DETECCIÓN Y BIOMARCADORES

NLP and sentiment analysis for mental health screening: mapping conversation to PHQ-9 with LLMs

SentirIA Research Papers · 2026

Infraestructura de detección temprana en salud mental

Este documento es parte de la base científica de SentirIA,
plataforma de detección temprana y monitoreo continuo de deterioro en salud mental.
No constituye diagnóstico clínico. La evaluación es responsabilidad del profesional.

NLP and sentiment analysis for mental health screening: mapping conversation to PHQ-9 with LLMs

Natural language processing can now map free-form conversation to individual PHQ-9 depression screening items with clinically meaningful accuracy, opening the door to invisible, longitudinal screening via conversational AI. GPT-4o achieves 75.9% binary accuracy across PHQ-8 items in zero-shot clinical transcript analysis (Teferra et al., 2025), while domain-specific models like MentalBERT reach F1 scores above 0.89 on Reddit depression detection (Ji et al., 2022). Combined with vocal biomarkers — where reduced pitch variability and increased pause duration yield AUCs exceeding 0.90 (Menne et al., 2024) — multimodal systems now approach the sensitivity-specificity profile of the PHQ-9 itself. This report provides the technical and scientific foundation for SentirIA, a WhatsApp-based AI companion for early depression detection, covering PHQ-9 item-level mapping, linguistic markers, LIWC methodology, transformer architectures, conversational screening design, and vocal biomarkers.

---

1. Mapping natural language to each PHQ-9 item with large language models

The Patient Health Questionnaire-9 (Kroenke, Spitzer & Williams, 2001) assesses nine DSM-5 depression criteria on a 0–3 Likert scale over two weeks, yielding scores from 0 to 27 with clinical cutoffs at 5 (mild), 10 (moderate), 15 (moderately severe), and 20 (severe). The core challenge for SentirIA is extracting item-level signals from unstructured conversation without asking clinical questions directly.

Linguistic signatures for each PHQ-9 domain

Item 1 — Anhedonia ("Little interest or pleasure in doing things" / "Poco interés o placer en hacer cosas"): Detection requires identifying the *absence* of positive affect rather than the presence of negative emotion. Key signals include negated positives ("Family gatherings should be fun, yet..."), withdrawal language ("I used to love painting but now I just don't see the point"), and reduced activity descriptions. In Spanish: "Ya no me interesa nada" or "Antes me gustaba salir, pero ahora no tengo ganas." Teferra et al. (2025) found Llama 3-8B showed superior detection of anhedonia compared to other LLMs, achieving 52% Likert-scale accuracy — the best among seven models tested. Teng et al. (2025) identified negated positives as the critical signal distinguishing anhedonia from depressed mood.

Item 2 — Depressed mood ("Feeling down, depressed, or hopeless" / "Sintiéndose decaído(a), deprimido(a), o sin esperanzas"): Sadness words combined with first-person pronouns and hopelessness language provide the strongest signals. Example: "I feel like a dark cloud is always

over me." Spanish: "Me siento triste todo el tiempo" or "No tengo esperanza de que las cosas mejoren." GPT-4o achieved 49–55% Likert accuracy on this item, which correlates strongly with negative emotional polarity in LLM-based text analysis.

Item 3 — Sleep disturbance ("Trouble falling or staying asleep, or sleeping too much" / "Dificultad en caer o permanecer dormido(a), o dormir demasiado"): References to insomnia, oversleeping, disrupted routines, and fatigue-related sleep talk are key markers. "I can't fall asleep until 4 AM, and then I sleep through my alarm." Spanish: "No puedo dormir bien, me despierto a media noche." FLAN-T5 achieved F1 = 0.92 for sleep problem extraction from pediatric EHRs (arXiv 2501.17510), making this the most reliably detected item.

Item 4 — Fatigue ("Feeling tired or having little energy" / "Sintiéndose cansado o teniendo poca energía"): Exhaustion language, inability-to-complete-tasks framing, and lethargy descriptions. "Even getting out of bed feels impossible." Spanish: "Estoy agotado(a), no tengo fuerzas para nada." Nguyen et al. (2022, ACL) mapped posts like "I'm overweight and I never succeed in..." to fatigue/energy items.

Item 5 — Appetite changes ("Poor appetite or overeating" / "Pobre de apetito o comer en exceso"): Food-related language, weight change mentions, and binge eating references. This is among the hardest items to detect passively, with Cohere and Gemini models achieving only 40% accuracy (Teferra et al., 2025). Spanish: "No tengo apetito, nada me provoca comer."

Item 6 — Guilt/worthlessness ("Feeling bad about yourself — or that you are a failure"): Self-blame, burden language, and self-deprecation. "I'm such a burden to everyone around me." Spanish: "Soy un fracaso, no sirvo para nada." FLAN-T5 achieved F1 = 0.80 for self-loathing extraction.

Item 7 — Concentration difficulty: Cognitive fog descriptions, distractibility, memory complaints. "I read the same paragraph five times." Spanish: "No puedo concentrarme en nada." GPT models performed consistently best on concentration items across all metrics.

Item 8 — Psychomotor changes: References to slowed movement, restlessness, or agitation observed by others. "My wife says I'm moving like a zombie." Notably, Cohere LLM showed dramatically superior detection for this item (binary F1 = 0.93, accuracy 94%), despite it being the most imbalanced item in the dataset.

Item 9 — Suicidal ideation ("Thoughts that you would be better off dead"): Death wishes, self-harm language, passive suicidal ideation ("I just want it to end"), and burden-to-others language. Spanish: "A veces pienso que sería mejor estar muerto(a)." This item requires continuous monitoring independent of any scheduled assessment cycle.

Prompt engineering for item-level prediction

The most rigorous evaluation of LLM-based PHQ item prediction comes from Teferra et al. (2025, PLOS Digital Health), who tested seven LLMs on 100 DAIC-WOZ transcripts using the RISEN framework (Role, Instruction, Steps, End goal, Narrowing). GPT-4o achieved the best mean

performance: 75.9% binary accuracy, F1 = 0.74, MCC = 0.25 across all eight PHQ-8 items. The framework assigns the LLM the role of a mental health screening assistant, instructs it to predict item scores from transcript data, specifies steps for highlighting symptoms aligned with each PHQ-8 question, and narrows the focus to clinically relevant responses.

Chain-of-thought (CoT) prompting substantially improves performance. Teng et al. (2025, arXiv 2502.05879) developed a four-stage CoT framework: emotion analysis (type, intensity, polarity, source), binary classification, causal reasoning, and severity assessment. This achieved CCC = 0.732 and MAE = 3.37 on E-DAIC, outperforming traditional multimodal methods like CubeMLP (CCC = 0.583). Shi et al. (2024) confirmed that zero-shot CoT ("Let's think step by step") yields PHQ-8 scores consistently closer to ground truth than direct scoring.

Few-shot prompting also shows promise. Lorenzoni et al. (2024/2025) found that including four representative examples (two positive, two negative) with an elaborate clinical prompt at temperature 0.0–0.2 produced optimal results. Lamichhane (2023) achieved F1 = 0.86 for depression detection with GPT-3.5 Turbo using few-shot strategies.

A critical finding from Teferra et al. is that no single LLM excels across all items. The authors recommend an ensemble approach: GPT-4o for cognitive/emotional items, Llama for anhedonia, Cohere for psychomotor changes. This has direct implications for SentirIA's architecture.

Multi-task architectures grounded in PHQ-9

Nguyen et al. (2022, ACL) established the foundational architecture for PHQ-9-grounded depression detection. Their PHQ9 Model uses a BERT encoder feeding nine parallel symptom classification heads, each corresponding to a PHQ-9 item, with an aggregation layer producing a final depression prediction. In cross-dataset transfer experiments across RSDD, eRisk, and TRT datasets, PHQ-9-grounded models substantially outperformed vanilla BERT in out-of-distribution settings, demonstrating that clinical structure improves generalizability.

PhqMML (2025, arXiv 2501.16106) extends this with multimodal multi-task learning. It identifies PHQ-8 symptoms for each utterance, generates dialogue-level symptom summaries, and predicts depression severity jointly. Wang et al. (2023) cast symptom identification as a span-prediction problem using PHQ-aware self-guided cross-attention, achieving macro-F1 = 68.84% and micro-F1 = 75.92% on the PRIMATE dataset.

The D2S dataset (Yadav et al., 2020, COLING) provides 12,000 tweets annotated with PHQ-9 symptom labels, enabling per-symptom classification at scale. Milintsevich et al. (2023) pioneered symptom-level severity estimation using hierarchical neural architectures on DAIC-WOZ, providing more fine-grained clinical information than global scores.

Bilingual and cross-lingual considerations

The Spanish PHQ-9 has been validated across multiple Latin American populations. Moreno et al. (2023, JAMA Network Open) conducted a systematic review of 10 cross-sectional studies (5,164 Spanish-speaking adults) finding pooled sensitivity and specificity of 0.89 each for the PHQ-2,

though optimal cutoff scores may be lower for Spanish-speaking populations than the standard English cutoffs.

Cross-lingual NLP for depression detection remains underdeveloped. Coello-Guilarte et al. (2019) proposed English→Spanish transfer using bilingual word embeddings and LIWC-based psycholinguistic features. MentalRiskES (Mármol-Romero et al., 2023) established the first CLEF shared task for early detection of depression and suicide in Spanish. Multilingual transformers like XLM-RoBERTa and mBERT enable cross-lingual transfer without translation, and a 2025 survey (arXiv 2505.15556) catalogued 108 data collections across languages for mental health NLP.

Bilingual speakers may report symptom severity differently depending on assessment language (PLOS ONE, 2024). Spanish speakers tend toward more somatic descriptions of depression ("me duele el cuerpo," "siento un peso en el pecho") and use cultural idioms of distress like "nervios," "susto," and "ataque de nervios" that map to multiple PHQ-9 items simultaneously.

---

2. Linguistic markers that predict depression: what the evidence shows

A large body of research has identified specific linguistic patterns associated with depression, though effect sizes vary and context-dependency is a critical caveat. A 2025 Nature study (randomized cross-over trial, N = 218) found no linguistic features were consistently associated with depression across different text tasks, underscoring that markers must be interpreted in context rather than as absolute indicators.

Negative emotion, absolutist language, and the I-word hypothesis

Negative emotion words represent the most intuitive marker. Tølbøll's (2019) meta-analysis found a group difference of $d = 0.72$ (depressed vs. controls) and a severity correlation of $r = 0.12$ — a medium group-level effect but small individual-level predictor. The LIWC negemo category captures words like "hurt," "ugly," "hate," while the sad subcategory ("crying," "grief") is more specific to depression. Positive emotion words show a complementary decrease ($d = -0.38$). Trifu et al. (2024) confirmed these patterns in clinical MDD samples ($p < 0.001$).

Absolutist language may be the single most powerful linguistic marker. Al-Mosaiwi and Johnstone (2018, Clinical Psychological Science) analyzed 63 internet forums with over 6,400 members and found that anxiety, depression, and suicidal ideation forums contained more absolutist words ("always," "never," "nothing," "completely," "every") than control forums, with effect sizes exceeding $d = 3.14$ — among the largest in the linguistic markers literature. Suicidal ideation forums showed even higher rates ($d > 1.71$ vs. depression forums). Critically, absolutist words tracked severity of affective disorder forums more faithfully than negative emotion words, making this marker particularly valuable for SentirIA. Recovery forums showed persistent elevation of absolutist language even after symptom remission, suggesting absolutist thinking is

a vulnerability factor that persists beyond acute episodes. Bathina et al. (2021, Nature Human Behaviour) replicated and extended these findings, showing cognitive distortions including all-or-nothing thinking were more than twice as prevalent in depressed Twitter users.

First-person pronoun overuse ("I-talk") reflects heightened self-focused attention per Pyszczynski and Greenberg's (1987) theory. Rude, Gortner, and Pennebaker (2004, Cognition & Emotion) found depressed students used significantly more first-person singular pronouns. Three meta-analyses quantify the effect: Edwards and Holtzman (2017) found $r = 0.13$ ($k = 21$, $N = 3{,}758$); Tackman et al. (2019, JPSP) found $r = 0.10$ ($N = 4{,}754$ across 6 labs); Tølbøll (2019) found $d = 0.44$. However, Tackman et al.'s critical finding was that when controlling for negative emotionality, the depression-I-talk effect was substantially reduced — I-talk is better construed as a marker of general distress rather than depression specifically. The effect was present for subjective ("I") and objective ("me/myself") pronouns but NOT possessive ("my/mine"). Burkhardt et al. (PMC10785936) found that contextualized MentalBERT embeddings of I-words outperformed simple frequency counting for predicting depression severity.

Vocabulary, syntax, timing, and temporal orientation

Reduced vocabulary diversity (measured via Type-Token Ratio, MATTR, or MTLD) is theoretically linked to depression but remains empirically underexplored. Schneider et al. (2023, Schizophrenia/Nature) found dramatic lexical reduction in schizophrenia but only marginal reduction in MDD. Trifu et al. (2024) described MDD patients communicating in "short sentences" with "limited exchange of ideas," but quantification is sparse. For implementation, MATTR with a 25–50 word window or MTLD is recommended over raw TTR, which confounds with text length.

Shorter sentences and reduced syntactic complexity are subtle in depression compared to schizophrenia. Schneider et al. (2023) found marginally reduced syntactic complexity in MDD versus controls, but the difference was not statistically significant after multiple comparison correction. The pattern manifests as shorter utterances, more sentence fragments, fewer subordinate clauses, and reduced spontaneous speech output rather than dramatic simplification.

Slower response patterns reflect psychomotor retardation in the digital domain. The BiAffect study (Zulueta et al., 2018, JMIR) found accelerometer displacement and interkey delay correlated with depression scores in bipolar subjects. Nature Digital Medicine (2024) reported that less phone movement while typing was associated with more anhedonia ($\beta = -0.12$, $p = 0.00030$). MDPI (2025) achieved 82.5% accuracy classifying youth depression from handwriting kinematics. For SentirIA, response latency, message length trends, and time-of-day messaging patterns serve as behavioral biomarkers beyond text content.

Temporal orientation shift toward past-focus is one of the more robust markers. Trifu et al. (2024) found MDD patients scored significantly higher on past focus ($M = 10.79$) compared to controls ($M = 6.47$, $p < 0.001$) — a 67% increase. Park et al. (2017, Journal of Personality) analyzed 1.3 million Facebook messages from 5,372 participants, finding more future-oriented people were less depressed and more satisfied with life.

Social referencing decline manifests as reduced use of first-person plural pronouns ("we," "us"), reflecting social withdrawal. Burkhardt et al. (2021) found first-person plural pronoun use associated with *decreased* depression, while Trifu et al. (2024) showed significantly lower LIWC affiliation scores in MDD patients. The I-to-we ratio serves as a quantifiable index of social disengagement.

Cognitive complexity and hedging reveal a counterintuitive finding. Trifu et al. (2024) found MDD patients used less tentative language (M = 4.72) compared to controls (M = 5.30, p = 0.045). Rather than showing more hedging, depressed individuals display cognitive rigidity and categorical, black-and-white thinking — consistent with elevated absolutist language. The decrease in tentative language and increase in absolutist language together form a cognitive rigidity index implementable by computing the absolutist-to-tentative word ratio.

Marker	Effect size	Source
Absolutist words (dep. vs. control)	d > 3.14	Al-Mosaiwi & Johnstone 2018
Negative emotion words (group diff.)	d = 0.72	Tølbøll 2019 meta-analysis
First-person pronouns (group diff.)	d = 0.44	Tølbøll 2019
Positive emotion words (group diff.)	d = −0.38	Tølbøll 2019
Positive emotion words (severity)	r = −0.21	Tølbøll 2019
First-person pronouns (severity)	r = 0.10–0.19	Tackman 2019; Edwards & Holtzman 2017
Past focus (MDD vs. controls)	~67% higher	Trifu et al. 2024
Cognitive distortions (dep. vs. control)	~2× prevalence	Bathina et al. 2021

---

3. LIWC-22: architecture, depression-relevant dimensions, and limitations

Linguistic Inquiry and Word Count, created by James W. Pennebaker at the University of Texas at Austin, has evolved through five major versions: the original (1993, ~2,290 words, ~64 categories), LIWC2001 (~2,300 words, ~74 categories), LIWC2007 (~4,500 words, ~80 categories), LIWC2015 (~6,400 words, ~93 categories), and the current LIWC-22 (Boyd, Ashokkumar, Seraj & Pennebaker, 2022), which contains over 12,000 words, word stems, phrases, and emoticons across approximately 117 categories. Over 20,000 scientific articles have used LIWC.

How the dictionary-based architecture works

LIWC processes text word-by-word, matching each token against a hierarchically organized dictionary using word stem matching (e.g., "certain*" matches "certainty," "certainly"). Multi-

word phrase entries were introduced in LIWC-22. Each matched word increments counters for all associated categories — the word "cried," for instance, maps to ten categories including affect, emo_neg, emo_sad, verbs, focuspast, and communication. Output for most categories is computed as (matching words / total words) × 100. Four proprietary summary variables — Analytical Thinking, Clout, Authenticity, and Emotional Tone — use algorithms combining multiple categories, output as percentiles (1–99).

LIWC-22 introduced a critical distinction between sentiment (tone_pos, tone_neg — broader valence) and emotion (emo_pos, emo_neg — strictly emotional words), addressing longstanding conflation concerns. New depression-relevant categories include fatigue, mental, wellness, illness, reward, and allnone (all-or-nothing language).

Depression-relevant dimensions in detail

The dimensions most relevant to SentirIA span affective, cognitive, social, and temporal categories. Negative emotion (emo_neg) and its subcategory sadness (emo_sad) capture the core affective signature, with Eichstaedt et al. (2018, PNAS) finding sadness words predicted future depression from Facebook posts ($\beta = 0.17$, $p < 0.001$). I-words (first-person singular pronouns) were equally predictive ($\beta = 0.19$, $p < 0.001$) in the same study, which analyzed 524,292 posts from 683 patients and detected depression as early as three months before clinical diagnosis.

The cognitive categories — insight ("think," "know," "consider"), cause ("because," "effect"), tentative ("maybe," "perhaps," "guess"), and certitude — together capture ruminative thinking and cognitive rigidity patterns. Discrepancy words ("should," "would," "could") may indicate self-critical thinking (Grant, 2010); Eichstaedt et al. found discrepancy words significantly predicted depression.

Temporal categories focuspast and focusfuture capture the past-oriented rumination and reduced future planning characteristic of depression. The social category and its subcategories (prosocial, family, friend) reflect social withdrawal when decreased. The newer reward category captures diminished reward sensitivity linked to anhedonia.

Strengths, limitations, and the Spanish dictionary

LIWC's strengths are considerable: 100% algorithmic reproducibility, transparent interpretability, a massive validation literature across thousands of studies, and rapid scalability to large text corpora. However, its bag-of-words architecture creates fundamental limitations. It cannot handle negation ("I am not sad" scores as containing sadness), sarcasm ("What a wonderful day" used sarcastically scores as positive), or polysemy ("blue" as color vs. emotion). The fixed dictionary cannot adapt to evolving language, slang, or domain-specific terminology. Reliable trait inference requires approximately 4,000–5,000 words of text, and LIWC-based classifiers are consistently outperformed by contextual models.

Hur et al. (2024, PNAS) delivered the most striking finding: LLM-based sentiment analysis (ChatGPT) and human raters predicted depression symptom changes at three-week follow-up,

while LIWC-22 sentiment did not. Zero-shot LLM classifiers achieved 65% overall accuracy versus LIWC+ML models yielding the lowest accuracy for social media depression classification (2025).

The Spanish LIWC dictionary was developed by Ramírez-Esparza, Pennebaker, García, and Suriá (2007, Revista Mexicana de Psicología). It was based on the LIWC2001 version (~2,300 words), overseen by a native Mexican Spanish speaker with Colombian and Peninsular Spanish input. Validation showed high correlations with English categories and similar discrimination between depression and cancer chat rooms across languages. However, it lacks newer LIWC-22 categories critical for mental health research (fatigue, mental, wellness, reward, allnone). The LIWC-22 manual now recommends translating texts to English and running the English LIWC rather than using translated dictionaries, noting that Google Translate paired with LIWC produces results with very small effect size differences from human translation.

---

4. Transformer-based approaches and the state of the art (2020–2025)

eRisk shared tasks and the evolution of early detection

The eRisk lab at CLEF, organized by David E. Losada (Univ. Santiago de Compostela), Javier Parapar (Univ. A Coruña), and Fabio Crestani (USI, Switzerland), has run since 2017 as the primary benchmark for early risk detection from internet text. Tasks evolved from binary depression detection (2017) to symptom-level BDI-II sentence ranking (2023–2024) and, in 2025, introduced a groundbreaking conversational depression detection task using LLM personas guided by BDI-II. This evolution mirrors the field's shift from classification to interpretable, interactive screening.

BERT-based classifiers became dominant from 2020 onward, achieving precision up to 91.3% for self-harm detection (Martínez-Castaño et al., 2020). The central metric, ERDE (Early Risk Detection Error), penalizes late detection — quantifying the tradeoff between accuracy and timeliness that remains the field's core tension. By 2024, 84 teams registered with 17 submitting runs, indicating growing but still modest community engagement.

MentalBERT and domain-specific pre-training

Ji et al. (2022, LREC) released MentalBERT and MentalRoBERTa, domain-adaptive continuations of BERT-Base and RoBERTa-Base pre-trained on 13.67 million sentences from Reddit mental health subreddits (r/depression, r/SuicideWatch, r/Anxiety, r/bipolar, r/mentalillness, r/mentalhealth, r/offmychest). Training ran for 624,000 iterations on 4× NVIDIA Tesla V100 GPUs (~8 days). The models are publicly available on HuggingFace as `mental/mental-bert-base-uncased`.

Domain-specific pre-training consistently improved performance over general BERT, BioBERT, and ClinicalBERT across all mental health tasks. On the Depression Reddit dataset,

MentalRoBERTa achieved F1 = 89.01%; on CLPsych15, F1 = 72.16%. A 2025 benchmark study found MentalBERT achieved F1 = 97.30% on multi-level depression severity classification, outperforming all other transformers tested. When further fine-tuned with triplet-loss on clinical notes (ScienceDirect, 2023), the model achieved F1 = 0.99 for anhedonia and 0.94 for suicidal ideation with plan.

The MentaLLaMA family (Yang et al., WWW 2024) represents the next evolution: instruction-following LLMs built on LLaMA2, trained on the IMHI dataset (105K instruction samples). MentaLLaMA-chat-13B surpasses or approaches state-of-the-art discriminative methods on 7 of 10 test sets while generating human-interpretable explanations — a critical capability for clinical applications.

Zero-shot and few-shot performance of frontier LLMs

The most rigorous head-to-head evaluation is from Ohse et al. (2024, Computer Speech & Language), who tested BERT, Llama2-13B, GPT-3.5, and GPT-4 on 82 clinical interview transcripts. GPT-4 zero-shot achieved F1 = 0.73 — outperforming fine-tuned BERT without any training data. Its PHQ-8 predictions correlated r = 0.71 with true severity scores. GPT-3.5 achieved only F1 = 0.34 in zero-shot but reached F1 = 0.82 after fine-tuning. A follow-up study (Ohse et al., 2025) found GPT-4 and GPT-4o reached "at least moderate" interrater reliability with human clinicians on depression rating scales.

Performance on simulated datasets is even more striking. Taylor & Francis (2024) found GPT-4 outperformed all models across three datasets, achieving perfect performance on a simulated dataset. Shin et al. (2024, JMIR) achieved accuracy = 0.902 and specificity = 0.955 using GPT-3.5 with CoT prompting on diary-based depression detection (428 diaries from 91 users). DeepSeek V3 (2025) emerged as the most cost-effective option, maintaining high AUC in complex diagnostic scenarios with zero-shot prompting.

However, LLMs consistently overestimate severity/urgency compared to human raters (Italian social media triage study, 2024). GPT-4o and Claude 3.5 Sonnet showed the highest agreement with human judgments but still demonstrated a false-positive bias. Temperature sensitivity is another concern — small parameter changes (0.0 vs. 0.3) cause large performance shifts (Lorenzoni et al., 2025).

Multimodal architectures combining text and audio

The DAIC-WOZ corpus (Gratch et al., 2014) — 189 clinical interviews with virtual interviewer "Ellie," labeled with PHQ-8 scores — remains the primary multimodal benchmark. DepAudioNet (Ma et al., 2016, AVEC), using CNN + LSTM on mel-spectrograms, established the audio baseline at F1 ≈ 0.61. Since then, performance has climbed dramatically:

Architecture	Year	Modality	F1 / Key metric
DepAudioNet (Ma et al.)	2016	Audio	~0.61

Architecture	Year	Modality	F1 / Key metric
Al Hanai et al.	2018	Audio+Text	~0.77
SVM (10 acoustic features, Menne et al.)	2024	Audio	AUC = 0.93
wav2vec 2.0 fine-tuned	2024	Audio	Acc = 0.9649
IntervoxNet (Ding et al.)	2024	Audio+Text	0.90
Multi-MTRB (MIL)	2025	Text	0.88
BERT + BiLSTM + deep GCN	—	Text+Audio+Video	0.963
Multimodal LLM (Qwen2-Audio-7B)	2024	Text+Audio+Video	SOTA at 7B

Fusion strategies matter significantly. Early fusion (concatenating feature vectors before classification) captures cross-modal interactions, achieving F1 = 0.93 in one comparison study (MDPI, 2025). Late fusion (combining separate model predictions) is more robust to missing modalities but typically underperforms. Cross-modal attention — where text guides audio feature reconstruction or vice versa — represents the current state of the art. Multimodal systems consistently outperform unimodal ones, with the complementary information from acoustic (prosodic/physiological) and linguistic (cognitive/semantic) channels being difficult to replicate with either modality alone.

A critical methodological caution: Burdisso et al. (2024) discovered that interviewer prompts in DAIC-WOZ contain bias, allowing models to exploit shortcuts. Models using only Ellie's questions can distinguish depressed from control participants, inflating reported performance. Recent ICMI 2025 work suggests state-of-the-art models may capture generic distress rather than depression-specific signals.

The 2024–2025 landscape

Five trends define the current moment. First, the field is shifting from binary classification to interpretable, item-level prediction — MentaLLaMA and PhqMML now generate explanations alongside predictions. Second, conversational and agentic paradigms are emerging, with eRisk 2025–2026 introducing LLM personas for interactive screening. Third, fairness and bias auditing has become essential, with growing attention to gender bias in DAIC-WOZ, interviewer bias, and cultural bias. Fourth, no LLM-based mental health tool has received FDA clearance or CE marking for clinical diagnosis — the FDA Digital Health Advisory Committee (November 2025) specifically cited risks of "missed crisis cues, inadequate response, AI hallucinations, sycophancy, and parasocial relationships." Fifth, the clinical validation gap persists: models perform well on benchmarks but cross-dataset and cross-disorder generalization remains poor.

5. Designing invisible conversational screening over a 7-day cycle

From clinical questionnaires to naturalistic conversation

The Perla chatbot (Schick et al., 2022) provides the most direct validation that conversational wrapping of the PHQ-9 preserves psychometric properties. It converted PHQ-9 items into a conversational interview and achieved Cronbach's α = 0.81, sensitivity 96%, specificity 90% — comparable to the traditional questionnaire — while achieving 2.5× greater reach than form-based screening. DEPRA (Almusharraf et al., 2023, PLOS ONE) eliminated dependence on multiple-choice replies entirely, using open-ended responses to allow spontaneous expression.

The design philosophy for SentirIA should follow Ecological Momentary Assessment (EMA) principles. EMA outperforms retrospective questionnaires — a study of 67 distressed older adults (Shiffman et al.) found EMA measures "substantially outperformed paper-and-pencil measures" with 25–50% lower NNTs. Moskowitz et al. (2021) found 91.7% concordance between twice-daily EMA reports and clinician-rated treatment responses. Key advantages include reduced recall bias, capture of within-day fluctuations, and higher sensitivity to change.

The invisible screening approach maps clinical domains to life domains: ask about work/school (concentration, fatigue), relationships (guilt/worthlessness, social withdrawal), sleep habits (sleep disturbance), eating patterns (appetite), hobbies (anhedonia), physical sensations (psychomotor changes), and general outlook (mood, hopelessness). Each naturally elicits PHQ-9-relevant information without clinical language. Varying question modality — mixing open-text, emoji responses, voice notes, and simple check-ins — reduces assessment fatigue while collecting multimodal data.

A 7-day cycle mapping PHQ-9 domains to daily themes

The PHQ-9's two-week timeframe can be addressed by running two 7-day cycles. Cycle 1 establishes a baseline; Cycle 2 completes the full two-week window. Within each cycle, items distribute across daily themes:

Day 1 (Interest & mood): Opens the cycle by exploring what the user enjoyed recently and their general emotional state, covering Items 1 and 2. "Hey! Starting a new week — anything you're looking forward to?" In Spanish: "¡Hola! ¿Cómo empiezas la semana? ¿Hay algo que te entusiasme?" If anhedonia signals appear, subsequent days probe deeper with follow-ups like "You mentioned not enjoying things as much — has that changed how you spend your time?"

Day 2 (Rest & sleep): Focuses on sleep quality and routines, covering Item 3. "How did you sleep? Tell me about your night." Spanish: "¿Cómo dormiste anoche? ¿Descansaste bien?"

Day 3 (Energy & activity): Explores fatigue and daily productivity, covering Item 4. "How's your energy today? What's on your plate?" Spanish: "¿Cómo te sientes de energía hoy?"

Day 4 (Daily functioning): Covers appetite and concentration through questions about eating and

focus, covering Items 5 and 7. "Have you eaten well today? How's your focus been?" Spanish: "¿Has podido comer bien hoy? ¿Cómo ha estado tu concentración?"

Day 5 (Self-reflection): Explores self-worth and relationships, covering Item 6. "Sometimes weeks can feel heavy. How are you feeling about things?" Spanish: "A veces la semana se siente pesada. ¿Cómo te sientes contigo mismo(a)?"

Day 6 (Physical experience): Assesses psychomotor changes and reviews prior signals, covering Item 8. "How's your body feeling? Any restlessness or slowness?" Spanish: "¿Cómo sientes tu cuerpo hoy?"

Day 7 (Weekly reflection): Integrates all signals and assesses functional impact. "Looking back on the week, how have things been overall?" Spanish: "Mirando atrás esta semana, ¿cómo han estado las cosas en general?"

Item 9 (suicidal ideation) is never assigned to a specific day. Instead, continuous passive monitoring runs across every interaction, with escalation triggered by risk signals detected at any point.

Adaptive branching and Bayesian probability updating

Adaptive branching should follow a traffic-light model. Green (low risk): standard conversational flow. Yellow (emerging concern): increased probe depth, supplementary questions, shortened intervals between relevant domain assessments. Orange (moderate concern): psychoeducation, coping resources, recommendation for professional consultation. Red (high risk/crisis): immediate crisis protocol activation.

Multi-session integration uses Bayesian updating of depression probability. Starting with the population base rate (~7% MDD prevalence), each daily session updates prior estimates. Items 1 and 2 (anhedonia and depressed mood) carry the highest discriminative weight, consistent with their use in the PHQ-2 as an initial screen. Research on Bayesian networks for depression prescreening (PMC11258528) achieved sensitivity 0.717–0.741 while reducing screening interviews by up to 52%.

Ethics, consent, and crisis protocols

SentirIA faces several ethical imperatives. Informed consent must transparently disclose the AI nature of the interaction, the embedded screening function, and data usage — avoiding the trap of presenting as a "lifestyle app" while performing psychological assessment (ResearchGate, 2025). Layered consent at onboarding with dynamic adaptation is recommended. False positive management should use conversational screening as a first-pass filter with professional evaluation referral for positive screens, framing results as "It might be helpful to talk to someone about how you're feeling" rather than diagnostic language.

Crisis protocols are non-negotiable. A 2025 study (Nature Scientific Reports) testing 29 AI chatbots on C-SSRS scenarios found that none met criteria for "adequate" crisis response.

SentirIA must implement multi-signal detection (direct statements, indirect cues, behavioral changes), tiered responses following the ACT model (Assessment, Crisis Intervention, Trauma Treatment), and locale-specific emergency resources (Línea de la Vida in Mexico: 800-911-2000; Crisis Text Line Spanish: text AYUDA to 741741). The system must never continue automated interaction during active crisis or de-escalate response to increasing risk.

Regarding regulatory positioning, the FDA has not authorized any generative AI-based device for mental health as of late 2025. If SentirIA is positioned as a wellness/screening tool without diagnostic claims, it may fall under FDA enforcement discretion. However, crisis escalation features may trigger SaMD (Software as a Medical Device) classification requiring premarket review.

Check-ins should be kept to 3–5 minutes (Bartal et al., 2024, J Affective Disorders) with warm, varied openers and easy opt-out. WhatsApp's technical affordances — rich media, voice notes, emojis, end-to-end encryption, no app installation barrier, dominant platform in Latin America — make it ideal for this population, though the Business API's rate limits, template message requirements, and 24-hour messaging windows impose design constraints.

---

6. Vocal biomarkers: acoustic features that reveal depression

The acoustic signature of depression

Depressed speech exhibits a distinctive acoustic profile driven by psychomotor retardation, reduced respiratory support, and decreased neuromuscular control. Reduced F0 (fundamental frequency) variability — flattened pitch contour producing "monotonic prosody" — is the single most replicated finding, with effect sizes of $\eta^2 = 0.183–0.3$ (Menne et al., 2024, BMC Psychiatry). Di et al.'s large multisite study found these associations even in individuals with genetic predisposition but no active symptoms, suggesting acoustic changes may precede clinical depression.

Speech rate decreases in depression, with fewer words per minute and fewer syllables per second reflecting cognitive slowing and psychomotor retardation. Critically, Cummins et al. (2023, J Affective Disorders) found reduced speaking rate was the most robust cross-linguistic marker in the RADAR-MDD study (585 participants across UK, Spain, and Netherlands), maintaining significance across English, Spanish, and Dutch — directly relevant to SentirIA's bilingual design.

Pause patterns — increased duration, increased frequency, longer response latency — are among the most reliably replicated findings. Mundt et al. (2012, Biological Psychiatry) demonstrated association with both depression severity and treatment response, showing that within-person changes in pause patterns correlate with symptom changes during treatment. This makes pause analysis particularly valuable for longitudinal monitoring.

Jitter (cycle-to-cycle frequency variation) and shimmer (cycle-to-cycle amplitude variation) reflect reduced neuromuscular control and respiratory support. Healthy jitter is ≤0.5% (relative); healthy shimmer is <3–5%. Both generally increase in depression, producing the breathy, hoarse quality characteristic of depressed speech. Harmonics-to-noise ratio (HNR) decreases (more noise, less clarity), typically from healthy values of 15–20 dB. Vocal energy/loudness decreases reflecting reduced subglottal pressure, with large effect sizes (η^2 = 0.183–0.3).

Spectral features provide additional discrimination. MFCCs (Mel-Frequency Cepstral Coefficients) are among the most discriminative features: Zhao et al. (2022, Frontiers in Psychiatry) found MFCC7 predicted PHQ-9 scores (β = 0.90, p = 0.01), with classification accuracy reaching 89.66%. The alpha ratio, Hammarberg Index, and spectral slope capture spectral tilt changes associated with reduced vocal effort in depression.

OpenSMILE and pyAudioAnalysis: practical toolkits

OpenSMILE (Eyben, Wöllmer & Schuller, 2010, ACM Multimedia) is the standard toolkit for depression-relevant acoustic feature extraction, developed at TU Munich and maintained by audEERING GmbH. Its eGeMAPS feature set (Eyben et al., 2016, IEEE Trans. Affective Computing) extracts 88 functionals from 25 low-level descriptors designed specifically for clinical and affective computing applications. These include F0 statistics (mean, coefficient of variation, percentiles), jitter, shimmer, loudness, HNR, spectral slopes, formant frequencies, MFCCs 1–4, and temporal features (rate of loudness peaks, mean length of unvoiced regions). The larger ComParE_2016 set extracts 6,373 features for brute-force analysis. Python integration is straightforward via the `opensmile` package, requiring three lines of code to extract 88 eGeMAPS features from an audio file.

pyAudioAnalysis (Giannakopoulos, 2015, PLOS ONE) offers 34 short-term features (68 with deltas) including ZCR, energy, spectral centroid, 13 MFCCs, and 12 chroma coefficients. It is pure Python with Apache licensing (free for commercial use, unlike OpenSMILE). However, it lacks jitter, shimmer, HNR, F0, and formant extraction — the very features most relevant to depression. It achieved F1 = 0.59 on DAIC-WOZ with minimal tuning. For SentirIA, OpenSMILE's eGeMAPS is the recommended primary toolkit, with pyAudioAnalysis as a supplementary tool for rapid prototyping.

Integrating voice with text for multimodal screening

The most relevant practical demonstration for SentirIA comes from Otani et al. (2026, PLOS Mental Health), who analyzed WhatsApp audio messages from 160 Brazilian Portuguese speakers. Peak accuracy reached 91.67% (women) and 80% (men), with AUC = 91.9% for women. Spontaneous speech outperformed structured counting tasks, validating that ecological voice messages contain sufficient information for depression detection. Key considerations for WhatsApp voice notes include compression artifacts (Opus codec), variable microphone quality, background noise, and typically short message duration.

For the multimodal integration pipeline, the recommended approach is: Voice Activity Detection

(VAD) → logMMSE noise reduction → silence removal → OpenSMILE eGeMAPS feature extraction → fusion with NLP text features. Early fusion and cross-modal attention architectures generally outperform late fusion. Minimum audio duration for reliable analysis is approximately 30 seconds (Sonde Health), though 2–5 minutes is preferred. On-device feature extraction with transmission of feature vectors only (not raw audio) provides privacy preservation.

A key finding for cross-cultural deployment: acoustic markers appear more language-robust than linguistic markers. The RADAR-MDD study confirmed this across English, Spanish, and Dutch, though models trained on English drop to ~48.72% accuracy on tonal languages like Mandarin (arXiv 2025). Within Western languages, >80% accuracy is transferable. Gómez-Zaragozá et al. (Interspeech 2025) conducted the first evaluation of speech foundation models specifically for Spanish depression detection using the DEPTALK dataset, finding no significant gender differences in speech-only models — a positive finding for equitable deployment.

Conclusion: from evidence to architecture

This research reveals both the promise and the boundaries of NLP-based depression screening. Five insights should guide SentirIA's design. First, item-level ensemble approaches outperform monolithic models — different LLMs excel at different PHQ-9 items (Teferra et al., 2025), suggesting SentirIA should route item-specific predictions to the best-performing model for each domain. Second, absolutist language ($d > 3.14$) is a far stronger signal than negative emotion words ($d = 0.72$) or first-person pronouns ($d = 0.44$), yet most existing systems weight emotion words most heavily. Implementing Al-Mosaiwi and Johnstone's absolutist word dictionary alongside the LIWC-based approach could substantially improve signal detection.

Third, acoustic biomarkers — particularly speech rate and pause patterns — transfer robustly across languages (Cummins et al., 2023), making them especially valuable for SentirIA's bilingual Spanish-English design where linguistic markers may require separate calibration. Fourth, conversational wrapping of clinical instruments preserves psychometric validity (Perla: sensitivity 96%, specificity 90%) while dramatically increasing reach, confirming the viability of invisible screening. Fifth, the field is converging on a critical truth: no AI system should operate without crisis protocols and human oversight, and the regulatory landscape is tightening. SentirIA should be designed as a screening-and-referral tool positioned under FDA enforcement discretion, with robust crisis escalation, transparent informed consent, and physician-supervised interpretation of results.

The technical foundation exists. GPT-4o achieves $F1 = 0.74$ on item-level PHQ-8 prediction from transcripts; MentalBERT reaches $F1 = 0.99$ on anhedonia detection with clinical fine-tuning; multimodal text+audio systems achieve $F1 > 0.90$ on DAIC-WOZ; and WhatsApp voice note analysis reaches $AUC = 0.92$ in real-world conditions. The challenge now is clinical validation — bridging from benchmark performance to real-world screening with diverse Spanish-speaking

populations, longitudinal follow-up, and integration into care pathways. SentirIA, grounded in PHQ-9 item-level mapping, ecological momentary assessment, and multimodal signal fusion, is well-positioned to address this gap.