To Deceive or Self-Deceive?

Framing Language to Discourage Deception in Diabetes Lifestyle Management Systems

Marina Mădăraș

Responsible Professor: Prof. dr. C.M. Jonker
Supervisor: J.D. Top, MSc
Examiner: Dr. Avishek Anand

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Abstract

Deceptive self-reporting in diabetes lifestyle management (DLM) systems limits their ability to offer meaningful and accurate support. Deception can function as a self-protective mechanism, driven by factors such as low self-esteem or the desire to protect self-image. This research builds on CHIP, a chatbot-based DLM prototype, to explore whether the language framing of its responses can influence the psychological determinants of deception. Two framing strategies, empathic and affirming, were implemented and evaluated through a pilot user study, which offers insights for refining the intervention and experimental design in future research.

1 Introduction

Diabetes is one of the most prevalent chronic illnesses worldwide [1]. However, access to healthcare is highly unequal, with nearly two billion people experiencing financial hardship due to medical expenses [2]. In parallel, the widespread availability of intelligent technologies has reshaped how health support can be delivered. By 2024, approximately 71% of the global population was estimated to own a smartphone [3], which facilitates access to a vast ecosystem of mHealth applications [4, 5]. These tools offer scalable support for fitness, mental health, and chronic disease management, including diabetes, and make healthcare resources more accessible to a wider population.

This kind of digital accessibility becomes particularly valuable amid rising medication costs. For example, one study [6] found that a hybrid approach combining access to an eHealth application with medical supervision led to a 71% reduction in required insulin units. Such outcomes show that eHealth tools can make a tangible difference in complementing care and reducing both medical and financial burdens.

Diabetes management requires consistent and long-term commitment to lifestyle changes, including diet, physical activity, and regular monitoring, to reduce dependence on medication and maintain control of the condition. Many individuals struggle to adhere to such treatment plans [7, 8]. While eHealth tools can offer valuable support in this process, they may not suffice on their own to sustain engagement over time.

In addition, users have been found to provide inaccurate reports of their blood glucose levels [9]. This behavior may not be intended to deceive the system per se. Instead, it may serve to protect oneself from feelings of guilt [10] or distress [11]. In this sense, the deception is directed inward, as a form of self-deception. From a research and design perspective, it becomes essential to understand and address these patterns.

This concern is echoed in a recent study on deception in human-AI interaction [12], which compares honesty in reports to chatbots versus humans, and notes that most AI systems still lack a theory of mind (ToM), the capacity to reason about others' internal, invisible mental states [13]. This gap presents opportunities to incorporate ToM in the design of support systems, which can serve as a stepping stone toward more psychologically informed eHealth tools.

To this end, and as part of the broader Hybrid Intelligence (HI) Project [14], this research contributes to the overarching goal of fostering sustained adherence to self-treatment with the help of diabetes lifestyle management (DLM) systems [15]. It builds on CHIP (Computer Human Interaction Project), a research prototype developed within the HI Project in collaboration with TNO (The Netherlands Organisation for Applied Scientific Research), to support DLM. CHIP is a chatbot-based DLM tool that helps users through conversations. The interactions are enabled by a structured series of processing layers that perform Natural Language Processing (NLP) tasks, reasoning, and decision-making. The final stage is response generation, which transforms structured data into coherent, user-friendly messages.

The semantic content of these messages is fundamental to communication. However, their mode of delivery, such as tone and style, can influence how users perceive and engage with the information. A mode of delivery can be linguistic framing (i.e, *how* messages are worded and what aspects are *emphasized*). For instance, the same guidance may be expressed as "Log your blood sugar now" or as "Let's check your blood sugar to keep your levels on track", with each framing likely to elicit a different user response. Building on this understanding, the question guiding this research is:

> *How does the framing of responses in a diabetes lifestyle management system influence the behavioral drivers behind users' deceptive self-reports?*

The framing of responses is considered a potential intervention to reduce deception and possibly increase adherence. However, to make such strategies more effective, it is important to move beyond targeting the behavior itself and address the deeper psychological and contextual factors that shape it [16]. For instance, someone might under-report their sugar intake not with the intention to deceive, but as a way to avoid feelings of guilt or shame [11]. As such, a literature review was first conducted to gather a theoretical understanding of deception in the context of diabetes self-management. Section 2 outlines its underlying drivers and uses them as the basis for designing two language framing interventions: one grounded in empathy and the other in self-affirmation.

A prototype of these interventions was implemented by extending the reasoner and response generator modules of CHIP, as explained in Section 3. To explore their potential effectiveness, a pilot user study was conducted. The experimental setup and methodology of this study are described in Section 4, followed by the preliminary findings in Section 5, and limitations in Section 6. Together, these insights informed a set of proposed improvements for a second iteration of the study, to both the prototype and the evaluation setup, which are presented in the future work section (Section 7). Finally, the responsible research section (Section 9) reflects on the ethical implications of this study, including the potential risks associated with using large language models in health-related contexts. It also discusses how concerns about reproducibility were addressed throughout the work.

2 Theoretical Foundation

To design an effective behavior change intervention, a clear understanding of the underlying behavior is essential. This section outlines the theoretical basis of this work and explains how a behavior change framework was applied to analyze deceptive behavior and inform the design of the intervention.

2.1 From Behavior to Intervention

The work was guided by the Behavior Change Wheel (BCW), which provides a structured approach for analyzing the target behavior and identifying appropriate intervention strategies [17]. The BCW was selected due to its strong empirical foundation, developed through the evaluation and synthesis of existing intervention frameworks based on their assessed usefulness. This foundation is further supported by initial testing across two health-related behavior interventions: tobacco control and obesity guidance, and by the framework's widespread use in later research [18–20].

The BCW begins with an analysis of the drivers behind behavior, using the COM-B system, which explains Behavior as an outcome of Capability, Opportunity, and Motivation [17]. Each component includes distinct subcategories (e.g., physical vs. social opportunity, or reflective vs. automatic motivation). Although behaviors typically involve a combination of COM-B components, one often plays a dominant role.

Once the relevant COM-B components are identified, the framework offers a structured mapping to intervention types that are suited to influencing them (e.g., persuasion, education) [17]. Moreover, an intervention type may target multiple behavioral components and is implemented through one or more *Behavior Change Techniques* (BCTs), which serve as the practical means of delivering the intervention. Although the BCW also links intervention functions to broader policy categories (such as legislation and fiscal) intended to support implementation at a higher level [17], this additional step falls outside the scope of the present study.

2.2 Modeling of Deceptive Behavior

This study is centered around deception. Recognizing that behavior is shaped by its context [16, 17], I situate deception in the everyday reality of managing diabetes. Since no studies were identified that link deception to diabetes self-management, I explored the drivers behind both behaviors.

Modeling Approach

My approach involved conducting a literature review focusing on two topics: deception and poor diabetes self-management. A fully systematic review was not feasible within the scope of this research, which also included the development of a pilot study. Papers were included based on the following criteria: they were either meta-analyses exploring the behaviors and their underlying drivers or empirical studies presenting experimental evidence.

The findings were annotated under the relevant COM-B sub-components (Automatic Motivation, Reflective Motivation, Social Opportunity) based on the dimension that best represented each factor. During this process, several deception-related factors, such as white lies intended to protect loved ones [10], were excluded, as they were not relevant to the specific context of deception directed at conversational agents. Capability and Physical Opportunity, the remaining COM-B factors, played a minimal role in the annotation process. This is due to the assumption that users are inherently capable of deception and that the DLM software interaction itself offers sufficient physical opportunity.

Table 1: Taxonomy of factors driving deception, grouped by COM-B sub-components. This subset includes only factors considered relevant in the context of human-computer interaction, such as chatbot-based diabetes lifestyle management systems.

COM-B component	Framing of Factor	Contributing Factor
Automatic Motivation	Avoidance/ Regulation of [21]	*negative emotions*: shame [11], guilt [11], discomfort, distress [10]
	Enabled by [22]	low self-control low self-awareness [23] (weak) internal reward system [23]
Reflective Motivation	Protection of [10, 11]	identity self-esteem [21]
	Avoidance of [10]	responsibility negative consequences [24]
Social Opportunity	Protection from [11, 21]	loss of face looking bad embarrassment
	Enabled by	low external cost [22, 24]

The annotation process was conducted to develop a contextual framework that would inform how I apply the BCW in my design decisions. Given the exploratory rather than comprehensive nature of this process, I did not attempt to systematically code the findings, such as by using a multi-rater annotation study. Nonetheless, I created two non-exhaustive taxonomies from the annotated findings: one for factors behind deception (Table 1) and one for poor diabetes self-management (Table 2). To ensure contextual alignment between the two, the diabetes taxonomy was annotated using only the COM-B sub-factors also used for deception.

Behavioral Drivers

Most findings on deception suggest it is primarily driven by intrinsic motives, as a means of protecting the self and managing one's perceived social image, particularly to avoid psychological loss or preserve emotional well-being [10, 21, 25].

This drive to protect the self also underlies *self-deception*, where one selectively distorts or rejects information to preserve emotional stability, maintain self-image, and avoid psychological discomfort [10]. As such, much of the deception directed at intelligent agents may be better understood as self-deception, aimed more at sustaining internal comfort rather than any intention to influence the system itself.

Table 2: Taxonomy of factors driving poor diabetes self-management, grouped by COM-B sub-components. This subset includes only factors that could be coded using the same COM-B sub-components identified in the deception taxonomy (see Table 1), to maintain consistency in the comparison.

COM-B component	Framing of Factor	Contributing Factor
Automatic Motivation	Enabled by	negative emotions [1, 28]: distress [26] frustration [1] low self-control [26]
Reflective Motivation	Enabled by	low self-efficacy [26–28] low self-esteem [28] low self-compassion [28] low perceived (behavioral) control [26, 27]
Social Opportunity	Enabled by	poor communication [1, 26]

Poor diabetes self-management is influenced by both structural and psychological factors. Structural factors, such as high costs [26] and limited access to glucose monitors [27], although highly relevant, are not modifiable through behavioral interventions [28] and are therefore not addressed in this study. Instead, I focus on psychological drivers, given that diabetes is widely recognized as a chronic stressor [26], placing patients in a well-defined situational context. Within this context, psychological factors, such as self-efficacy, depressive symptoms, diabetes distress, and self-esteem, have been shown to be statistically correlated with one another [28].

Several factors, such as *self-esteem*, *negative emotions*, and *low self-control*, appear in both, while a wider range of constructs related to the self, identity, and perceived capabilities function as central psychological drivers of both deception and poor diabetes self-management. These findings carry meaningful implications: by addressing the psychological drivers of deception, the support-system not only discourages deceptive user self-reporting, but also strengthens its core function of supporting effective diabetes self-management. This, in turn, underscores the broader potential of psychologically informed design in the development of eHealth tools.

2.3 Intervention Design

Selection of Intervention Targets

Since both taxonomies revealed similar themes, I focused on two recurring factors as primary targets for the intervention: *self-esteem* and *self-control*. Although *negative emotions* were also among the shared behavioral drivers, they were deemed too broad for addressing in the intervention design. In addition, I selected a third target specific to deception: *protection from loss of face and looking bad*, which I refer to in this report as *self-image protection*. This choice is grounded in the exploratory aim of designing the diabetes support chatbot as a safe, judgment-free space, where users are not triggered into managing their self-presentation.

The selection of these targets also guided the experiment design, as they are measurable constructs with validated scales [29–31]. This aligns with recent recommendations to assess not only behavioral outcomes, but also changes in the determinants driving those behaviors, which are often overlooked in evaluations of intervention effectiveness [16].

Framing Language as a Behavioral Intervention

The three key targets correspond to Reflective Motivation, Automatic Motivation, and Social Opportunity. To address these, the BCW recommends as interventions *persuasion*, which uses communication to influence attitudes and emotions, and *enablement*, which involves removing barriers to make behavior change more achievable [17]. The framework also identifies other intervention types such as *coercion* and *restriction*, which rely on penalties or limiting access to discourage behaviors [17]. However, in the context of digital health tools that patients use voluntarily, these approaches may be counterproductive to fostering long-term adherence.

Framing the chatbot's language is a concrete way to design the intervention. It can act as *persuasion* by shaping how users interpret and emotionally respond to messages, and as *enablement* by fostering an environment where non-adherence and struggle are met with supportive, non-judgmental responses. For example, rather than stating "You missed your medication again", the chatbot could frame the message more constructively as "You're making progress! Adding consistency with your medication can help you feel even better". Thus, the underlying assumption is that a system perceived as attentive and other-oriented may help reduce the emotional barriers contributing to deceptive behavior.

Expressing Empathy

In the context of language-based interventions, empathic expressions have been found to positively influence users' perceptions of an agent's trustworthiness, supportiveness, and care [32]. These effects, however, are shaped by the context in which the interaction occurs.

In therapy settings, empathy has been shown to support better self-treatment, including improved self-esteem [33]. Moreover, empathy in clinical contexts has been linked to good effects on patients' anxiety levels and enablement [34]. Since patient enablement entails supporting individuals in their ability to manage their health [35], it can be seen as closely related to self-control.

Based on this evidence, one version of the intervention involves generating empathic responses that show understanding of the user's situation. These responses aim to support self-esteem and self-control, but also to reduce the perceived need to protect one's self-image by signaling acceptance. When users perceive that their experiences are acknowledged with care rather than evaluation, they may feel less compelled to present themselves in a particular way. This may foster a sense that even in low moments, they will be met with recognition and support, not criticism or emotional disregard.

Affirming the Self

Another approach to framing the chatbot's language involves affirming the user's values or efforts. Self-affirmations are acts that reinforce an individual's sense of adequacy [36]. This is particularly relevant in the context of diabetes, which functions as a chronic psychological stressor [26]. Admitting to lapses in self-management, such as non-adherence to medication, may trigger feelings of inadequacy or failure. Likewise, receiving feedback that highlights these lapses can be experienced as threatening. Timely self-affirmation has been shown to reduce defensiveness, enabling individuals to process such information more openly by lowering the threat it poses to their concept of self [37, 38].

This study builds on the premise that strengthening the sense of self may improve self-esteem and self-control, while also reducing the perceived need to protect one's self-image when facing challenges. Because self-affirmations are internally generated, their impact is understood to rely on individuals actively engaging with their own values and self-concept [38, 39]. However, affirming language from the system can complement this process by validating users' efforts and aligning responses with the values they have expressed through self-affirmations themselves.

3 Intervention Implementation

This section details the implementation of the intervention described in Section 2.3. CHIP was extended to interpret user input, reason over it, and generate responses according to different framing strategies. These functionalities were implemented by prompting a large language model (LLM) to perform natural language processing (NLP) tasks.

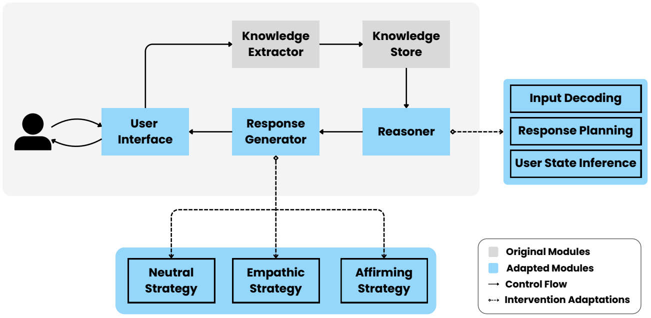
3.1 Foundations of the Implementation

The implementation of the interventions builds on CHIP[1], an open-source system developed within the Hybrid Intelligence (HI) project to support research into AI-driven DLM [15].

Structured for experimental use, the system adopts a modular architecture, as illustrated in Figure 1. The system decomposes the human-system interaction pipeline into distinct modules, each responsible for a key function such as interpreting user input and intent, reasoning over prior patient and domain knowledge, and generating responses. This modularity provides a strong foundation for implementing the interventions, as it reduces the need for low-level system design and allows the focus to remain on the research objectives.

The primary component extended to implement the intervention was the *Response Generator*. Although the Reasoner module within CHIP included a core structure for enabling reasoning functionality, its implementation remained limited to a minimal use case. As a result, it did not support more nuanced reasoning required to enable the intervention. To address this, I introduced a lightweight reasoning module. Additionally, minor modifications were made to the front-end module to adopt a minimal white-and-blue color scheme. This design choice was informed by the association of both blue and white with calmness [40]. The aim was to minimize the likelihood of visual elements interfering with the effectiveness of the language intervention and to contribute to a calm, non-judgmental perception of the interaction style.

Figure 1: Modular architecture of the CHIP system, with the adapted modules shown in blue. The reasoner module has three new processing sub-steps, and the response generator was modified to allow the use of three generation strategies.



3.2 Language Model Integration

The intervention prototype relies on NLP tasks for dialogue management, including intent detection and response generation. Several approaches were considered, and prompt engineering using an LLM was selected as a middle ground between rule-based techniques and model fine-tuning.

Rule-based methods (e.g., predefined input-output mappings using conditional logic) can be overly rigid and insufficiently expressive. They are unlikely to account for the diversity of user responses anticipated in real-world interactions, such as those expected during the pilot study. Conversely, fine-tuning language models was not pursued due to several practical constraints: limited development time, the exploratory nature of the study, and the large amount of data typically required to fine-tune such models effectively.

For the implementation, I selected `gemini-2.0-flash`, a commercial model. This decision was based on its strong performance in publicly available benchmarks [41], ease of integration, and a substantial free usage quota compared to other commercial providers such as OpenAI. Given the limited time resources, I was unable to experiment with open-source alternatives such as LLaMA. While the chosen model was appropriate for prototyping, it posed a limitation during the pilot study, as it occasionally became overloaded and failed to respond to requests. This was later identified as a known issue affecting the API [42, 43].

To enhance reproducibility and encourage controlled LLM outputs, two measures were applied. The `temperature` was set to 0 to promote deterministic behavior, and a fixed `seed` was used to minimize variability across calls. Although these steps reduce randomness considerably, achieving complete determinism in LLM outputs remains challenging [44].

As an additional step to guide the model's behavior, each sub-component that called the LLM used a system prompt [45, 46]. These prompts defined the model's role and context, specified the expected input format, and included detailed instructions to shape the model's reasoning and output.

[1] https://github.com/Archer6621/chip-demo

3.3 Reasoner

The *Reasoner*[2] module was extended to contextually interpret user input and perform reasoning that guides the system's responses. This is achieved through a three-step process, each executed by prompting the LLM with detailed task instructions. The expected output of the LLM for each step was defined using `JSON` schemas to reduce the ambiguity in the model's responses [45] and to promote their interpretability. Additionally, a preliminary rule-based dialogue state tracker has been implemented to preserve user context across dialogue turns. As such, the three steps in this module are:

1. **Input Decoding**: This step processes the raw user input by performing *intent classification* (e.g., identifying that the user is *logging* a meal) and *slot filling* (e.g., extracting the food item). It uses a schema tailored to diabetes-related interactions, informed by the structure of CHIP's patient knowledge graph, to keep the dialogue within the scope of the diabetes use case. The extracted data is used to update the conversational context incrementally.
2. **Response Planning**: This step uses the updated conversational context to determine the system's communicative intent, such as asking a question, which is recorded in the `response_type` field. It also identifies a minimally sufficient `value` that captures the main content of the response to be generated (e.g., `ask_portion_size`), along with a corresponding `reason` that justifies its relevance (e.g., the user logged a meal but did not specify its size).
3. **User State Inference**: This step leverages the two previously developed taxonomies (see Tables 1, 2) of factors underlying deception and poor diabetes self-management to infer whether such elements (e.g., lack of self-control) are indirectly present in the raw user input. These inferences are stored in the `soft_self_management_indicators` field and provide supplementary context for adapting the tone used during response generation.

At the end of this process, the Reasoner combines the outputs of the second and third steps with a relevant field from the conversational context to generate a structured dictionary (see Figure 2). This output contains the communicative intent (`response_type`), the minimal content needed in the response (`value`), and a justification for its relevance (`reason`). Additionally, it includes information to guide the behavioral interventions, such as `soft_self_management_indicators` and previously indicated `personal_values`.

Figure 2: Example of *Reasoner* output as a structured data dictionary.

```
{
  "response_type": "question",
  "value": "ask_portion_size",
  "reason": "user logged sugary cereals for breakfast
             but did not specify portion size",
  "soft_self_management_indicators": [
    "low self-control"
  ],
  "personal_values": ["hope"]
}
```

To provide the Response Generator with meaningful data for its framing strategies, I iteratively improved the prompts used for each step. Early prompts aimed to generate fully atomic data, but this stripped away important contextual cues. The instructions were revised to produce data that is as concise as possible while retaining context. Nevertheless, semantic loss remains a key limitation of the prototype, which is amplified by the simplicity of the dialogue state tracking.

3.4 Response Generator

The second adaptation done to CHIP to support the intervention prototype involved the Response Generator module[3]. CHIP's existing modular architecture inspired the implementation of the response framing, which follows a Strategy Design Pattern [47]. Each framing implementation is encapsulated as a separate strategy that can be selected dynamically at runtime through a configuration file. This includes both intervention strategies and a control strategy, which serves as a baseline for comparison during evaluations.

For each strategy, a system prompt and intent-specific task prompts were defined. The system prompt ensures consistency with the intervention's goals in tone, style, and framing. Each task prompt includes the specific objective, references the Reasoner output, and provides examples of the expected input-output behavior. Providing such examples is a prompt engineering technique known as few-shot prompting [45]. This method has been shown to improve language model performance compared to zero-shot prompting [48], where no exemplars are given.

The generation process entails transforming the data dictionary received from the Reasoner module (e.g., Figure 2) into a framed response expressed in natural language. The `response_type`, which captures the intent of the response to be created, is first used to delegate the generation to specific methods. Methods for intents such as *acknowledgement*, *answer*, and *question* prompt the LLM with their corresponding task prompts. In contrast, the remaining intents, *greeting* and *closing*, involve minimal variation and are implemented using hard-coded templates, without invoking the LLM.

The content of the responses primarily draws from the `value` field, and when appropriate, is contextualized using the `reason` field to enhance transparency. The indicators for self-management are used in the two intervention strategies to soften the tone. Additionally, the affirming intervention uses the `personal_values` field to align the response with what the user finds important. More details about the strategies are outlined below, with example responses in Table 3.

Neutral Framing: This *control* condition was designed to present only the value and its justification, without adapting to the user's emotional state or affirming their personal values.

Empathic Framing: This framing aims to express other-oriented empathy by conveying understanding toward the user, rather than mirroring their emotions. Such an approach has been found to elicit more positive user responses [32, 49]. The examples and generation rules I used were informed by research on empathy in human-intelligent system interac-

[2] https://github.com/marinamadaras/CHIPxDeception/tree/master/modules/reasoning-intervention

[3] https://github.com/marinamadaras/CHIPxDeception/tree/master/modules/response-generator-intervention

Table 3: Example responses to two user inputs, showing how different framings (neutral, empathic, and affirming) influence the tone and emphasis of the messages. Affirming responses also incorporate users' previously expressed self-affirmations.

Framing	**User Input 1:** *Work has been making it super hard for me to exercise regularly*	**User Input 2:** *I ate too many snacks*
NEUTRAL	Noted: your work schedule is a potential barrier.	What time did you eat?
EMPATHIC	I understand that your work schedule is affecting your physical activity. Has that been a long-term challenge?	To make sure I'm on the right track, when did you have that snack?
AFFIRMING[A]	It sounds like your work schedule can get in the way. It's still meaningful that you're thinking about how to make space for physical activity.	What time did you have your snack? Sometimes noticing the details can help you understand your own patterns a little better.

[A] the users' previously expressed core values, *friendship* and *discipline* (Input 1), and *passion* and *hope* (Input 2).

tion [49]. Additionally, in cases where the Reasoner selects `acknowledgement` as a semantic intent, typically because no additional user input is required, empathy is reinforced by complementing the acknowledgement with an empathic question that signals active listening.

Affirming Framing: This strategy begins conversations with a question that encourages users to reflect on their personal values, thereby prompting self-affirmation. This technique was shown to reduce defensiveness in response to challenging health information [38]. The interaction then continues in a supportive tone, with messages framed to affirm the user and align with their expressed values.

4 Pilot Study Methodology

This section outlines the methodology of an exploratory pilot study conducted to investigate whether different response framings of a diabetes lifestyle management (DLM) system influence users' tendency to provide deceptive self-reports. Participants of the study interacted with a DLM prototype in a simulated diabetes self-management scenario, where they role-played as patients struggling with the condition.

4.1 Study Design

The pilot study hypothesized that *participants exposed to the empathic or affirming intervention would report higher self-esteem and lower self-presentation concerns than those exposed to the neutral intervention*, as these factors have been identified as potential drivers of deception (see Section 2.2).

To evaluate the hypothesis, the study was set up as a controlled experiment, allowing for a structured comparison of different framings. Participants were randomly assigned to one of three framing conditions in a between-subjects design. A within-subjects approach was less suitable, as the brief interaction with the prototype could have led to carry-over effects if participants were exposed to multiple framings. Given this setup, the following variables were controlled:

1. **Independent Variable**: The type of intervention, with three framing strategies: neutral, empathic, and affirming.
2. **Dependent Variables**: The measured outcomes were self-esteem and self-image protection. Further details on the measurement methods are provided in Section 4.4.
3. **Confounding Variables**: The identified sources of bias and their mitigation strategies are listed in Table 4.

Table 4: Confounding factors identified in the context of the pilot study and their corresponding mitigation strategies.

Confounding Factor	**Mitigation Strategy**
Non-diabetic participants	Provided background information on diabetes management before the interaction.
Reasoner limitations	Instructed participants to be patient and to phrase their input clearly and expressively.
Familiarity with researcher	Emphasized that responses were anonymous and interactions were not stored.

4.2 Participants

The study aimed to recruit 12 participants per condition, following recommendations for pilot study sample sizes in clinical research [50]. In practice, however, the prototype suffered from stability and performance issues, which limited the depth and consistency of user interactions. These technical shortcomings, combined with the previously identified confounding factors (see Table 4), were expected to impact the reliability of the collected measures.

Consequently, the study proceeded with a reduced sample of four participants per condition (12 in total), recruited through personal networks. This approach still allowed for exploratory outcomes under the given limitations. Since the study was not designed to produce broadly generalizable results, demographic data were not collected, and gender balance was not controlled for.

4.3 Procedure

To ensure consistency across participants, I followed a fixed procedure, detailed below. All necessary materials and technical instructions are available in the project documentation.[4]

1. **Prototype setup**: The prototype was run locally for each session using the setup described in the project documentation. The framing condition (`empathic`, `affirming`, or `neutral`) was manually set in the configuration file before the participants interacted with CHIP.
2. **Informed consent**: Participants received a consent form and could ask questions before signing.
3. **Participant Instructions**: Participants received a User Context sheet (see Appendix A) describing the diabetes self-management scenario and instructing them to role-play as a patient struggling with the condition.

[4] https://github.com/marinamadaras/CHIPxDeception/blob/master/documentation

4. **Interaction**: Participants greeted CHIP and engaged in a conversation for at least five turns, with a maximum time limit of ten minutes. The interaction followed the steps outlined in the User Context sheet.
5. **Questionnaire**: After the interaction, participants completed a digital questionnaire via Microsoft Forms. The content was identical across conditions, but three separate instances were used to distinguish responses by condition.

4.4 Measures

The study aimed to investigate the effects of framing interventions on deceptive self-reporting. However, actual deception was not directly measurable: participants only simulated the role of a diabetic patient, and therefore could not reflect on genuine deceptive intent. Moreover, deception is inherently difficult to verify, given the lack of universal cues and its dependence on individual and contextual factors [25].

Instead, two subjective measures were chosen: self-esteem and self-image protection, which align with two of the three intervention targets defined in Section 2.3. These measures can reveal whether the intervention affected participants' attitudes or self-presentation, even without directly capturing actual deceptive behavior. The third target, self-control, was excluded to prevent the post-interaction questionnaire from becoming disproportionately long relative to the brief session, which could have introduced participant fatigue.

To measure self-esteem, the *Brief Rosenberg Self-Esteem Scale* (B-RSES) [30] was used. This is a validated short form of the original *Rosenberg Self-Esteem Scale* [51, 52], which has been widely used to assess self-esteem [53]. The State Self-Esteem Scale [54] was considered as an alternative but excluded, as its phrasing was found to be more personal and less suited for the simulated nature of the interaction.

To measure self-image protection, the *Balanced Inventory of Desirable Responding (BIDR)* Scale [55] was considered, as it captures two dimensions particularly relevant to this construct: *self-deceptive positivity* and *impression management*. To reduce participant burden, the validated short form BIDR-16 [29] was used, which preserves both dimensions.

Alongside the two scales, the questionnaire included an open-ended question to gather qualitative feedback on the interaction, as well as two control questions to assess whether any technical issues occurred.

4.5 Data Preparation and Analysis

Before analysis, the questionnaire data were pre-processed to ensure participant anonymity. Any potentially identifying information, such as timestamps or responses that could indirectly reveal participants' identities, was removed. This step was carried out in accordance with the ethical protocol approved by the Human Ethics Comittee (HREC) of the Delft University of Technology for the study. During this process, it was found that the B-RSES questions used in this study followed a 5-point Likert scale, contrary to the intended 4-point format. To maintain consistency with the scoring guidelines, the responses were normalized to align with a 4-point scale.

Quantitative scores were calculated for each scale using continuous scoring, with certain items reverse-scored according to official guidelines [29, 52]. To support interpretation, descriptive statistics such as the mean and standard deviation were computed for each scale. To compare these means, an appropriate statistical test given the study design would have been MANOVA. However, the dataset in this pilot was not sufficiently large and may have been influenced by confounding variables, which limited the suitability of such analyses. As a result, the hypothesis was not formally tested, and the analysis focused on identifying general trends in the data.

In addition to the scaled items, the questionnaire included one open-ended question. A qualitative analysis was conducted on the responses to identify recurring patterns. Finally, the control questions were aggregated to assess the technical stability of the prototype and were used to support the interpretation of both the quantitative and qualitative findings.

5 Pilot Study Results and Discussion

The questionnaire scores from a sample of 12 participants were summarized using descriptive statistics, as shown in Table 5. On the Brief Rosenberg Self-Esteem Scale (B-RSES), where higher scores indicate greater self-esteem, participants in the *affirming* condition had the highest mean score, followed by those in the *empathic* and *neutral* conditions. For the BIDR-16 scale, where lower scores are associated with less socially desirable responding, the *empathic* condition showed the lowest average score, while the *neutral* condition and the *affirming* condition had slightly higher means.

Table 5: Descriptive statistics by condition of self-esteem (B-RSES) and social desirability (BIDR-16). The results are exploratory (n = 12).

Condition	**B-RSES**		**BIDR-16**	
	M	**SD**	**M**	**SD**
Empathic	**2.16**	0.41	**3.50**	0.11
Affirming	2.61	0.36	4.38	0.46
Neutral	1.86	0.31	4.14	0.49

The analysis of the qualitative data suggests that the perception of the language used by CHIP largely aligned with the intended mechanisms of each framing strategy. This supports the idea that the linguistic interventions were perceptible and had the potential to serve as explicit persuasive techniques (see Section 2.3). In the neutral condition, the language was described as impersonal and emotionally flat. The empathic framing was experienced as gentle and non-judgmental, which appeared to help some participants feel more comfortable opening up about their experiences. Similarly, the affirming strategy was perceived as kind and friendly, encouraging users to share their struggles.

However, these impressions were not consistent across all participants. The flow of the interactions was disrupted when CHIP became unresponsive, which occurred in 8 out of the 12 cases. Additionally, many participants noted that the conversations lacked a coherent thread or sense of continuity. Finally, several participants mentioned during the session that the questionnaires felt somewhat out of place, particularly given the brief nature of the interaction and the limitations of simulating an unfamiliar health context.

These results tentatively suggest that empathic framing may enhance self-esteem while reducing self-image protection relative to neutral framing, in line with the intervention's hypothesized effect. While the affirming condition also showed increased self-esteem, it was linked to higher levels of self-image protection. This may hint that by framing the interaction as emotionally open from the start, self-affirmations can make participants more aware of how they are perceived. However, given the sample size and exploratory nature of the study, it is not possible to determine whether these patterns reflect meaningful effects or are due to outliers.

While the hypothesis was not validated, the results should not be interpreted as evidence against the study's premise. When considered in light of the strong theoretical background and feasibility constraints, they *support the recommendation for a second iteration of the study*. To this end, Section 7 outlines improvements proposed for a follow-up study.

6 Limitations

This study has limitations that should be acknowledged. The literature review was not conducted systematically, and a single annotator coded the taxonomies derived from it. This may have introduced bias into the theoretical background and led to the omission of essential aspects. A second limitation is the participant sample: without diabetic patients, the study could not assess how those affected by the condition would engage with the intervention or interpret the questionnaires. Additionally, this study evaluated factors driving deceptive behavior rather than the behavior itself, due to the inherent challenges of measuring deception directly [25, 56].

Another limitation was the prototype's lack of robustness, which reduced interaction quality and made interpreting measures more difficult across framings. When CHIP's responses lacked contextual relevance, it was unclear whether questionnaire answers reflected the framing's effect or the system's perceived misunderstanding. Finally, although the Gemini model typically performs well in latency benchmarks [41], it was occasionally unavailable during the study [42, 43].

7 Future Work

The results of the pilot study led to the recommendation of carrying out a second iteration for a more effective evaluation of the intervention strategies. This section outlines specific improvements to support that goal in future work.

Prototype Improvements The prototype could be improved by refining the logic used to track the dialogue context, so that it follows the conversation thread more effectively. Additionally, the rules used to guide the large language model in choosing the semantic intent of the system's response could be made more robust. These changes would keep the chatbot's responses more contextually relevant.

Participant Recruitment A second iteration of the user study should exclusively recruit individuals diagnosed with diabetes. Additionally, since the study explores differences in language framing, only English-speaking participants should be included to avoid confounds related to language comprehension. Based on an expected medium effect size, a power analysis conducted using the G*Power tool [57] indicates that a minimal sample of 126 participants would be needed to detect statistically significant differences (see Appendix B).

Procedure The procedure could be expanded upon by adopting a longitudinal design. Participants would first complete the same questionnaire, which includes the B-RSES and BIDR-16 scales. Their responses would serve as a baseline for measuring any changes over time. Participants would then be asked to interact with CHIP independently, outside of a controlled setting, four times over the course of one week. This frequency is suggested to ensure feasibility and reduce concerns about long-term adherence. After one week, participants would complete the same questionnaire to evaluate short-term changes. To explore whether the effects persist over time, they could be asked to fill out the questionnaire again, two weeks after the initial interaction period.

Data Analysis To analyze differences in responses across conditions and over time, a *repeated measures MANOVA* (Multivariate Analysis of Variance) [58] could be carried out. This method is appropriate given the study design, which includes one independent variable (framing condition), two dependent variables (self-esteem and self-image protection), and three measurement points (pre-, post-, and follow-up).

8 Conclusions

This research focused on deceptive behavior in diabetes lifestyle management (DLM) systems, aiming to identify, implement, and evaluate mechanisms that might help prevent it. To explore the underlying causes of such behavior and what might drive patients to struggle with self-management, the Behavior Change Wheel (BCW) framework was used as a methodological approach. Notably, several overlapping factors were identified, including low self-esteem, negative emotions, and lack of self-control. These insights informed the design of two interventions, empathic and affirming, which were implemented as response-framing strategies in CHIP, a chatbot-based DLM prototype. For this, a large language model (LLM) was used to interpret user input and generate responses that aligned with the framing strategies.

A pilot user study was conducted to evaluate the effectiveness of the interventions. Although the small sample size limited statistical power, exploratory findings suggest that empathic language can help boost self-esteem and reduce users' need to protect their self-image. Participants also noted that the tone of the messages made the system feel friendly and non-judgmental. This suggests that using softer and more supportive language could help create a space where users feel more comfortable being open. These results underscore the need for a second iteration to assess the intervention's effectiveness more accurately, which in turn informs directions for improving CHIP and refining the experiment.

Thus, the question remains: *to deceive or self-deceive?* This work, however, moves us toward the possibility that, when support is framed in the right way, patients may feel there is no need for either.

9 Responsible Research

To support responsible research practices, this section addresses the ethical considerations surrounding this work, its compliance with the European Union Artificial Intelligence Act (EU AI Act), the measures taken to improve reproducibility, and a disclaimer on the use of generative AI tools.

9.1 Ethical Considerations

Accessibility

Digital health tools, such as CHIP, can support diabetes self-management and have a positive impact on patients. Despite this, using them requires access to digital devices and stable internet connections, which are unavailable to millions worldwide [59, 60]. As a result, such interventions risk widening the digital divide and excluding those who may already face barriers to healthcare. While this cannot be directly addressed within the scope of this work, it emphasizes the need to consider broader structural barriers when developing such tools.

Moreover, general and health literacy levels among individuals managing diabetes can also act as barriers to effective self-management [26]. While the implemented prototype did not explicitly address these challenges, they may limit users' ability to benefit from language-based interventions. Future research should consider making interventions adaptive to users' literacy levels, for example by adjusting language complexity or providing additional support based on individual needs, in order to improve accessibility.

Integration of Large Language Models

Integrating large language models (LLMs) into tools raises important ethical concerns. In this case, the model I used (Gemini) is proprietary, which offers limited control over how user data is handled. To mitigate this during the study, participants were advised not to share personal information. If the system were to be used in practice, this decision would need to be re-evaluated to ensure transparency and to protect user privacy through greater control over data handling.

In addition to data concerns, it must be acknowledged that LLMs are prone to hallucinations, potentially generating inaccurate or misleading information. In the context of digital health tools, this can be particularly harmful if users take the advice at face value, as it may unintentionally reinforce unhealthy behaviors. During the user study, a disclaimer was included to make participants aware of this risk.

Furthermore, the implementation of the reasoner involves prompting the LLM to justify its choices, but the soundness of the explanations cannot be assumed. While the prompts were structured to encourage consistent outputs and limit randomness, the underlying decision process is not fully observable or interpretable, as explainable AI remains an open challenge.

User Study

Before the pilot user study was conducted, an ethics protocol was approved by the Human Research Ethics Committee (HREC) of the Delft University of Technology (approval number *5739*). This included the creation of a risk assessment plan, an informed consent form, and a data management plan. As such, multiple decisions regarding data privacy and potential risks associated with the study were explicitly considered.

Some of these considerations included allowing participants to withdraw from the study at any time, not storing any conversation data from interactions with the chatbot, and ensuring that all collected data was anonymized. During the study, participants were presented with the informed consent form before interacting with CHIP. They were allowed to ask questions before signing it.

9.2 Compliance with EU AI Act

The EU Artificial Intelligence Act [61] defines a set of rules for the responsible development and deployment of AI systems. Given that this work implemented an AI-based system, it is important to consider its compliance with the Act. According to Article 2, since the CHIP prototype was part of a research project, the full scope of the regulations does not apply. Nonetheless, an analysis was conducted to better understand the current and longer-term implications of the work.

The CHIP prototype does not explicitly fall under the AI Act's definition of a high-risk system; however, some of its functionality reflects criteria treated with particular caution in the regulation. As a DLM system, the prototype *profiles* [62] users based on self-reported behaviors to adapt its support, and infers the potential emotional states of users to enable the behavioral interventions. The interventions can raise concerns under Article 5 of the AI Act [61], which prohibits AI systems that manipulate users through subliminal techniques or exploit vulnerabilities in a harmful manner. The article also restricts systems that infer or identify emotional states in educational or workplace settings. However, the goals of this work, to prevent deception and improve adherence, aim to encourage behaviors that are considered *beneficial* within a *medical* context. Accordingly, Recitals 29 and 44 clarify that such medical and therapeutic practices are not prohibited if they comply with legal requirements, medical standards, and are carried out with the user's explicit consent.

While the current implementation is a proof of concept, its possible integration with other modules developed by peers as part of the same research project, which focus on detecting deception and non-adherence in DLM systems, would add additional layers of user profiling. A strong medical basis and legal considerations should guide such integration. Moreover, if deployed outside the research context, the system's transparency toward users must be prioritized, in line with the requirements of Article 50. This includes clearly informing users about the nature and purpose of the *profiling* used by the DLM tool.

9.3 Reproducibility

Reproducibility of Literature Review and Taxonomies

The literature review that informs the theoretical foundation of this work is not systematic and thus not fully reproducible. While a systematic review was not feasible due to time and resource constraints, it is important to acknowledge that the selection of sources might have introduced bias. Furthermore, the interpretation of the reviewed literature and the coding process may reflect individual perspectives. To address this, I have aimed to provide a transparent explanation of the choices made throughout in Section 2.

Reproducibility of Implementation

The implementation of the CHIP prototype is open source[5], with all necessary setup and execution instructions provided in the documentation. To support reproducibility, the configuration of the LLM calls within the implementation included a `temperature` value set to 0 and a fixed `seed`. Although these settings were intended to reduce variability in the chatbot's responses, large language models are not fully deterministic, and the underlying model may change without notice; thus, variation can still occur.

Reproducibility of User Study

Section 4 details the methodology of the pilot user study, including the procedure, data processing, and analysis steps. All necessary materials and steps were made available to enable others to reproduce the setup. However, because the study relied on self-reported measures, which are inherently subjective and sensitive to contextual factors, a second run of the pilot study will not yield identical outcomes.

9.4 Use of Generative AI

To increase the transparency about the research and writing process, I acknowledge that a generative AI tool (ChatGPT, `GPT-4o`) was used to assist with editing the language of content that I had already written. This included prompting the tool to rephrase words such that they are not repetitive (e.g., "*What is a synonym for ... in this context?*") or to help with rephrasing parts of existing sentences to improve their clarity (e.g., "*How can I make this sound more clear?*"). Before using them, the model's outputs were critically evaluated to ensure that they remained factually correct and contextually appropriate. No suggestions that modified the intended meaning were used, and all ideas, arguments, and claims in this work are entirely my own. The tool was not used to generate any new content.

It should be noted, however, that the developed prototype is based on an LLM, and as such, its outputs are inherently AI-generated language. The examples of CHIP's responses are only included to illustrate the system's behavior, while the prompt design and research were my own work.

References

[1] A. Wilkinson, L. Whitehead, and L. Ritchie, "Factors influencing the ability to self-manage diabetes for adults living with type 1 or 2 diabetes," *International Journal of Nursing Studies*, vol. 51, no. 1, pp. 111–122, 2014. DOI: 10.1016/j.ijnurstu.2013.01.006.

[2] World Health Organization, *Universal health coverage (UHC)*, Mar. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc) (visited on 04/22/2025).

[3] GSMA, *Global smartphone penetration rate as share of population from 2016 to 2024 [Graph]*, Feb. 2025. [Online]. Available: https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/ (visited on 04/21/2025).

[4] Appfigures, *Number of mHealth apps available in the Google Play Store from 1st quarter 2015 to 2nd quarter 2024 [Graph]*, Sep. 2024. [Online]. Available: https://www.statista.com/statistics/779919/health-apps-available-google-play-worldwide/ (visited on 06/15/2025).

[5] Appfigures, *Number of mHealth apps available in the Apple App Store from 1st quarter 2015 to 2nd quarter 2024 [Graph]*, Sep. 2024. [Online]. Available: https://www.statista.com/statistics/779910/health-apps-available-ios-worldwide/ (visited on 05/23/2025).

[6] L. P. A. Simons, H. Pijl, J. Verhoef, H. J. Lamb, B. v. Ommen, B. Gerritsen, M. B. Bizino, M. Snel, R. Feenstra, and C. M. Jonker, "Intensive Lifestyle (e)Support to Reverse Diabetes-2," in *BLED 2016 Proceedings*, 2016. [Online]. Available: https://aisel.aisnet.org/bled2016/24.

[7] V. Mogre, N. A. Johnson, F. Tzelepis, J. E. Shaw, and C. Paul, "A systematic review of adherence to diabetes self-care behaviours: Evidence from low-and middle-income countries," *Journal of Advanced Nursing*, vol. 75, no. 12, pp. 3374–3389, 2019, Publisher: Wiley. DOI: 10.1111/jan.14190.

[8] S. R. Shrivastava, P. S. Shrivastava, and J. Ramasamy, "Role of self-care in management of diabetes mellitus," *Journal of Diabetes & Metabolic Disorders*, vol. 12, p. 14, 2013, Publisher: BioMed Central. DOI: 10.1186/2251-6581-12-14.

[9] R. S. Mazze, H. Shamoon, R. Pasmantier, D. Lucido, J. Murphy, K. Hartmann, V. Kuykendall, and W. Lopatin, "Reliability of blood glucose monitoring by patients with diabetes mellitus," eng, *The American Journal of Medicine*, vol. 77, no. 2, pp. 211–217, 1984. DOI: 10.1016/0002-9343(84)90693-4.

[10] E. Armas-Vargas, R. J. Marrero, and J. Hernández-Cabrera, "Psychometric properties of the CEMA-A questionnaire: Motives for lying," *Frontiers in Psychology*, vol. 14, 2023. DOI: 10.3389/fpsyg.2023.1289209.

[11] D. Buller and J. K. Burgoon, "Interpersonal Deception Theory," *Communication Theory*, vol. 6, no. 3, pp. 203–242, 1996. DOI: 10.1111/j.1468-2885.1996.tb00127.x.

[12] C. Biener and A. Waeber, "Would I lie to you? How interaction with chatbots induces dishonesty," *Journal of Behavioral and Experimental Economics*, vol. 112, p. 102 279, 2024. DOI: 10.1016/j.socec.2024.102279.

[13] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978. DOI: 10.1017/S0140525X00076512.

[14] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Ver-

[5] https://github.com/marinamadaras/CHIPxDeception

heij, P. Vossen, and M. Welling, "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence," *Computer*, vol. 53, no. 8, pp. 18–28, 2020. DOI: 10.1109/MC.2020.2996587.

[15] B. J. W. Dudzik, J. S. v. d. Waa, P.-Y. Chen, R. Dobbe, Í. M. D. R. d. Troya, R. M. Bakker, M. H. T. d. Boer, Q. T. S. Smit, D. Dell'Anna, E. Erdogan, P. Yolum, S. Wang, S. B. Santamaria, L. Krause, and B. A. Kamphorst, "Viewpoint: Hybrid Intelligence Supports Application Development for Diabetes Lifestyle Management," en, *Journal of Artificial Intelligence Research*, vol. 80, pp. 919–929, 2024. DOI: 10.1613/jair.1.15916.

[16] A. M. van Valkengoed, W. Abrahamse, and L. Steg, "To select effective interventions for pro-environmental behaviour change, we need to consider determinants of behaviour," eng, *Nature Human Behaviour*, vol. 6, pp. 1482–1492, 2022. DOI: 10.1038/s41562-022-01473-w.

[17] S. Michie, M. M. van Stralen, and R. West, "The behaviour change wheel: A new method for characterising and designing behaviour change interventions," *Implementation Science*, vol. 6, p. 42, 2011. DOI: 10.1186/1748-5908-6-42.

[18] C. Mangurian, G. C. Niu, D. Schillinger, J. W. Newcomer, J. Dilley, and M. A. Handley, "Utilization of the Behavior Change Wheel framework to develop a model to improve cardiometabolic screening for people with severe mental illness," *Implementation Science*, vol. 12, p. 134, 2017. DOI: 10.1186/s13012-017-0663-z.

[19] E. M. Murtagh, A. T. Barnes, J. McMullen, and P. J. Morgan, "Mothers and teenage daughters walking to health: Using the behaviour change wheel to develop an intervention to improve adolescent girls' physical activity," *Public Health*, vol. 158, pp. 37–46, 2018. DOI: 10.1016/j.puhe.2018.01.012.

[20] D. Y. Wang, E.-Y. Wong, A.-L. Cheung, Z.-Y. Tam, K.-S. Tang, and E.-K. Yeoh, "Enhancing implementation of information and communication technologies for post-discharge care among hospitalized older adult patients: Development of a multifaceted implementation intervention package using the behavior change wheel and implementation research logic model," *Implementation Science Communications*, vol. 6, 2025. DOI: 10.1186/s43058-025-00739-4.

[21] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, "Lying in Everyday Life," *Journal of Personality and Social Psychology*, vol. 70, pp. 979–95, 1996. DOI: 10.1037/0022-3514.70.5.979.

[22] A. Błachnio, "Be Happy, Be Honest: The Role of Self-Control, Self-Beliefs, and Satisfaction with Life in Honest Behavior," *Journal of Religion and Health*, vol. 60, no. 2, pp. 1015–1028, 2021, Publisher: Springer Nature. DOI: 10.1007/s10943-019-00956-x.

[23] N. Mazar and D. Ariely, "Dishonesty in Everyday Life and Its Policy Implications," *Journal of Public Policy & Marketing*, vol. 25, no. 1, pp. 117–126, 2006. DOI: 10.1509/jppm.25.1.117.

[24] U. Gneezy, "Deception: The Role of Consequences," *American Economic Review*, vol. 95, no. 1, pp. 384–394, 2005. DOI: 10.1257/0002828053828662.

[25] J. Meibauer, Ed., *The Oxford Handbook of Lying*. Oxford University Press, 2018, ISBN: 978-0-19-873657-8. DOI: 10.1093/oxfordhb/9780198736578.001.0001.

[26] J. S. Gonzalez, M. L. Tanenbaum, and P. V. Commissariat, "Psychosocial factors in medication adherence and diabetes self-management: Implications for research and practice," *American Psychologist*, vol. 71, no. 7, pp. 539–551, 2016. DOI: 10.1037/a0040388.

[27] S. Gatt and R. Sammut, "An exploratory study of predictors of self-care behaviour in persons with type 2 diabetes," *International Journal of Nursing Studies*, vol. 45, no. 10, pp. 1525–1533, 2008. DOI: 10.1016/j.ijnurstu.2008.02.006.

[28] S. Heo, J. Kang, T. Barbé, J. Kim, T. F. Bertulfo, P. Troyan, D. Stewart, and E. Umeakunne, "Relationships of multidimensional factors to self-management in patients with diabetes: A Cross-sectional, correlational study," *Geriatric Nursing*, vol. 55, pp. 270–276, 2024. DOI: 10.1016/j.gerinurse.2023.11.020.

[29] C. M. Hart, T. D. Ritchie, E. G. Hepper, and J. E. Gebauer, "The Balanced Inventory of Desirable Responding Short Form (BIDR-16)," *SAGE Open*, vol. 5, no. 4, 2015. DOI: 10.1177/2158244015621113.

[30] R. P. Monteiro, G. Lins de Holanda Coelho, P. Hanel, E. Medeiros, and P. da Silva, "The Efficient Assessment of Self-Esteem: Proposing the Brief Rosenberg Self-Esteem Scale," *Applied Research in Quality of Life*, 2021. DOI: 10.1007/s11482-021-09936-4.

[31] P. D. Manapat, M. C. Edwards, D. P. MacKinnon, R. A. Poldrack, and L. A. Marsch, "A Psychometric Analysis of the Brief Self-Control Scale," *Assessment*, vol. 28, no. 2, pp. 395–412, 2021. DOI: 10.1177/1073191119890021.

[32] S. Brave, C. Nass, and K. Hutchinson, "Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent," *International Journal of Human-Computer Studies*, Subtle expressivity for characters and robots, vol. 62, no. 2, pp. 161–178, 2005. DOI: 10.1016/j.ijhcs.2004.11.002.

[33] J. C. Watson, P. L. Steckley, and E. J. McMullen, "The role of empathy in promoting change," *Psychotherapy Research*, vol. 24, 2013. DOI: 10.1080/10503307.2013.802823.

[34] F. Derksen, J. Bensing, and A. Lagro-Janssen, "Effectiveness of empathy in general practice: A systemati review," *British journal of general practice*, vol. 63, no. 606, e76–e84, 2013. DOI: 10.3399/bjgp13X660814.

[35] C. Hudon, D. St-Cyr Tribble, F. Légaré, G. Bravo, M. Fortin, and J. Almirall, "Assessing enablement in clinical practice: A systematic review of available instruments," *Journal of Evaluation in Clinical Practice*, vol. 16, no. 6, pp. 1301–1308, 2010. DOI: 10.1111/j.1365-2753.2009.01332.x.

[36] C. M. Steele, "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed., vol. 21, 1988, pp. 261–302. DOI: 10.1016/S0065-2601(08)60229-4.

[37] W. M. P. Klein, P. R. Harris, R. A. Ferrer, and L. E. Zajac, "Feelings of vulnerability in response to threatening messages: Effects of self-affirmation," *Journal of Experimental Social Psychology*, vol. 47, no. 6, pp. 1237–1242, 2011. DOI: https://doi.org/10.1016/j.jesp.2011.05.005.

[38] G. L. Cohen and D. K. Sherman, "The Psychology of Change: Self-Affirmation and Social Psychological Intervention," *Annual review of psychology*, vol. 65, pp. 333–371, 2014. DOI: 10.1146/annurev-psych-010213-115137.

[39] D. K. Sherman, K. A. Hartson, K. R. Binning, V. Purdie-Vaughns, J. Garcia, S. Taborsky-Barba, S. Tomassetti, A. D. Nussbaum, and G. L. Cohen, "Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat," *Journal of Personality and Social Psychology*, vol. 104, no. 4, pp. 591–618, 2013. DOI: 10.1037/a0031495.

[40] D. Jonauskaite and C. Mohr, "Do we feel colours? A systematic review of 128 years of psychological research linking colours and emotions," *Psychonomic Bulletin & Review*, 2025. DOI: 10.3758/s13423-024-02615-z.

[41] *Gemini 2.0 Flash: API Provider Performance Benchmarking & Price Analysis — Artificial Analysis*, en. [Online]. Available: https://artificialanalysis.ai/models/gemini-2-0-flash (visited on 06/08/2025).

[42] *Model is overloaded - - Gemini API*, Section: Gemini API, Jan. 2025. [Online]. Available: https://discuss.ai.google.dev/t/model-is-overloaded/59817 (visited on 06/08/2025).

[43] *Error: The model is overloaded - Gemini API*, Section: Gemini API, Nov. 2024. [Online]. Available: https://discuss.ai.google.dev/t/error-the-model-is-overloaded/48410 (visited on 06/08/2025).

[44] B. Atil, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, G. R. Rajagopal, A. Sloan, T. Tudrej, F. Ture, Z. Wu, L. Xu, and B. Baldwin, *Non-Determinism of "Deterministic" LLM Settings*, _eprint: 2408.04667, 2025. [Online]. Available: https://arxiv.org/abs/2408.04667 (visited on 05/28/2025).

[45] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. Da Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, and P. Resnik, *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*, arXiv:2406.06608 [cs], 2025. DOI: 10.48550/arXiv.2406.06608.

[46] L. Zhang, T. Ergen, L. Logeswaran, M. Lee, and D. Jurgens, *SPRIG: Improving Large Language Model Performance by System Prompt Optimization*, arXiv:2410.14826 [cs], 2024. DOI: 10.48550/arXiv.2410.14826. [Online]. Available: http://arxiv.org/abs/2410.14826 (visited on 06/08/2025).

[47] *Strategy*, en. [Online]. Available: https://refactoring.guru/design-patterns/strategy (visited on 06/06/2025).

[48] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language Models are Few-Shot Learners*, arXiv:2005.14165 [cs], 2020. DOI: 10.48550/arXiv.2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165 (visited on 06/06/2025).

[49] J. Urakami, B. A. Moore, S. Sutthithatip, and S. Park, "Users' Perception of Empathic Expressions by an Advanced Intelligent System," in *Proceedings of the 7th International Conference on Human-Agent Interaction*, ser. HAI '19, event-place: Kyoto, Japan, New York, NY, USA: Association for Computing Machinery, 2019, pp. 11–18, ISBN: 978-1-4503-6922-0. DOI: 10.1145/3349537.3351895.

[50] S. A. Julious, "Sample size of 12 per group rule of thumb for a pilot study," *Pharmaceutical Statistics*, vol. 4, no. 4, pp. 287–291, 2005. DOI: 10.1002/pst.185.

[51] M. Rosenberg, "The Measurement of Self-Esteem," in *Society and the Adolescent Self-Image*, Princeton University Press, 1965, pp. 16–36, ISBN: 978-0-691-09335-2.

[52] M. Rosenberg, "Rosenberg Self-Esteem Scale (RSE)," en, 2006. [Online]. Available: https://www.apa.org/obesity-guideline/rosenberg-self-esteem.pdf (visited on 05/15/2025).

[53] J. Blascovich and J. Tomaka, "Measures of Self-Esteem," in *Measures of Personality and Social Psychological Attitudes*, Elsevier, 1991, pp. 115–160, ISBN: 978-0-12-590241-0. DOI: 10.1016/b978-0-12-590241-0.50008-3.

[54] T. F. Heatherton and J. Polivy, "Development and Validation of a Scale for Measuring State Self-Esteem," *Journal of Personality and Social Psychology*, vol. 60, no. 6, pp. 895–910, 1991. DOI: 10.1037/0022-3514.60.6.895.

[55] D. L. Paulhus, "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, J. P. Robinson, P. R. Shaver, and L. S.

Wrightsman, Eds., Academic Press, 1991, pp. 17–59, ISBN: 978-0-12-590241-0. DOI: https://doi.org/10.1016/B978-0-12-590241-0.50006-X.

[56] V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer, "Linguistic Cues to Deception Assessed by Computer Programs: A Meta-Analysis," in *Proceedings of the Workshop on Computational Approaches to Deception Detection*, E. Fitzpatrick, J. Bachenko, and T. Fornaciari, Eds., Avignon, France: Association for Computational Linguistics, 2012, pp. 1–4. [Online]. Available: https://aclanthology.org/W12-0401/.

[57] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," en, *Behavior Research Methods*, vol. 41, pp. 1149–1160, 2009. DOI: 10.3758/BRM.41.4.1149.

[58] R. G. O'Brien and M. K. Kaiser, "MANOVA method for analyzing repeated measures designs: An extensive primer," *Psychological Bulletin*, vol. 97, no. 2, pp. 316–333, 1985. DOI: 10.1037/0033-2909.97.2.316.

[59] *What Is the Digital Divide?* 2022. [Online]. Available: https://ctu.ieee.org/blog/2022/12/14/what-is-the-digital-divide/ (visited on 06/12/2025).

[60] *Digital Around the World.* [Online]. Available: https://datareportal.com/global-digital-overview (visited on 06/12/2025).

[61] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*, Legislative Body: CONSIL, EP, Jun. 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj/eng (visited on 06/20/2025).

[62] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, Legislative Body: OP_DATPRO, May 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng (visited on 06/20/2025).

Appendices

A User Context Sheet

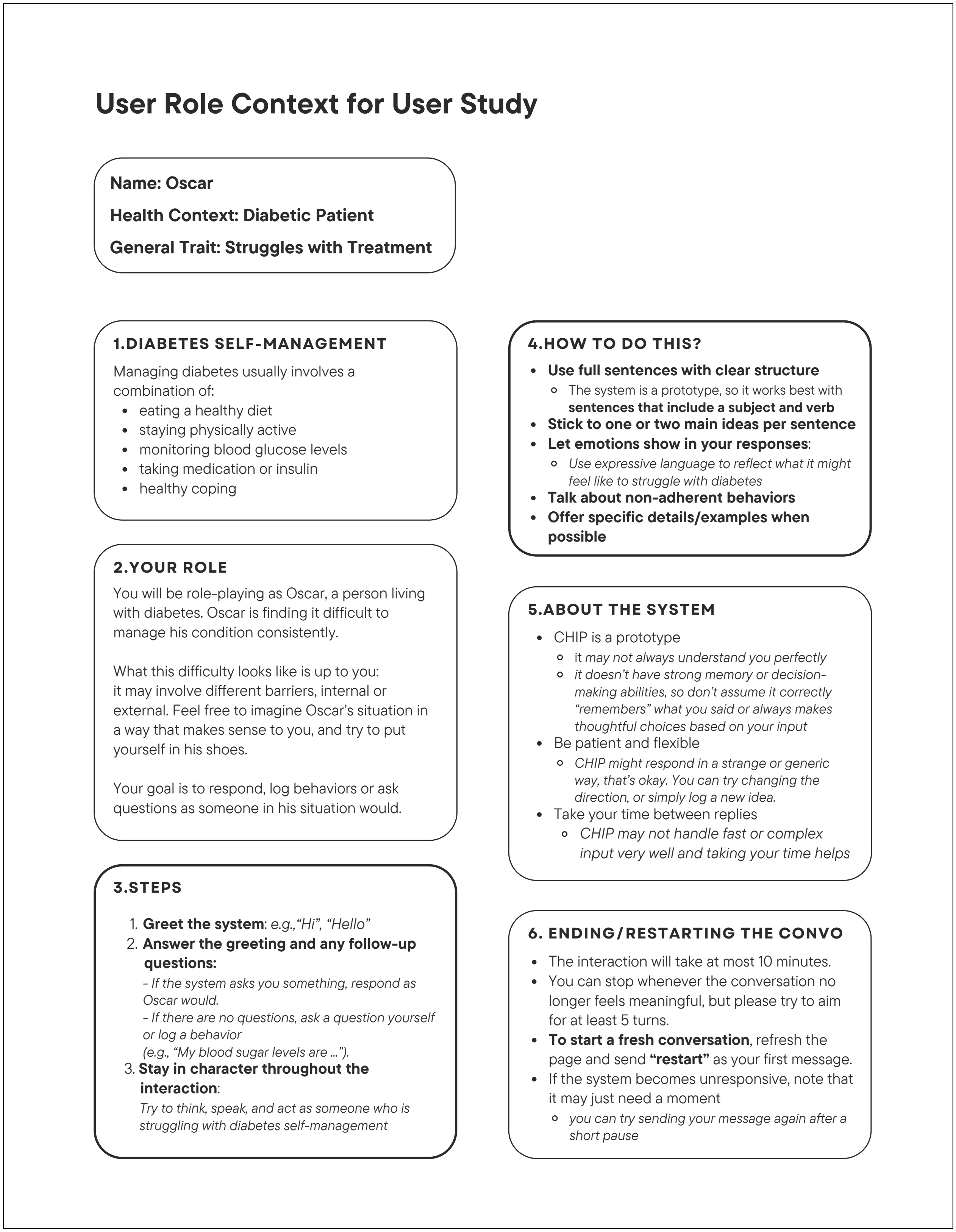
Figure 3 displays the User Context Sheet used during the pilot study.

B Power Analysis

Table 6: G*Power settings for sample size estimation.

Parameter	Value
Test family	F-tests
Analysis type	MANOVA: Repeated measures, within-between interaction
Type of power analysis	A priori: Compute required sample size - given α, power, and effect size
Effect size ($f(V)$)	0.25
Significance level (α)	0.05
Statistical power ($1-\beta$)	0.90
Number of groups	3
Number of measurements	3
Total sample size	126

Figure 3: Instruction sheet used by participants in the pilot study.



User Role Context for User Study

Name: Oscar

Health Context: Diabetic Patient

General Trait: Struggles with Treatment

1.DIABETES SELF-MANAGEMENT

Managing diabetes usually involves a combination of:

- eating a healthy diet
- staying physically active
- monitoring blood glucose levels
- taking medication or insulin
- healthy coping

2.YOUR ROLE

You will be role-playing as Oscar, a person living with diabetes. Oscar is finding it difficult to manage his condition consistently.

What this difficulty looks like is up to you:
it may involve different barriers, internal or external. Feel free to imagine Oscar's situation in a way that makes sense to you, and try to put yourself in his shoes.

Your goal is to respond, log behaviors or ask questions as someone in his situation would.

3.STEPS

1. **Greet the system**: *e.g.,"Hi", "Hello"*
2. **Answer the greeting and any follow-up questions:**
   - If the system asks you something, respond as Oscar would.
   - If there are no questions, ask a question yourself or log a behavior
   (e.g., "My blood sugar levels are ...").
3. **Stay in character throughout the interaction**:
   Try to think, speak, and act as someone who is struggling with diabetes self-management

4.HOW TO DO THIS?

- **Use full sentences with clear structure**
  - The system is a prototype, so it works best with **sentences that include a subject and verb**
- **Stick to one or two main ideas per sentence**
- **Let emotions show in your responses**:
  - *Use expressive language to reflect what it might feel like to struggle with diabetes*
- **Talk about non-adherent behaviors**
- **Offer specific details/examples when possible**

5.ABOUT THE SYSTEM

- CHIP is a prototype
  - *it may not always understand you perfectly*
  - *it doesn't have strong memory or decision-making abilities, so don't assume it correctly "remembers" what you said or always makes thoughtful choices based on your input*
- Be patient and flexible
  - *CHIP might respond in a strange or generic way, that's okay. You can try changing the direction, or simply log a new idea.*
- Take your time between replies
  - *CHIP may not handle fast or complex input very well and taking your time helps*

6. ENDING/RESTARTING THE CONVO

- The interaction will take at most 10 minutes.
- You can stop whenever the conversation no longer feels meaningful, but please try to aim for at least 5 turns.
- **To start a fresh conversation**, refresh the page and send **"restart"** as your first message.
- If the system becomes unresponsive, note that it may just need a moment
  - *you can try sending your message again after a short pause*