CHAT SUMMARIES FOR DECEPTION AND ADHERENCE DETECTION

Author Hugo Koot | **Email** h.m.g.koot@student.tudelft.nl | **Date:** June 25th, 2025 | **Supervisors:** Prof. Catholijn Jonker, J.D. Top, Msc

1. INTRODUCTION

Patients interacting with diabetes support chats can struggle with accurately reporting behaviors such as diet, medication intake, and blood sugar levels. This can be due to **deceptive reporting** or **non-adherence** to their prescribed program. Deceptive or inaccurate reporting can hinder effective diabetes management. This project explores using **AI-generated summaries** to detect patient deception and adherence within the **CHIP** system (a diabetes support chat application).

2. OBJECTIVE

- **Develop and implement** AI-based chat summary pipeline to capture patient-chatbot interactions as an extension of the current CHIP system.
- **Measure** the effect of the AI-generated summary on deception and adherence detection.
- **The goal** is to enable doctors of diabetes patients to easily identify deceptive behaviors or adherence issues in diabetes support apps using the summary.

3. METHODOLOGY

Literature study:
- Identify relevant patient behavior indicators from research.
- Research relevant prompt engineering techniques.

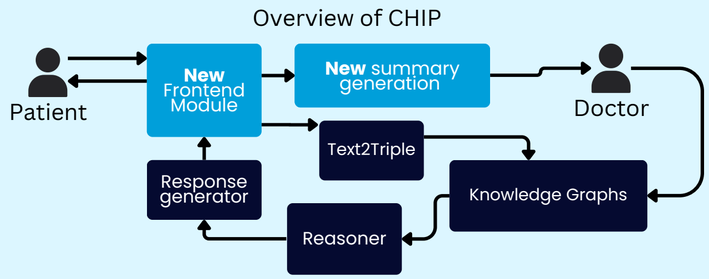
Implementation:
- Extend the CHIP system with the chat summarization functionality.
- Create summarization prompts focusing on key deception and adherence indicators.
- Create a web application in which the annotation experiment can take place.

Annotation experiment:
- Generate synthetic data with a ground truth for both deception and adherence.
- Measure the effect of the summaries on the annotation accuracy, speed, and the overall inter-annotator agreement.
- Data to annotate will be presented in one of two conditions:
  - Chat logs only
  - Chat logs + AI-generated summary

4. KEY INDICATORS AND PROMPT ENGINEERING TECHNIQUES

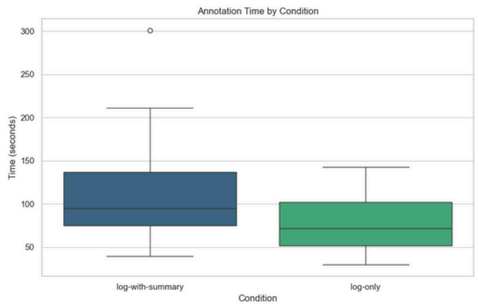
Key indicators:
- Inconsistencies
- Vague or evasive language
- Engagement level
- Gaming the system

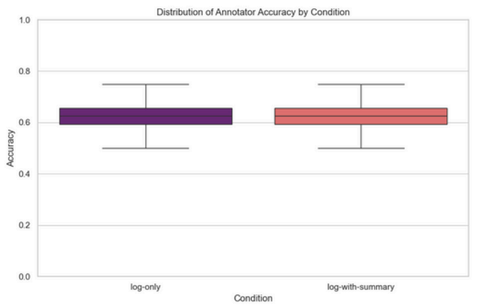
Prompt engineering techniques:
- Implicit Retrieval-Augmented Generation (Implicit RAG)
- Annotation-Guideline Prompting
- Hallucination Rail

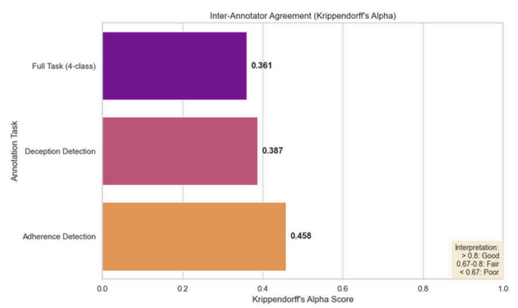


5. ANNOTATION EXPERIMENT RESULTS

Based on the experimental findings, the AI-generated summaries were **not a successful intervention**. The summaries **increased decision time**, likely due to an increased cognitive load as the annotators tried to process more information. Furthermore, the summaries **failed to improve the accuracy**. In fact, the accuracy stayed exactly the same. Lastly, we measured a **poor inter-annotator agreement**. This signals that the task was highly subjective and the results depended on the individual annotators. This means a low reliability of the results as running the experiment with different annotators would likely produce different results.







6. LIMITATIONS

- Simulated chat data may not fully represent real patient interactions.
- The absence of a summary-only condition meant the independent effect of the summary was not fully explored
- LLMs are black boxes, which limits the explainability of the summaries.
- Real-world effectiveness and acceptance by patients and medical professionals remain to be validated.
- The small sample size of four in the annotation experiment restricts the generalization and reliability of the findings.
- The study did not systematically compare the individual components of the prompt, meaning the chosen behavioral indicators and engineering techniques may not be the most effective ones.

7. FUTURE WORK

- **Refined experimental design:** Experiment with a summary-only condition to isolate the effect of the summary.
- **Systematic component evaluation:** The key indicators and prompt engineering techniques chosen, should be evaluated against other possible key indicators or prompt engineering techniques to determine the most effective ones.
- **Real-world validation:** The annotation experiment should be validated with real-world data and the annotation should be done by doctors to evaluate effectiveness in an actual usecase.
- **Validating utility:** This experiment was done under the assumption that summaries for deception and adherence detection would be valuable to doctors. This assumption should be tested and doctors should be surveyed about what key indicators they would find useful.