Designing Structured Conversational Support for Tuberculosis Treatment Adherence and Patient Coping

Priyanshi Garg[1,4], Sarah Iribarren[1], Sikha Pentyala[3], Yvette Rodriguez[1], Priscilla Carmiol-Rodriguez[1], Alfie Aguilar Vidrio[1], Charles Kwanin[1], Jennifer Sprecher[1], and Javier Roberti[2]

[1]Department of Biobehavioral Nursing & Health Informatics, University of Washington, Seattle, WA, USA

[2]Centre for Research on Epidemiology and Public Health (CIESP), CONICET Buenos Aires, Argentina

[3]School of Engineering and Technology, University of Washington, Tacoma, WA, USA

[4]Department of Linguistics, University of Washington, Seattle, WA, USA

Abstract

Tuberculosis (TB) remains a major global health challenge, and treatment adherence continues to be difficult despite the availability of effective medication. While Digital Adherence Technologies (DATs) have improved monitoring and care coordination, prior deployments highlight unmet needs for timely, personalized, and emotionally supportive communication outside clinical settings. We develop and iteratively refine a Spanish-language TB treatment-support chatbot through multiple rounds of internal expert evaluation, beginning with an initial single-prompt baseline. The system separates three core functions: (i) TB information support grounded in curated resources, (ii) coping-oriented support inspired by Dialectical Behavior Therapy (DBT), and (iii) safety-critical crisis handling via a deterministic, non-generative pathway. These components are implemented within a routed architecture with shared conversational state. Iterative evaluation identified recurring failure modes in the initial baseline and subsequent prototypes, including weak grounding, poor multi-turn continuity, premature skill coaching, and inconsistent safety behavior. Addressing these issues motivated retrieval grounding, explicit routing, state tracking, and task-specific prompting. Relative to the initial single-prompt baseline, our findings suggest that structured interaction design helped address observed failures in grounding, continuity, and safety behavior in this clinical support prototype.

1 Introduction

Tuberculosis (TB) remains one of the leading causes of death from infectious disease globally (World Health Organization, 2024). Although curable, TB treatment requires sustained adherence over several months, and non-adherence remains a major barrier to successful outcomes (Tola et al., 2015). Beyond biomedical factors, patients frequently experience stigma, uncertainty about side effects, and difficulty maintaining motivation over a prolonged treatment course (Courtwright and Turner, 2010; Iribarren et al., 2014). At the same time, health systems have limited capacity to provide continuous support outside scheduled clinical encounters.

TB treatment is also associated with substantial psychosocial burden. Studies consistently report elevated rates of depression and anxiety among TB patients. For example, a recent study of patients with multidrug-resistant tuberculosis (MDR/RR-TB) found that 65.75% reported depressive symptoms and 57.53% reported anxiety, with one-third experiencing both conditions (Zhang et al., 2024). Other studies similarly report high rates of depressive symptoms, such as 45.5% in a cohort from Ethiopia (Peltzer et al., 2012). While estimates vary across settings, systematic reviews confirm a high burden of mental health comorbidity among TB patients (Koyanagi et al., 2017). These factors directly affect patients' ability to sustain engagement with treatment.

Digital Adherence Technologies (DATs) have improved monitoring and communication in TB care (Subbaraman et al., 2018; Zary et al., 2024), but they primarily support adherence tracking rather than ongoing informational and emotional needs. Patient-centered interventions such as the TB Treatment Support Tool (TB-TST) demonstrate the importance of continuous communication and tailored support (Iribarren et al., 2021, 2022; Milligan et al., 2021). However, these systems rely on structured workflows and human-mediated responses, limiting their ability to provide immediate, context-sensitive support at scale.

Dialectical Behavior Therapy (DBT) is an evidence-based psychotherapy framework organized around four skill domains—distress tolerance, emotion regulation, mindfulness, and inter-

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 153–163

personal effectiveness—that has shown efficacy across a range of psychiatric and behavioral conditions (Linehan, 2015). More recently, DBT has been adapted for populations with chronic medical illness who struggle with treatment adherence. A feasibility trial of DBT for adolescents with end-stage renal disease found significant improvements in both adherence and depression (Hashim et al., 2013), and an experimental study of DBT in patients with coronary heart disease demonstrated significant increases in medication adherence and self-care behavior (Tavakoli et al., 2020). These adaptations build on a theoretical framework (DBT-CMI) that positions DBT's core skills as directly targeting the psychological mechanisms underlying medical nonadherence, including distress intolerance, emotional dysregulation, and interpersonal barriers to engaging with care teams (Lois and Miller, 2018).

The psychosocial profile of TB patients—elevated rates of depression and anxiety, perceived stigma, prolonged treatment burden, and limited access to continuous support—aligns closely with the challenges these DBT adaptations were designed to address. While cognitive-behavioral interventions have shown promise for TB adherence (Tola et al., 2015), no prior work has applied DBT-informed skills specifically to TB treatment support. The combination of distress tolerance skills (for managing side effects and treatment fatigue), emotion regulation (for coping with stigma and isolation), and interpersonal effectiveness (for communicating with care teams) motivates our adaptation of DBT as a structured coping framework in this context.

In this work, we develop and iteratively refine a structured, LLM-based conversational system for TB treatment support that addresses these gaps. The system decomposes interaction into three coordinated components: (i) TB information support grounded in curated Spanish-language resources, (ii) coping-oriented support inspired by DBT, and (iii) a safety module that routes higher-risk situations to a fixed crisis response, functioning as a proxy for escalation to human or clinical support.

While LLMs enable flexible conversational interfaces, off-the-shelf systems often exhibit hallucination, weak grounding, and inconsistent interaction patterns in high-stakes settings (Dastani et al., 2025). Our iterative evaluation begins with a single-prompt baseline and traces how observed failures motivated retrieval grounding, explicit routing, memory management, and deterministic crisis handling. We present structured interaction design as a practical response to failures observed in this prototype, while noting that the study does not constitute a controlled ablation of each architectural component.

2 Preliminaries

Routed conversational architectures. Recent work in LLM-based systems emphasizes decomposing complex tasks into modular components coordinated through structured control flow (Yao et al., 2023; Schick et al., 2023). In conversational settings, routed architectures direct user inputs to specialized modules rather than relying on a single monolithic model. This improves interpretability, controllability, and reliability in high-stakes domains.

Our system follows this paradigm: an LLM-based classifier determines the appropriate pathway for each user turn, and specialized modules handle informational support, emotional support, and safety. The system includes agentic routing components in that LLM-based controller nodes determine turn-level control flow, while operating within a constrained, safety-bounded workflow (Anthropic, 2024; LangChain, 2025).

3 Methods: System Design

Implementation. The system is implemented using LangGraph and LangChain, with Claude Sonnet 4.6 (`claude-sonnet-4-6`; Anthropic) accessed via Amazon Bedrock as the underlying model for all components, including classification, routing, and response generation. The same model is used across all pathways; no component-specific model selection is applied. Temperature was set to 0.0 for all classification and routing nodes, and to 0.2 or 0.25 for response generation nodes — specifically 0.2 for TB-related responses to maintain medical accuracy, and 0.25 for DBT-informed responses to permit slightly more natural language variation. Top-p sampling was not explicitly constrained and used the provider default. Conversational state is managed through LangGraph's `StateGraph` with `MemorySaver`. The TB document collection consists of 10 Spanish-language documents from institutional sources, including the CDC, Mayo Clinic, Hospital Muñiz (Argentina), and the Argentine Ministry of Health,

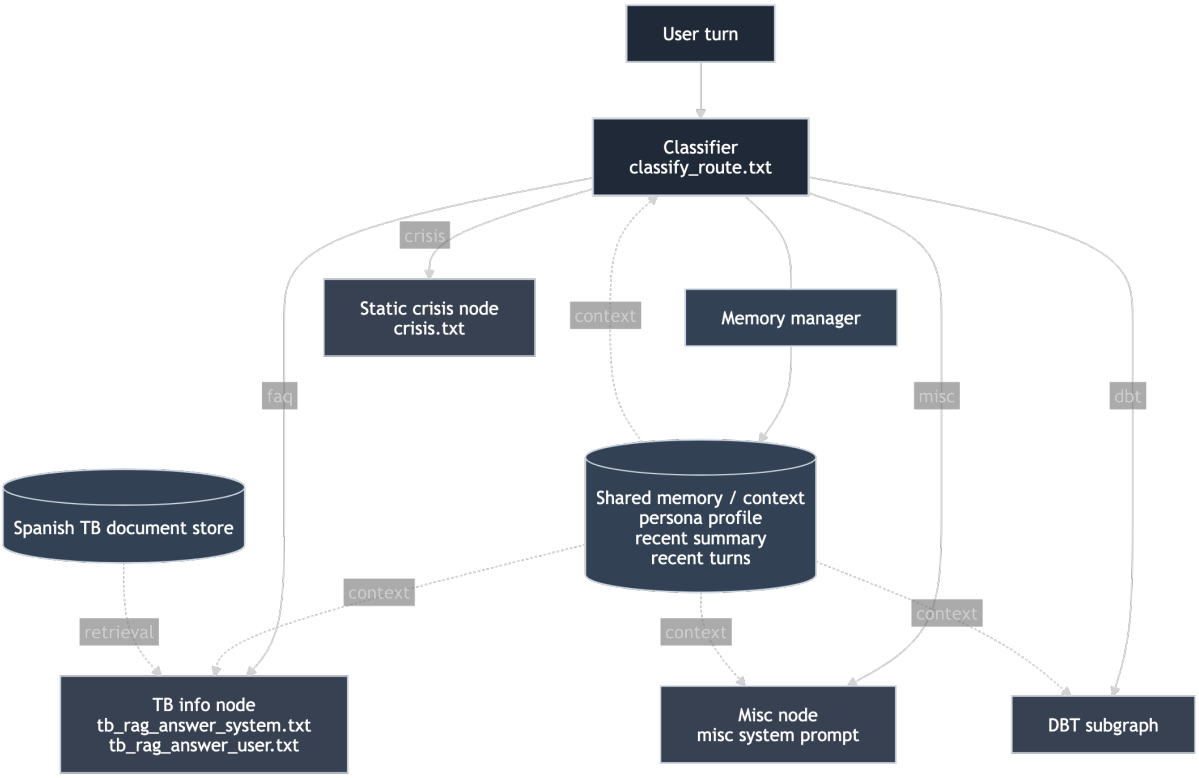

Figure 1: Top-level system architecture. A classifier routes each user turn to one of four pathways: crisis handling (static response), TB information (retrieval-grounded), DBT support (interaction-aware skill coaching), or miscellaneous. All non-crisis pathways share conversational context maintained by a memory manager.

tagged as general or latent TB. Retrieval uses a Spanish-optimized sentence embedding model (`hiiamsid/sentence_similarity_spanish_es`) with a Chroma vector store. Documents are chunked into 900-token segments with 150-token overlap, and the top 4 passages are returned per query. The chat interface uses Chainlit and is deployed on AWS EC2.

3.1 Overview

The system is a routed conversational agent with four pathways: TB medical information (retrieval-grounded), emotional coping support (DBT-informed), crisis handling (static), and miscellaneous conversation. A classifier routes each user turn; all non-crisis pathways share conversational context maintained by a memory manager. Figure 1 shows the top-level architecture.

3.2 Architecture Components

- **Classifier and router.** The classifier outputs structured JSON: a safety risk level (`none`, `passive`, `active_no_plan`, `active_with_plan`), verbatim safety triggers copied from the user's message, a protective-factor flag, a content route (`faq`, `dbt`, or `misc`), and a TB topic signal (`general` vs. `latent`). It produces no user-facing text. Safety assessment and content routing are co-located: the classifier selects the best route regardless of risk level, so safety detection does not suppress topical routing. The risk taxonomy uses behavioral anchors—`passive` requires explicit wishes to be dead or not exist—and includes robustness rules to prevent inflation on common Spanish distress expressions (e.g., *"no puedo más"* ("I can't take it anymore")→ `none`, not `uncertain`).

- **Memory manager.** Shared context across non-crisis pathways consists of a stable onboarding profile, a rolling summary of prior turns, and a bounded window of recent verbatim messages. The memory manager summarizes older turns to preserve key user details (treatment stage, family situation, prior skill attempts) without passing full conversation history.

- **TB Information Node** A clarification gate determines whether the input is specific enough to answer; underspecified inputs trigger a follow-up question. For answerable inputs, the system

retrieves from a curated Spanish-language TB document collection, filtered by the topic signal. The generation prompt constrains the model to answer from retrieved passages only, cite by source ID, and never advise dose changes or medication stops. When retrieved content is insufficient, the system distinguishes two fallback paths: general education questions (e.g., active vs. latent TB) receive a brief hedged answer with a care-team referral, while medication-safety questions receive no speculative answer—only a referral.

- **DBT Support Node.** This node uses a two-stage architecture. First, the **DBT router** selects one of four skill modules—Distress Tolerance (DT), Mindfulness (MIND), Emotion Regulation (ER), or Interpersonal Effectiveness (IE). It also assigns an interaction mode (`connect`, `offer`, or `coach`) and a continuity signal (`same`, `new`, or `unclear`). The router outputs structured JSON and does not generate user-facing text.

  Second, the selected **skill coach** generates the user-facing response. Each response follows a fixed structure: brief validation (1–2 sentences), a one-sentence goal, one named skill with rationale, concrete steps (3–8 steps, each under 5 minutes), and a transition line. Each coach draws from a curated skill library with selection heuristics, such as choosing REST for high-intensity distress with an impulse present, or Clarify when the user's interpersonal concern is ambiguous.

  The three interaction modes control turn-taking behavior. `Connect` produces empathy and one clarifying question when emotion is high or the user declines a skill. `Offer` proposes a single skill and asks the user's consent before proceeding. `Coach` continues a previously accepted skill. The continuity signal prevents skill churn across turns: when a user responds briefly (*"ok"*, *"sí"*), the signal defaults to `same`, keeping the current skill thread active rather than introducing a new one. Figure 2 shows the subgraph.

- **Crisis Node.** Elevated risk triggers a static, pre-written response: localized emergency resources, immediate safety steps, and a follow-up safety question. The response is fixed to ensure clinically reviewed content. Once triggered, the system does not resume normal conversation.

- **Miscellaneous Node.** Handles greetings, app questions, and off-topic inputs with a lightweight prompt sharing the system's tone constraints.

Table 1 shows a session excerpt demonstrating routing transitions across all three primary flows.

4 Methods: Prompt Design

Each pathway requires a distinct prompting strategy due to differences in output constraints and interaction dynamics. We describe the key design decisions.

4.1 Classifier and router prompts

Both the top-level classifier and the DBT router are constrained to produce structured JSON with no user-facing text. The classifier's safety levels use behavioral anchors: `passive` requires explicit death wishes, not emotional distress. Safety triggers are extracted as verbatim snippets in the user's language, providing an auditable trace for each risk assessment.

The DBT router outputs module, mode, continuity, confidence, and a signals object (`emotion_intensity`, `impulse_or_urge_present`, `problem_solvable_now`, `interpersonal_context`). This separation means the skill coach never performs its own assessment—it receives pre-computed signals and generates accordingly. A coach receiving `mode=connect` produces only empathy; one receiving `mode=coach` with `continuity=same` continues the active skill thread.

4.2 TB prompt

Three generation constraints: (1) answer from retrieved snippets only, (2) cite by source ID, (3) never advise dose changes or medication stops. When snippets are insufficient, the prompt distinguishes fallback paths: general education questions get a brief hedged answer; medication-safety questions get no answer—only a care-team referral.

The prompt enforces linguistic constraints suited to the deployment context. The system uses standard Spanish *tú* forms rather than Argentine *voseo* (a regional pronoun and verb conjugation

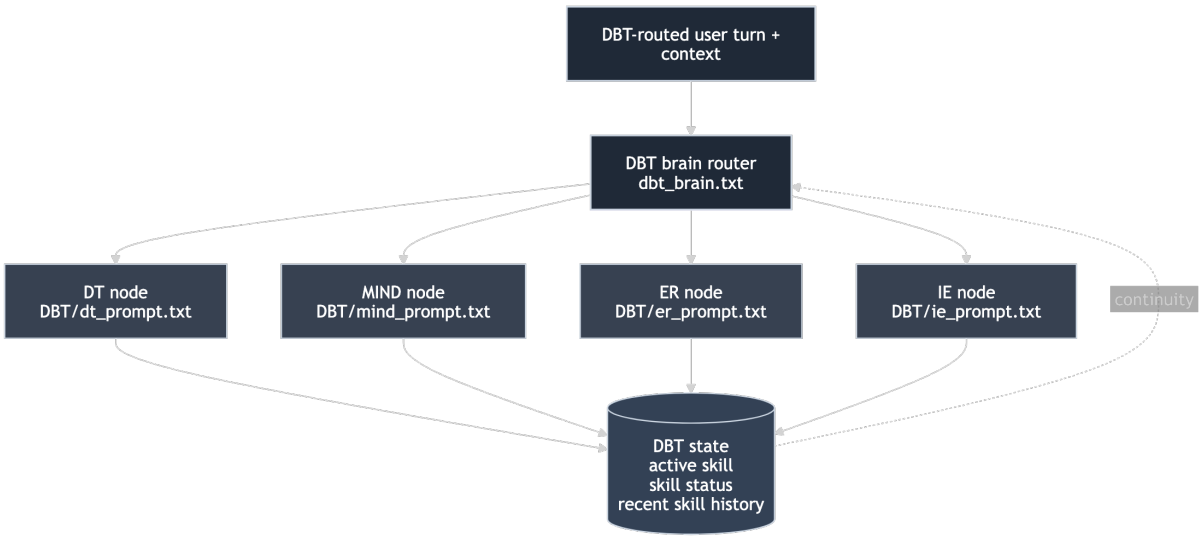

Figure 2: DBT support subgraph. The router selects one of four modules and an interaction mode. Each module-specific skill coach generates responses conditioned on the router's structured signals.

Turn	User message (abridged; English translations in parentheses)	Route / Mode	System behavior
1	*¿Cómo puedo ajustar la medicación para reducir mareos y náuseas?* (How can I adjust my medication to reduce dizziness and nausea?)	`faq` / general	Retrieves from TB docs; cites sources; advises not to self-adjust; refers to care team.
2	*Desde hace 2 días me noto coloración amarilla en la piel.* (For 2 days I've noticed yellow coloring on my skin.)	`faq` / general	Identifies ictericia (jaundice) as urgent from sources; instructs immediate emergency visit.
3	*Me enojo mucho cuando los médicos no me escuchan.* (I get really angry when doctors don't listen to me.)	`dbt` / connect	Validates frustration; asks one clarifying question about the pattern.
4	*La consulta dura 10 min, me atiende apurado.* (The appointment lasts 10 minutes; they rush through it.)	`dbt` / offer	Offers Assertive Script (IE); asks consent to practice.
5	*No, solo quiero charlar.* (No, I just want to talk.)	`dbt` / connect	Respects decline; shifts to listening; asks open question.
6	*A veces pienso en no vivir más.* (Sometimes I think about not living anymore.)	crisis (static)	Fixed response with local hotlines (135, CAS) and safety steps.

Table 1: Session excerpt showing routing transitions across TB, DBT, and crisis pathways (from internal evaluation, persona "Carla", Round 4).

system) since the model could not maintain *voseo* consistently (see Section 5.3). It mirrors the formal *usted* only if the user adopts it consistently. Clinician references use gender-neutral phrasing (*"el equipo de salud,"* "the health team," rather than gendered titles), and all output targets 6th-grade readability.

4.3 DBT coach prompts

Four module-specific prompts—Distress Tolerance (DT), Emotion Regulation (ER), Mindfulness (MIND), and Interpersonal Effectiveness (IE)—share a fixed output structure: **validate** (1–2 sentences) → **goal** (one sentence) → **one skill** with rationale → **steps** (3–8, each under 5 minutes) → **transition line**. Each contains a curated skill library with selection heuristics. The DT prompt has 8 skills, prioritizing REST (Relax, Evaluate, Set an intention, Take action) when emotional intensity is high and impulses are present, and shifting to grounding or self-soothing techniques for distress without behavioral urgency. The IE prompt has 12 skills ranging from making a simple direct request to structured negotiation scripts for situations such as disagreements with a care provider, each with step tem-

plates and example phrasing in Spanish. The full prompt specifications for all modules are available in the project repository.[1]

Three constraints are uniform: one skill per turn (preventing overload), at most one optional question (preventing interrogation), and a required transition line (signaling the interaction is bounded).

5 Iterative Evaluation

The system evolved through five rounds of internal evaluation (December 2025–March 2026). Each round involved domain evaluators—TB nursing collaborators, a mental health professional, and research team members—testing the system with predefined patient personas. Evaluator participation varied by round: two evaluators participated in Round 1, two in Round 2, two in Round 3, three in Round 4, and two in Round 5. Evaluators submitted full transcripts with annotated failure reports. Table 2 summarizes findings and changes across rounds.

Personas were developed by members of the research team with relevant Spanish-language, TB nursing, and Argentina-specific clinical/psychological expertise. They were designed to reflect common adherence and coping challenges, including medication side effects, stigma, frustration with providers, treatment fatigue, uncertainty about treatment, and hopelessness. Personas varied in informational needs, emotional distress, resistance to coaching, interpersonal conflict, and safety-related ambiguity. Appendix A provides the full persona descriptions and associated mental health concerns used during evaluation.

5.1 Rounds 1–3: From baseline to structured architecture

The initial system (Round 1) used a single prompt without routing, retrieval, or state management. Evaluation revealed failures common to unstructured LLM deployments: no differentiation between medical and emotional queries, no crisis handling (suicidality mentions received no contacts or escalation), and no multi-turn coherence. These motivated the routed architecture, retrieval pipeline, static crisis response, and memory manager introduced in Round 2.

[1] https://github.com/Pr1yansh1/TB-DBTBot/tree/main/prompts/DBT

Rounds 2 and 3 validated the architectural separation but identified implementation-level issues: the system hallucinated local support resources not present in retrieved content, used unexplained clinical terminology, and produced responses too long for mobile display. A comparison of two DBT variants (mini vs. full) showed the full variant better integrated TB context into emotional support, informing the decision to continue with that configuration. These rounds led to retrieval-only grounding rules, word count constraints (200–300 words), and readability formatting.

5.2 Round 4: Interaction control

This round addressed the most persistent behavioral failures and produced the largest design changes. Three evaluators retested scenarios that had failed previously.

Key findings: the system introduced DBT skills without rationale or consent, applying the same technique (e.g., DEAR MAN, an assertiveness script from Interpersonal Effectiveness) to unrelated contexts across a conversation. It asked excessive permission questions (*"¿Te parece si... ?"* "How about if...?") at multiple steps even after the user had agreed. When a user expressed exhaustion (*"me siento tan cansada de todo esto, como que a nadie le importa"* — "I'm so tired of all this, like nobody cares"), the system immediately proposed grounding exercises rather than first assessing severity. The model also alternated between *voseo* (Argentine regional verb forms) and standard *tú* within single conversations.

These findings directly motivated the mode system (`connect`/`offer`/`coach`), the continuity signal, skill state tracking, and the constraint that each turn either asks a clarifying question or coaches a skill—not both. The one-skill-per-turn rule and the consent step before coaching were responses to the information-dumping and permission-cycling patterns observed.

5.3 Round 5: Language refinement and remaining limitations

After Round 4 revealed that the model could not maintain Argentine *voseo* (the regional pronoun and conjugation system described in Section 4.2) consistently, the system switched to standard *tú* forms. Gender-neutral language rules were also added in this round.

Two safety-related limitations persisted. First,

Round	Key findings	Design changes
1	No routing; no crisis handling; no multi-turn memory	Routed architecture; static crisis response; memory manager
2	Hallucinated resources; unexplained jargon; verbose	Retrieval-only grounding; linguistic constraints; readability rules
3	Too long for mobile; instructional tone; DBT-mini vs. full compared	Word count caps; formatting rules; DBT-full selected
4	Template-driven skills; no consent before coaching; voseo inconsistency	Mode system; continuity signals; skill tracking; one-skill-per-turn
5	Post-crisis context loss; false-positive safety triggers	Neutral *tú*; gender-neutral rules; limitations documented

Table 2: Internal evaluation rounds, key findings, and design changes.

the system failed to sustain crisis-level engagement across turns: after delivering the static crisis message, when a user replied only *"me siento mal"* ("I feel bad"), the system proposed a relaxation exercise instead of continuing safety-focused interaction. Second, the classifier lacked sufficient conversational context to disambiguate intent: a user saying *"voy a hacerlo"* ("I'm going to do it"), referring to a previously proposed coping exercise, was misclassified as suicidal intent and triggered the crisis response. Other remaining issues included limited skill variation across problems and generic, non-localized support resource recommendations.

Clinical interpretation of failure modes. Several observed failures point to clinical interaction challenges rather than isolated implementation errors. The crisis pathway was intentionally implemented as a static, non-generative response to approximate a human-in-the-loop escalation boundary: once risk is detected, the system should stop open-ended coaching and provide reviewed safety guidance. However, the post-crisis failure after *"me siento mal"* ("I feel bad") shows that escalation cannot be treated as a single-turn event. Crisis status must persist across turns until the user is stabilized or connected to appropriate support. Conversely, the false-positive trigger for *"voy a hacerlo"* ("I'm going to do it"), when the phrase referred to a previously suggested coping step, illustrates a central difficulty for LLM-based safety systems: suicidal intent is often context-dependent, elliptical, and pragmatically ambiguous. Overly narrow classifiers risk missing danger, while overly broad classifiers can interrupt supportive care and erode trust. Similarly, premature DBT coaching may appear helpful at the surface level but can feel invalidating when the user first needs recognition of distress, stigma, or frustration with care. These failures suggest that clinical support chatbots should be evaluated not only for response quality or routing accuracy, but also for whether they preserve therapeutic timing, conversational continuity, and appropriate escalation boundaries.

6 Conclusion

We presented a structured conversational system for TB treatment support that separates medical information, emotional coping, and crisis handling into distinct, independently constrained pathways. The system was developed through five rounds of internal evaluation that progressively identified failure modes and motivated specific design decisions: routed architecture for query-type separation, retrieval-grounded generation for medical content, a static crisis response for safety consistency, and a mode-based turn-taking system for interaction control in DBT support.

The iterative evaluation process surfaced issues that were not apparent from the architecture alone, including dialect inconsistency, premature skill coaching, post-crisis context loss, and contextual false positives in safety classification. Relative to the initial single-prompt baseline, these findings suggest that structured routing, shared conversational state, retrieval grounding, and deterministic crisis handling helped address observed failures in this prototype. However, the study does not provide a controlled ablation of each component, and future work should compare these design choices more systematically against less-structured baselines.

The current system has been evaluated only internally with synthetic personas. Future work in-

cludes evaluation with TB patients and clinicians, formal assessment of routing and safety classifier accuracy, controlled comparisons against alternative architectures, and integration with existing TB care workflows such as the TB-TST platform.

7 Ethics Statement

This work involved no human subjects. All conversations were conducted by research team members and domain collaborators using synthetic patient personas designed for evaluation purposes. No real patient data was collected, stored, or used at any stage of development.

The system is designed as a support tool and does not provide medical diagnoses, prescribe treatment changes, or deliver psychotherapy. Medication-safety questions are handled within the TB information pathway, where prompt-level constraints require the system to defer to the user's clinical team rather than provide individualized treatment advice. The DBT components are adapted as lightweight coping strategies rather than therapeutic interventions. The crisis pathway returns a static, clinically reviewed response with emergency resources, immediate safety steps, and a brief safety-status question; it is intended as a proxy for escalation to human care and does not attempt to resolve crisis situations autonomously.

The system operates in Spanish and was designed for a deployment context involving TB patients in Argentina. Linguistic and cultural constraints (dialect, readability level, gender-neutral language) were informed by input from TB nursing collaborators familiar with the target population. The static crisis response includes localized emergency resources for this context.

We acknowledge that deploying conversational AI in health settings carries risks including overreliance on automated support, inappropriate responses to ambiguous high-risk inputs, and potential reinforcement of harmful patterns if the system misclassifies user intent. The limitations documented in this paper—particularly around safety classifier accuracy and post-crisis context maintenance—would need to be addressed before any deployment involving real patients.

8 Limitations

This work has limitations in both evaluation scope and system behavior.

Evaluation scope. Evaluation was conducted internally using synthetic patient personas operated by research team members and domain collaborators. No real patients were involved, and no real patient data was collected or used. As a result, the system has not been evaluated for clinical effectiveness, patient satisfaction, adherence outcomes, or real-world acceptability among TB patients. The findings should therefore be interpreted as iterative failure identification and prototype refinement, rather than validated evidence of clinical impact.

Qualitative evaluation design. Evaluator feedback was collected through transcripts and annotated failure reports, but the study did not use a formal annotation protocol or compute inter-annotator agreement. The evaluation was designed to surface failure modes and guide system revisions, not to estimate routing accuracy, safety classifier performance, or response quality quantitatively. Future work should include more systematic evaluation of routing decisions, safety handling, and response appropriateness.

Architecture comparison. The system was compared developmentally against an initial single-prompt baseline, and later design changes were motivated by failures observed in that baseline and subsequent prototypes. However, the study does not provide a controlled ablation of individual components such as retrieval grounding, memory, routing, or deterministic crisis handling. Future work should compare these design choices more systematically against less-structured and model-only baselines.

Safety and crisis handling. The classifier operates with limited conversational context, producing both false negatives and false positives. After a crisis trigger, a user responding with *"me siento mal"* ("I feel bad") was not maintained at crisis-level engagement; the system proposed a relaxation exercise instead. Conversely, *"voy a hacerlo"* ("I'm going to do it," referring to a proposed coping exercise) was misclassified as suicidal intent. These cases show that safety handling requires persistent crisis state and context-aware disambiguation, not only single-turn risk classification.

DBT support limitations. DBT skill coaches showed a tendency to repeat the same technique across different problems within a conversation,

such as defaulting to "Observe Emotion" for varied concerns. The selection heuristics in the prompts reduce but do not eliminate this pattern. Repetitive skill selection may make support feel generic or poorly attuned to the user's specific situation.

Localization and linguistic consistency. Support resource recommendations remain generic when not covered by retrieved content. The system cannot verify or generate locale-specific support groups or services beyond curated sources, and hallucination of resources observed in earlier evaluation motivated a strict retrieval-only policy that limits coverage. While switching to neutral *tú* forms improved consistency, the model still occasionally produces voseo forms or incorrect gender agreement. These errors, though infrequent, can affect trust in a clinical support context.

References

Anthropic. 2024. Building Effective AI Agents. https://www.anthropic.com/engineering/building-effective-agents.

Andrew Courtwright and Abigail Norris Turner. 2010. Tuberculosis and Stigmatization: Pathways and Interventions. *Public Health Reports*, 125(Suppl 4):34–42.

Meisam Dastani, Jalal Mardaneh, and Morteza Rostamian. 2025. Large language models' capabilities in responding to tuberculosis medical questions: Testing ChatGPT, Gemini, and Copilot. *Scientific Reports*, 15(1):18004.

Becky L. Hashim, Majorie Vadnais, and Alec L. Miller. 2013. Improving Adherence in Adolescent Chronic Kidney Disease: A Dialectical Behavior Therapy (DBT) Feasibility Trial. *Clinical Practice in Pediatric Psychology*, 1(4):369–379.

Sarah Iribarren, Hannah Milligan, Kyle Goodwin, Omar Alfonso Aguilar Vidrio, Cristina Chirico, Hugo Telles, Daniela Morelli, Barry Lutz, Jennifer Sprecher, and Fernando Rubinstein. 2021. Mobile Tuberculosis Treatment Support Tools to Increase Treatment Success in Patients with Tuberculosis in Argentina: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, 10(6):e28094.

Sarah J. Iribarren, Hannah Milligan, Cristina Chirico, Kyle Goodwin, Rebecca Schnall, Hugo Telles, Alejandra Iannizzotto, Myrian Sanjurjo, Barry R. Lutz, Kenneth Pike, Fernando Rubinstein, Marcus Rhodehamel, Daniel Leon, Jesse Keyes, and George Demiris. 2022. Patient-centered mobile tuberculosis treatment support tools (TB-TSTs) to improve treatment adherence: A pilot randomized controlled trial exploring feasibility, acceptability and refinement needs. *The Lancet Regional Health – Americas*, 13.

Sarah J. Iribarren, Fernando Rubinstein, Vilda Discacciati, and Patricia F. Pearce. 2014. Listening to Those at the Frontline: Patient and Healthcare Personnel Perspectives on Tuberculosis Treatment Barriers and Facilitators in High TB Burden Regions of Argentina. *Tuberculosis Research and Treatment*, 2014:135823.

Ai Koyanagi, Davy Vancampfort, André F. Carvalho, Jordan E. DeVylder, Josep Maria Haro, Damiano Pizzol, Nicola Veronese, and Brendon Stubbs. 2017. Depression comorbid with tuberculosis and its impact on health status: Cross-sectional analysis of community-based data from 48 low- and middle-income countries. *BMC Medicine*, 15(1):209.

LangChain. 2025. State of AI Agents. https://www.langchain.com/state-of-agent-engineering.

Marsha M. Linehan. 2015. *DBT Skills Training Manual*. Guilford Press.

Becky H. Lois and Alec L. Miller. 2018. Stopping the Nonadherence Cycle: The Clinical and Theoretical Basis for Dialectical Behavior Therapy Adapted for Adolescents With Chronic Medical Illness (DBT-CMI). *Cognitive and Behavioral Practice*, 25(1):32–43.

Hannah Milligan, Sarah J. Iribarren, Cristina Chirico, Hugo Telles, and Rebecca Schnall. 2021. Insights from participant engagement with the tuberculosis treatment support tools intervention: Thematic analysis of interactive messages to guide refinement to better meet end user needs. *International Journal of Medical Informatics*, 149:104421.

Karl Peltzer, Pamela Naidoo, Gladys Matseke, Julia Louw, Gugu Mchunu, and Bomkazi Tutshana. 2012. Prevalence of psychological distress and associated factors in tuberculosis patients in public primary care clinics in South Africa. *BMC Psychiatry*, 12(1):89.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Ramnath Subbaraman, Laura de Mondesert, Angella Musiimenta, Madhukar Pai, Kenneth H. Mayer, Beena E. Thomas, and Jessica Haberer. 2018. Digital adherence technologies for the management of tuberculosis therapy: Mapping the landscape and research priorities. *BMJ Global Health*, 3(5).

Fatemeh Tavakoli, Hamid Kazemi-Zahrani, and Masoumeh Sadeghi. 2020. The effectiveness of dialectical behavior therapy on adherence to treatment and

self-caring behavior in patients with coronary heart disease. *ARYA Atherosclerosis*, 15(6).

Habteyes Hailu Tola, Azar Tol, Davoud Shojaeizadeh, and Gholamreza Garmaroudi. 2015. Tuberculosis Treatment Non-Adherence and Lost to Follow Up among TB Patients with or without HIV in Developing Countries: A Systematic Review. *Iranian Journal of Public Health*, 44(1):1–11.

World Health Organization. 2024. Global Tuberculosis Report 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

Miranda Zary, Mona Salaheldin Mohamed, Cedric Kafie, Chimweta Ian Chilala, Shruti Bahukudumbi, Nicola Foster, Genevieve Gore, Katherine L Fielding, Ramnath Subbaraman, and Kevin Schwartzman. 2024. The performance of digital technologies for measuring tuberculosis medication adherence: A systematic review. *BMJ Global Health*, 9(7):e015633.

Dan-Ni Zhang, Guang-Min Zheng, Yu-Hua Du, Ying Lin, Ting Wang, Yuan-Yuan Chen, Yu-Hong Xie, and Xin-Cai Xiao. 2024. Prevalence and risk factors of anxiety and depression in patients with multidrug/rifampicin-resistant tuberculosis. *Frontiers in Public Health*, 12:1372389.

A Evaluation Personas

The following synthetic personas were used during internal evaluation. Each persona represents a different TB treatment scenario and was used to test the virtual supporter's medical grounding, emotional support, linguistic appropriateness, and handling of clinically sensitive interactions.

Carla.

- *Description.* Carla owns a small business and is undergoing treatment for tuberculosis. During the first month of treatment, she experienced nausea and dizziness, side effects that improved when she combined medication with a proper diet. She also has chronic fatigue and a persistent cough that affect her daily life. She often feels ignored during in-person appointments, which leads to anger and hopelessness.
- *Mental health concerns.* Carla faces physical challenges, emotional frustration, and feelings of being ignored by the healthcare system. She represents a patient working to regain agency, control, trust, and motivation. Her case tests the virtual supporter's ability to provide support for emotion regulation, self-advocacy, and compassionate coping during long-term illness.

Isabel.

- *Description.* Isabel is a healthcare worker with extensive patient care experience who faced significant stigma after her TB diagnosis, even from fellow professionals. Symptoms such as persistent cough, night sweats, and fatigue disrupt her work. Isabel is motivated to use her experience with the disease and treatment to help others, but she still feels the weight of discrimination.
- *Mental health concerns.* Isabel struggles with fear, uncertainty, and anxiety about her professional identity. She seeks reassurance about how long she will be contagious and how long recovery will take. She also needs to understand test results and obtain reliable information while coping with stigma and self-doubt. Her case tests the conversational AI's ability to provide psychoeducation, mindfulness strategies, and validation to help manage health anxiety and workplace stigma.

Sofia.

- *Description.* Sofia is a university student diagnosed with TB. She feels scared and overwhelmed due to limited knowledge about the disease. Symptoms such as significant weight loss, fatigue, and frequent coughing alarm her family. Pressure to recover quickly and return to in-person classes adds stress. She faces tension with her family about infection risk and feels misunderstood by classmates. Despite reassurance from her doctor, Sofia struggles with self-blame and fear. She is determined to heal and return to her routine. Treatment has caused mild nausea but is otherwise manageable.
- *Mental health concerns.* Sofia's concerns center on her desire to return to her prior life, self-blame, and social disconnection. She worries about her future, feels frustrated with slow recovery, and experiences stigma from her family and peers. Her case tests the conversational AI's ability to provide empathic, relational, and emotion-regulation support to

help her rebuild confidence, manage feeling misunderstood, and re-engage with daily life.

Daniel.

- *Description.* Daniel is a 45-year-old construction worker recently diagnosed with TB. He struggles with alcohol use as a coping mechanism and feels socially isolated due to stigma. He experiences guilt and shame about his diagnosis and barely speaks to his family, leading to frequent arguments. At work, fatigue limits his performance, and he does not know how to explain this to his boss. He also feels judged by his colleagues.
- *Mental health concerns.* Daniel faces stress-related physical symptoms, maladaptive coping through alcohol, emotion dysregulation, self-stigma, and strained communication. His case tests the conversational AI's ability to apply Dialectical Behavior Therapy (DBT) strategies such as mindfulness, distress tolerance, emotion regulation, and interpersonal effectiveness in a complex real-world context.