INTERVENCIÓN Y RESPUESTA

# Motivational interviewing for AI companions: a comprehensive technical and clinical guide for SentirIA

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.

# Motivational interviewing for AI companions: a comprehensive technical and clinical guide for SentirIA

MI-based conversational agents show moderate clinical efficacy (Hedge's g = 0.64 for depression) and can achieve 98% MI fidelity using LLMs, but SentirIA faces a critical regulatory decision: its "early depression detection" framing places it squarely in clinical intervention territory under FDA, ANVISA, and COFEPRIS frameworks, requiring either regulatory authorization or careful repositioning as a wellness companion. The evidence supports hybrid architectures combining rule-based safety guardrails with generative MI responses, and culturally adapting for Latin American populations demands integration of familismo, personalismo, respeto, and simpatía values into every conversational design decision. This report synthesizes the full research landscape — from NLP techniques for change talk detection to crisis escalation protocols — into an actionable blueprint for building SentirIA.

---

## 1. How OARS translates into working chatbot code

The OARS framework (Open-ended questions, Affirmations, Reflective listening, Summarizing) is the operational backbone of MI, and each component now has documented AI implementation patterns.

Open-ended questions that explore ambivalence can be generated through two proven approaches. MIBot (Brown et al., 2023, *JMIR Mental Health*) uses a "running head start" technique with five scripted questions that move from exploring what the user likes about a behavior to what they'd change and what steps they need to take. The more advanced CAMI system (ACL 2025) uses a STAR framework (State-Topic-Action-Response) where GPT-4o first infers the client's readiness state (precontemplation, contemplation, or preparation), then navigates a hierarchical topic tree of 59 fine-grained motivation topics across five domains (health, economy, relationships, law, education) to generate contextually appropriate open-ended questions. For SentirIA, this means questions should be dynamically selected based on both the user's stage of change and the specific life domain most salient to them.

Reflective listening is where AI implementation has made the most dramatic progress. MIBot evolved across four versions, moving from scripted responses to a GPT-2 XL fine-tuned model that generates reflections specific to user statements. Kumar et al. found that 88% of AI-generated reflections met predefined MI criteria. The MITI 4.2.1 framework distinguishes simple reflections (restatement or rephrasing) from complex reflections (adding meaning, feeling, or metaphor beyond what the client explicitly stated). Double-sided reflections — "On one hand you enjoy X, and on the other you're concerned about Y" — emerge naturally when chatbots structure

conversations to elicit both pros and cons sequentially. Can, Marín, Georgiou, Imel, Atkins, and Narayanan (2016, *Journal of Counseling Psychology*) developed an NLP approach using Maximum Entropy Markov Models with linguistic features — including n-grams, semantic similarity between counselor and client utterances, and contextual features — to automatically detect counselor reflections with strong performance.

Affirmation requires distinguishing genuine recognition of client strengths from empty praise. The MITI 4.2.1 coding manual operationalizes "Affirm" as recognizing specific client strengths, values, or efforts rather than generic praise. Welivita and Pu (2023, *Findings of ACL*) demonstrated that 92.86% of peer-generated advice on support platforms was MI-nonadherent (typically "Advise without permission"), and fine-tuned BlenderBot and GPT-3 to rephrase non-adherent responses into MI-adherent forms — essentially building an affirmation quality filter. For SentirIA, this means every affirmation should reference something specific the user said or did, never defaulting to "great job!" patterns.

Summarizing techniques include collecting summaries (tying together themes from multiple user statements), linking summaries (connecting different motivation areas), and transitional summaries (synthesizing earlier input to guide the conversation forward). CAMI's topic exploration mechanism tracks discussed topics and generates linking summaries connecting different motivation areas, while MIBot employs what Brown et al. call "backward-looking reflections" that synthesize prior user input.

---

## 2. MI spirit and rolling with resistance in AI design

The MI spirit — partnership, acceptance, compassion, evocation — presents the deepest design challenge for AI systems. Partnership, rated on MITI's 1-5 scale, requires what Moyers et al. (2014, *Journal of Substance Abuse Treatment*) describe as appearing to "dance" with the client. A critical finding from Jörke et al. (2024) with GPTCoach is that LLMs' instruction-following objectives inherently conflict with MI's facilitative approach — they default to providing solutions rather than asking questions. This means system prompts must explicitly counteract the LLM's tendency toward advice-giving.

Evocation — drawing out the client's own motivations — is implemented in CAMI through topic exploration that searches for motivation topics likely to evoke change talk. Acceptance translates to "Emphasize Autonomy" behaviors in MITI coding, which in chatbot design means permission-asking before giving information. Brown et al. measured compassion through the CARE (Consultation and Relational Empathy) scale, finding that enhanced generative reflections significantly improved CARE scores (P=.004).

Rolling with resistance is operationalized at the response-generation level. CAMI includes "Softening Sustain Talk" as a key evaluation metric. Welivita and Pu (2023) demonstrated rephrasing MI-nonadherent responses (direct confrontation, unsolicited advice) into MI-adherent

forms using fine-tuned LLMs, effectively automating the "roll with resistance" technique. Karve et al.'s 2025 JMIR scoping review found users perceived AI MI systems as "judgment-free, supportive, and easier to engage with than human counselors, particularly in stigmatized contexts" — a finding directly relevant to SentirIA's Latin American deployment, where mental health stigma is pervasive.

---

## 3. Detecting change talk with NLP: from topic models to transformers

Automated change talk detection has evolved through three generations of NLP approaches. The DARN-CAT framework (Miller & Rollnick, 2013) classifies change talk into Desire ("I want to..."), Ability ("I can..."), Reason ("If I... then..."), Need ("I must..."), Commitment ("I will..."), Activation ("I'm ready to..."), and Taking Steps ("I went out and did..."). DARN represents preparatory change talk; CAT represents mobilizing change talk — the distinction matters because mobilizing change talk is a stronger predictor of actual behavior change.

The pioneering computational work began with Atkins, Steyvers, Imel, and Smyth (2014, *Implementation Science*), who used Latent Dirichlet Allocation topic models for automated MI fidelity coding across 148 participants from 5 RCTs, achieving session-level agreement with human coders comparable to inter-rater reliability ($\kappa > .75$ for open questions). Tanana et al. (2016, *Journal of Substance Abuse Treatment*) advanced this by comparing Discrete Sentence Features models with Recursive Neural Networks on 341 sessions (78,977 talk turns), finding the RNN model outperformed for utterance-level predictions. Pérez-Rosas, Mihalcea, Resnicow, Singh, and An (2017) built a dataset of 22,719 counselor utterances from 277 sessions annotated with 10 MITI behavioral codes.

The current state of the art combines LLMs with probabilistic models. Lim et al. (2025, *BMC Psychiatry*) used GPT-4o to classify client utterances into change talk, sustain talk, or neutral categories, assigning strength scores from -5 to +5, then combined these with Hidden Markov Models to model motivational state transitions. The LLM-only approach achieved 0.67 accuracy; the LLM+HMM pipeline reached 0.86 accuracy. High-quality MI sessions showed fluid transitions between motivational states, while low-quality sessions showed persistence in resistance states.

The most comprehensive automated MI coding system is Flemotomos et al. (2021, *Behavior Research Methods*), deployed on 5,000+ recordings via the Lyssn.io platform. Its pipeline runs: Voice Activity Detection → Speaker Diarization → ASR → Speaker Role Recognition → Behavior Code Prediction (BiLSTM with attention) → Session-Level Metrics (Reflection-to-Question Ratio, MI-Adherent percentage). For SentirIA, this pipeline could be adapted to analyze both text conversations and transcribed voice notes.

Publicly available datasets for training change talk detectors include AnnoMI (Wu et al., 2022/2023) with 133 expert-annotated MI conversations available on GitHub and HuggingFace,

and Welivita and Pu's (2022) dataset of ~17K responses annotated using MITI-derived labels.

---

## 4. Stages of change shape the entire conversation flow

The transtheoretical model integration is not optional — it fundamentally determines what MI techniques the chatbot should deploy at any given moment. CAMI (ACL 2025) explicitly implements stage inference, using LLM prompting to classify client state and then selecting strategies conditionally. MIBot uses "readiness rulers" (0-10 scales measuring readiness, confidence, and importance) that map to TTM stages.

For SentirIA, the stage-adaptive flow should follow established principles. In precontemplation, the chatbot focuses on consciousness raising and exploring ambivalence — asking exploratory questions and avoiding action planning. In contemplation, it deploys decisional balance exercises, exploring pros and cons. In preparation, it transitions to planning questions about specific steps. In action, it supports commitment and problem-solves barriers. In maintenance, it reinforces strategies and addresses relapse prevention. MHC-Coach (Nature, npj Cardiovascular Health, 2025) demonstrated this approach with a fine-tuned LLaMA3-70B that autonomously generates stage-matched messages, evaluated with N=632 participants.

---

## 5. The evidence base: what MI chatbots actually achieve

The most rigorous evidence comes from a convergence of individual studies and meta-analyses. Li et al. (2024, *npj Digital Medicine*) meta-analyzed 15 RCTs and found chatbot interventions produced Hedge's g = 0.64 (95% CI: 0.17–1.12) for depression and g = 0.70 (95% CI: 0.18–1.22) for distress. Zhao et al. (2024) found smaller but significant effects across 18 RCTs (N=3,477): g = -0.26 for depression and g = -0.19 for anxiety, with benefits most pronounced after 8 weeks but no sustained effects at 3-month follow-up. Zhang et al. (2025), focusing on generative AI chatbots across 14 RCTs (N=6,314), found an overall effect size of 0.30 (P=.047).

Specific MI-focused systems show promising results. The landmark Therabot RCT (Heinz et al., 2025, *NEJM AI*, N=210) — though CBT-focused rather than MI-specific — demonstrated 51% PHQ-9 symptom reduction at 8 weeks (d=0.845–0.903) and 31% GAD-7 reduction (d=0.794–0.840), with therapeutic alliance scores comparable to human therapists. Tess (Fulmer et al., 2018, *JMIR Mental Health*, N=75) achieved Cohen's d = 0.68 for depression reduction with a remarkable 0% attrition in test groups. MIBot v6.3A (Mahmood et al., 2025), a fully generative GPT-4o MI chatbot tested on 106 smokers, achieved 98% MI adherence — higher than human counselors — and increased quit confidence by 1.7 points on a 0-10 scale.

Head-to-head comparisons between MI and non-MI chatbots reveal nuanced findings. He et al.

(2024, *J Med Internet Res*, N=229) compared MI versus confrontational counseling chatbots and found the MI chatbot produced significantly higher user experience outcomes (engagement, therapeutic alliance, perceived empathy, satisfaction), though cessation-related outcomes were similar. Meyer and Elsweiler (2025, *International Journal of Human-Computer Studies*) showed that MI principles effectively mitigated potential harms in LLM outputs, with MI-adapted GPT-4 successfully increasing readiness to change in a pre-registered RCT.

The Karve et al. (2025, JMIR) scoping review of 15 AI-MI studies found 60% used rule-based chatbots, 27% were LLM-based, and 87% reported positive feasibility/acceptability — but only 20% showed significant behavioral changes. Only 40% formally assessed MI fidelity. The progression from rule-based to fully generative systems is clear: Brown et al.'s MIBot series demonstrated that adding generative reflections consistently improved outcomes across four iterations.

LLM benchmarking reveals that current models achieve strong MI competence. Shukla et al. (2026) tested 10 LLMs and found all achieved fair (MITI > 3.5) to good (> 4) MI competence, with LLMs outperforming human experts in Complex Reflection percentage (39% vs 96%) and Reflection-to-Question ratio (1.2 vs > 2.8). Psychiatrists could identify LLM-generated responses with only 56% accuracy. GPT-4o scored 0.95 accuracy on MI knowledge tests, though ethical classification accuracy remained lower (best model at 0.56).

---

## 6. The reassurance trap and why empathic calibration matters most

A 2025 JMIR Mental Health study (Scholich et al.) comparing LLM chatbots and licensed therapists found the most concerning divergence: chatbots reassured 1.6 times per interaction versus therapists' 0.2 times (U=23; P=.02; r=–0.47), gave suggestions 7.4 times versus therapists' 3 times (r=–0.61), and asked fewer clarifying questions. Human therapists evoked more elaboration (U=9; P=.001). This systematic over-reassurance creates what clinicians call the "reassurance trap" — reinforcing avoidance and dependency rather than building genuine coping capacity.

The APA's November 2025 Health Advisory explicitly warns that for people with anxiety or OCD, chatbots can reinforce reassurance-seeking compulsions. The advisory states: "Developers must reduce high-risk features by implementing safety mechanisms (e.g., reducing overly human-like chatbot qualities, limiting sycophancy, and preventing the system from validating or promoting delusional thoughts)." A Nature Human Behaviour study (2025, n=6,282) found that human-attributed empathic responses were rated as more empathic than AI-attributed ones even when the text was identical — suggesting transparency about AI nature may inherently reduce therapeutic impact, creating a fundamental tension.

Practical design principles to avoid paternalism without sacrificing warmth include: asking before

advising (therapists asked more clarifying questions while chatbots jumped to solutions), acknowledging limitations openly (Wysa's approach of immediately changing course when users object, emphasizing it is "a bot that is still learning"), and communicating uncertainty rather than projecting false confidence. The HAILEY system (Sharma et al., 2023, *Nature Machine Intelligence*), co-developed by Adam Miner at Stanford, demonstrated that AI-augmented human responses achieved a 20% increase in expressed empathy and 39% increase for those who struggle with empathy — suggesting the optimal model may be AI-assisted human care rather than fully autonomous delivery.

---

## 7. Crisis detection: a four-level escalation architecture

Crisis detection must be treated as foundational infrastructure, not an edge case. The composite escalation algorithm, synthesized from MIND-SAFE (Boit & Patil, 2025), Wysa's global deployment data, and clinical guidelines, operates across four levels.

Level 0 (Routine Monitoring) maintains regular mood check-ins using PHQ-2 screening, engagement tracking, and sentiment analysis baselines. Level 1 (Elevated Concern) triggers on persistent negative mood, social withdrawal language, sleep disturbance mentions, or hopelessness themes — actions include increased check-in frequency, psychoeducation, and PHQ-9 administration. Level 2 (Moderate Risk) activates when PHQ-9 Item 9 scores ≥ 1, with vague references to death, expressions of burdensomeness, or increased substance use language — the chatbot administers C-SSRS screener questions, provides empathetic acknowledgment, presents crisis resources, and flags for clinical review. Level 3 (High Risk) triggers on C-SSRS questions 4-6 positive (intent, plan, or behavior), explicit suicidal statements, or PHQ-9 Item 9 score of 3 — and critically, bypasses the LLM entirely, switching to pre-scripted, clinically validated crisis protocols. Level 4 (Imminent Danger) directs users to emergency services with clear, calm instructions.

The PHQ-9 Item 9 validation data (Na et al., 2018, *Journal of Affective Disorders*) reveals important limitations: while sensitivity is 87.6% (good at catching true positives), specificity is only 66.1% and positive predictive value is 28.6% — meaning most positive screens are false alarms. The negative predictive value of 97.2% makes it excellent for ruling out risk but insufficient as a standalone assessment. This is why the C-SSRS screener must follow any positive Item 9 result.

Wysa's global crisis data (2024) shows that 82% of crisis instances were detected by AI and confirmed by the user, but only 2.4% of users in crisis chose to call helplines when encouraged — underscoring the need for alternative support pathways like in-app safety planning rather than relying solely on hotline referrals. For SentirIA's Latin American deployment, country-specific crisis lines must be integrated: Línea de la Vida (800-911-2000) in Mexico, CVV (188) in Brazil, Línea 106 in Colombia, and Centro de Asistencia al Suicida (135) in Argentina.

## 8. Cultural adaptation for Latin American populations goes beyond translation

Only 35.1% of Hispanic/Latinx adults with mental illness receive treatment annually versus 46.2% of the U.S. average, driven by stigma, familial pressure, and structural barriers. SentirIA's chatbot design must integrate four core cultural values that profoundly shape therapeutic communication.

Familismo — the centrality of family loyalty, support, and closeness — means behavior change should be framed in terms of family benefit. The saying "La ropa sucia se lava en casa" (don't air dirty laundry in public) actually creates an opportunity: chatbots offer private, anonymous support that bypasses the shame of disclosing family problems to strangers. Personalismo — prioritizing personal over institutional connections — requires building rapport before therapeutic content, using the user's name, and demonstrating curiosity about their background. Latinos may interpret neutral or businesslike tone as negative. Respeto — deference to authority and sensitivity to perceived slights — means offering formal address options (usted vs. tú) and avoiding language that could seem dismissive. Simpatía — valuing smooth relationships and politeness — argues against confrontational MI techniques and favors collaborative, warm language even when exploring difficult topics.

Lee et al.'s culturally adapted MI for Latino heavy drinkers (2011-2016) demonstrated this in practice. Their RCT (n=57) produced significant decreases in heavy drinking days and drinking consequences ($p<.001$), with CAMI showing greater reductions at 2 months ($p=.009$). Key adaptations included training interventionists to elicit cultural and social influences on behavior, adapting the "Typical Day" strategy to acknowledge acculturation stress, and advertising the intervention as a "health education study" to avoid stigmatizing stereotypes. 95% of participants reported that understanding their culture was important to understanding their behavior.

Language considerations extend beyond simple translation. Dialectal variation across Mexican, Caribbean, Central American, and South American Spanish affects vocabulary, idiom, and register. The formal/informal address distinction (usted/tú) carries significant cultural weight. Some users may speak indigenous languages (Quechua, Nahuatl) or Portuguese. SAMHSA (2022) recommends incorporating culturally resonant formats like cuento (storytelling) and fotonovelas, which could inform SentirIA's psychoeducation delivery.

## 9. The hybrid architecture SentirIA needs

The evidence overwhelmingly supports a hybrid architecture combining rule-based safety with generative MI responses. Pure rule-based systems (used by 60% of MI chatbots in the Karve

review) produce repetitive, generic responses that reduce engagement over time. Pure generative systems risk hallucination, MI-nonadherent behaviors, and safety failures. The sweet spot is demonstrated by MIBot's architecture, which uses an intent classifier, yes/no classifier, and content/no-content classifier for NLU, a dialogue management engine controlling conversation state, and GPT-2 XL (now upgradeable to GPT-4o) for generating MI reflections.

The recommended architecture for SentirIA has four layers:

Input layer handles proactive risk detection using crisis keyword matching (pre-LLM), voice note processing (WhatsApp voice note → Whisper API transcription → openSMILE/eGeMAPS feature extraction → depression classifier), and user state assessment from conversation history.

Dialogue engine manages session phase tracking across MI's four processes (Engaging → Focusing → Evoking → Planning), stage-of-change inference using LLM classification, topic selection from a motivation domain tree, and user state persistence across sessions via an external database (since WhatsApp lacks persistent session state).

Generation layer wraps user messages in MI-specific system prompts with few-shot examples, uses two-step prompting (dialogue act prediction → response generation, per Sun et al., 2025), and applies post-generation filters for MI adherence, toxicity, and clinical appropriateness. The system prompt structure should include role definition, current MI session phase, OARS instructions, change talk cultivation directives (DARN-CAT), safety constraints, and few-shot examples of good MI responses.

Safety layer uses pre-scripted, clinically validated responses that bypass the LLM entirely when crisis is detected, implements the four-level escalation protocol described above, and integrates country-specific crisis resources.

For prompt engineering, the MICA study (Borsari et al., 2025) demonstrated that iterative prompt refinement improved CEMI fidelity scores from M=69.6 to M=81.3, achieving the MI fidelity benchmark. Key strategies include chain-of-thought prompting (Huang et al., 2025, showed this improved GPT-4's MI performance by reducing inappropriate advice while enhancing reflections), two-step dialogue act prediction followed by response generation (Sun et al., 2025), and explicit instructions to counteract LLMs' default advice-giving tendency.

MI fidelity measurement should adapt the MITI 4.2.1 framework for continuous AI monitoring. Key metrics to track: Reflection-to-Question Ratio (target ≥ 1:1 for competency, ≥ 2:1 for proficiency), Percent Complex Reflections (target ≥ 40% for competency, ≥ 50% for proficiency), MI-Adherent vs. MI-Nonadherent behavior ratio, and global scores for Cultivating Change Talk, Softening Sustain Talk, Partnership, and Empathy. The LLM+HMM pipeline (Lim et al., 2025) provides a validated automated approach, achieving 0.86 accuracy for MI quality evaluation.

---

## 10. Vocal biomarkers add a critical detection layer

Depression-associated vocal features include lower fundamental frequency (F0), reduced pitch variability, slower articulation rate, longer pause durations, increased jitter and shimmer, reduced harmonic-to-noise ratio, and differences in MFCCs. The eGeMAPS feature set (88 features, Eyben et al., 2015, IEEE TAFFC) provides the recommended minimal set for voice research, extractable via openSMILE. More recent approaches using Wav2Vec2/WavLM self-supervised speech representations have achieved 89.2-97.4% accuracy on specific tasks (Attas, 2026).

For SentirIA's WhatsApp voice note processing, the pipeline would be: receive voice note → download via WhatsApp Business API → transcribe using Whisper → extract acoustic features using openSMILE (eGeMAPS) → run depression classifier → fuse with text-based signals and PHQ-9 scores. Key challenges include variable audio quality from WhatsApp's compression, background noise in real-world recording conditions, and the 16MB file size limit for voice notes. The DAIC-WOZ corpus (Distress Analysis Interview Corpus) and AVEC challenge datasets provide training data labeled with PHQ-8 scores.

Conversational PHQ-9 integration should balance natural dialogue flow with psychometric validity. Perla (Arrabales, 2020) achieved Pearson correlation of 0.91 between conversational PHQ-9 scores and standard form administration, with mean absolute error of 1.88 points. Tess achieved 99.82% completion rates for chatbot-delivered PHQ-9 with internal consistency $\alpha$=0.896. However, psychometricians caution against excessive rephrasing that could compromise psychometric validity — SentirIA should weave PHQ-9 themes into natural conversation while maintaining sufficient fidelity to the validated instrument.

---

## 11. SentirIA's regulatory crossroads: wellness companion or clinical tool

SentirIA's self-description as a "companion for early depression detection" creates immediate regulatory tension. Under the FDA framework, any software intended to detect or screen for depression performs a diagnostic function, making it a regulated Software as Medical Device (SaMD). The word "detection" in marketing materials would likely trigger FDA oversight. Similarly, deploying MI systematically with clinical intent and using validated instruments like the PHQ-9 places SentirIA in clinical intervention territory.

The regulatory landscape includes FDA SaMD classification (Class II for most mental health applications, requiring 510(k) or De Novo), EU MDR Rule 11 with AI Act "high-risk" classification, and country-specific frameworks in Latin America. Woebot's shutdown on June 30, 2025 — citing inability to navigate regulatory requirements for LLM-based evolution — serves as a cautionary tale. Wysa has pursued a hybrid model with both a freely accessible wellness chatbot and FDA Breakthrough Device Designation for specific clinical indications.

SentirIA has two positioning options. Option A (lower regulatory burden): position as a "wellness companion that supports emotional self-awareness and encourages help-seeking," avoiding any claims about detection, diagnosis, or treatment. Use MI-aligned conversational style without naming it as MI therapy. Under this option, PHQ-9 should not be used as a core screening feature. Option B (higher regulatory burden, stronger positioning): pursue regulatory authorization as SaMD for depression screening and support, requiring clinical validation, premarket review, and ongoing surveillance across ANVISA (Brazil), COFEPRIS (Mexico), INVIMA (Colombia), and ANMAT (Argentina).

Data privacy compliance requires country-by-country analysis. Brazil's LGPD classifies health data as sensitive personal data requiring explicit consent, with fines up to 2% of annual revenue (capped at R$50M). Mexico's Federal Law on Protection of Personal Data requires opt-in consent for health data, with INAI having imposed $16.7M USD in fines. Colombia's Law 1,581/2012 requires express authorization for sensitive health data. Argentina's Law 25,326 includes a constitutional "habeas data" right. WhatsApp's end-to-end encryption provides baseline protection, but metadata and usage patterns may be accessible per Meta's policies — SentirIA should implement additional encryption and maintain data residency compliance for each target country.

The APA's November 2025 Health Advisory states bluntly: "Even generative AI tools developed with high-quality psychological science and using best practices do not have enough evidence to show they are effective or safe to use in mental health care." WHO's mhGAP guidelines support digital interventions as adjuncts but not replacements for functioning health systems. No Latin American psychological association has issued specific guidelines on AI-delivered interventions, though the field is nascent — a 2025 scoping review found only 15 digital mental health intervention studies across Brazil, Chile, Colombia, Mexico, and Peru combined.

---

## Conclusion: what SentirIA should build and what it should avoid

The research reveals several clear conclusions that should guide SentirIA's development. First, LLM-based MI delivery has crossed the competence threshold — GPT-4o achieves 98% MI adherence and produces reflections that psychiatrists cannot reliably distinguish from human-generated responses. The hybrid architecture (rule-based safety + generative MI) is the only defensible approach given current evidence on both efficacy and safety.

Second, the effect sizes are real but modest. Meta-analytic estimates for chatbot mental health interventions range from g = 0.26 to g = 0.70 for depression, with effects most pronounced after 8 weeks but fading at 3-month follow-up. This means SentirIA should be positioned as a bridge to professional care and an engagement tool for ongoing monitoring, not as a standalone treatment.

Third, cultural adaptation is not cosmetic. The 95% of Latino participants who said cultural

understanding mattered for behavior change represent a design mandate. SentirIA's conversational design must weave familismo, personalismo, respeto, and simpatía into every interaction pattern — from the onboarding flow (building rapport before any assessment) to the reflection style (warm, collaborative, never confrontational) to the framing of behavior change (in terms of family benefit and community wellness rather than individual pathology).

Fourth, the regulatory decision is existential. SentirIA cannot simultaneously claim "early depression detection" and avoid clinical intervention classification. The team must choose a path and design accordingly — either building the clinical evidence base and pursuing regulatory authorization across target Latin American countries, or repositioning as a wellness companion that supports emotional self-awareness without making detection or diagnostic claims. Given the absence of specific Latin American AI mental health regulation and the rapid evolution of FDA's approach (evidenced by the November 2025 DHAC meeting), engaging regulatory counsel in each target country immediately is not optional but essential.

Finally, crisis safety is non-negotiable infrastructure. The four-level escalation protocol — with Level 3 and 4 bypassing the LLM entirely for pre-scripted clinical responses — must be built first, not last. Wysa's finding that only 2.4% of users in crisis actually call helplines means SentirIA must offer alternative safety planning tools (warning signs identification, coping strategy lists, trusted contact activation) rather than simply displaying a phone number.