To deceive or self-deceive?

Framing Language to Discourage Deception in Diabetes Lifestyle Management Systems

Marina Mădăraș
m.madaras@student.tudelft.nl

Supervised by:
Prof. dr. C.M. Jonker, J.D. Top, MSc

TU Delft

1. Background

Diabetes requires ongoing self-management through sustained lifestyle changes. **Diabetes lifestyle management** (DLM) tools aim to support patients in this process.

Non-adherence is common, and DLM tools rely on self-reported input from patients, which may be **misleading** or **inaccurate**.

To **discourage** users from lying, a behavioral intervention can be designed to target the factors that drive deception.

CHIP is a **chatbot**-based research prototype of a DLM system, extended in this work to explore **language-framing** interventions.

2. Research question

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

3. Understanding Behavior

Goal:
Understand what drives **deception** and **poor diabetes self-management** to design interventions that support behavior change.

Method:
Performed a **literature review**. From the findings, used an intervention design framework, the **Behavior Change Wheel** (BCW), to categorize drivers and align them with effective intervention functions.

Key Findings:
Deception is often a means to **protect the self**, and has **overlapping psychological drivers** with poor diabetes self-management.

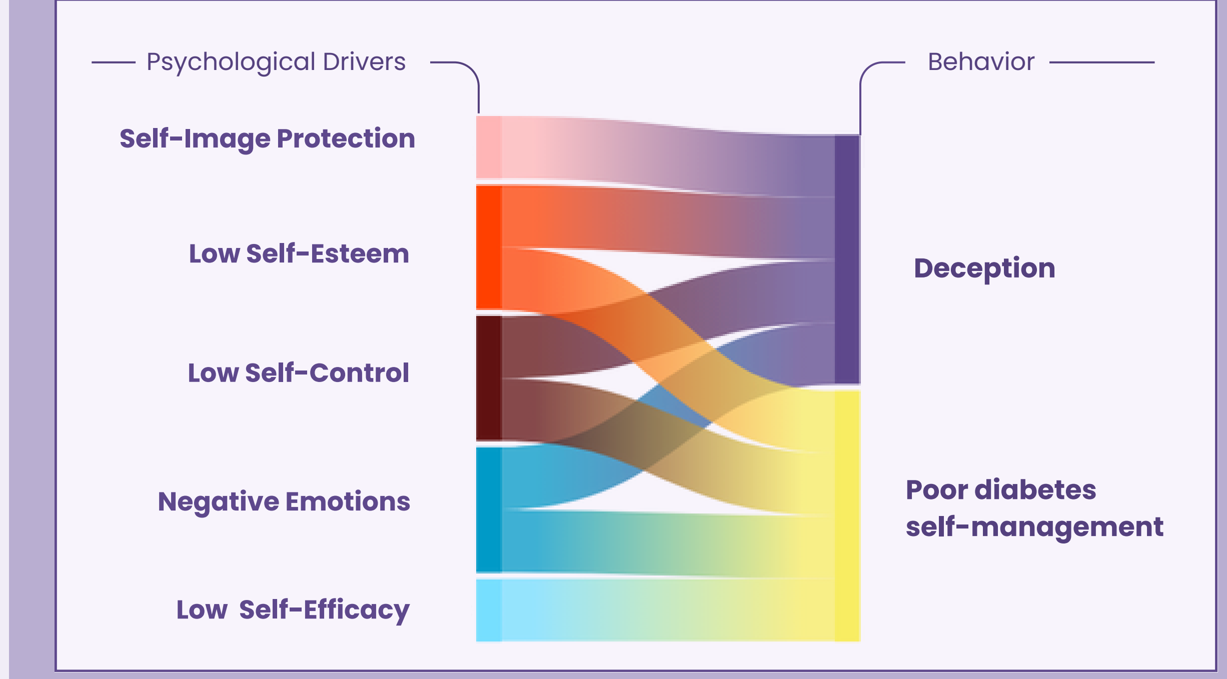


Figure 1: Psychological drivers of deception and poor diabetes self-management.

4. Intervention Design

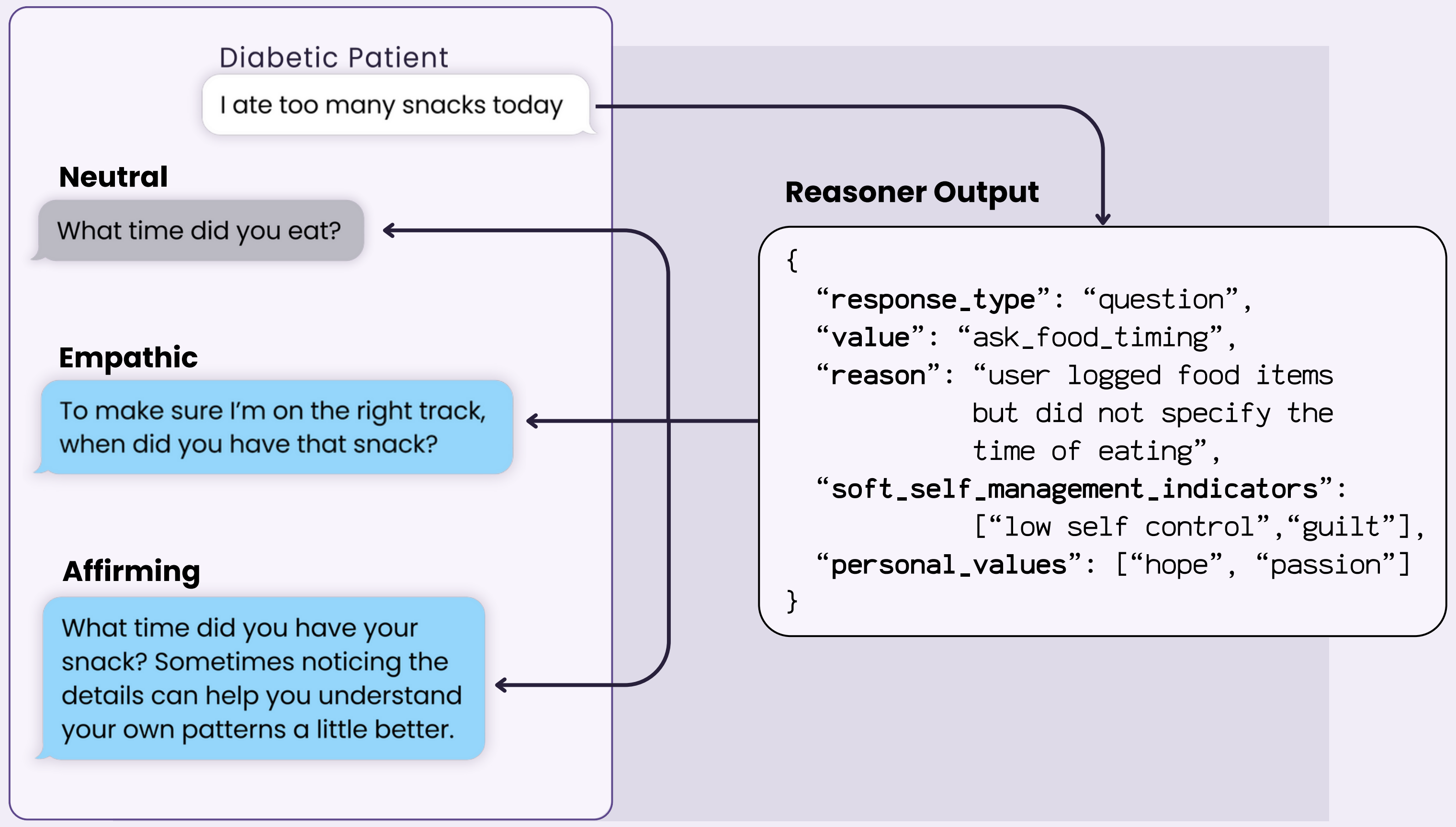
Intervention functions

- Shape how users interpret and emotionally respond to messages.
- Foster a supportive, non-judgemental environment.

Delivered through language framing

Empathic language communicates an effort to understand the user's experience

Affirming language reinforces the user's values and identity



Diabetic Patient

I ate too many snacks today

Neutral

What time did you eat?

Empathic

To make sure I'm on the right track, when did you have that snack?

Affirming

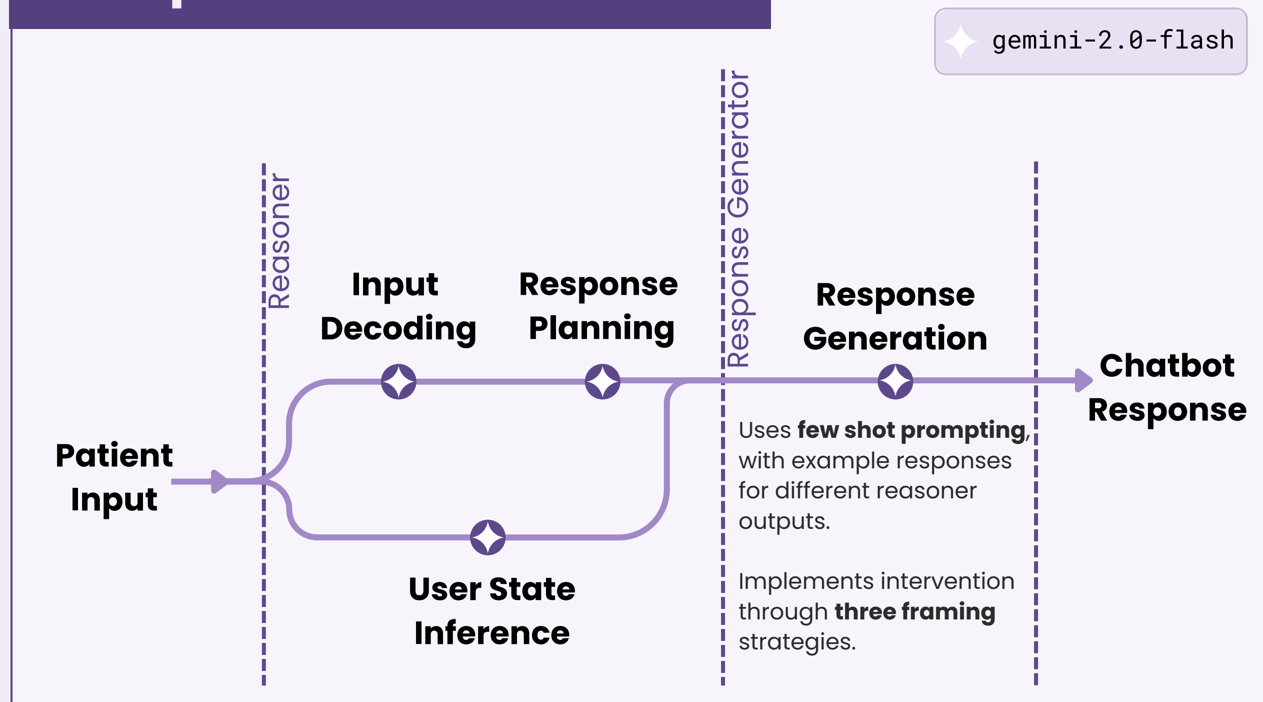
What time did you have your snack? Sometimes noticing the details can help you understand your own patterns a little better.

Reasoner Output

```
{
  "response_type": "question",
  "value": "ask_food_timing",
  "reason": "user logged food items
            but did not specify the
            time of eating",
  "soft_self_management_indicators":
            ["low self control","guilt"],
  "personal_values": ["hope", "passion"]
}
```

Figure 2. CHIP's Reasoner interprets user input and produces a structured data dictionary, which the Response Generator uses to frame the message using **Neutral**, **Empathic**, or **Affirming** strategy. The *personal_values* field is derived from previous interactions.

5. Implementation in CHIP

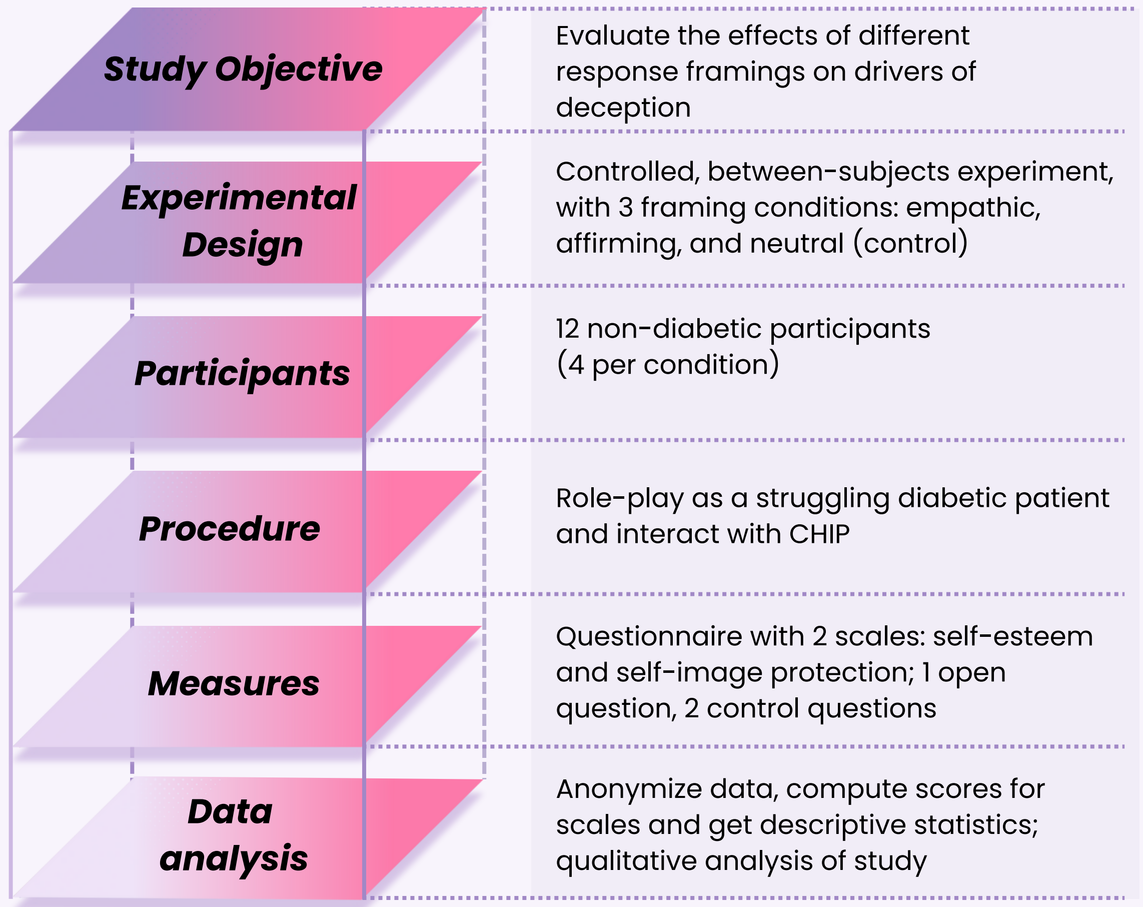


Uses **few shot prompting**, with example responses for different reasoner outputs.

Implements intervention through **three framing** strategies.

6. Pilot Study

A **user study** was conducted to explore whether response framings influence drivers of deception. This diagram provides an overview of the **study's methodology**.



Study Objective: Evaluate the effects of different response framings on drivers of deception

Experimental Design: Controlled, between-subjects experiment, with 3 framing conditions: empathic, affirming, and neutral (control)

Participants: 12 non-diabetic participants (4 per condition)

Procedure: Role-play as a struggling diabetic patient and interact with CHIP

Measures: Questionnaire with 2 scales: self-esteem and self-image protection; 1 open question, 2 control questions

Data analysis: Anonymize data, compute scores for scales and get descriptive statistics; qualitative analysis of study

7. Results

⚠ Results are exploratory:
Participants: 12 non-diabetic participants
LLM: unresponsive in **8/12** interactions (due to the model being overloaded)
Reasoner: output sometimes lacked contextual coherence

Empathic: lowest self-image protection, perceived as gentle, non-judgemental; *most aligned with hypothesis*
Affirming: highest self-esteem, perceived as kind and supportive
Neutral: seen as emotionally flat or impersonal

Table 1. Average B-RSES (self-esteem scale) and BIDR-16 (self-image protection scale) scores across conditions. Results are **not** significant, due to small sample size (n=12) and confounding factors.

B-RSES: self-esteem
higher score means
higher self-esteem

BIDR-16: self-image protection
lower score means
lower need to protect self-image

Condition	*B-RSES*	*BIDR-16*
Empathic	**2.16**	**3.50**
Affirming	2.61	4.38
Neutral	1.86	4.14

8. Future Work

Prototype:
Improve dialogue context-tracking and enhance response planning to enable CHIP to carry out more coherent conversations.

Experiment:
Repeat the study with diabetic patients (approximately 126), using a longitudinal design with pre-, post-, and follow-up phases to assess the effectiveness of language-framing interventions.