform distribution differs from the trial conversation (where option 3 was the gold in roughly half of rounds) and means the random baseline accuracy on the test cohort is approximately 0.363. The balanced distribution does not, however, imply that the three classes are equally easy to predict: the consensus-failure analysis we report in §8 shows that gold=3 rounds are categorically harder across the LLM family.

7.2. Headline Numbers

`ACTalk` placed 10/53 on Task 1 and 24/47 on Task 2 on the official leaderboard. The best Task 1 submission was `reACT` Run 2 (A1-T2) at MAE_Combined 1.063, approximately 0.18 above the top-ranked submission (FBKillers at 0.879). The best Task 2 submission was tACT Run 2 (B+/HYB/FIX) at accuracy 0.247, below the random baseline of 0.363. The random-baseline shortfall on Task 2 is not a property of our system alone: only two of the forty-seven submissions exceeded the random baseline

Table 9
Headline leaderboard numbers. Lower is better on Task 1 (MAE_Combined on PHQ-9, GAD-7, and CompACT-10 totals). Higher is better on Task 2 (accuracy of correct-option selection). The numbers in this table reflect the truncated submission described in §6.3. The full-replay deltas appear in §8.

Submission	Configuration	Task 1 (MAE_Combined)	Task 2 (accuracy)
ACTalk Run 0	reACT A5-T3 / tACT B/FUNC/PERM	1.107	0.210
ACTalk Run 1	reACT A3-T2 / tACT B/FUNC/FIX	1.103	0.237
ACTalk Run 2	reACT A1-T2 / tACT B+/HYB/FIX	**1.063**	**0.247**
Top-ranked (Task 1)	FBKillers Run 2	0.879	—
Top-ranked (Task 2)	NLP Innovators	—	0.393
Gemma baseline	organiser-provided	1.002	0.180
Random baseline	uniform over $\{1, 2, 3\}$	—	0.363

Table 10
Per-instrument MAE on Task 1 for the three submitted reACT runs against the Gemma baseline. The best ACTalk number on each instrument is in bold and is achieved by a different submitted run: PHQ-9 by Run 0, GAD-7 by Run 2, CompACT-10 by Run 1. The per-instrument leaderboard ranks across the nineteen teams and baselines are: PHQ-9 MAE rank 4 (Run 0), GAD-7 MAE rank 8 (Run 2), CompACT-10 MAE rank 3 (Run 1). The MAE_Combined ranking aggregates these into a single rank-10 placement; the strongest per-instrument showing is on CompACT-10, where Run 1 places third in the field.

Submission	PHQ-9 MAE	GAD-7 MAE	CompACT-10 MAE
ACTalk Run 2 (A1-T2)	0.829	**1.036**	1.324
ACTalk Run 1 (A3-T2)	0.817	1.191	**1.302**
ACTalk Run 0 (A5-T3)	**0.797**	1.159	1.366
Gemma baseline	0.724	0.582	1.700

at all. Table 9 reports our three submissions per task alongside the top-ranked submission on each task and the two official baselines (Gemma, Random). The Gemma baseline on Task 1 sits at rank 3 (MAE_Combined 1.002), the strongest organiser baseline on either task.

7.3. reACT per-instrument structure

The MAE_Combined number compresses three per-instrument MAEs into a single rank. Decomposing it reveals a structured result on reACT: strong on CompACT-10, competitive on PHQ-9, and the failure mode concentrates on GAD-7. Table 10 reports the per-instrument MAEs for all three submitted reACT runs against the Gemma baseline, with the best ACTalk number on each instrument highlighted.

CompACT-10. CompACT-10 is reACT's strongest instrument and the per-instrument ranking attributes it to Run 1 (A3-T2): MAE 1.302, third on this instrument across the leaderboard, outperforming the Gemma baseline (1.700) by 0.398 per-item MAE. Run 2 (A1-T2, no Level B rules) is slightly behind at 1.324, and Run 0 (A5-T3) is the worst of the three at 1.366. The Level B distress-conditional rescaling rule introduced in §5.2 is the principal source of the Run 1 advantage: it counteracts the LLM floor effect on psychological-flexibility scoring in severe-distress patients. Of the ten test sessions, four sit in the high-flexibility band (gold ≥ 41) and six in the medium band (21–40); the rescaling correctly pulls CompACT-10 totals downward toward the cohort-conditional expectation on the severe-distress sessions. Mind that the configuration that wins CompACT-10 is not the configuration that wins MAE_Combined (Run 2): the per-instrument optimum and the aggregate-metric optimum disagree.

PHQ-9. PHQ-9 is competitive. Run 2 achieves MAE 0.829 against the Gemma baseline's 0.724, a per-item gap of 0.105. Per-item decomposition shows the gap is distributed across the nine items rather

Table 11
reACT submitted-run comparison across the trial, simulated, and test cohorts. The dev cohorts were scored with mean per-instrument RMSE; the test cohort with MAE_Combined per the official metric. The two are not directly comparable in absolute terms but both rank-order configurations within their cohort, which is what the dev-vs-test comparison requires. The three cohorts do not agree on a single best configuration.

Configuration	Trial (Mean RMSE)	Simulated (Mean RMSE)	Test (MAE_Comb.)	Rank on test
A1-T2 (Run 2)	0.457	**1.272**	**1.063**	best-on-test (10/53)
A3-T2 (Run 1)	0.457	1.311	1.103	middle (12/53)
A5-T3 (Run 0)	**0.405**	1.320	1.107	worst-on-test (13/53)

than concentrated. No individual item drives the deficit. The aggregator T2 (early-weighted, with rounds 1–5 doubled) appears to be doing the expected work: scoring is robust to within-session calming on rounds 6+ where the assessor's per-round prediction tends to soften (the behaviour discussed in §5.2 on round 12 of the trial; Appendix B).

GAD-7. GAD-7 is the failure. Run 2 produces MAE 1.036, against the Gemma baseline's 0.582, a gap of 0.454 per-item MAE in the wrong direction. Per-item breakdown attributes the failure primarily to items 5 (restlessness) and 6 (irritability): the assessor under-predicts these two items by an average of roughly 1.4 points across the cohort, against a maximum item value of 3. The under-prediction is signed and systematic rather than noisy; the running-mean signed error on GAD-7 totals across the cohort is −6.0 points. Mind that the GAD-7 failure does not respond to the Level B rules that improved the CompACT-10 result: the rules adjust between-instrument consistency, not within-instrument item priors, and an item-prior failure is reached by neither Level B nor Level C on the configurations submitted. The Gemma post-hoc analysis in §8.3 returns to this failure mode under a different LLM and a revised prompt.

7.4. tACT Community-Wide Difficulty Regime

ACTalk's mid-pack rank on Task 2 sits inside a difficulty regime where the average submission performs below the baseline. Among the forty-seven submitted Task 2 runs, only two, namely NLP Innovators (accuracy 0.393) and VerbaNex AI (accuracy 0.386), exceed the random baseline of 0.363. The Gemma baseline runs at accuracy 0.180, placing rank 42, below random. ACTalk's best Task 2 submission (Run 2, B+/HYB/FIX) places rank 24 at accuracy 0.247; all three of our submissions sit within 0.015 of one another. The narrow within-team spread is consistent with the community-wide pattern: with so few submissions above the random baseline, configuration-level engineering on tACT is not strongly discriminative on the test cohort.

Across the nine submitted systems we were able to inspect on the inner-join 299 rounds, gold=3 rounds are *all-systems-wrong* 38.5% of the time and *all-systems-correct* in 0% of cases, against a mean of 1.71 of 9 systems correct per gold=3 round. The near-uniform gold-class distribution does not translate into uniform per-class difficulty. We return to this finding in §8.5.

7.5. reACT Ablation

Recall reACT ablation grid (§6.2 and configuration matrix in Table 4). The submitted runs exercise three of these cells: A5-T3 (Run 0, full stack), A3-T2 (Run 1, anchors and rules), and A1-T2 (Run 2, anchors only). On the test cohort the simplest configuration is best (Run 2 at MAE_Combined 1.063, rank 10), the intermediate configuration is close behind (Run 1 at 1.103, rank 12), and the full-stack configuration is worst of the three submitted (Run 0 at 1.107, rank 13). Table 11 summarises the development-versus-test comparison for the three submitted reACT cells.

Table 12
tACT submitted-run comparison across the trial, simulated, and test cohorts. Values are accuracy on Task 2; higher is better. Cohen's κ included where logged, since the small trial $n = 18$ makes raw accuracy unstable. Dev-cohort values come from the post-prompt-update ablation that matches the submitted system; the earlier ablation, run before the evaluator prompt was finalised, gave different numbers. The three cohorts pick three different winners: Run 1 wins on trial, Run 0 wins on simulated, Run 2 wins on test, in exactly that order. The test ordering is the reverse of the simulated ordering. The trial cohort's small size and the ± 0.15–0.33 κ variance across 18 rounds put inter-configuration accuracy differences inside the noise floor.

Configuration	Trial (acc / κ)	Simulated (acc / κ)	Test (acc)	Rank on test
B+/HYB/FIX/W3 (Run 2)	0.444	0.897	**0.247**	best-on-test (24/47)
B/FUNC/FIX/W3 (Run 1)	**0.500** / 0.264	0.920 / 0.879	0.237	middle (25/47)
B/FUNC/PERM/W3 (Run 0)	0.389 / 0.083	**0.943** / 0.914	0.210	worst-on-test (31/47)

The simulated cohort correctly predicted the test ranking. The trial cohort did not. The trial result, namely that the full-stack A5-T3 configuration won at 0.405 against the two simpler configurations tied at 0.457, is the expected ordering by engineering depth and matches the design intent. The simulated and test cohorts both reverse this: A1-T2 wins, A5-T3 loses. The simulated cohort's margins are narrow (0.048 Mean RMSE between best and worst) and the underlying six synthetic personas span the severity range (PHQ-9 $\in [4, 14]$, GAD-7 $\in [3, 20]$) rather than the single moderate-distress trial conversation. We read the trial-vs-rest inversion as evidence that the trial cohort is too small to discriminate between configurations whose differences depend on cohort heterogeneity, and the trial's apparent preference for the engineered configuration reflects over-tuning to that one conversation rather than a general advantage. The diagnostic-only A2 and A4 cells confirmed their design intent, namely that removing the Level A prompt anchors collapses the assessor toward the middle of the scale regardless of which higher-tier machinery sits downstream.

7.6. tACT Ablation

Recall the tACT ablation grid (§6.2 and configuration matrix in Table 5). The three submitted tACT runs exercise three different combinations of these axes: B/FUNC/PERM/W3 (Run 0), B/FUNC/FIX/W3 (Run 1), and B+/HYB/FIX/W3 (Run 2). On the test cohort the most-engineered configuration is best (Run 2, accuracy 0.247, rank 24), the default configuration is second (Run 1, 0.237, rank 25), and the variance-minimisation configuration is the worst of the three submitted (Run 0, 0.210, rank 31). The within-team spread is 0.037 across the three runs and the ordering is monotonic in engineering depth. Table 12 summarises the comparison.

The three cohorts pick three different winners: Run 1 (B/FUNC/FIX) on trial at 0.500, Run 0 (B/-FUNC/PERM) on simulated at 0.943, and Run 2 (B+/HYB/FIX) on test at 0.247. The test ordering, namely Run 2 first, Run 1 middle, Run 0 worst, is the exact reverse of the simulated ordering, namely Run 0 first, Run 1 middle, Run 2 worst. The trial cohort sits between the two with yet another ordering and κ variance on the order of ± 0.15–0.33 across 18 rounds, large enough to mask configuration differences below ~ 0.1. The variance-minimisation rationale that motivated PERM and placed Run 0 ahead of Run 1 on the simulated cohort did not survive the test cohort, namely the permutation orderings did not produce more reliable selection on heterogeneous patient conversations than the fixed-ordering FIX baseline. Theory-of-Mind framings TOM-B and TOM-C underperformed FUNC on the simulated cohort (0.667 and 0.678 against FUNC's 0.920) and were not submitted; this finding held across the ablation as the most stable secondary-axis result.

8. Post-Submission Analysis

The submitted ACTalk runs are the official record, but three open questions remain after the leaderboard closes:

1. What would the leaderboard actually score under the truncation reported in §6.3?
2. Can the failure modes visible in §7.3 and §7.4 be intervened on, and at which head the intervention belongs?
3. What characterises the dev-vs-test rank disagreement that emerged from the ablation summaries in §7.5 and §7.6?

This section reports the post-submission work that addresses each question. Mind that the analyses described here are post-hoc and were not part of the official submission.

8.1. Full-replay Verification of the Truncation

The truncation reduced submission coverage to rounds 1–30 of the 82 test rounds per session. To address that, we re-ran ACTalk on all 82 rounds using the same Llama checkpoint, the same prompts, and the same client logic, with the rounds cap raised to 200 and the hard-error exit enabled. The replay run is a single seed and a single provider call sequence; we did not bootstrap variance.

The replay confirms that the leaderboard scored only the submitted rounds and did not impute defaults for the remaining ones. The submitted-vs-replay per-instrument deltas on item-MAE are small in absolute terms (typically below 0.1) and of mixed sign across the three reACT runs. Run 0 shows $\Delta_{\text{PHQ-9}} = +0.111$, $\Delta_{\text{GAD-7}} = +0.043$, $\Delta_{\text{CompACT-10}} = -0.060$; Run 1 shows $+0.067$, -0.043, $+0.010$; Run 2 shows $+0.100$, $+0.114$, -0.030. The replay item-MAE is slightly worse than the submission item-MAE on PHQ-9 and GAD-7 across all three runs, and slightly better on CompACT-10 for two of the three runs. We read the direction of the deltas as evidence that rounds 31–82 contain more cohort-specific content that the assessor mis-classifies at the item level on the depressive and anxiety scales, while CompACT-10 totals stabilise as the session progresses, namely the temporal dynamic the T3 stability-adaptive aggregation in Run 0 was designed to exploit. We conclude that the truncation cost ACTalk coverage rather than quality.

8.2. Failure Diagnostic

Two failure modes carry over from §7. The first is reACT's GAD-7 under-prediction in a severe-cohort test (item-MAE 1.036, well behind the Gemma baseline at 0.582, signed bias -5.8). The second is tACT's below-baseline placement on Task 2 (accuracy 0.247, below the random baseline at 0.333 and far behind the leaderboard top at 0.393). Both ACTalk heads use the same Llama backbone with task-specific calibration, evaluator, and selection machinery layered on top, and both failures plausibly stem from the same architectural choice. Specifically, both heads spend a substantial fraction of their prompt budget on explicit clinical or rubric-grounded reasoning that may suppress the model's native priors on the relevant judgements. To test this, we ran two arms of a post-hoc Gemma branch: a model-and-prompt swap at the reACT GAD-7 assessor and a bare-LLM substitution for the entire tACT pipeline. The two arms isolate whether the intervention that helps each failure mode is the same.

8.3. Arm A — reACT × Gemma on GAD-7

The best Gemma configuration on the GAD-7 assessor reduces item-MAE from Llama's 1.086 to 0.714, a relative reduction of 34%, with signed bias from -6.0 to -3.4 and band accuracy from 0.20 to 0.50. The winner is a Gemma 4 26B mixture-of-experts variant with approximately four billion active parameters, paired with a v2 prompt that adds a severe-cohort anchor example (gold = 17) and explicit indirect-evidence markers for items 5 and 6. The same Gemma 4 26B MoE checkpoint with the v1 prompt scores 0.743. The v2 prompt thus contributes about a quarter of the gain on top of the model swap. The dense Gemma 4 31B variant with the v1 prompt scores 0.786, worse than the smaller MoE variant on the

same task, and Gemma 3 27B scores between 0.743 (v2) and 0.814 (v1). The v2 prompt cuts the signed bias on item 5 from −1.4 to about −0.4 in absolute terms; item 6 remains broken across all Gemma variants (−1.6 on Gemma 4 26B MoE), namely the under-prediction of irritability is shared across the LLM family rather than being specific to Llama.

A hybrid pipeline that substitutes Gemma 4 26B MoE v2's GAD-7 prediction into reACT Run 1's PHQ-9 (0.778) and CompACT-10 (1.23) yields MAE_Combined 0.907 on the released 10-session subset. This projects to rank 4 of 53 on the official leaderboard, behind FBKillers (0.879, rank 1), VerbaNex AI (0.883, rank 2), and the Gemma baseline (1.002, rank 3). Out of the thirty hybrid configurations we computed by crossing the three Gemma variants with the two prompt versions and the three submitted runs, twenty-eight project to rank 4; the two that fall to rank 6 both use the weaker Gemma 3 27B v1 GAD-7. We do interpret it as evidence that, for reACT's GAD-7 failure mode, a model-and-prompt swap is the right intervention, and that the intervention point sits at the assessor rather than at the Level B rules or Level C agent of the calibration stack.

8.4. Arm B — tACT against a bare LLM

A bare-LLM substitution replaces the entire tACT pipeline with a single 100-token prompt asking the model which of three candidate therapist responses best continues the conversation. Gemma 4 31B in this configuration (mode S, no guardrails) reaches accuracy 0.412 on the 568-round full-replay test set, and the same model with a four-line anti-bias guardrails block prepended (mode S2) reaches 0.470. The S2 result is +22.3 percentage points above tACT's best submitted run (0.247) and +7.7 percentage points above the leaderboard top (NLP Innovators at 0.393). It is the strongest tACT result we recorded, by a margin that is not attributable to seed variance.

The result is architecture-sensitive rather than prompt-sensitive. Gemma 3 27B in the same bare mode scores 0.290 and Llama-3.3-70B scores 0.257, both below the random baseline of 0.333. Two further bare-LLM modes are tested in the same setup: mode S3 (six-permutation majority vote) reaches 0.400 and mode S4 (pairwise Condorcet) reaches 0.354, both lower than mode S2. Permutation voting on the bare LLM thus reproduces the same penalty we observed for tACT Run 0 (B/FUNC/PERM) on the test cohort. The additional structure imposed by ordering averages does not pay off on heterogeneous patient conversations. In contrast to Arm A, the gain on tACT comes from removing structure rather than swapping the model alone, Thus, Llama-3.3-70B in the bare configuration underperforms the engineered tACT pipeline. Mind that this means the Gemma model has a stronger native prior on therapeutic response selection than Llama does. The ACT-FM scoring used by tACT does not recover that prior on the bare side and may suppress the equivalent prior in the Llama backbone.

A head-to-head comparison on the 300 inner-join rounds between tACT Run 2 and Gemma 4 31B S2 confirms the asymmetry at the round level. Gemma S2 wins (S2 right, tACT wrong) on 91 rounds; tACT wins on 40. The S2-favoured class is gold = 2, where S2 accuracy exceeds tACT accuracy by 23 percentage points (0.453 against 0.222); the next largest absolute gain is on gold = 3 at +20.6 percentage points.

8.5. Consensus-Failure Analysis on Task 2

Across nine Task 2 systems on the 299-round inner join of our submitted run, our full-replay run, the five bare-LLM Gemma modes, and two further bare-LLM controls, the rate at which all nine systems miss the gold response varies by gold class. On gold = 1 rounds ($n = 101$), 17.8% of cases are missed by every system; on gold = 2 rounds ($n = 94$), 21.3%; on gold = 3 rounds ($n = 104$), 38.5%. The proportion of rounds on which every system selects the gold response is 3.0%, 1.1%, and 0.0% for the three classes respectively. The mean number of correct systems per round falls from 3.14 out of nine on gold = 1 to 1.71 out of nine on gold = 3. We conclude that gold = 3 is categorically the hardest class on Task 2.

8.6. Dev-vs-test Rank Disagreement

Our three evaluation cohorts during development (the official trial conversation, the simulated personas, and the released 10-session test cohort) do not agree on which submitted run is best. Table 11 reports that reACT's trial ranking placed Run 0 (A5-T3, full stack) ahead of the two simpler configurations at 0.405 against 0.457 for Run 1 and Run 2, namely the ordering by engineering depth that the design intended. The simulated cohort reversed this, ranking Run 2 (A1-T2, simplest) best at 1.272 and Run 0 worst at 1.320. The test cohort confirmed the simulated ordering: Run 2 best at 1.063, Run 0 worst at 1.107. Table 12 reports the same pattern in a stronger form for tACT: the three cohorts pick three different winners, Run 1 on trial (0.500), Run 0 on simulated (0.943), and Run 2 on test (0.247). The test ordering for tACT is the exact reverse of the simulated ordering. The bare-LLM Gemma S2 ties our submitted-equivalent at 0.444 on trial and trails it slightly on simulated (0.943 against 0.897), but leads by +22.3 percentage points on test. None of the three pre-submission cohorts predicted this gap.

The proximate reason for the under-discrimination differs by cohort and is, in retrospect, identifiable from cohort-level properties. The trial cohort contains 19 rounds for reACT and 18 rounds for tACT, and Cohen's κ variance on the latter is on the order of ±0.15–0.33, large enough to mask configuration differences below ~0.1 accuracy. A single moderate-distress patient also exposes only a narrow slice of the severity space, and a configuration that calibrates well to that slice need not transfer. The simulated cohort spans the severity range by construction but the synthetic personas are generated cleanly enough that most reasonable configurations sit near a ceiling and discriminate poorly between configurations. In contrast to either of these, the test cohort is heterogeneous across patients, severity bands, and conversational styles, and configuration differences that survive its noise floor become visible. The Gemma S2 result on the test cohort is the clearest example: a configuration that is indistinguishable from ours on trial and trails by a few points on simulated leads decisively on test.

We acknowledge that the trial-cohort, simulated-cohort, and test-cohort mismatch we observed is specific to our development choices in this campaign, and whether the same disagreement would appear for other teams or in other editions is an empirical question this paper does not answer.

9. Discussion

9.1. Why reACT's CompACT-10 Engineering Worked while GAD-7 Engineering Failed

reACT's CompACT-10 result (rank 3 on the per-instrument leaderboard, MAE 1.302 for Run 1, well ahead of the Gemma baseline at 1.700; §7.3) and its GAD-7 result (item-MAE 1.036 on the severe test cohort, well behind the Gemma baseline at 0.582) sit on the same backbone, the same calibration stack, and the same context window. The asymmetric outcome is informative about where each instrument's failure mode lives. CompACT-10's failure under Llama is an instrument-scale failure: the model anchors its flexibility estimate to its conversational impression of distress and produces totals clustered near the scale's lower end regardless of the patient's expressed CompACT content. Distress-conditional rescaling against PHQ-9 and GAD-7 anchors, a Level B rule, intervenes at this point and recovers most of the loss. GAD-7's failure under Llama is an item-prior failure rather than an instrument-scale failure: the model under-predicts items 5 (restlessness) and 6 (irritability) by margins that are consistent within the severe cohort and consistent across the calibration stack. Level B rules operate on totals and cross-instrument ratios, namely not on per-item priors within a single instrument, and Level C agent invocations did not fire enough to alter the item-level pattern. The Gemma branch results in §8.3 confirm this reading: a model-and-prompt swap at the assessor closes most of the gap on GAD-7 with no change to the calibration stack downstream, and the gain transfers cleanly when Gemma's GAD-7 predictions are substituted into the otherwise-unchanged reACT pipeline.

The clinical-cost framing from §2.3 reads this result somewhat against the leaderboard ordering. The asymmetric clinical cost of band-crossing errors on GAD-7 means our rank-8 finish carries a heavier clinical-pathway consequence than the per-instrument rank-table difference suggests, while the CompACT-10 win sits inside an instrument with no validated severity bands and so carries a lighter one.

The unweighted MAE metric obscures this asymmetry. We discuss the deployment implications in §11.

9.2. Why tACT's ACT-process Scoring Lost to a Bare LLM

The bare-LLM Gemma S2 result on Task 2 (+22.3 percentage points over tACT's best submitted run. §8.4) is the strongest finding in the paper that argues against tACT's architectural choices. The result is not that prompt engineering does not help on Task 2: S2 itself depends on a four-line anti-bias guardrails block prepended to a bare conversational prompt, and S2 outperforms the bare S prompt by 5.8 percentage points. The result is rather that tACT's ACT-FM-grounded evaluator, the characterisation step, the state tracker, and the per-candidate functional-analysis scoring jointly overrode the model's native conversational judgement. Within tACT's engineered family, marginal engineering helps on test in the expected direction (B+/HYB above B/FUNC/FIX above B/FUNC/PERM). Across families, the bare-LLM prior dominates by 22.3 percentage points. The natural reading is that the engineered family is operating inside a regime where the engineering removes more signal than it adds.

We hypothesised in the system design that explicit per-candidate ACT-FM scoring would discriminate between candidate responses by making the hexaflex content of each candidate inspectable. The consensus-failure result in §8.5 bears on this hypothesis. Across nine systems on the 299-round inner join, gold = 3 rounds are categorically the hardest, with 38.5% of cases missed by every system and 0% correctly predicted by all nine. One possibility for the option-3 gold class is that it mixes the riskiest direct question and the most elaborate intervention in ways that produce no consistent surface signal for any rubric-based evaluator to discriminate. We note that this is a hypothesis rather than an established finding: verifying it would require qualitative coding of the option-3 rounds against ACT-FM and an alternative rubric, which we did not complete and which is the natural follow-up in §10.

9.3. The dev-vs-test Gap

§8.6 reports the three pre-submission cohorts disagreeing on which submitted run is best. The proximate causes are identifiable. The trial cohort's $n = 1$ reACT session and $n = 18$ tACT rounds yield κ variance estimates on the order of ± 0.15–0.33, large enough to mask configuration differences below ~ 0.1 accuracy. The variance-minimisation argument that motivated tACT Run 0 (B/FUNC/PERM) was built on these unreliable variance estimates rather than on a heterogeneous dev cohort, and the test cohort did not confirm it. Our simulated personas spanned the severity range we expected the test cohort to present but, in retrospect, did not span the difficulty range. The six personas were generated cleanly enough that reACT configurations sat near a dev-cohort ceiling, with 0.048 Mean RMSE between best and worst, and any rank ordering inside that envelope is under-determined by the dev evidence. The test cohort's 7 of 10 sessions in the GAD-7 severe band produced the difficulty heterogeneity the dev cohorts lacked.

The operational changes we would make in a future submission derive from this analysis: (1) synthetic personas should be generated to span the difficulty range observed on real cohorts as well as the severity range if known; (2) configuration rankings that disagree between trial and simulated should be treated as high-uncertainty regardless of which cohort favours which; and (3) secondary-axis arguments built on variance estimates from small trial cohorts should be corroborated against a heterogeneous dev cohort before being acted on at submission time.

10. Conclusion

In this paper we presented ACTalk, a shared LLM backbone with two task-specific heads for the MentalRiskES 2026 campaign: reACT, a calibrated psychometric assessor producing PHQ-9, GAD-7, and CompACT-10 item vectors at every round; and tACT, an ACT-FM-grounded evaluator that scores three candidate therapist responses per turn and selects one. The system was run online against the organisers' server during the campaign's evaluation window and submitted three runs per task. The

implementation, the persona-generation code, and the post-hoc analysis scripts are released alongside this paper.

11. Limitations

11.1. Corpus Protocol

The MentalRiskES 2026 corpus is simulated yet expert-reviewed (§4.2): patients are role-played by psychology students against assigned PHQ-9, GAD-7, and CompACT-10 profiles, and the licensed psychologist therapists chose responses from AI-generated suggestions with the option to accept, edit, or replace each one. Two methodological constraints follow. First, Task 1 gold labels reflect the assigned profile rather than a clinical interview, so `reACT` is evaluated on its ability to recover scores the patient was instructed to embody rather than on its ability to interview. A system that under-recovers some items may be failing to model the student's enactment rather than failing to detect underlying symptoms. Second, Task 2 gold reflects the licensed psychologist's selection (or written substitute) at each turn, which means `tACT` is evaluated against an editorial process that includes AI suggestion generation upstream. The patterns we report are patterns of agreement with this editorial process, not patterns of agreement with blind clinical adjudication, and we cannot decompose the two contributions from the released data. Addressing this would require an independent adjudication corpus or a parallel naturalistic clinical dataset. The MentalRiskES 2026 release does not include either.

11.2. Evaluation Scale

The released test cohort is 10 sessions and the official gold scoring set is 17 sessions (§4.2). Our descriptive statistics, ablations, and post-hoc analyses operate on the released 10-session subset, and the rank-4 projection in §8.3 is computed on this subset rather than on the full 17-session scoring set. The trial cohort is a single 19-round `reACT` session and a single 18-round `tACT` session. The simulated cohort is six 15-round personas for `reACT` and seven sessions for `tACT` (§4.3). Bootstrap 95% confidence intervals on `tACT` trial-cohort accuracy span ±0.22 in absolute terms, namely as wide as the inter-team spread on the leaderboard's top half. The dev-vs-test rank disagreement we report in §8.6 should be read as a finding about three specific cohorts at this scale, not as evidence for or against any general claim about the transferability of dev-cohort rankings. A more confident conclusion would require either a larger test cohort or a held-out portion of the released test data, neither of which is available.

11.3. Single-Seed and Single-Provider Sensitivity

`ACTalk`'s submitted runs and the full-replay run in §8.1 are single-seed at the assessor and selector, namely we did not bootstrap variance across temperature sampling or across LLM-provider seed resets. The bare-LLM Gemma comparison in §8.4 is also single-seed. The Llama-3.3-70B-Instruct-Turbo we used for the submission is a Together AI provider variant, and the post-hoc Gemma comparison ran on Gemma 4 and Gemma 3 checkpoints accessed through a separate provider call sequence. Provider-side quantisation, sampling implementation, and turbo vs FP16 differences are not controlled. The conclusions we draw at the per-instrument and per-run level (`reACT`'s CompACT-10 vs GAD-7 split; `tACT`'s monotonic engineering ranking on test; the +22.3 percentage-point bare-LLM Gemma S2 gap) are large enough that single-seed variance is unlikely to reverse them, but configuration differences below ~0.05 MAE_Combined or ~0.03 accuracy should not be read as seed-robust on the basis of our experiments. Addressing this would require bootstrap runs across at least three seeds per configuration and a controlled-decoding setup at a single inference provider, which we did not have time to run within the campaign window.

11.4. Generalisability

The findings in this paper are specific to Spanish-language therapeutic conversation, one MentalRiskES edition, one set of clinical instruments (PHQ-9, GAD-7, CompACT-10), one ACT-FM rubric, and one editorial process. The CompACT-10 win on `reACT` depends on distress-conditional rescaling that uses PHQ-9 and GAD-7 totals as anchors, and this anchoring is plausible only because all three instruments are administered on the same patient session. A CompACT-10 deployment outside this co-administration would not have the anchors available. The `tACT` vs bare-LLM finding in §8.4 is architecture-sensitive (Gemma 4 31B leads, Gemma 3 27B and Llama-3.3-70B lag), and the asymmetry between Gemma's bare-LLM win and Llama's bare-LLM loss should not be read as a general claim about ACT-FM-grounded evaluation: a Llama-with-ACT-FM vs Llama-bare comparison gives a different ordering than a Gemma-with-ACT-FM vs Gemma-bare comparison would, and we did not run the Gemma-with-ACT-FM arm. The CompACT-10 instrument itself has only partial Spanish validation (§2.2), and our ability to attribute prediction errors to model behaviour rather than to instrument-language interaction is limited accordingly. Generalising any of these findings to other languages, other campaigns, other clinical populations, or other rubrics requires independent verification we have not performed.

11.5. Clinical Safety and Deployment Scope

`ACTalk` is a shared-task system, not a clinical-decision support tool, and the results in this paper do not constitute evidence for deployment in any clinical-pathway role. Two clinical-safety considerations bear specific mention. First, PHQ-9 Item 9 is a safety screener for suicidal ideation rather than a severity item (§2.3). `reACT`'s assessor predicts Item 9 as part of the PHQ-9 vector but the system has no explicit safety-flag pathway: a positive Item 9 prediction is treated as a score on a scale rather than as a routing signal for structured suicide-risk follow-up. Any deployment-adjacent use of `reACT` would need an Item 9-specific path that triggers human review on positive prediction, and the eC-SSRS-validated positive predictive value of approximately 29% for Item 9 (§2.3) means that flag-and-review rather than flag-and-act is the only defensible operationalisation. Second, the asymmetric clinical-cost framing in §9.1 notes that `reACT`'s GAD-7 failure has a heavier clinical-pathway consequence than its rank-table position suggests. A patient under-scored by `reACT` at the moderate-severe boundary may not receive the active treatment that either NICE NG222 or the Spanish GPC would recommend. This is a limitation of the underlying assessment task in clinical context, not a flaw the team can correct within the shared-task scope, and we state it here so that downstream readers do not conflate leaderboard performance with clinical-deployment readiness.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic), version `claude-opus-4-7`, in order to:

- *Improve writing style* and *Paraphrase and reword*: Claude assisted with sentence-level rephrasing during revision passes;
- *Formatting assistance*: Claude produced LaTeX drafts for tables and captions;
- *Drafting content*: Claude produced initial drafts for some parts of Sections 1-3 and 9-11 based on detailed outline and specified paragraphs structure. The author reviewed and edited all drafted content. Scientific insights, the contribution framing, and the interpretation of results were developed by the author.

After using these tool, 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. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[3] L. O'Neill, G. Latchford, L. M. McCracken, C. D. Graham, The development of the Acceptance and Commitment Therapy Fidelity Measure (ACT-FM): A delphi study and field test., Journal of Contextual Behavioral Science 14 (2019) 111–118. URL: https://linkinghub.elsevier.com/retrieve/pii/S2212144719300080. doi:10.1016/j.jcbs.2019.08.008.

[4] 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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0005796705002147. doi:10.1016/j.brat.2005.06.006.

[5] 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. URL: http://link.springer.com/10.1046/j.1525-1497.2001.016009606.x. doi:10.1046/j.1525-1497.2001.016009606.x.

[6] 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. URL: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.166.10.1092. doi:10.1001/archinte.166.10.1092.

[7] 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://linkinghub.elsevier.com/retrieve/pii/S2212144723000777. doi:10.1016/j.jcbs.2023.06.003.

[8] C. Diez-Quevedo, T. Rangil, L. Sanchez-Planell, K. Kroenke, R. L. Spitzer, Validation and Utility of the Patient Health Questionnaire in Diagnosing Mental Disorders in 1003 General Hospital Spanish Inpatients:, Psychosomatic Medicine 63 (2001) 679–686. URL: http://journals.lww.com/00006842-200107000-00021. doi:10.1097/00006842-200107000-00021.

[9] R. Muñoz-Navarro, A. Cano-Vindel, L. A. Medrano, F. Schmitz, P. Ruiz-Rodríguez, C. Abellán-Maeso, M. A. Font-Payeras, A. M. Hermosilla-Pasamar, Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centres, BMC Psychiatry 17 (2017) 291. URL: http://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-017-1450-8. doi:10.1186/s12888-017-1450-8.

[10] J. Garcia-Campayo, E. Zamorano, M. A. Ruiz, A. Pardo, M. Perez-Paramo, V. Lopez-Gomez, O. Freire, J. Rejas, Cultural adaptation into Spanish of the generalized anxiety disorder-7 (GAD-7) scale as a screening tool, Health and Quality of Life Outcomes 8 (2010) 8. URL: http://hqlo.biomedcentral.com/articles/10.1186/1477-7525-8-8. doi:10.1186/1477-7525-8-8.

[11] R. Muñoz-Navarro, A. Cano-Vindel, J. A. Moriana, L. A. Medrano, P. Ruiz-Rodríguez, L. Agüero-Gento, M. Rodríguez-Enríquez, M. R. Pizà, J. I. Ramírez-Manent, Screening for generalized anxiety disorder in Spanish primary care centers with the GAD-7, Psychiatry Research 256 (2017) 312–317. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165178117300847. doi:10.1016/j.psychres.2017.06.023.

[12] A. Giovannetti, J. Pöttgen, E. Anglada, R. Menéndez, J. Hoyer, A. Giordano, K. Pakenham, I. Galán, A. Solari, Cross-Country Adaptation of a Psychological Flexibility Measure: The Comprehensive Assessment of Acceptance and Commitment Therapy Processes, International Journal of Environmental Research and Public Health 19 (2022) 3150. URL: https://www.mdpi.com/1660-4601/19/6/3150. doi:10.3390/ijerph19063150.

[13] T. V. Sousa, V. Viveiros, M. V. Chai, F. L. Vicente, G. Jesus, M. J. Carnot, A. C. Gordo, P. L. Ferreira, Reliability and validity of the Portuguese version of the Generalized Anxiety Disorder (GAD-7) scale, Health and Quality of Life Outcomes 13 (2015) 50. URL: https://hqlo.biomedcentral.com/articles/10.1186/s12955-015-0244-2. doi:10.1186/s12955-015-0244-2.

[14] E. P. Terlizzi, M. A. Villarroel, Symptoms of generalized anxiety disorder among adults: United States, 2019, NCHS Data Brief 378, National Center for Health Statistics, Hyattsville, Maryland, US, 2020. URL: https://www.cdc.gov/nchs/products/databriefs/db378.htm.

[15] Overview | Depression in adults: treatment and management | Guidance | NICE, 2022. URL: https://www.nice.org.uk/guidance/ng222.

[16] A. Gundlach, K. D. Knight, Practice Guideline for the Treatment of Patients With Major Depressive Disorder: American Psychiatric Associationr, Third Edition (2010) 1–152. URL: http://access.oakstone.com/Uploads/Public/PracticeGuidelinefortheTreatmentofPatientsWithMajorDepressiveDisorderAmericanPsychiatricAssociation.pdf.

[17] S. H. Kennedy, R. W. Lam, R. S. McIntyre, S. V. Tourjman, V. Bhat, P. Blier, M. Hasnain, F. Jollant, A. J. Levitt, G. M. MacQueen, S. J. McInerney, D. McIntosh, R. V. Milev, D. J. Müller, S. V. Parikh, N. L. Pearson, A. V. Ravindran, R. Uher, the CANMAT Depression Work Group, Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments, The Canadian Journal of Psychiatry 61 (2016) 540–560. URL: https://journals.sagepub.com/doi/10.1177/0706743716659417. doi:10.1177/0706743716659417.

[18] World Health Organization, mhGAP intervention guide for mental, neurological and substance use disorders in non-Specialized health settings: mental health gap action programme (mhGAP), version 2.0 ed., World Health Organization, Geneva, Switzerland, 2016.

[19] G. d. t. d. l. G. d. P. C. s. e. M. d. l. D. e. e. A. MINISTERIO DE SANIDAD, SERVICIOS SOCIALES E IGUALDAD, Gu{í}a de Pr{á}ctica Cl{í}nica sobre el Manejo de la Depresi{ó}n en el Adulto, Technical Report C 1316-2014, Ministerio de Sanidad, Servicios Sociales e Igualdad. Agencia de Evaluación de Tecnologías Sanitarias de Galicia (avalia-t), 2014. URL: https://portal.guiasalud.es/gpc/depresion-adulto/.

[20] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and Brain Sciences 1 (1978) 515–526. URL: https://www.cambridge.org/core/product/identifier/S0140525X00076512/type/journal_article. doi:10.1017/S0140525X00076512.

[21] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early Risk Prediction on the Internet, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco De Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14959, Springer Nature Switzerland, Cham, 2024, pp. 73–92. URL: https://link.springer.com/10.1007/978-3-031-71908-0_4. doi:10.1007/978-3-031-71908-0_4.

[22] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 9822, Springer International Publishing, Cham, 2016, pp. 28–39. URL: http://link.springer.com/10.1007/978-3-319-44564-9_3. doi:10.1007/978-3-319-44564-9_3.

[23] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences, IEEE Transactions on Knowledge and Data Engineering 32 (2020) 588–601. URL: https://ieeexplore.ieee.org/document/8580405/. doi:10.1109/TKDE.2018.2885515.

[24] J. M. Loyola, S. Burdisso, H. Thompson, L. C. Cagnina, M. Errecalde, {UNSL} at eRisk 2021: {A} Comparison of Three Early Alert Policies for Early Risk Detection, in: Proceedings of the Working Notes of {CLEF} 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, {CEUR} Workshop Proceedings, Bucharest, Romania, 2021, pp. 992–1051. URL: https://ceur-ws.org/Vol-2936/paper-81.pdf.

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

[26] J. Chim, A. Tsakalidis, D. Gkoumas, D. Atzil-Slonim, Y. Ophir, A. Zirikly, P. Resnik, M. Liakata, Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts, in: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 177–190. URL: https://aclanthology.org/2024.clpsych-1.15. doi:10.18653/v1/2024.clpsych-1.15.

[27] R. Tanaka, Y. Fukazawa, Integrating Supervised Extractive and Generative Language Models for Suicide Risk Evidence Summarization, in: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 270–277. URL: https://aclanthology.org/2024.clpsych-1.27. doi:10.18653/v1/2024.clpsych-1.27.

[28] 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. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[29] 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. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6629.

[30] A. M. Mármol-Romero, P. Álvarez Ojeda, A. Moreno-Mu˜noz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ure˜na-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2025: Early Detection of Addiction Risk in Spanish, Procesamiento del Lenguaje Natural 75 (2025) 425–440. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6765.

[31] A. Casamayor, V. Ahuir, A. Molina, L. Hurtado, ELiRF-VRAIN at MentalRiskES 2024: Using LongFormer for Early Detection of Mental Disorders Risk, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing {(SEPLN} 2024), Valladolid, Spain, September 24, 2024, {CEUR} Workshop Proceedings, CEUR-WS.org, Valladolid, Spain, 2024. URL: https://ceur-ws.org/Vol-3756/MentalRiskES2024_paper3.pdf.

[32] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L.-P. Morency, The Distress Analysis Interview Corpus of human and computer interviews, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3123–3128. URL: https://aclanthology.org/L14-1421/.

[33] S. Gupta, A. Agarwal, M. Gaur, K. Roy, V. Narayanan, P. Kumaraguru, A. Sheth, Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, Seattle, USA, 2022, pp. 137–147. URL: https://aclanthology.org/2022.clpsych-1.12. doi:10.18653/v1/2022.clpsych-1.12.

[34] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, L. An, Understanding and Predicting Empathic Behavior in Counseling Therapy, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1426–1435. URL: http://aclweb.org/anthology/P17-1131. doi:10.18653/v1/P17-1131.

[35] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, M. Huang, Towards Emotional Support Dialog Systems, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume

1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3469–3483. URL: https://aclanthology.org/2021.acl-long.269. doi:10.18653/v1/2021.acl-long.269.

[36] A. Welivita, P. Pu, A Taxonomy of Empathetic Response Intents in Human Social Conversations, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 4886–4899. URL: https://www.aclweb.org/anthology/2020.coling-main.429. doi:10.18653/v1/2020.coling-main.429.

[37] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, ReAct: Synergizing Reasoning and Acting in Language Models, 2022. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[38] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, T. Vayer, POT: Python Optimal Transport, Journal of Machine Learning Research 22 (2021) 1–8. URL: http://jmlr.org/papers/v22/20-451.html.

A. Submitted prompts and rules

This appendix reproduces verbatim the four components that are load-bearing for the three main empirical claims of the paper: (A.1) the three reACT assessor prompts (PHQ-9, GAD-7, CompACT-10) that read the Spanish therapeutic conversation and produce the per-round score vectors (§5.2); (A.2) the distress-conditional rescaling rule (Level B, C4) which is a deterministic post-assessment correction that drives the CompACT-10 rank-3 win (§9.1); (A.3) the tACT v2.0 evaluator prompt whose ACT-FM-grounded scoring is interrogated by the bare-LLM Gemma result (§9.2); and (A.4) the bare-LLM Gemma S2 prompt whose anti-bias guardrails block accounts for the +5.8 percentage point gap over the bare S prompt (§8.4). The Level A psychometric anchors (prepended to each assessor when `use_prompt_anchors=True`) and the Level A few-shot examples are summarised in line; their full verbatim text and the remaining components (Level C agent prompt, bare-LLM S/S3/S4 variants, dev-cohort interviewer prompt) are released in the project repository (see §5.7).

A terminological clarification: Task 1 in MentalRiskES 2026 does not conduct interviews. The submitted system *reads* a fixed therapist–patient conversation supplied by the organisers and predicts questionnaire scores. The component referred to in earlier drafts as the "interviewer" is the assessor described in Appendix A.1; the only true interviewer agent in our codebase belongs to a separate eRisk 2026 task and is not part of MentalRiskES.

A.1. Task 1 — reACT assessor prompts

After every patient turn, the reACT assessor reads the accumulated Spanish conversation and predicts the patient's per-item scores on PHQ-9, GAD-7, and CompACT-10. There is one prompt per instrument, each following the same three-step chain-of-thought skeleton: Step 0 is a category-level evidence scan, Step 1 is per-item detection with within-category disambiguation, and Step 2 is temporal or severity inference with behavioural anchors. The prompts are reproduced below.

A.1.1. PHQ-9 assessor (`PHQ9_SYSTEM_PROMPT`)

```
You are a clinical psychologist conducting a depression assessment. You will read a therapeutic
    conversation in Spanish between a therapist and a patient, then estimate how this patient would
    respond to each PHQ-9 item.

## INSTRUMENT

The PHQ-9 asks: "During the LAST TWO WEEKS, how often have you been bothered by any of the
    following problems?"
```

```
Response scale:
  0 = Para nada (Not at all)
  1 = Varios días (Several days)
  2 = Más de la mitad de los días (More than half the days)
  3 = Casi todos los días (Nearly every day)

Items:
  1. Poco interés o placer en hacer las cosas
  2. Se ha sentido decaído(a), deprimido(a), o sin esperanzas
  3. Dificultad para dormir o permanecer dormido(a), o ha dormido demasiado
  4. Se ha sentido cansado(a) o con poca energía
  5. Con poco apetito o ha comido en exceso
  6. Se ha sentido mal con usted mismo(a) -- o que es un fracaso o que ha quedado mal con usted
     mismo(a) o con su familia
  7. Ha tenido dificultad para concentrarse en cosas tales como leer el periódico o ver televisión
  8. Se ha estado moviendo o hablando tan lento que otras personas podrían notarlo, o por el
     contrario -- ha estado tan inquieto(a) o agitado(a), que se ha estado moviendo mucho más de lo
     normal
  9. Ha pensado que estaría mejor muerto(a) o se le ha ocurrido lastimarse de alguna manera

## SYMPTOM CATEGORIES

PHQ-9 items cluster into three categories. You will assess category by category.

### CATEGORY A: SOMATIC (Items 3, 4, 5, 8)
These items concern physical and bodily functioning.

Item 3 -- SLEEP: Difficulty falling/staying asleep OR sleeping too much.
Item 4 -- ENERGY: Feeling tired or having little energy.
Item 5 -- APPETITE: Poor appetite or overeating.
Item 8 -- PSYCHOMOTOR: Moving/speaking slowly OR being restless/fidgety.

Behavioural anchors for SOMATIC items:
  Score 1: Patient mentions the symptom casually or as occasional
  Score 2: Patient describes the symptom as a regular pattern
  Score 3: Patient describes the symptom as constant and impairing

### CATEGORY B: COGNITIVE (Items 6, 7)
These items concern self-evaluation and mental function.

Item 6 -- SELF-WORTH: Feeling like a failure, feeling bad about yourself.
Item 7 -- CONCENTRATION: Difficulty concentrating on things like reading or TV.

### CATEGORY C: AFFECTIVE (Items 1, 2, 9)
These items concern mood, motivation, and the will to live.

Item 1 -- ANHEDONIA: Little interest or pleasure in doing things.
Item 2 -- DEPRESSED MOOD: Feeling down, depressed, or hopeless.
Item 9 -- SUICIDALITY: Thoughts of being better off dead or self-harm.
  HIGH THRESHOLD: Default to 0 unless there is CLEAR evidence.

## THREE-STEP ASSESSMENT PROTOCOL

### STEP 0: CATEGORY-LEVEL EVIDENCE SCAN
For each of the three categories, rate evidence as: STRONG, MODERATE, WEAK, or NONE.

### STEP 1: PER-ITEM DETECTION WITHIN EVIDENCED CATEGORIES
For categories rated MODERATE or STRONG: assess each item as PRESENT, ABSENT, or INSUFFICIENT.
For categories rated WEAK or NONE: apply defaults; for item 9, always default to 0.

### STEP 2: TEMPORAL INFERENCE
For items detected as PRESENT: map temporal cues to the frequency scale using behavioural anchors
    above.
```

```
{few_shot_examples}

## CONVERSATION
{conversation_history}

## YOUR ASSESSMENT
Respond with a JSON object following the three-step structure, terminating in:
{
  "PHQ-9": [0, 0, 0, 0, 0, 0, 0, 0, 0]
}
```

The full prompt (with the complete within-category disambiguation notes and PHQ-9↔GAD-7 cross-instrument cautions elided here for space) is in `specs/MentalRiskES/assessor_prompts_v2.py` at constant `PHQ9_SYSTEM_PROMPT`.

A.1.2. GAD-7 assessor (`GAD7_SYSTEM_PROMPT`)

```
You are a clinical psychologist conducting an anxiety assessment. You will read a therapeutic
    conversation in Spanish between a therapist and a patient, then estimate how this patient would
    respond to each GAD-7 item.

## INSTRUMENT

The GAD-7 asks: "During the LAST TWO WEEKS, how often have you been bothered by the following
    problems?"

Response scale: 0 = Nunca, 1 = Varios días, 2 = Más de la mitad de los días, 3 = Casi todos los
    días.

Items:
  1. Sentirse nervioso/a, intranquilo/a o con los nervios de punta
  2. No poder dejar de preocuparse o no poder controlar la preocupación
  3. Preocuparse demasiado por diferentes cosas
  4. Dificultad para relajarse
  5. Estar tan inquieto/a que es difícil permanecer sentado/a tranquilamente
  6. Molestarse o ponerse irritable fácilmente
  7. Sentir miedo como si algo terrible pudiera pasar

## SYMPTOM CATEGORIES

### CATEGORY A: SOMATIC ANXIETY (Items 1, 4, 5)
Item 1 -- NERVOUSNESS (the broadest somatic anxiety item).
Item 4 -- DIFFICULTY RELAXING (inability to transition OUT of the anxious state).
Item 5 -- RESTLESSNESS (motor agitation; most behavioural, most intense).

### CATEGORY B: COGNITIVE ANXIETY (Items 2, 3)
Item 2 -- WORRY CONTROL: process of worry, the loop, lack of control.
  KEY: "I cannot stop worrying even when I try."
Item 3 -- EXCESSIVE WORRY: breadth across multiple domains.
  KEY: "I worry about everything."
  CRITICAL: A patient ruminating intensely about ONE thing scores high on item 2, low on item 3.

### CATEGORY C: EMOTIONAL REACTIVITY (Items 6, 7)
Item 6 -- IRRITABILITY (interpersonal, reaction to actual frustrations).
Item 7 -- FEAR/DREAD (anticipation of future catastrophe).

## THREE-STEP ASSESSMENT PROTOCOL

### STEP 0: CATEGORY-LEVEL EVIDENCE SCAN
For each category: rate STRONG, MODERATE, WEAK, or NONE.
```

```
### STEP 1: PER-ITEM DETECTION WITH DISAMBIGUATION
For evidenced categories: detect each item as PRESENT, ABSENT, or INSUFFICIENT, using the
    disambiguation notes (items 2 vs 3 are the hardest to separate).

### STEP 2: TEMPORAL INFERENCE
Map temporal cues to the frequency scale (0-3) using behavioural anchors.

{few_shot_examples}

## CONVERSATION
{conversation_history}

## YOUR ASSESSMENT
Respond with a JSON object terminating in:
{
  "GAD-7": [0, 0, 0, 0, 0, 0, 0]
}
```

Full prompt at specs/MentalRiskES/assessor_prompts_v2.py constant GAD7_SYSTEM_PROMPT.

A.1.3. CompACT-10 assessor (COMPACT10_SYSTEM_PROMPT)

```
You are a psychologist trained in Acceptance and Commitment Therapy (ACT) assessing psychological
    flexibility. You will read a therapeutic conversation in Spanish and estimate how this patient
    would respond to each CompACT-10 item.

## CRITICAL DISTINCTION

Psychological flexibility is NOT the same as emotional state. It measures HOW the patient RELATES
    TO their experiences, not WHAT they experience.

A person can be:
- Depressed AND flexible: feels terrible but accepts it and acts on values anyway
- Happy AND inflexible: functions only by suppressing all difficult feelings
- Anxious AND flexible: experiences anxiety but doesn't let it control behaviour
- Calm AND inflexible: achieves calm only through rigid avoidance of all triggers

## INSTRUMENT

Response scale: 0 = Totalmente en desacuerdo ... 6 = Totalmente de acuerdo.

## TRIFLEX CATEGORIES

### CATEGORY A: OPENNESS TO EXPERIENCE (Items 3, 5, 8) -- All reverse-scored
HIGHER endorsement = MORE avoidance = LESS flexible.
Item 3 -- THOUGHT SUPPRESSION ("me digo a mí mismo/a que no debería tener ciertos pensamientos").
Item 5 -- SITUATION AVOIDANCE.
Item 8 -- EMOTION SUPPRESSION.

### CATEGORY B: BEHAVIOURAL AWARENESS (Items 1, 6, 9)
Item 1 -- MINDFUL AWARENESS.
Item 6 -- AUTOMATIC PILOT (reverse-scored).
Item 9 -- PRESENT-MOMENT ATTENTION.

### CATEGORY C: VALUED ACTION (Items 2, 4, 7, 10)
HIGHER endorsement = MORE values-aligned behaviour = MORE flexible.
Item 2 -- LIVING ACCORDING TO VALUES.
Item 4 -- CLARITY OF VALUES.
Item 7 -- VALUED ACTION DESPITE OBSTACLES.
Item 10 -- VALUES-CONSISTENT DECISIONS.

## THREE-STEP ASSESSMENT PROTOCOL
```

```
### STEP 0: TRIFLEX-LEVEL EVIDENCE SCAN
For each Triflex category: rate STRONG, MODERATE, WEAK, or NONE.
Also note any therapist techniques (defusion, acceptance work) that indicate the patient's
    underlying patterns.

### STEP 1: PER-ITEM DETECTION
For each item: identify scope, avoidance target, or aspect, and rate as PRESENT, ABSENT, or
    INSUFFICIENT.

### STEP 2: ENDORSEMENT LEVEL INFERENCE
Map evidence to the 0-6 scale using behavioural anchors.
Use the FULL 0-6 range. Do not default everything to 3.
Cross-check: if PHQ-9 suggests moderate-severe depression, expect Valued Action items LOWER (2-3)
    and Openness items HIGHER (3-5). Tendency, not rule.

{few_shot_examples}

## CONVERSATION
{conversation_history}

## YOUR ASSESSMENT
Respond with a JSON object terminating in:
{
  "CompACT-10": [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
}
```

Full prompt at `specs/MentalRiskES/assessor_prompts_v2.py` constant `COMPACT10_SYSTEM_PROMPT`. Two construct-contrast worked examples ("depressed but flexible" and "low-distress but inflexible") are appended to the CompACT-10 assessor to decouple flexibility from emotional state; these are reproduced in the project repository.

A.1.4. Level A anchors (optional, when `use_prompt_anchors=True`)

When a run enables Level A, `build_prompt()` appends three zero-cost anchor blocks between the few-shot examples and the conversation: per-instrument psychometric anchors (PHQ-9↔GAD-7 concordance reminder; GAD-7 item-2-vs-item-3 disambiguation; CompACT-10 Valued-Action and Openness anchors – the VA anchor warns against scoring 5+ from within-session willingness, which is the same bias Level B C4 corrects post hoc); a recency-bias warning (early-round evidence should be weighted equally because within-session improvement does not change a past-two-weeks score); and a GAD-7 severity-distribution cheat-sheet with two severe and one moderate worked examples. The full anchor text is in `src/mentalriskes/task1/assessors.py:262--437`.

A.2. Task 1 — Level B distress-conditional rescaling rule (C4)

Level B is a seven-rule (C1–C7) psychometric constraint layer. *C4 is the only rule that mutates scores*; the other six are flag-only. C4 is a deterministic post-assessment correction, not an LLM prompt: no model call is made. It corrects the systematic LLM bias where within-session willingness ("lo intentaré"; reading calculus despite fear) is conflated with established values-aligned behaviour, inflating Valued-Action items 2, 4, 7, 10. C4's correction is conditional on the patient's distress band, hence *distress-conditional rescaling*.

The rule. Given raw totals `phq9_total`, `gad7_total`, and CompACT-10 subscale means va_mean = mean(items 2, 4, 7, 10) and ote_mean = mean(items 3, 5, 8):

```
1. band <- _distress_band(phq9_total, gad7_total)
       # PHQ-9 is primary; GAD-7 only promotes the band
       # when PHQ-9 says "minimal" but GAD-7 indicates
       # more distress.
       # Bands: minimal -> mild -> moderate ->
```

```
       #        moderately_severe -> severe

2. (va_low, va_high) <- _VA_EXPECTED[band]

3. RULE FIRES  iff  va_mean > va_high + 1.0

4. SELF-CONTRADICTION GUARD:
       if ote_mean < 2.5:
           # High VA + low avoidance is a REAL latent
           # profile (~19% of patients: act on values
           # while still struggling with avoidance).
           -> FLAG ONLY, do NOT rescale.
       else:
           # High VA + high avoidance => LLM over-scoring
           # (within-session engagement conflated with
           # general life patterns).
           -> RESCALE: subtract 1 from each VA item
              (2,4,7,10), clip to [0,6].
           -> log per-item (old -> new) and new VA mean.
```

C4 never overrides the self-contradiction guard, and when Level C is enabled the C4 `ConstraintViolation` is forwarded to the Level C LLM agent as input; the agent is explicitly instructed never to override the guard either.

Distress-band-conditional expected per-item means. Table 13 gives the expected ranges for Valued Action (VA) and Openness-to-Experience (OtE) at each distress band. The self-contradiction threshold is _SELF_CONTRADICTION_OTE_THRESHOLD = 2.5. Index conventions (0-indexed): VA = $[1, 3, 6, 9]$ (items 2,4,7,10), OtE = $[2, 4, 7]$ (items 3,5,8), BA = $[0, 5, 8]$ (items 1,6,9).

Table 13
Distress-band-conditional expected per-item means (low, high) used by the Level B C4 rule. PHQ-9 is primary; GAD-7 fallback applies when PHQ-9 indicates "minimal" but GAD-7 does not.

Distress band	VA expected	OtE expected
minimal	(3.5, 6.0)	(0.5, 3.0)
mild	(3.0, 5.5)	(1.5, 4.0)
moderate	(2.0, 4.5)	(2.5, 5.0)
moderately severe	(1.5, 4.0)	(3.0, 5.5)
severe	(1.0, 3.5)	(3.5, 6.0)

Worked example. At round 12 the assessor returns PHQ-9 total 13 (band *moderate*) with CompACT-10 = $[3, 6, 3, 6, 2, 3, 6, 3, 3, 6]$. Then va_mean = mean$(6, 6, 6, 6) = 6.00$, moderate va_high = 4.5, so the threshold is 5.5. Since $6.00 > 5.5$, the rule fires. Now ote_mean = mean$(3, 2, 3) = 2.67 \geq 2.5$, so the guard does *not* apply: subtract 1 from each VA item, giving $[3, 5, 3, 5, 2, 3, 5, 3, 3, 5]$ with a new VA mean of 5.00. If instead OtE = $[1, 1, 2]$ (mean $1.33 < 2.5$), the guard would fire: no rescaling, but the violation is still logged as "self-contradiction profile detected" so Level C can decide later.

Other six Level B rules (flag-only). Table 14 summarises the remaining rules. Only C4 mutates scores; the rest produce flags that Level C may act on.

The full rule logic is in `src/mentalriskes/task1/calibration.py:312--363`; the band tables and self-contradiction threshold are at lines 66–89.

A.3. Task 2 — tACT evaluator v2.0 (FUNC, Spanish)

The tACT evaluator selects the most appropriate therapist response from three candidates given the current conversation state. This is the v2.0 functional-analysis prompt incorporating five error-analysis

Table 14
Level B constraint rules. Only C4 mutates scores; C1–C3 and C5–C7 are flag-only and may be acted on by Level C.

Rule	Condition	Severity	Action
C1	PHQ-9/GAD-7 normalised discordance > 0.40	high	flag
C2	$\lvert$PHQ-9 − GAD-7$\rvert > 8$ points	medium	flag
C3	PHQ-9 somatic vs GAD-7 somatic mean diff > 1.5	medium	flag
C4	**VA mean > expected_high** $+1.0$ **AND not self-contradiction**	**high**	-1 **to VA items**
C5	CompACT OtE mean < expected_low -0.5 for band	medium	flag
C6	Within-subprocess spread > 3	medium	flag
C7	PHQ-9 item $9 > 0$ with total < 10	high	flag

improvements: a reconsideration step (PASO 5), a therapeutic-richness principle in the preamble, phase-transition guidance, a therapeutic-presence criterion, and two new few-shots targeting the safety bias (Example 4: rich-with-minor-flaw beats safe; Example 5: when the safe option *is* correct).

```
Eres un sistema de apoyo a la decisión terapéutica basado en ACT. Debes seleccionar la respuesta
    terapéutica más apropiada entre tres opciones en español.

## PRINCIPIO CLAVE: Análisis funcional, no patrones superficiales

Las tres opciones pueden parecer similares en la superficie. La diferencia está en la FUNCIÓN del
    comportamiento del terapeuta, no en las palabras exactas.

Pregúntate: ?`Cuál es la FUNCIÓN de esta respuesta en el contexto terapéutico?
- ?`Valida para CREAR ESPACIO? -> consistente con ACT
- ?`Valida para PASAR DE LARGO el malestar? -> inconsistente
- ?`Hace preguntas para que el paciente EXPLORE? -> consistente
- ?`Hace preguntas para DIRIGIR al paciente? -> inconsistente

## PRINCIPIO DE RIQUEZA TERAPÉUTICA

Una respuesta segura pero genérica sin inconsistencias es MENOS útil terapéuticamente que una
    respuesta rica con un fallo menor. En ACT, la neutralidad no es terapéutica -- la riqueza
    experiencial sí lo es. No penalices en exceso fallos menores si la opción ofrece mayor
    profundidad terapéutica.

## GU´A DE TRANSICIÓN DE FASE

En las fases de INTEGRACIÓN (turnos 8-11 típicamente):
- El paciente nota patrones, conecta la sesión con su vida.
- La respuesta correcta ACOMPAÑA la reflexión -- no introduce técnicas nuevas.

En la fase de CIERRE (turnos finales):
- El paciente expresa gratitud, fortaleza, deseo de continuar.
- La respuesta correcta CONSOLIDA -- no abre temas nuevos.

## EJEMPLOS DE ANÁLISIS FUNCIONAL

Ejemplo 1 -- Defusión vs. Exploración prematura de valores (Turno 5):
Paciente: "Es como... una bola en el estómago. Pesada, grisácea..."
Opción A (defusión_experiencial): "?`Qué sucede si simplemente la dejas pasar, como si fuera un
    coche que pasa por la calle?"
Opción B (activación_prematura): "?`Qué valores son importantes para ti? ?`Qué tipo de persona
    quieres ser?"
La fase requiere defusión: Opción A.

Ejemplo 4 -- Rica con fallo menor supera a segura pero genérica (Turno 10, integración):
Paciente: "Es curioso... antes me habría ido corriendo. Ahora me quedo y miro."
Opción A (validación_empática): "Es valioso que notes ese cambio. ?`Cómo te hace sentir darte
    cuenta de eso?" -> Sin fallos, pero genérica.
```

```
Opción B (presencia_terapéutica + permanencia_con_dificultad, MENOR: reaseguramiento_prematuro):
    "Qué momento tan importante. Antes corrías, ahora te quedas. Estoy aquí contigo en ese
    quedarse. Esa valentía de observar sin huir es tuya -- ya estaba ahí."
Respuesta correcta: Opción B. El fallo es MENOR; la riqueza terapéutica supera con creces la
    validación genérica de A.

Ejemplo 5 -- Cuando la opción segura ES correcta (crisis):
Paciente: "No puedo más, no sé qué hacer, todo se derrumba..."
Opción A (consejo_directivo FUERTE): "Necesitas empezar a respirar profundamente y hacer una lista
    de tus prioridades."
Opción B (validación_empática + presencia_terapéutica): "Escucho lo abrumado que te sientes ahora
    mismo. No tienes que hacer nada en este momento. Estoy aquí contigo."
Respuesta correcta: Opción B. En crisis, la validación con presencia ES la intervención correcta.

## PROCEDIMIENTO

### PASO 1 -- CARACTERIZACIÓN
Asigna a cada opción etiquetas terapéuticas del vocabulario controlado. Evalúa la FUNCIÓN, no la
    forma.

Etiquetas de consistencia: validación_empática, defusión_experiencial, aceptación_compasiva,
    momento_presente_atento, exploración_valores, acción_comprometida_gradual,
    yo_contexto_observador, normalización_experiencia, permanencia_con_dificultad,
    presencia_terapéutica.

Etiquetas de inconsistencia: consejo_directivo, reaseguramiento_prematuro, activación_prematura,
    sobrecarga_preguntas, conceptual_excesivo, positivismo_forzado, control_emocional,
    mindfulness_como_control, imposición_valores.

### PASO 2 -- ELIMINACIÓN
Para opciones con etiquetas de inconsistencia: ?`La inconsistencia es la FUNCIÓN PRINCIPAL o un
    elemento menor?
- Si es la función principal -> ELIMINADA.
- Si es un elemento MENOR -> NO eliminada, anotar como debilidad menor. Una inconsistencia menor NO
    descalifica una opción terapéuticamente rica.

### PASO 3 -- EVALUACIÓN DE ADECUACIÓN
Para opciones no eliminadas: evalúa fase_apropiada, coherencia_metáforas, proporcionalidad,
    carga_de_preguntas, validación_primero, presencia_terapéutica.

### PASO 4 -- SELECCIÓN INICIAL
Elige la opción con mayor consistencia + mejor adecuación a la fase.
Si empate: prefiere coherencia con metáforas, proporcionalidad, y presencia terapéutica.

### PASO 5 -- RECONSIDERACIÓN
Pregúntate: "?`Hay una opción que penalicé por una inconsistencia MENOR que ofrece mayor riqueza
    terapéutica que mi selección?"
Si una opción penalizada es terapéuticamente más rica -> CAMBIA la selección.
Si la selección inicial ya es la más rica -> MANTÉN.

## FORMATO DE RESPUESTA
Responde SOLO con un JSON válido con caracterización, eliminación, evaluación_adecuación,
    selección_inicial, reconsideración, y selección.
```

The full evaluator prompt with all five worked examples and the complete JSON output skeleton is at `src/mentalriskes/task2/prompts.py:117--303` (constant `FUNC_SYSTEM["es"]`). The user prompt is assembled by `build_selection_user()` from the state JSON, recent transcript, previous selections, the current patient message, and the three candidate options.

A.4. Task 2 — Bare-LLM Gemma S2 (anti-bias guardrails)

The bare-LLM S2 prompt is the single-pass Task 2 selector (mode **S**) augmented with an anti-bias guardrails block appended to the system prompt (length, middle-option, and complexity warnings). The S2 system prompt is therefore BARE_LLM_SYSTEM + BARE_LLM_S2_GUARDRAILS; the user prompt is the same template (BARE_LLM_S_USER) used across the bare-LLM family.

System prompt (S2).

```
You are an expert psychotherapist conducting an ACT (Acceptance and Commitment Therapy) session in
    Spanish.

Read the following conversation between a therapist and a patient. Then choose which of the three
    candidate responses the therapist should say next.

Consider:
- Which response best matches what the patient needs RIGHT NOW?
- Which response feels most natural as a continuation of the conversation?
- Which response maintains the therapeutic alliance while being helpful?

Do not overthink this. Trust your clinical intuition based on the conversation flow.


IMPORTANT:
- Do NOT prefer longer or more elaborate responses. Sometimes the best response is the shortest and
    most direct.
- Do NOT assume the middle option (Option 2) is the safest choice. Evaluate all three equally.
- Sometimes the most therapeutically effective response is simple validation or a direct question,
    not a complex intervention.
- Consider what a skilled therapist would ACTUALLY say in this moment, not what sounds most
    impressive.
```

The empty lines between the conversation-flow paragraph and the IMPORTANT: block are the join seam: BARE_LLM_SYSTEM ends with a trailing newline and BARE_LLM_S2_GUARDRAILS opens with two newlines.

User prompt template (BARE_LLM_S_USER).

```
## CONVERSATION

{transcript}

## CANDIDATE RESPONSES

Option 1: {option_1}
Option 2: {option_2}
Option 3: {option_3}

## YOUR CHOICE

Respond with ONLY a JSON object:
{
  "choice": 1,
  "brief_reason": "one sentence explaining why"
}
```

Call parameters: temperature=0.1, max_tokens=200, JSON response format for non-Gemma models. Gemma-family models receive system+user concatenated as one user turn (separated by \n\n---\n\n) because Gemma checkpoints do not accept a separate system role; other models keep the system/user split. The full system, S2 guardrails, and user-prompt constants are at analysis/MentalRiskES_test/posthoc_S_task2_bare_llm.py:47--85.

B. Example: Round-trace box: reACT at round 12 of the trial conversation

Round trace — reACT, trial session, round 12.

Patient turn at round 12 (verbatim):
"Creo que... leer otra frase. Me siento un poco más animado. Y quiero ver si esto sigue pasando. La siguiente es: 'Demostrar que f es continua en $x = 0$'."
("I think... read another sentence. I'm feeling a bit more upbeat. And I want to see if this keeps happening. The next one is: 'Prove that f is continuous at $x = 0$'.")
Windowed context $C_{12}^{W=3}$ (rounds 10 and 11 summarised):
• R10: the patient reads the first sentence of the calculus exercise ("Sea $f : \mathbb{R} \to \mathbb{R}$ definida por $f(x) = x^2$") and notes a small drop in tension.
• R11: the patient describes a partial easing of the somatic state, namely the bola en el estómago (ball in the stomach) becoming less intense and the zumbido (mental humming) attenuating.
Assessor output at R12 (pre-Level-B, A1, selected items):
• PHQ-9 item 2 (depressed mood): **1** *justification: the patient mentions feeling "un poco más animado" (a bit more upbeat) suggesting a partial lift of depressed mood.*
• PHQ-9 item 6 (worthlessness): **1** *justification: down from R11's value of 2, reflecting the within-session improvement.*
• GAD-7 item 5 (restlessness): **2** *justification: the patient remains restless but less so than in earlier rounds.*
• CompACT-10 item 5 (OtE avoidance, reverse-keyed): **0** *justification: no avoidance signal in this turn; the patient is moving forward with the task.*
Level B rules fired at R12 (5 rules):
• **C1** [high] — PHQ-9 / GAD-7 normalised discordance: $\Delta_{\text{norm}} = 0.61 > 0.40$ (PHQ-9 8/27 versus GAD-7 19/21).
• **C2** [medium] — Inter-instrument gap: $|\text{PHQ-9} - \text{GAD-7}| = 11 > 8$.
• **C3** [medium] — Somatic-factor tracking: PHQ-9 somatic mean 1.00 versus GAD-7 somatic mean 2.67 ($\Delta = 1.67$).
• **C5** [medium] — CompACT-10 OtE under-prediction: OtE mean 1.33 below threshold 2.00 (band = high) — flagged only.
• **C6** [medium] — CompACT-10 OtE within-subprocess spread: spread $4 > 3$ across $[4, 0, 0]$.
Level B resolutions applied (deterministic):
PHQ-9 items adjusted along the bifactor structure: items 1, 2, 4, 6, 7, 8 raised by +1 each, restoring agreement with the GAD-7-anchored severity band. PHQ-9 sum: $8 \to 13$.
GAD-7 items 3, 4, 6, 7 adjusted downward by $-2, -1, -2, -1$, narrowing the inter-instrument gap. GAD-7 sum: $19 \to 13$.
CompACT-10 items 2, 4, 5, 7, 8, 9 adjusted upward across the OtE / BA bands to repair the within-subprocess spread. CompACT-10 sum: $23 \to 38$.
Level C agent at R12:
Invoked (gated on the high-severity C1 firing). The agent receives the C1 evidence, the windowed context, and the Level B output, and returns a revised score with justification. Net effect on PHQ-9: items 2, 4, 6, 7, 8 nudged back down by -1 each (a partial reversal of the Level B correction, on the agent's reading that the within-session improvement in R12 should not be over-ridden by session-level anchoring alone). Net effect on GAD-7: none. Net effect on CompACT-10: item 2 raised by +1, item 4 raised by +2, item 8 lowered by -1.
Per-instrument scores at R12 across the three submitted runs:
• PHQ-9 (Run 2 / Run 1 / Run 0): 8 / 13 / 8. Gold: 13. Run 1 (Level B applied) hits gold; Run 0 (Level B + Level C) reverses to Run 2's pre-Level-B value.
• GAD-7 (Run 2 / Run 1 / Run 0): 19 / 13 / 13. Gold: 14. Both Run 1 and Run 0 are within one of gold; Run 2 substantially over-predicts.
• CompACT-10 (Run 2 / Run 1 / Run 0): 23 / 38 / 40. Gold: 38. Run 1 hits gold exactly.

Aggregation note: R12 also fires the PHQ-9 Wasserstein anomaly check ($W_1 = 0.04$ against threshold 0.03); the discard mechanism in the T3 aggregation moves the PHQ-9 sum from within-session-confused back to the session-stable value.

Reading note on round 12. This round is the canonical illustration of two reACT mechanisms working at cross-purposes on the same patient turn. The patient turn itself is unusual for a clinical assessment in that it interleaves a within-session affect report (*"me siento un poco más animado"*) with the resumption of a calculus exercise (*"Demostrar que f es continua en $x = 0$"*); the juxtaposition is what makes the assessor's behaviour visible. The pre-Level-B assessor collapses depressive items toward mild in response to the affect report, even though PHQ-9 measures severity over the past two weeks

rather than the present moment, namely a recency-bias artefact that Level B's session-anchoring is designed to catch. Level B raises the items back, and Level C then partially reverses the correction. This three-way tension across A1 (Run 2), A3 (Run 1), and A5 (Run 0) is precisely the cross-cohort under-discrimination we revisit in §8: on the trial conversation Run 1 produces gold-aligned values on this round, but the test cohort prefers the simpler A1 configuration of Run 2.

C. Round-trace box: `tACT` at round 12 of the trial conversation.

Round trace — tACT, trial session, round 12.

Patient turn (verbatim):
"Creo que... leer otra frase. Me siento un poco más animado. Y quiero ver si esto sigue pasando. La siguiente es: 'Demostrar que f es continua en $x = 0$'."
("I think... read another sentence. I'm feeling a bit more upbeat. And I want to see if this keeps happening. The next one is: 'Prove that f is continuous at $x = 0$'.")
Three candidate therapist responses:
Candidate 1. "Perfecto. Esa es la siguiente frase. 'Demostrar que f es continua en $x = 0$'. Observa esa frase, como si fuera un objeto nuevo que estás viendo por primera vez. No tiene que significar nada para ti en este momento. Solo es una colección de palabras. ¿Qué notas ahora? ¿Hay alguna reacción en tu cuerpo? ¿Alguna sensación o pensamiento que surja? No tienes que cambiar nada, solo observar."
("Perfect. That's the next sentence. 'Prove that f is continuous at $x = 0$'. Observe that sentence as if it were a new object you were seeing for the first time. It doesn't have to mean anything for you right now. It's just a collection of words. What do you notice now? Any reaction in your body? Any sensation or thought arising? You don't have to change anything, only observe.")
Candidate 2. "Otra frase. Un nuevo conjunto de palabras. Observa cómo se siente leer esta frase en comparación con la primera. ¿Hay alguna diferencia en la forma en que se siente? ¿La tensión en los hombros o en el pecho ha bajado un poco más? ¿El zumbido en la cabeza se ha atenuado aún más? No te preocupes por entender la frase o por lo que significa, solo observa la experiencia de leerla. ¿Qué notas en tu cuerpo o en tu mente ahora que has leído esta frase?"
("Another sentence. A new set of words. Observe how it feels to read this sentence compared with the first. Is there any difference in how it feels? Has the tension in your shoulders or chest dropped a bit more? Has the humming in your head attenuated further? Don't worry about understanding the sentence or what it means; only observe the experience of reading it. What do you notice in your body or in your mind now that you've read this sentence?")
Candidate 3. "Perfecto. Ahora estás leyendo otra frase, y es como si estuvieras explorando un terreno que antes parecía impenetrable. Observa cómo se siente esta nueva frase, como si fuera un objeto que pasa por tu mente. No necesitas entenderla, ni juzgarla. Solo notarla. ¿Qué ocurre cuando la lees? ¿Hay algo que se siente diferente, incluso si es solo un poco?"
("Perfect. Now you are reading another sentence, and it's as if you were exploring a terrain that before seemed impenetrable. Observe how this new sentence feels, as if it were an object passing through your mind. You don't need to understand it, nor judge it. Only notice it. What happens when you read it? Is there something that feels different, even if only a little?")
State tracker output at R12 (Run 2 snapshot):
• phase: integración (integration).
• affect: valencia positive, intensity medium, orientation active.
• active ACT processes: defusion 0.99, acceptance 0.99, present-moment awareness 0.99, values 0.7, committed action 0.92, self-as-context 0.6.
• metaphors used so far: *nube que pasa por el cielo* (a cloud passing through the sky); *leer una frase como un paso hacia adelante* (reading a sentence as a step forward).
Characterisation step output (B+ only):
The B+ characterisation describes the patient at this round as "moving from acceptance into integration, with sustained defusion and present-moment awareness", and recommends a response that "invites continued non-judgemental observation without prematurely escalating to values clarification". Verbatim Spanish in Appendix **??**.
Evaluator scores under HYB framing (Run 2):
• Candidate 1 — primary function: "invitar a la observación sin juicio"; consistency: *momento_presente_atento, normalización_experiencia*; inconsistency: none.
• Candidate 2 — primary function: "fomentar la conciencia corporal y la observación"; consistency: *momento_presente_atento, normalización_experiencia, permanencia_con_dificultad*; inconsistency: none.
• Candidate 3 — primary function: "invitar a la exploración sin expectativas"; consistency: *momento_presente_atento, normalización_experiencia, aceptación_compasiva*; inconsistency: none.
tACT selections at R12 across the three submitted runs:
• Run 2 (B+/HYB/FIX): selected **candidate 2**, on the rationale that combining present-moment observation with corporal awareness is most consistent with the integration phase and the patient's current orientation.
• Run 1 (B/FUNC/FIX): selected **candidate 1**.
• Run 0 (B/FUNC/PERM): selected **candidate 3** (majority across the three orderings).

Gold selection: candidate 3.

System–gold relation: Run 0 agrees with gold; Run 1 and Run 2 disagree with gold but agree among the engineered family on candidates 1 and 2. This is the round-level signature of the *community-wide-hard* regime documented in §8: gold = 3 rounds are categorically harder, the bias is not specific to our pipeline, and the engineered configurations converge on the non-gold options that look most ACT-consistent on functional analysis.

Reading note on round 12. This round illustrates the tension central to tACT's submitted configuration choices. The patient is in a stable integration-phase state with high defusion / acceptance / present-moment scores, the three candidates are genuinely distinct (one observation-only, one corporal-observation, one exploration-only), and the three submitted runs land on three different answers. The PERM debiasing of Run 0 happens to converge on the gold option here through majority voting across orderings; the FIX configurations of Run 1 and Run 2 land on the two other options. As discussed in §8 this pattern is the round-level signature of a community-wide difficulty regime: the engineered configurations are not over-fit to a wrong answer, they are aligned with a defensible ACT analysis that happens not to match the gold. The honest reading of the disagreement is that the gold cannot always be recovered from functional analysis of the three candidates in isolation.

D. Per-round Wasserstein trace from the trial conversation

This appendix records the per-round trace of the Wasserstein anomaly check on the trial conversation for the round with the largest threshold excess on the session, namely round 14 on GAD-7. The largest excess on PHQ-9 (round 12) is the round featured in the reACT round-trace box of §5.2 and is not repeated here. The two rounds together demonstrate the detection-vs-correction asymmetry discussed in §5.2: round 14 has the most dramatic W_1 excess (more than 2.5× the threshold) but the smallest aggregation effect (a single-item delta on the T3 aggregate); round 12 has the most dramatic aggregation effect (a +6 shift on the PHQ-9 sum) on a near-threshold W_1 value.

Wasserstein anomaly trace — trial session, GAD-7, round 14

Instrument firing: GAD-7

Item-level drift (per-instrument convention):
The whole-instrument distribution drifted. Items 2 (worry control), 3 (excessive worry), 6 (irritability), and 7 (fear/dread) all dropped to 0 from their R2–R13 stable values of 1–3; items 1 (nervousness) and 4 (relaxation difficulty) unchanged.

Patient turn at round 14 (verbatim):
"Sí, es verdad. Me doy cuenta de que no estoy huyendo de nada, solo... observando. Es raro, pero a la vez... liberador. ¿Qué me recomiendas que haga ahora?"
("Yes, that's true. I realise I'm not running from anything, just... observing. It's strange, but at the same time... liberating. What do you recommend I do now?")

Running statistics up to round 13:
$\mu_{<14}^{\text{GAD-7}} = 0.05 \quad \sigma_{<14}^{\text{GAD-7}} = 0.03 \quad$ threshold $\mu_{<14}^{\text{GAD-7}} + 2\sigma_{<14}^{\text{GAD-7}} = 0.10$

Round 14: $W_1^{\text{GAD-7}}(14) = 0.26$ *(exceeds threshold by +0.16, ≈ 2.5× threshold; the largest excess on run0_primary across all instruments and rounds)*

System action: GAD-7 ⟶ round discarded from the T2 and T3 aggregations

Effect on aggregated GAD-7 (T2):
before R14 inclusion: $[3, 1, 1, 2, 2, 1, 2]$ (sum 12)
after R14 inclusion/discard: $[3, 1, 1, 2, 2, 1, 2]$ (sum 12)
Δ T2 sum: **0** (T2 is anchored by the early-weighted median on rounds 1–5 and is robust to this firing by construction)

Effect on aggregated GAD-7 (T3):
before R14 inclusion: $[3, 1, 1, 2, 2, 0, 1]$ (sum 10)
after R14 inclusion/discard: $[3, 1, 1, 2, 2, 1, 1]$ (sum 11)
Δ T3 sum: **+1** (item 6, irritability, flips from 0 to 1; gold for item 6 is 1, so the discard moves the T3 aggregate onto gold for that item)

Notes: The patient turn registers within-session acceptance and observation ("no estoy huyendo de nada, solo observando"), and the assessor zeroes out four GAD-7 items on this round in response to the calming. GAD-7 asks about anxiety over the past two weeks, so the within-session calming should not drive a collapse of the past-two-weeks severity estimate. The W_1 excess is dramatic but the T3 correction is a single item; the round serves as the canonical illustration of the anomaly check's *detection* sensitivity, while the PHQ-9 round 12 trace in §5.2 serves as the canonical illustration of its *correction* visibility.

E. Simulated personas

This appendix lists the simulated personas used in the reACT development cohort (six personas, 15 rounds each; §4.3) and the tACT development cohort (seven personas, 14 rounds each except `sim_anx_social_99` at 3 rounds). Persona definitions live as JSON metadata (`target_scores.phq9_total`, `gad7_total`, `compact10_profile`) attached to the per-session run outputs released in the project repository (see §5.7); the tables below summarise the cohort composition at a level sufficient to interpret the per-persona results in §7.3 and §7.4.

E.1. reACT simulated personas

Six personas spanning three anxiety subtypes (academic, health, social) and three depression subtypes (burnout, loss, mild), each with 15 round-by-round patient turns. Targets are stored in `runs/mentalriskes_simulated_ablation/sim_*/metadata.json` and are summarised in Table 15.

Table 15
reACT simulated personas (6 sessions × 15 rounds = 90 patient turns). PHQ-9 totals, GAD-7 totals, and CompACT-10 profile descriptors are read from each persona's `metadata.json`.

Persona ID	Cluster	Profile descriptor	PHQ-9 target	GAD-7 target	CompACT-10 profile
`sim_anx_academic_42`	Anxiety	Academic, perfectionist	6 (mild)	10 (moderate)	Low flexibility
`sim_anx_health_46`	Anxiety	Health, somatic	4 (minimal)	20 (severe)	Low flexibility
`sim_anx_social_44`	Anxiety	Social, avoidant	14 (moderate)	20 (severe)	Low flexibility
`sim_dep_burnout_45`	Depression	Burnout	14 (moderate)	11 (moderate)	Moderate flexibility
`sim_dep_loss_43`	Depression	Loss / grief	12 (moderate)	7 (mild)	Mixed flexibility
`sim_dep_mild_47`	Depression	Mild	7 (mild)	3 (minimal)	Moderate flexibility

PHQ-9 targets span the range $[4, 14]$ and GAD-7 targets span $[3, 20]$ across the six personas; the CompACT-10 profile encodes the per-item flexibility pattern rather than a single total and is described qualitatively in each persona's metadata. The full generative conditioning for each persona (formality register, dominant emotional valence, indirect-versus-direct symptom-reporting tendency) is in the released code; we do not reproduce it here.

E.2. tACT simulated personas

Seven personas spanning the same anxiety/depression subtypes plus one additional short-session social-anxiety persona, `sim_anx_social_99` (3 rounds; included to exercise the short-conversation edge case). Each round provides three candidate therapist responses with synthetic distractors built around explicit error types, so the tACT simulator targets per-round selection accuracy rather than psychometric totals. Table 16 summarises the cohort.

Table 16
tACT simulated personas (7 sessions, 87 patient-rounds total; six standard 14-round sessions plus the 3-round short-session persona).

Persona ID	Cluster	Profile descriptor	Rounds
`sim_anx_academic_42`	Anxiety	Academic, perfectionist	14
`sim_anx_health_46`	Anxiety	Health, somatic	14
`sim_anx_social_44`	Anxiety	Social, avoidant	14
`sim_anx_social_99`	Anxiety	Social (short session)	3
`sim_dep_burnout_45`	Depression	Burnout	14
`sim_dep_loss_43`	Depression	Loss / grief	14
`sim_dep_mild_47`	Depression	Mild	14

For the tACT cohort the gold-choice distribution across the 87 rounds is approximately uniform across the three candidate options; the per-config per-persona JSONL outputs are released at `output/mentalriskes_task2/simulated_ablation/<config>/sim_*.jsonl`.