



UNIVERSITY OF CALGARY

ENGG 519: Special Topics in Engineering

Maham Jamal (30153574)
Chloe Villaranda (30097691)
Khadiza Binte Ahsan (30150381)

Final Report

Title: Designing a Conversational Agent to Enhance
Emotional Awareness in Individuals with Alexithymia

Table of Contents

Abstract	2
Keywords	2
1. Background Research	3
1.1 Scope and Intended Use	4
Materials and Methodology	4
2. Methodology	4
2.1 Conceptual Foundations: Integration of ENGG 519 Course Concepts	6
3. Modeling and Prototyping	7
3.1 Design	7
3.2 Final Prototype	9
3.3 Nielsen's 10 Heuristics	15
4. Modified Trust-Reliance Model for Alexithymia	16
4.1 Shift From Affective Trust to Cognitive-Behavioral Reliance Readiness	17
4.2 Modified Pre-Interaction Factors	18
4.3 Modified Situational and Learned Components	18
4.4 Functional Reliance Over Emotional Reliance	18
4.5 Role of Design Features	19
4.6 Situational Factors Not Related to Trust	19
4.7 Link to the LLM Implementation Plan	19
5. Final System Implementation	20
5.1 System Architecture	20
5.2 Prompting Framework	21
5.3 Trust Factor Model	21
5.4 Trust Responsive Behavior	22
5.5 Safety, Ethics, and Risk Controls	23
5.6 Summary of the Final System	23
Results	23
6. Testing	23
Stage 1: Override Testing Method	24
Stage 2: Trust Building	26
Stage 3: Simulated User Testing	28
Discussion	31
Distinct Adaptive Behaviours and Predictable Interaction Patterns	31
Natural Trust Evolution and Realistic Human-AI Dynamics	31
Usability, Clarity, and Cognitive Load Management	31
Safety, Boundaries, and Sociotechnical Considerations	32
Integrated Interpretation	32
Conclusion	32
References	33

Abstract

We designed a conversational agent that supports emotional awareness in individuals with alexithymia, a trait marked by difficulty identifying and describing emotions. Research shows that alexithymia affects a significant portion of the population and is linked to challenges in emotional regulation, communication, and mental health treatment. These findings highlight the need for structured, accessible tools that help users practice emotion labeling in a low-pressure environment.

Using a human-centered design approach, we have developed a rule-based prototype and a Figma interface that prioritise clarity, predictability, and supportive interaction. A key contribution of this project is the introduction of the Cognitive Behavioral Reliance Readiness model, which replaces traditional affect-based trust with a behavioral measure of user comfort and engagement. This model informs the planned LLM architecture, prompting framework, and trust-based adaptations that will guide future development.

Keywords

Word	Definition
Alexithymia	A personality trait involving difficulty identifying, describing, or interpreting one's own emotions.
Emotional Awareness	The ability to notice, understand, and name emotional and physiological states.
Conversational Agents	AI-driven systems are designed to communicate with users through natural language.
Human-Centered AI	An AI design approach that prioritizes user needs, safety, clarity, and accessibility.
Trust in Automation	The degree to which users rely on and feel confident in automated systems.
Cognitive-Behavioral Reliance Readiness	A behavioral model of trust that measures user comfort through observable actions rather than emotions.
Affective Computing	Technology that interprets or responds to human emotions and affective states.
Usability Testing	Evaluation of how effectively and comfortably users can interact with a system.
Safety-Critical AI	AI systems are designed with strict safeguards to

	prevent harm in sensitive contexts.
Structured Dialogue Systems	Conversational systems that use guided, rule-based, or stepwise prompts to support user interactions.
Explainable AI	AI methods that make system behavior understandable and predictable to users.
Adaptive Interaction Design	Designing systems that adjust their behavior based on user responses, preferences, or needs.

Table 0: Keywords

Introduction

Alexithymia is a multifaceted personality trait characterized by difficulties in identifying, describing, and interpreting emotional states, along with a tendency to focus on external rather than internal experiences. First conceptualised by Nemiah and Sifneos in the early 1970s, it has since become a central construct in clinical psychology and psychiatry [1]. Individuals with alexithymia often struggle to recognise emotional cues, translate these cues into verbal language, and engage in reflective emotional processing. These challenges influence emotional regulation, interpersonal communication, and overall well-being. To understand the broader impact of alexithymia and the populations most affected, it is important to review current research on its prevalence, associated mental health outcomes, and developmental patterns. This background establishes the foundation for designing supportive digital tools tailored to the needs of alexithymic users.

1. Background Research

Alexithymia affects approximately 10% of the general population, though prevalence varies across cultures and clinical groups [2]. Elevated rates are consistently observed in individuals with depression, anxiety, substance use disorders, post-traumatic stress disorder, and autism spectrum conditions, suggesting that alexithymia operates as a transdiagnostic vulnerability factor influencing emotional regulation and symptom severity across multiple mental health conditions [3]. Rather than a categorical disorder, alexithymia is best understood as a dimensional trait with varying levels of severity.

Alexithymia appears across the lifespan. In children and adolescents, higher alexithymia scores are associated with emotion-regulation challenges, peer difficulties, and increased internalising symptoms [4]. Young adults often report stress sensitivity, interpersonal strain, and reduced help-seeking due to uncertainty about their emotional states. Among adults, alexithymia correlates with chronic stress, somatic symptom burden, depression, and lower engagement in insight-oriented psychotherapy [3]. Older adults may show elevated levels due to cognitive changes or age-related shifts in emotional processing [5]. Although some studies note slightly higher average scores in men, the trait affects individuals across genders and cultural backgrounds [2].

These characteristics pose challenges for traditional mental health interventions. Difficulty identifying and articulating emotions can reduce the effectiveness of therapies that rely on verbal emotional insight, such as cognitive-behavioural or psychodynamic approaches. Instead, individuals with alexithymia often benefit more from structured, concrete, and stepwise interventions that guide emotional awareness and communication [6]. This has contributed to growing interest in AI-mediated tools that can support emotion labeling, reflective prompts, and adaptive scaffolding while integrating safety mechanisms such as crisis-related keyword detection and context-aware boundaries.

Trust plays a critical role in the usability and effectiveness of AI systems in sensitive domains. For emotionally focused tools, trust depends on predictable system behaviour, clear communication, accurate interpretation of user input, and transparency in how responses are generated. Automation trust research shows that users rely on a system only when its performance and behaviour align with their expectations [7], [8]. For individuals with alexithymia, who may already experience uncertainty in emotional communication, trust becomes even more important. Without sufficient trust, users may disengage before any reflective or learning benefits can occur.

1.1 Scope and Intended Use

Given the prevalence of alexithymia and its impact on emotional communication, this project focuses on designing a supportive conversational agent that helps users practice emotional recognition and reflection through guided, structured dialogue. The system does not function as a diagnostic tool, therapeutic intervention, or replacement for professional mental health care. Instead, it serves as a non-clinical wellness tool that encourages self-reflection, emotional articulation, and awareness of bodily and cognitive cues.

Although the chatbot draws on psychological concepts, it does not assess mental health conditions or deliver personalized therapeutic strategies. Any emotionally sensitive or distress-related content triggers supportive but non-clinical responses that redirect users toward appropriate professional resources rather than attempting intervention. These constraints ensure the system operates within safe boundaries, aligning with ethical standards for AI systems in emotionally sensitive contexts.

The following section outlines the human-centered design process, safety considerations, and iterative development methods used to implement and evaluate the system.

Materials and Methodology

2. Methodology

The goal of this project was to design a conversational chatbot that supports users in improving their ability to recognize and describe emotions, particularly individuals who experience alexithymia. The system provides a private and supportive environment where users can practice naming emotions, reflecting on experiences, and building awareness of emotional cues through short guided interactions. The chatbot is not intended to deliver therapy. Its purpose is to promote self-reflection and emotional learning within a non-clinical and user-friendly format.

The development process followed a human-centered design approach combined with safety-critical AI principles [3]. Early stages involved reviewing literature on empathy in conversational interfaces and identifying risks commonly reported in emotional support chatbots. Subsequent stages included prototyping interface elements, constructing preliminary dialogue flows, and evaluating design decisions based on clarity, predictability, and user comfort. The final design stage focused on refinement based on usability feedback and safety considerations. Throughout development, the chatbot's interaction style and communication boundaries were guided by human AI teaming concepts from ENGG 519, including calibrated trust, transparency, and ethical design practices [4].

The completed chatbot architecture contains five core components.

1. User Input Interface, which supports text and emoji-based communication.
2. Intent Recognition, which classifies user messages into categories such as emotional reflection, confusion, neutral statements, or mentions of physical or emotional discomfort.
3. Dialogue Management, which selects the next supportive prompt or question based on the intent and conversational context.
4. Safety Layer which detects crisis-related or sensitive content and restricts the system to predefined supportive responses.
5. Response Output, which delivers clear, empathetic, and context-appropriate feedback.

The system follows a structured interaction pipeline. User messages first pass through intent recognition and safety screening. The dialogue manager then selects an appropriate response, which is displayed after tone and safety checks. This pipeline ensures that all outputs respect safety constraints and maintain a predictable and supportive interaction style [2].

Safety considerations were integrated throughout the design process. If the system detects indications of distress or possible self-harm, it provides supportive statements such as, "It sounds like you may be going through something difficult. You are not alone. Please consider reaching out to a trusted person or a local helpline." When the system is uncertain about the meaning of a message, it returns a neutral and safe fallback response. All interactions are anonymous, and no personal data is stored. The interface also communicates clearly that the chatbot is not a substitute for professional mental health care.

Evaluation focused on usability, clarity, and conversational tone. Peer testers and instructors reviewed dialogue flows to assess whether responses felt supportive, non-judgmental, and easy to follow. Early testing also monitored points of disengagement, such as repeated clarifying questions or unclear transitions between prompts. Ethical review was conducted throughout the project to ensure informed consent, privacy protection, and appropriate handling of sensitive content [3].

2.1 Conceptual Foundations: Integration of ENGG 519 Course Concepts

The methodological choices in this project were directly shaped by concepts introduced throughout ENGG 519: Special Topics in Engineering. These concepts informed the system architecture, interaction style, safety constraints, and evaluation approach.

1. Human AI Teaming Principles (Lectures 1 and 2)

Lectures on human AI teaming emphasized the importance of calibrated reliance, predictable behavior, and clear communication in high-stakes or emotionally sensitive contexts. These principles guided the chatbot's dialogue style, which prioritizes literal and structured prompts that reduce ambiguity for users with alexithymia. The decision to avoid over-personalization and emotionally interpretive language was also based on the risks of miscalibrated trust highlighted in these lectures.

2. Mental Models and the Gulf of Evaluation (Lecture 2)

Lecture discussions about mental model mismatches motivated the use of highly transparent response patterns and consistent conversational structure. Users with alexithymia often have difficulty interpreting emotional nuance, so reducing the gulf of evaluation was essential. This led to the implementation of simple interface elements, predictable choice-based prompts, and clear explanations of the system's role and limitations.

3. Cognitive Load and Situation Awareness (Lectures 1 and 3)

Lecture material on cognitive load informed the design of low complexity text prompts, limited branching logic, and short message structures. The goal was to minimize extraneous cognitive load and support basic levels of situation awareness during emotional reflection. These lectures demonstrated how high cognitive load can impair decision-making, which is especially relevant for alexithymic users who experience internal uncertainty.

4. Trust Formation and Appropriate Reliance (Lectures 3 and 4)

Lectures on trust in automation and trust calibration shaped the development of the modified Cognitive Behavioral Reliance Readiness model. These lectures emphasized that trust must be aligned with system capability and user perception. This directly informed the methodological decision to model trust behaviorally rather than affectively, since individuals with alexithymia have limited access to emotional cues. These lectures also supported the idea that clear communication and consistent behavior are essential for appropriate reliance.

5. Safety Critical AI and Boundary Setting (Lecture 10)

Lecture discussions of safety failures in real-world deployed systems reinforced the need for strict boundary control and crisis diversion. As a result, the methodology included the construction of a safety layer, predefined crisis messages, refusal to provide therapeutic guidance, and explicit disclaimers regarding non-clinical intent. These decisions reflect safety principles taught in the lectures, including hazard identification, risk mitigation, and prevention of over-reliance.

6. Ethical and Sociotechnical Considerations (Lectures 9 and 10)

The ethical review process was shaped by lecture material on sociotechnical failures, user vulnerability, and accountability in emotionally sensitive systems. These lectures highlighted the

risks of anthropomorphism, harmful advice, and misinterpretation of model outputs. This directly influenced the decision to implement neutral fallback responses, avoid emotional inference, and restrict data logging to non-identifiable interaction metadata.

7. Human Centered Explainability (Lecture 7)

Lecture content on explainability influenced the emphasis on transparent system behavior. Although the chatbot does not generate technical explanations, it follows HCXAI principles through clear reasoning steps, explicit intentions behind questions, and avoidance of opaque or ambiguous probes. These lecture concepts justified the use of structured prompting over open-ended emotional exploration.

3. Modeling and Prototyping

3.1 Design

A low-fidelity prototype was created using Figma, outlining the conversation interface and flow of interaction. The visual design includes a minimalist chat layout with alternating message bubbles, neutral colors, and readable text. The layout contains three main screens: the primary chat window, an emotion-prompt page with three to five selectable options, and a reflection page where the user types short responses.

The following images showcase the prototype that was designed using Figma. Figma has an AI feature, and the chatbot's history, as well as interactive design, can be found here.

<https://www.figma.com/make/TsNDaNO33cVAITFASgbnas/Feelio-Chatbot-UI-Design?node-id=0-4&p=f&t=vFYIwX9viegLHJj3-0>

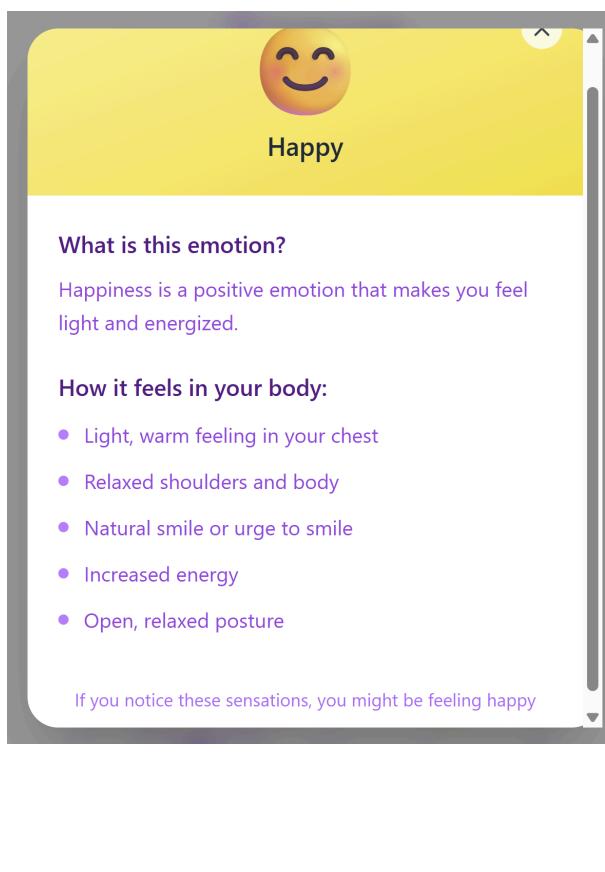
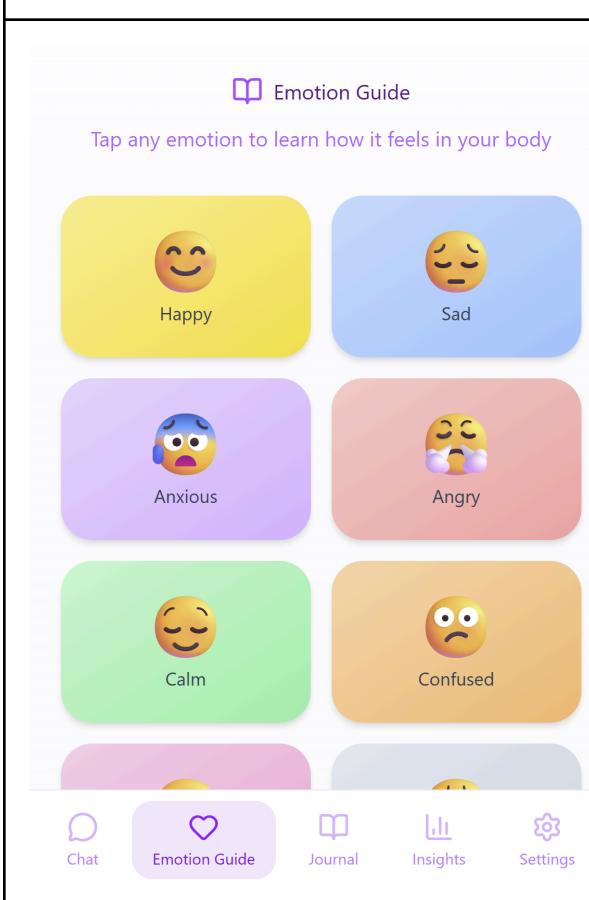
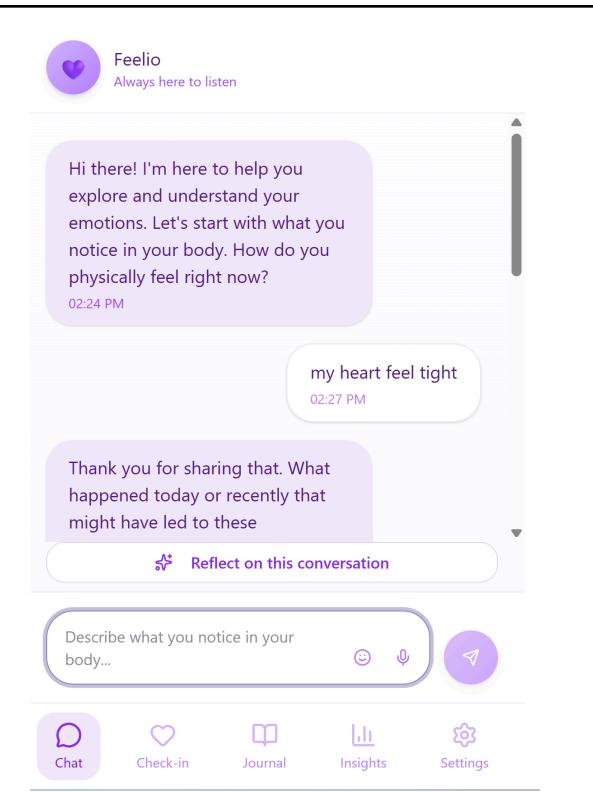
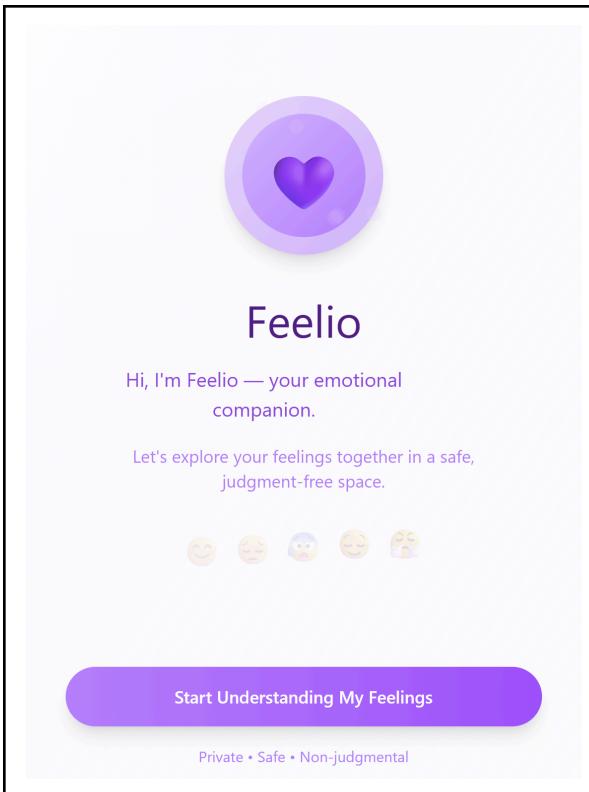




Figure 1: Figma Designs of the Chatbot Feelio

An initial Streamlit implementation simulates the logic using a rule-based structure. It allows users to type a message and receive predefined responses. The dialogue follows short emotional reflection cycles. For example:

User: "I guess I've been feeling kind of tense."

Chatbot: "Thanks, that's helpful. When you say tense, do you notice it more in your body—like your shoulders, jaw, or stomach—or more as a mental kind of tension, like your thoughts feel tight or pressured?"

User: "Maybe in my shoulders and neck."

Chatbot: "Got it. Physical tension can sometimes be a sign of stress or being on alert, even if we're not sure why. Has anything lately been taking up your time or energy—like responsibilities, school/work, social stuff, or even just too much noise or change?"

User: "Yeah, there's been a lot going on at school."

Chatbot: "That makes sense. When a lot is happening, even if you're handling it, your body can still carry the load. Would you like to talk a bit about what's been going on, or would you rather focus on ways to ease that tension first?"

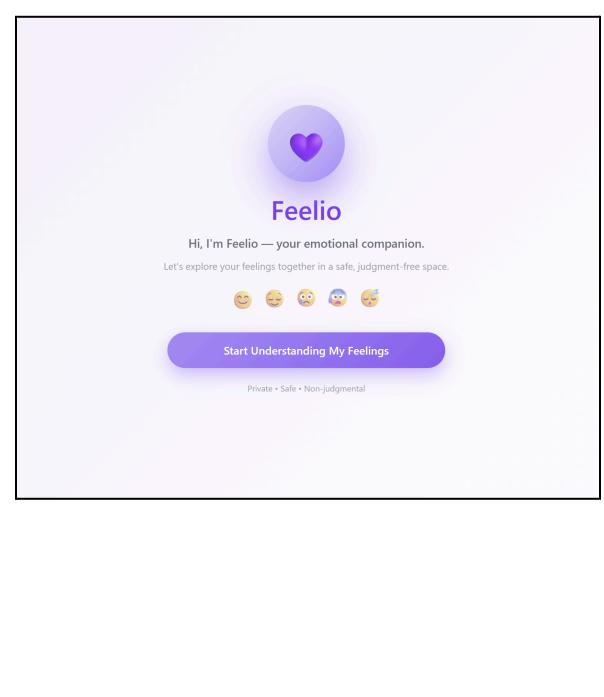
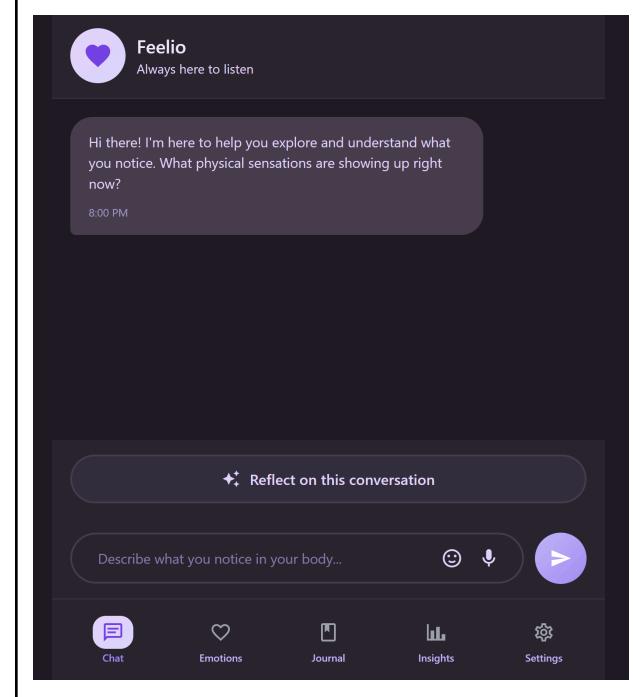
This early version demonstrates empathy and conversational pacing while still maintaining clear system boundaries. At present, the code for the front end of the application is complete and development for the LLM and chatbot integration is in progress.

3.2 Final Prototype

The prototype was developed using TypeScript for the front end, with Svelte as the primary JavaScript framework due to its lightweight structure and suitability for rapid prototyping. The

system integrates with a backend through custom API calls, and the chatbot functionality is powered by a GROQ-based LLM. The application was hosted on Vercel to enable seamless deployment and accessibility. Additionally, the interface was designed to be fully responsive, adapting to both desktop and mobile views. Throughout development, Nielsen's 10 Usability Heuristics guided the design process, ensuring a simple, intuitive, and user-friendly experience.

The images below showcase the final design of the Feelio application as hosted on Vercel.

	
Figure 2: Homepage	Figure 3: A Dark Mode has been implemented on all pages alongside the default theme.

Feelio
Always here to listen

Emotion Guide

Tap any emotion to learn how it feels in your body

Chat Emotions Journal Insights Settings

Happy

What is this emotion?
Happiness is a positive emotion characterized by feelings of joy, contentment, and satisfaction.

How it feels in your body:

- Lightness in chest or body
- Relaxed muscles
- Smiling or upturned mouth
- Increased energy
- Warm feeling throughout body

Chat Emotions Journal Insights Settings

Sad

What is this emotion?
Sadness is a natural response to loss, disappointment, or difficult situations.

How it feels in your body:

- Heavy feeling in chest or body
- Tight or lump in throat
- Low energy, feeling tired
- Tears or urge to cry
- Slumped posture, looking down

Chat Emotions Journal Insights Settings

Figure 4: The Emotion Guide page- Contains information about emotions and how they affect the human body physically.

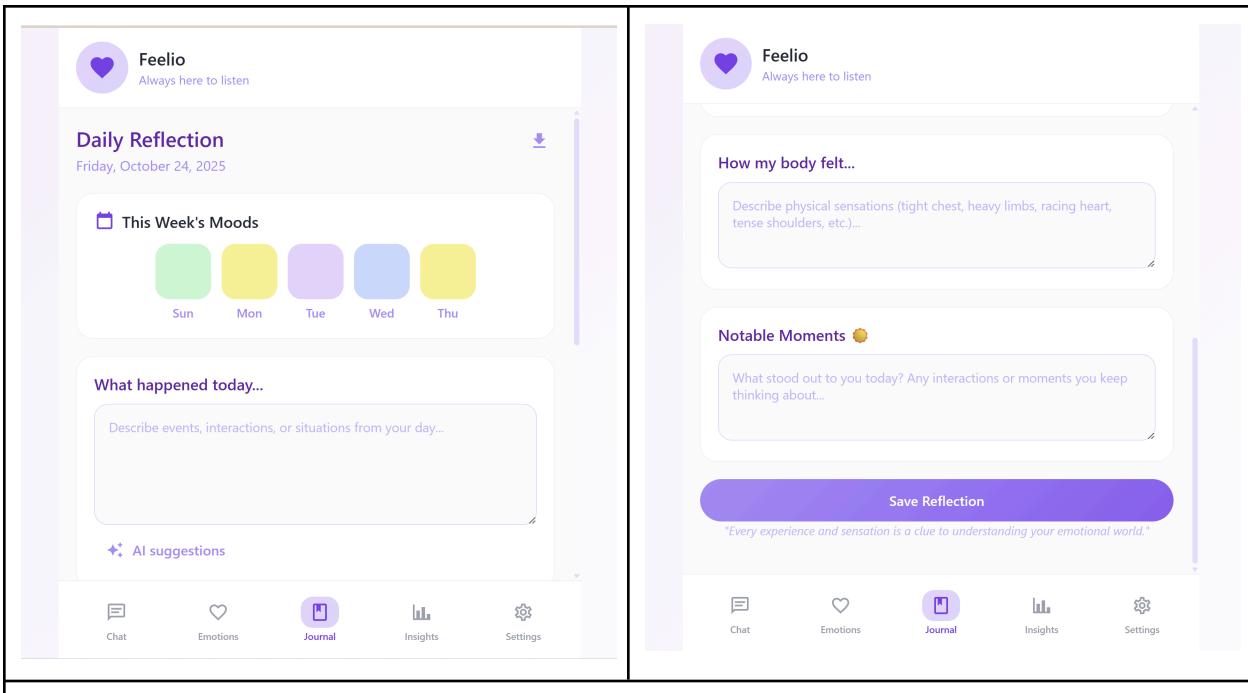
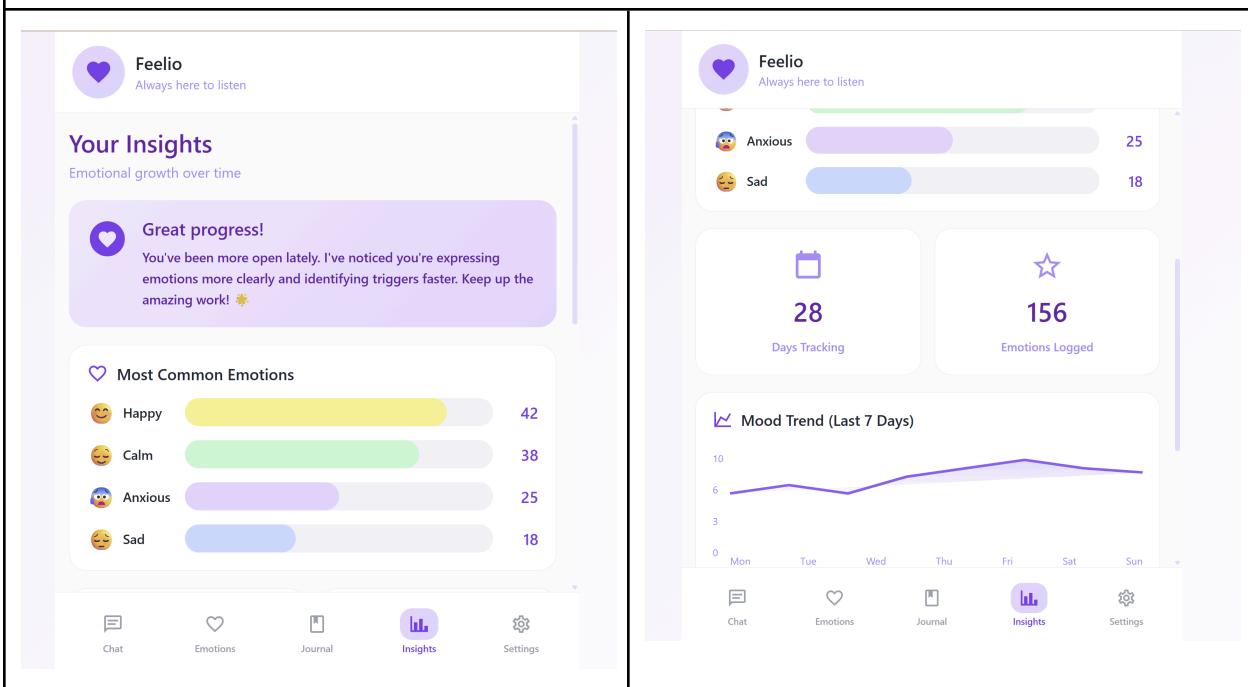


Figure 5: Daily Reflection page - To keep a record of weekly moods and a personal user journal.



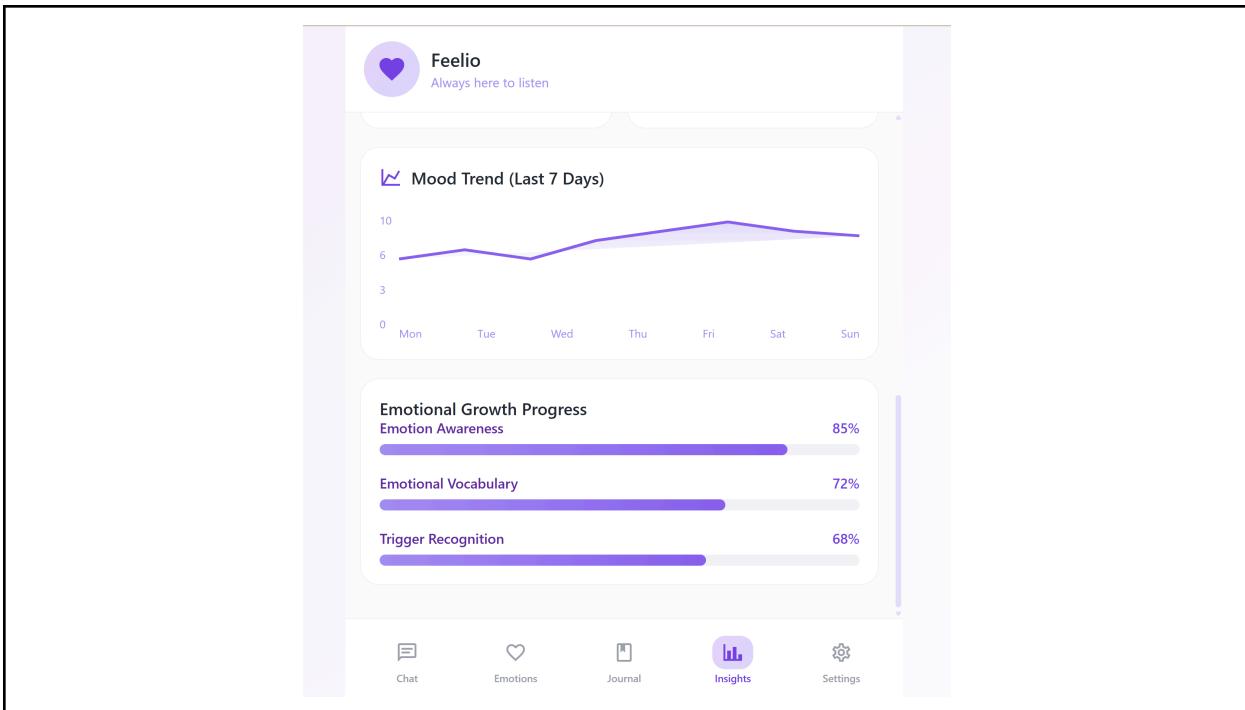
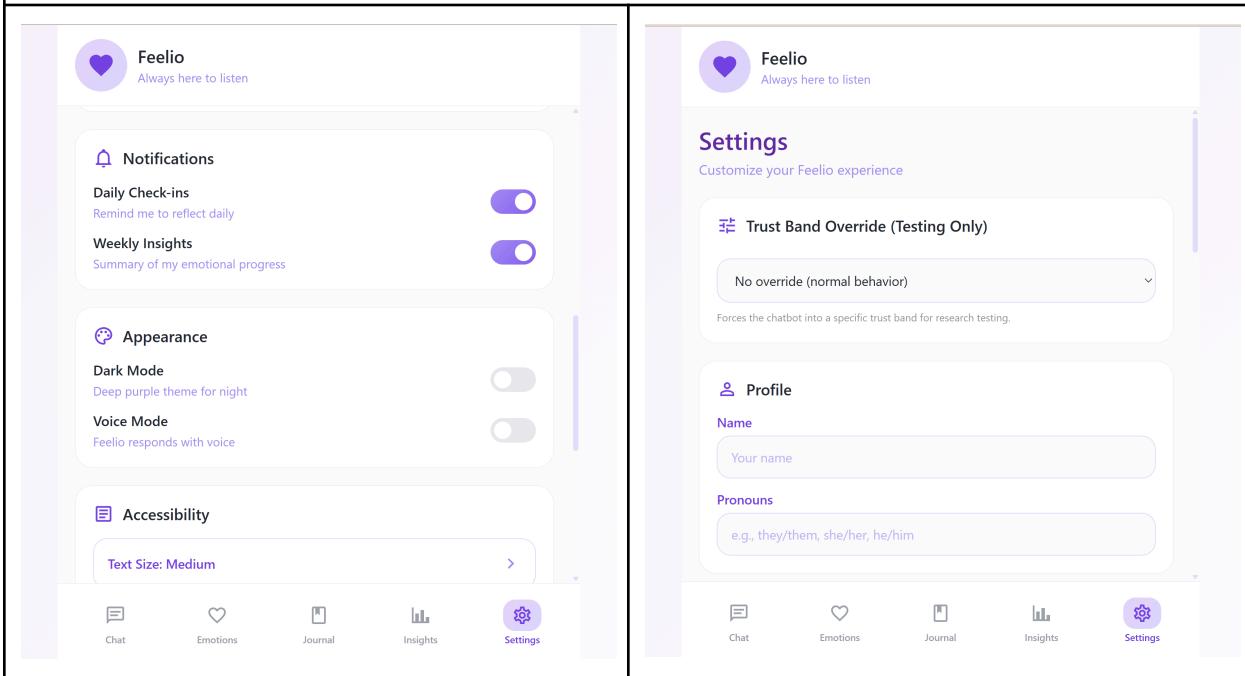


Figure 6 : Insights Page - To view user progress and track patterns.



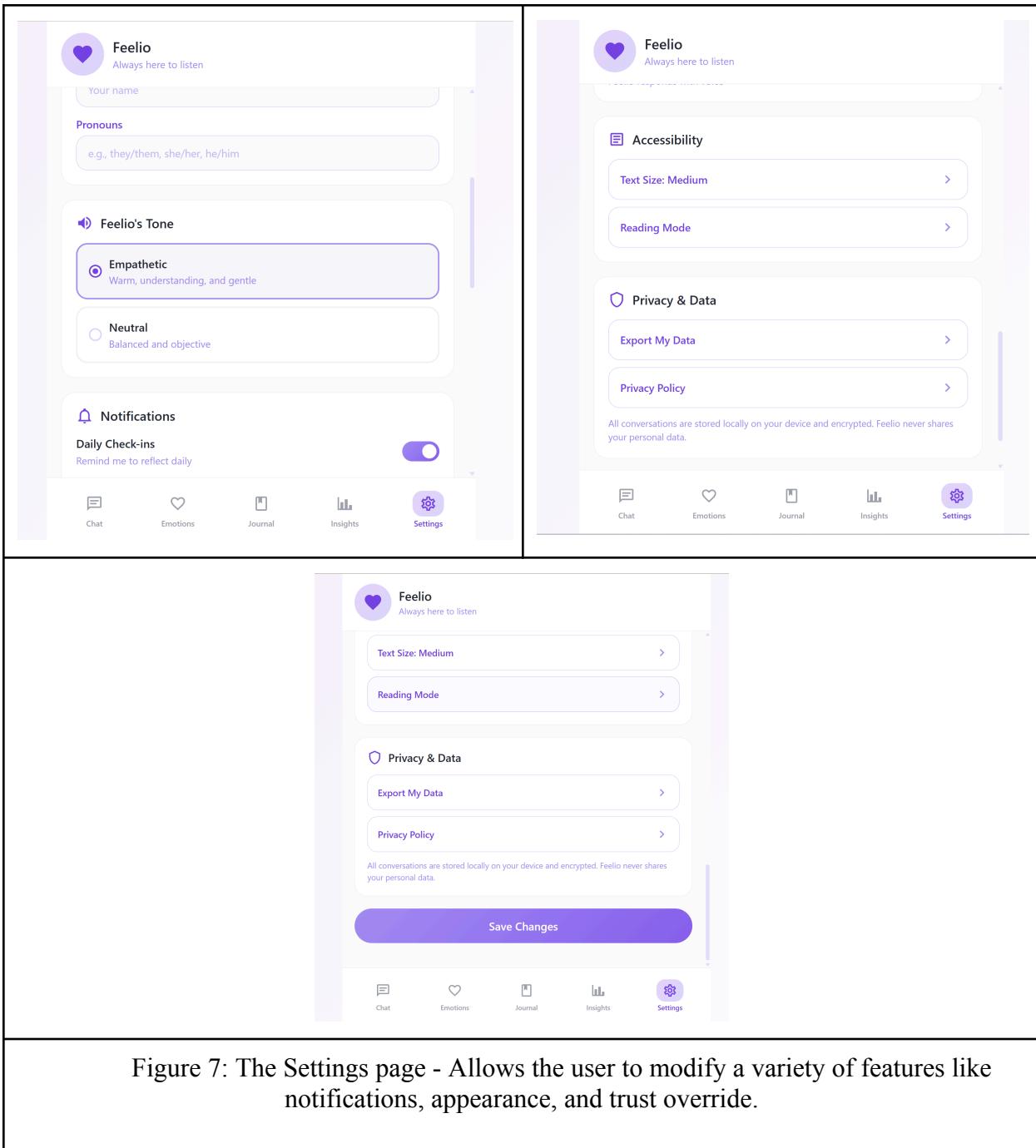


Figure 7: The Settings page - Allows the user to modify a variety of features like notifications, appearance, and trust override.

Compared to the initial design, the final prototype contains many similarities, and slight UI changes were made to better ease the eyes and make the design look cleaner.

The live application is accessible at: <https://lexifeel.vercel.app/>

3.3 Nielsen's 10 Heuristics

The final Feelio prototype was guided by Nielsen's 10 Usability Heuristics to ensure a clear, intuitive, and user-friendly experience. Key applications of these principles include:

1. **Visibility of System Status**
The chatbot clearly communicates its understanding of user input, current conversation stage, and any background processing to maintain transparency and build user confidence.
2. **Match Between System and the Real World**
Language and prompts were designed to reflect natural human conversation, while concrete, literal phrasing supports users with alexithymia in interpreting emotional content.
3. **User Control and Freedom**
Users can correct mistakes, skip questions, navigate between pages, or set user preferences in the settings, ensuring autonomy throughout their interaction.
4. **Consistency and Standards**
Layouts, buttons, color schemes, and message formatting are maintained consistently across all pages, reducing cognitive load and allowing users to quickly understand interface behavior. The design aligns with familiar patterns (such as the chatbot layout) so that users can navigate and interact with the system without encountering unnecessary complexity during the learning phase.
5. **Error Prevention**
Input options are constrained where possible (e.g., emotion-selection buttons in the guidebook), and the chatbot avoids ambiguous questions, reducing potential for misunderstandings.
6. **Recognition Rather than Recall**
Visual cues, prompts, and selectable options support users in making decisions without needing to memorize prior information.
7. **Flexibility and Efficiency of Use**
The interface accommodates both new and experienced users, providing simple guided paths as well as opportunities for deeper reflection once users are comfortable.
8. **Aesthetic and Minimalist Design**
The interface avoids unnecessary clutter, using neutral colors, clear typography, and structured layouts to reduce cognitive load and focus attention on reflection activities.
9. **Help Users Recognize, Diagnose, and Recover from Errors**
When the system cannot understand input, it provides clear feedback and guidance, allowing users to adjust their responses without frustration.

10. Help and Documentation

The Emotion Guide page and tooltips provide contextual information to support learning and understanding of emotions, reducing reliance on external guidance.

By incorporating these heuristics, the Feelio prototype balances functionality, accessibility, and cognitive support, providing a reliable environment for users to practice emotional recognition and reflection.

4. Modified Trust-Reliance Model for Alexithymia

Traditional trust models rely on affective trust - users' feelings of confidence or comfort with a system. However, people with alexithymia have difficulty identifying and expressing emotions, so these models don't apply. Instead, trust must be treated as a cognitive-behavioral construct, based on observable user behavior rather than feelings.

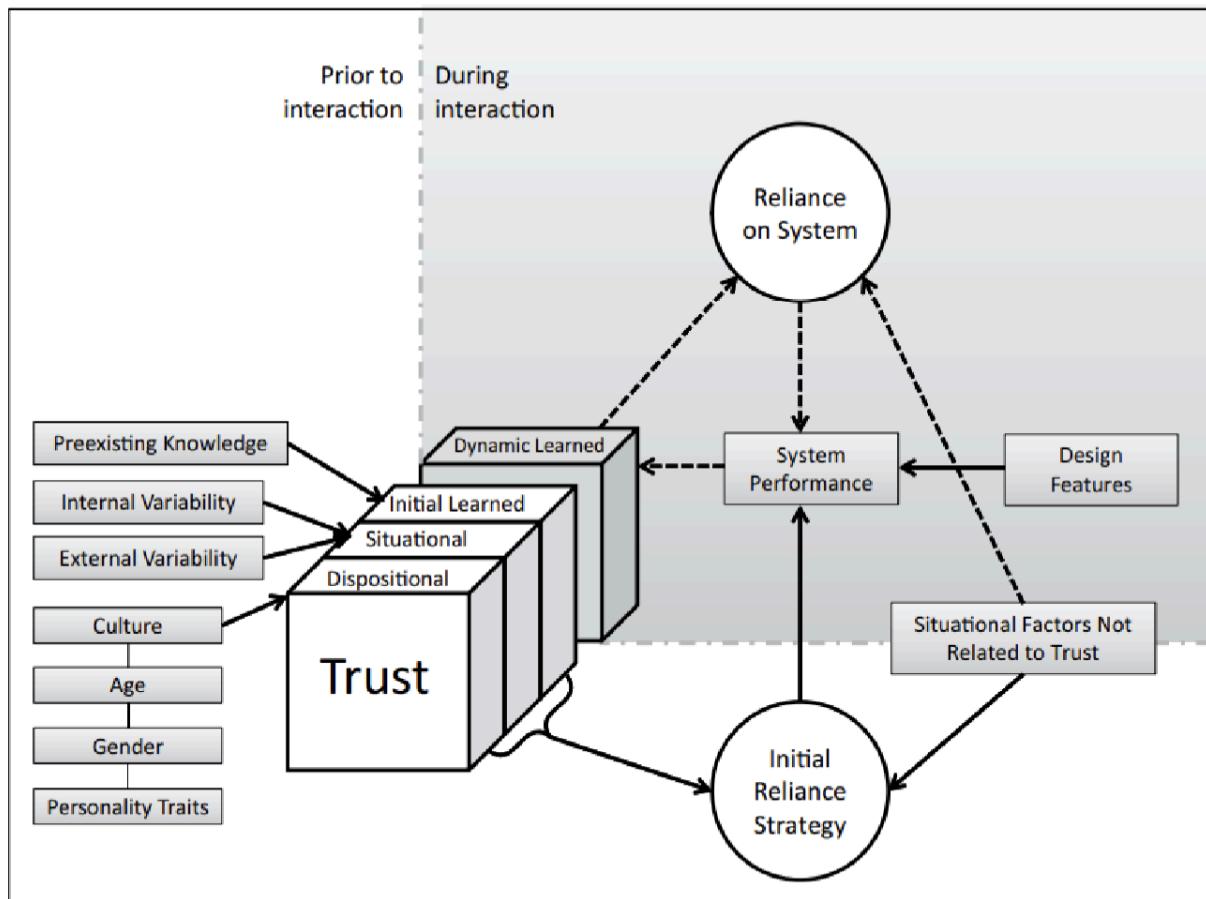


Figure 1. Full model of factors that influence trust in automation. The dotted arrows represent factors that can change within the course of a single interaction.

4.1 Shift From Affective Trust to Cognitive-Behavioral Reliance Readiness

We replace affective trust with Cognitive-Behavioral Reliance Readiness (CBRR): a user's willingness to rely on a system based on clarity, predictability, ease of use, and successful outcomes. CBRR is measured through behavior, not emotion. Key factors include:

- Clear system communication
- Reduced uncertainty or ambiguity
- Predictable system behavior
- Demonstrated usefulness and task success

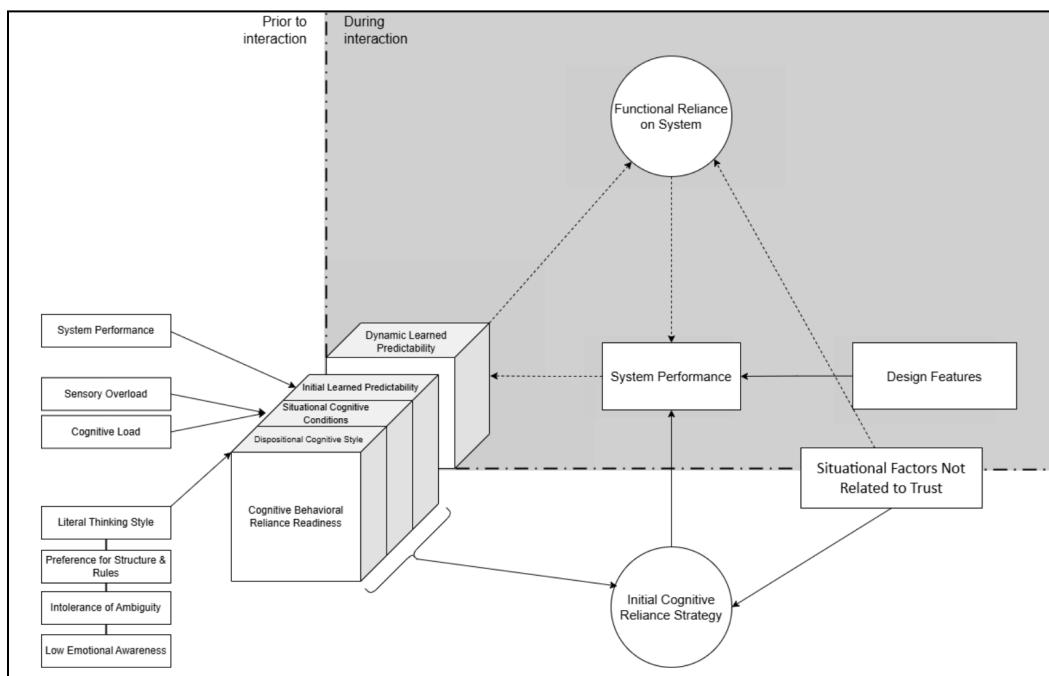


Figure 2. Cognitive-Behavioral Reliance Readiness Model. The dotted arrows represent factors that can change within the course of a single interaction.

4.2 Modified Pre-Interaction Factors

Before interaction, dispositional emotional trust is replaced with cognitive style, such as preference for structure, literal thinking, intolerance of ambiguity, and analytic processing. Environmental factors still matter, but influence understanding rather than comfort.

4.3 Modified Situational and Learned Components

During interaction, traditional learned trust becomes Dynamic Learned Predictability. Alexithymic users rely on consistent patterns rather than emotional cues. Repeated exposure builds stable expectations of system behavior.

4.4 Functional Reliance Over Emotional Reliance

The outcome behavior in this model is not emotional trust but Functional Reliance. Functional reliance is the user's tendency to utilize the system because it reduces cognitive burden and ambiguity. Reliance strengthens when the system:

- maintains literal and concrete language
- avoids metaphors and emotionally laden phrasing
- offers stepwise or choice-based guidance
- improves the user's ability to interpret internal states through scaffolding

This operational definition aligns more closely with the cognitive and experiential patterns of alexithymic users.

4.5 Role of Design Features

Design features act as major determinants of CBRR. For alexithymic users, the interface and communication style must emphasize:

- Predictability
- Transparency
- Structural consistency
- Explicit intent
- Avoidance of emotional assumptions

Because emotional intuition is limited, design clarity becomes a functional substitute for emotional reassurance.

4.6 Situational Factors Not Related to Trust

The model retains non-trust situational moderators, such as distraction, sensory overload, or fatigue. For alexithymic users, these factors especially influence comprehension and processing capacity, underscoring the importance of low cognitive load and clear interaction flow.

4.7 Link to the LLM Implementation Plan

The LLM will measure CBRR rather than emotional trust, using behavioral indicators like engagement, continuity, self-disclosure proxies, and response patterns. This ensures:

- Adaptive behavior aligns with cognitive trust formation
- System responses remain predictable, clear, and unambiguous
- Interactions are ethically and empirically suited for alexithymic users

This modified model forms the foundation for the LLM's adaptive trust mechanism.

5. Final System Implementation

Our Large Language Model (LLM) based conversational system includes a modular architecture, a structured prompting framework, an operational Trust Factor model, and fully integrated safety and ethical controls. All components described below are implemented and functional in the deployed version of the chatbot.

5.1 System Architecture

The final system follows a modular architecture that supports transparent operation, adaptation to user behavior, and safe management of sensitive interactions. The four major components were fully implemented as follows.

User Interface: We developed a simple web-based interface that prioritizes accessibility and clarity. The interface uses high contrast text, a clean message thread, and optional emotion selection tools that provide structured choices when users prefer them. The UI maintains predictable interaction patterns to reduce cognitive load for users who may struggle with ambiguity.

Backend Application Layer: The backend application manages the operational logic of Feelio. It performs the following functions.

- Maintains user sessions
- Stores short-term conversation memory
- Calculates and updates the Trust Factor
- Constructs dynamic prompts for the LLM
- Logs anonymized interaction metrics for system evaluation

This layer ensures stable and predictable system behavior and allows the chatbot to adapt its conversational depth based on user engagement.

LLM Integration: Feelio uses a hosted LLM accessed through a secure API. All responses are generated through a structured prompting pipeline that controls tone, boundaries, and level of emotional scaffolding. The model does not rely on fine-tuning. Instead, prompt engineering and backend logic shape the system's behavior. This allows the model to remain consistent and safe while offering natural language responses.

Data Storage and Logging: A secure database stores the following information:

- Session metadata
- Trust Factor history
- Non-identifiable user feedback
- Time-stamped logs for research and performance evaluation

No personal identifiers or sensitive data are stored, and all logs follow strict data minimization standards.

5.2 Prompting Framework

A layered prompting framework was implemented to ensure consistent behavior, clarity of communication, and adaptive guidance.

System Prompt: The system uses a stable instruction set that defines Feelio's supportive tone, non-judgmental style, and structured conversational approach. The prompt also enforces alexithymia-sensitive guidelines such as concrete language, step-based questions, and explanations that reduce ambiguity. Safety constraints are embedded to prevent clinical interpretation or diagnostic behavior.

Dynamic Context Prompt: The backend generates a context prompt on every turn that includes:

- The current Trust Factor value
- The corresponding trust band
- The immediate conversational goal
- Any recent user feedback or corrections

This allows the chatbot to adjust the level of emotional scaffolding and the type of guidance provided.

Conversation History: A limited number of recent turns are included to maintain context while controlling token use. This provides continuity for the user without overwhelming the model or compromising privacy.

5.3 Trust Factor Model

The Trust Factor model is fully implemented and provides real-time estimates of user comfort and readiness for deeper emotional scaffolding.

Purpose: The Trust Factor enables Feelio to scale its conversational depth in a controlled and appropriate manner. Users who appear uncertain receive simpler, more concrete prompts, while users who show readiness receive gradually increasing reflection support.

Metrics Used: The Trust Factor is calculated using normalized metrics that capture behavioral indicators of reliance.

Metric	Measurement	Interpretation
Feedback (F)	Thumbs up or down	Direct evaluation of usefulness
Engagement Depth (E)	Message length and reply frequency	Proxy for comfort and investment
Continuity (C)	Returns across sessions	Indicator of sustained reliance
Self Disclosure Proxy (D)	Use of personal language	Signal of emotional openness
Repair Success (R)	Improvement after misunderstandings	Indicator of resilience and trust

Table 1. Cognitive-Behavioral Reliance Readiness Model. Trust Factor metrics.

Each metric is scaled from 0 to 1 before weighting. Details on the calculation can be found in the appendix.

5.4 Trust Responsive Behavior

Once calculated, the Trust Factor adjusts Feelio's responses through four trust bands.

Band A: Very Low Trust (0.0 to 0.2)

- Very short and clear messages
- High levels of reassurance and explanation
- No emotional probing

Band B: Cautious Trust (0.2 to 0.4)

- Gentle questions about physical sensations
- Simple emotion choices when needed
- Clear rationales for each prompt

Band C: Moderate Trust (0.4 to 0.7)

- Gradual emotional scaffolding
- Short psychoeducational elements
- Balanced open and closed questions

Band D: High Trust (0.7 to 1.0)

- Deeper reflection prompts
- Multi-step explanations
- References to previous themes when appropriate

These behaviors are now fully implemented and functional within the live system.

5.5 Safety, Ethics, and Risk Controls

The deployed version of Feelio includes comprehensive safety mechanisms.

Crisis Detection: The system detects crisis-related language through keyword spotting. When triggered, Feelio provides supportive messages and directs users to professional or emergency resources. It does not attempt to address crises.

Boundary Enforcement: The model is constrained from performing the following actions.

- Clinical diagnosis
- Medical or legal advice
- Simulation of therapeutic roles

User Autonomy Tools: Users can skip questions or change topics at any time. These controls help support comfort, consent, and agency.

Data Minimization: The system stores only non-identifiable text logs and essential session metadata. No personal identifiers or sensitive data are collected.

5.6 Summary of the Final System

Feelio is now a fully implemented adaptive chatbot that uses a behavioral Trust Factor to adjust conversational depth for individuals with alexithymia. The architecture, prompting framework, trust model, and safety systems work together to provide structured emotional learning while maintaining strict non-clinical boundaries. The modular design supports ongoing refinement and aligns with human-centered and safety-critical principles.

Results

6. Testing

To evaluate the usability and safety of the Feelio chatbot, a three-stage testing framework was implemented. This framework was designed to assess both the technical performance of the trust-responsive model and the user experience components central to human-centered AI systems. The CBRR model plays a critical role in adapting the chatbot's depth and style of interaction. The evaluation process called for testing the following:

1. Whether the trust bands produce distinct conversational behaviors
2. Whether trust evolves appropriately during realistic interactions
3. How users perceive the system's clarity, emotional appropriateness, and overall usability

The following subsections describe each testing stage in detail, along with results. Together, the tests provide a holistic assessment of the chatbot's functionality. The multi-layered approach ensures that both the internal logic of the model and the external user-facing experience are systematically validated.

Stage 1: Override Testing Method

Method

Stage 1 of testing evaluated whether the chatbot's trust-responsive behavior changed meaningfully between the two extreme trust bands: Band A (lowest reliance) and Band D (highest reliance). The trust factor was manually overridden to isolate the system's adaptive behavior from the dynamic calculation pipeline. This allowed for a controlled comparison of the Chatbot's performance between the bands. This stage focused only on Band A and Band D as they represent the full range of the model, which makes them the clearest way to test whether the CBRR framework produces distinct conversational styles. The intermediate bands (B and C) were evaluated later during the natural trust-building test in Stage 2, where the trust factor changes dynamically during conversation.

In this stage, the chatbot began the conversation in the same way for both trust bands ("Hi there! I'm here to help you explore and understand what you notice. What physical sensations are showing up right now?"), And the first user response was also kept identical to ensure both tests started under the same conditions. After this initial turn, user replies were allowed to vary naturally so the interaction remained realistic. This setup made it possible to observe how each trust band influenced the chatbot's behaviour without forcing scripted or unnatural responses. Because the purpose of Stage 1 was to highlight the contrast between very low trust (Band A) and very high trust (Band D), we used a single shared starting prompt and one short conversation per band rather than testing multiple prompts or branches. This allowed us to clearly evaluate whether Band A remained simple and surface-level and whether Band D offered more depth, reflection, and guidance.

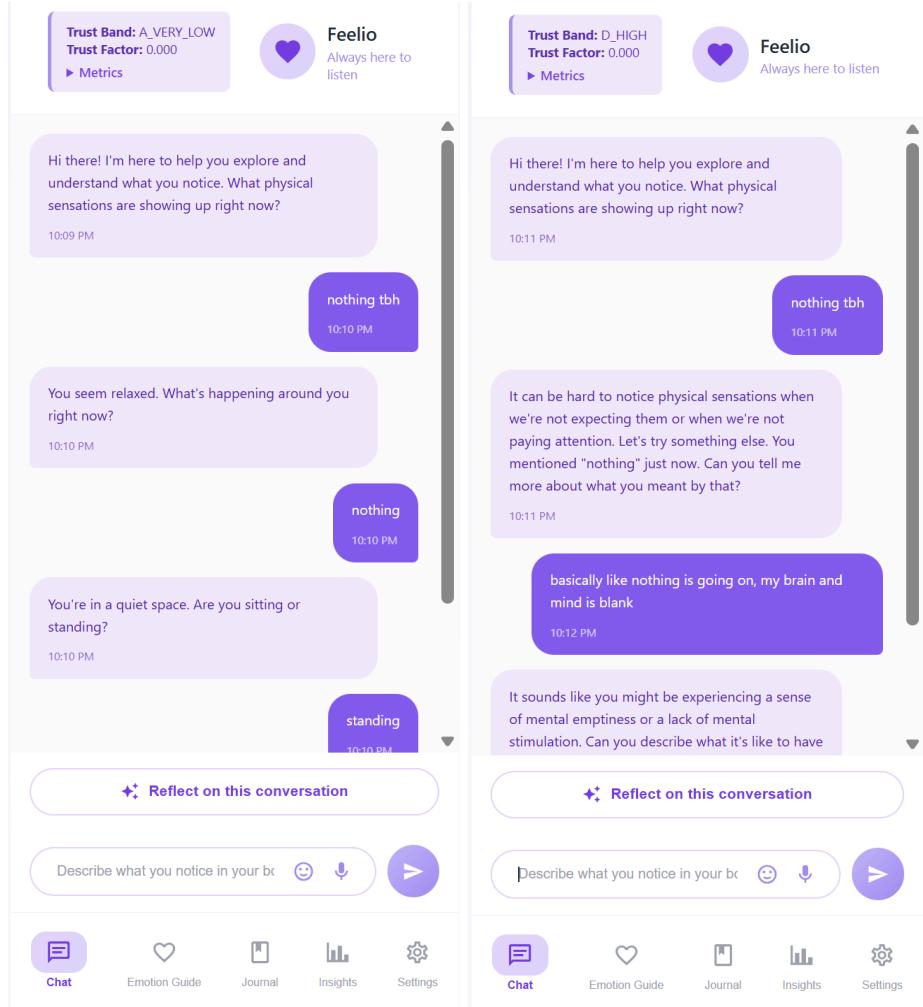


Figure 3. Example conversation from Stage 1 of Testing: Low vs. High Trust Conversations

Results

Feature	Band A (Low Trust)	Band D (High Trust)
Response length	Short, minimal guidance	Longer, more detailed
Tone	Neutral, observational	Warm, reflective, supportive
Question style	Simple, literal questions	Open-ended, exploratory questions

Emotional inference	None	Frequent, gentle inference
Conversational flow	Step-by-step	More holistic and relational

Table 2. Results from Stage 1 of Testing: Low vs. High Trust Qualitative Comparison

Average Chatbot Word Count

- Band A average: ≈21 words
- Band D average: ≈33 words

Example sentence complexity

- Band A: single-clause statements
- Band D: multi-clause reflective statements

The two override conversations showed clear differences in how the chatbot responded under the lowest and highest trust bands. In Band A, the chatbot stayed concrete and surface-level, asking simple factual questions about the user's environment ("Are you sitting on a chair or a bed?") and avoiding any form of emotional interpretation. Responses were brief, repetitive, and focused only on observable sensations and surroundings. In contrast, Band D produced longer and more reflective replies. The chatbot referenced earlier parts of the conversation, introduced mild interpretive statements ("It sounds like you're in a calm and private space"), and encouraged deeper exploration by asking open-ended questions about meaning and emotion ("What's the first thing you think about when you look at your cat?"). The overall conversational depth, use of context, and cognitive demand were noticeably higher in Band D. These differences confirm that the trust-band override mechanism successfully generated two distinct styles of interaction, consistent with the goals of the CBRR framework.

Stage 2: Trust Building

Method

Stage 2 evaluated how the chatbot adapted its behaviour when the trust factor was allowed to evolve naturally during a conversation. Unlike Stage 1, where trust bands were manually overridden, this stage began with a low initial trust value and allowed the system to update the trust factor after each user message using the CBRR weighting formula. As the interaction progressed, the chatbot could transition through the intermediate trust bands (B and C) before reaching higher levels, depending on the user's engagement and clarity signals.

The conversation was conducted using a simulated user who responded consistently and realistically to the chatbot's questions. User replies were not scripted, allowing the trust factor to change organically based on message length, continuity, and level of disclosure. This setup created a realistic scenario for testing whether the chatbot gradually increased conversational depth in response to rising trust, and whether these transitions remained smooth and appropriate. A single longer conversation was used for this stage, as the goal was to observe trust progression over time rather than compare isolated snapshots.

Results

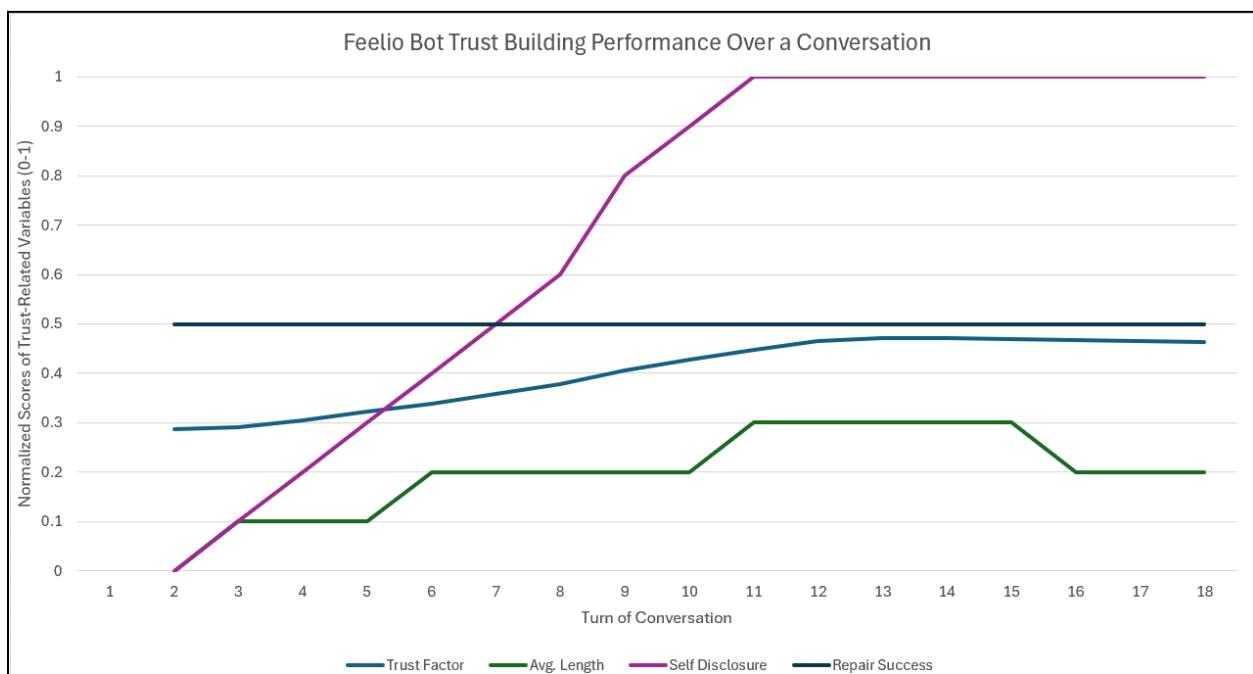


Figure 4. Feelio Bot Trust Building Performance Observed Over an 18 Turn Conversation

In Stage 2, the goal was to see how the trust score changed during a real conversation. The chatbot and user completed 18 turns, and after each turn, the system recorded the Trust Band, Trust Factor, message length, self-disclosure, and repair success.

The interaction started in Trust Band B, with a low Trust Factor of 0.287. At this point, the user gave short answers and did not share much about themselves. As the conversation continued, the user responded with slightly longer messages and started describing physical sensations (e.g., tension in the shoulders). Because of this, the Trust Factor slowly increased turn by turn. By Turn 8, the user mentioned that the breathing exercise "made me feel a bit calmer." This was the first clear sign of emotional openness, and the system moved the user from Band B to Band C. Trust Factor also rose to 0.379 at this point.

Between Turns 8-12, the user showed the most engagement in the entire conversation. The messages became longer, and the user openly described how they do not usually notice stress until it “piles up.” During this section, self-disclosure reached the maximum value (1.0). The Trust Factor increased to its highest point (0.471), showing that the user was interacting in a cooperative and reflective way.

A shift happened around Turn 13. The user’s responses became short and dismissive (“whatever. idk”). Shortly after, they also expressed frustration (“this isn’t really helping”). This marked the start of a rupture. Although the Trust Factor did not collapse, it began to decline slowly from 0.471 to 0.463 by the end of the session.

Throughout this period, the Repair Success score stayed at 0.5, which means the chatbot’s attempts to repair the conversation neither improved nor worsened trust. The user continued responding, but with less interest and less emotional involvement.

The overall pattern across the 18 turns was:

1. Low trust at the start: short answers, low self-disclosure.
2. Gradual increase: the user shared more sensations and thoughts.
3. Peak trust: user became reflective and open; Trust Factor reached 0.471.
4. Rupture: frustration and withdrawal appeared.
5. Slight decline and stabilization: trust dropped a bit but did not fall out of Band C.

The results show that the trust model reacts realistically to changes in user behaviour. When the user shared more about their internal experience, trust increased. When the user became frustrated and disengaged, trust decreased. This demonstrates that the model is sensitive to both building moments (e.g., successful exercises, emotional sharing) and ruptures (e.g., short, frustrated replies).

Stage 3: Simulated User Testing

Method

Stage 3 assessed the chatbot’s usability, clarity, and perceived safety using 5 distinct simulated user personas, shown in the table below. Each persona represented a different communication style and level of emotional awareness to reflect a range of potential user behaviours, such as brief responders, analytical thinkers, overwhelmed students, or highly engaged users. For each persona, a short conversation with the chatbot was conducted to evaluate how well the system handled varied interaction patterns and whether its tone and guidance remained appropriate across users.

Persona	Core Communication Style	Key Characteristics
A	Brief, reserved	Short replies, low emotional awareness, minimal self-disclosure
B	Engaged, expressive	Longer messages, willing to reflect, higher self-disclosure
C	Hesitant, inconsistent	Mix of short and detailed replies, easily overwhelmed, variable openness
D	Logical, analytical	Literal responses, prefers structure, avoids emotional language
E	Guarded, masking	Polite but cautious, slow to open up, moderate message length

Table 3. Persona Summary for Simulated Testing

After completing their conversation, each persona completed a System Usability Scale (SUS) questionnaire along with a small set of custom Likert-scale questions measuring clarity, comfort, and perceived emotional appropriateness. Open-ended feedback was also collected to capture qualitative impressions of the chatbot's performance. This stage focused on evaluating the overall user experience rather than testing specific trust-band behaviours. Using simulated personas allowed for consistent, repeatable tests while still capturing a variety of communication tendencies relevant to individuals with alexithymia.

Results

Persona	SUS Score	Trust Band Progression	Trust Factor (Min-Max)	Avg. Message Length (0-1)	Self-Disclosure (0-1)
A	77.5	B - C	0.340-0.417	0.5-0.6	0.10-0.40
B	78	B - C	0.417-0.525	0.6-0.7	0.40-0.90
C	65	B - C	0.333-0.469	0.4-0.5	0.10-0.90
D	80	B - C	0.340-0.417	0.5-0.6	0.10-0.40

E	76	B - C	0.346-0.430	0.5-0.6	0.20-0.70
---	----	-------	-------------	---------	-----------

Table 4. Summary of Metrics Collected from Simulated Persona Testing

The Stage 3 evaluation showed that the chatbot performed reliably across five simulated user personas with different communication styles. Usability scores were consistently strong, with a mean SUS score of 75.3 (SD = 5.76), which is above the commonly accepted usability benchmark of 68. This indicates that, overall, the chatbot was perceived as clear, easy to use, and functionally coherent.

Metric	Mean	SD
SUS Score	75.3	5.76
Trust Factor Min	0.355	0.035
Trust Factor Max	0.452	0.045
Avg. Message Length	0.55	0.07
Self-Disclosure	0.42	0.15

Table 5. Mean and Standard Deviations of Metrics Collected From Simulated Persona Testing

Trust development followed a similar pattern. All personas moved from Band B to Band C, suggesting that users steadily became more comfortable with the system during the conversation. The trust factor values also showed low variability (Min SD = 0.035, Max SD = 0.045), meaning that trust increased at a similar rate across very different user types.

Communication patterns varied between personas, but the chatbot handled these differences smoothly. Average message length remained moderately stable ($M = 0.55$, $SD = 0.07$), showing that the chatbot did not trigger unusually long or short responses. Self-disclosure, which can fluctuate widely across individuals, showed more variability ($M = 0.42$, $SD = 0.15$). This makes sense because some personas were designed to be brief or guarded, while others were more expressive. The important finding is that the chatbot maintained an appropriate tone regardless of disclosure level.

Taken together, the results suggest that the system delivers a consistent and emotionally safe experience across diverse user profiles. The narrow variability in usability and trust metrics indicates that the chatbot adapts without depending on a particular type of user behavior. This is especially valuable for real-world use with people who may struggle with emotional awareness, since the system appears robust to different communication styles and levels of engagement.

Discussion

The goal of this project was to design a conversational agent that supports emotional awareness for individuals with alexithymia by providing clear, predictable, and low-pressure dialogue. Unlike systems that rely on affective inference, this project introduced the Cognitive Behavioral Reliance Readiness (CBRR) model, which measures trust through observable behavioural cues rather than emotional signals that alexithymic users may struggle to express. The results from all three testing stages show that this approach is both technically functional and user-appropriate, and they demonstrate how the system's behaviour aligns with the human-centered principles explored throughout ENGG 519.

Distinct Adaptive Behaviours and Predictable Interaction Patterns

Stage 1 demonstrated that the trust bands produce meaningfully different conversational behaviours, confirming that the CBRR framework can shape the system's communication style in transparent and predictable ways. Band A responses were short, literal, and grounded in observable facts, while Band D responses showed richer context use and gentle reflection. This aligns closely with the course emphasis on calibrated reliance (Lectures 1-2), which stresses that users must be able to anticipate how a system will behave, especially in emotionally sensitive contexts. By producing distinct but stable interaction patterns at each trust band, the chatbot provides a predictable structure that helps users form accurate mental models of the system, reducing the “gulf of evaluation” discussed in Lecture 2.

Natural Trust Evolution and Realistic Human-AI Dynamics

Stage 2 showed that trust evolved in a realistic and psychologically coherent way during an extended conversation. Increases in message length and self-disclosure led to gradual rises in the trust factor, while a rupture in the interaction produced a small but meaningful decline. These results mirror patterns described in the course's material on trust calibration and appropriate reliance (Lectures 3-4): trust should increase when the system and user collaborate smoothly and decline when conversational strain appears, but without extreme volatility. The ability of the CBRR model to detect both building moments and ruptures indicates that behavioural trust modelling—rather than affective inference—is not only safer for alexithymic users but also capable of capturing important relational dynamics.

Usability, Clarity, and Cognitive Load Management

Stage 3 demonstrated that the system remains usable and emotionally appropriate across diverse communication styles. The SUS mean score of 75.3 ($SD = 5.76$) exceeds standard usability benchmarks, and trust trajectories were nearly identical across personas. These results reflect core ideas from cognitive load theory and situation awareness (Lectures 1 and 3): predictable structure, low-complexity dialogue, and consistent conversational scaffolding help users remain

oriented even when discussing emotional content. The system's design—particularly its step-by-step prompting style—appears to have successfully minimized extraneous cognitive load, a central need for individuals with alexithymia, who often experience uncertainty when interpreting internal states.

The stability of trust and usability across personas also indicates that the system's transparency and structured reasoning steps are functioning as intended, echoing Human-Centered Explainability (Lecture 7). Although the chatbot does not verbalize its internal mechanisms, its behaviour is sufficiently consistent and interpretable for users to understand its intentions and boundaries.

Safety, Boundaries, and Sociotechnical Considerations

Across all testing stages, the chatbot maintained a supportive but bounded tone, avoided inappropriate emotional inference, and did not escalate beyond its intended scope. These behaviours directly reflect the principles from Safety-Critical AI and sociotechnical risk (Lectures 9-10), which emphasize that systems interacting with vulnerable users must enforce strict limits to prevent over-reliance, anthropomorphism, or perceived therapeutic authority. The consistent movement from Band B to Band C across personas suggests that users experienced comfort and clarity, but did not enter levels of dependency that would raise safety concerns. The system's refusal mechanisms, crisis safeguards, and avoidance of clinical language also directly embody risk-mitigation strategies taught in the course.

Integrated Interpretation

Overall, the results show that the Feelio prototype is consistent with the project's goal of supporting emotional awareness in a low-pressure and structured environment. The system's behaviours reflect a thoughtful application of ENGG 519 concepts: calibrated reliance, mental model alignment, cognitive load reduction, trust calibration, explainability, and safety-critical design. The CBRR model proved effective in producing adaptive dialogue while remaining inclusive of users who struggle with emotional introspection. These findings suggest that the methodological and theoretical foundations of this project translate into measurable improvements in clarity, usability, and emotional safety.

Conclusion

This project establishes a foundation for creating an AI conversational agent tailored to the cognitive and emotional needs of individuals with alexithymia. By reframing trust as a behavioral construct and designing interaction patterns that minimise ambiguity, the system can provide structured guidance that supports emotional reflection. The planned LLM implementation builds on these principles and outlines a safe, adaptive, and ethically grounded framework. As development continues, usability testing and prototype refinement will help

ensure that the chatbot becomes a practical and supportive tool for improving emotional awareness.

References

- [1] P. E. Sifneos, “The prevalence of ‘alexithymic’ characteristics in psychosomatic patients,” *Psychotherapy and Psychosomatics*, vol. 22, no. 2, pp. 255-262, 1973. [Online]. Available: <https://doi.org/10.1159/000286529>
- [2] G. J. Taylor, R. M. Bagby, and J. D. A. Parker, “The 20-item Toronto Alexithymia Scale. IV. Reliability and factorial validity in different cultures,” *Journal of Psychosomatic Research*, vol. 55, no. 3, pp. 277-283, 2003. [Online]. Available: [https://doi.org/10.1016/S0022-3999\(02\)00601-3](https://doi.org/10.1016/S0022-3999(02)00601-3)
- [3] G. J. Taylor and R. M. Bagby, “Alexithymia and mental disorders: A review of the literature,” *Journal of Psychosomatic Research*, vol. 41, no. 6, pp. 535-544, 1996. [Online]. Available: [https://doi.org/10.1016/S0022-3999\(96\)00217-6](https://doi.org/10.1016/S0022-3999(96)00217-6)
- [4] J. G. Meganck, F. Vanheule, and E. Desmet, “Alexithymia in adolescents: A review of the literature,” *Journal of Adolescence*, vol. 35, no. 6, pp. 1417-1430, 2012. [Online]. Available: <https://doi.org/10.1016/j.adolescence.2012.06.004>
- [5] N. Suslow and H. Donges, “Alexithymia in the elderly: Associations with cognitive functioning, depression, and anxiety,” *Aging and Mental Health*, vol. 19, no. 6, pp. 1-7, 2015. [Online]. Available: <https://doi.org/10.1080/13607863.2014.1003284>
- [6] A. Ograniczuk, G. J. Taylor, and R. M. Bagby, “Psychotherapy for individuals with alexithymia: A review,” *Australian and New Zealand Journal of Psychiatry*, vol. 47, no. 10, pp. 1-12, 2013. [Online]. Available: <https://doi.org/10.1177/0004867413497990>
- [7] J. D. Lee and K. A. See, “Trust in Automation: Designing for Appropriate Reliance,” *Human Factors*, vol. 46, no. 1, pp. 50-80, 2004. [Online]. Available: https://doi.org/10.1518/hfes.46.1.50_30392
- [8] K. A. Hoff and M. Bashir, “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust,” *Human Factors*, vol. 57, no. 3, pp. 407-434, 2015. [Online]. Available: <https://doi.org/10.1177/0018720814547570>

Appendix

Trust Factor Calculation

The final Trust Factor is computed with the following formula:

$$\begin{aligned} \text{TrustRaw} = & 0.40 F_{\text{norm}} \\ & + 0.20 E_{\text{norm}} \\ & + 0.15 C_{\text{norm}} \\ & + 0.15 D_{\text{norm}} \\ & + 0.10 R_{\text{norm}} \end{aligned}$$

A smoothing step stabilizes fluctuations.

$$\text{TrustNew} = 0.7 \text{ TrustOld} + 0.3 \text{ TrustRaw}$$

This produces gradual changes that reflect the evolving interaction.

Simulation Testing ChatGPT Chats

<https://chatgpt.com/share/693a6872-0944-8004-9a81-ca4f217f8eba>
<https://chatgpt.com/share/693a6885-0488-8004-a520-a8b4de303480>
<https://chatgpt.com/share/693a6893-b3e4-8004-805f-177d5464b68f>
<https://chatgpt.com/share/693a68a2-a80c-8004-8796-2c89f7959d9e>
<https://chatgpt.com/share/693a6885-0488-8004-a520-a8b4de303480>